

Comme ci ou comme ça?

**A classroom-based evaluation of two approaches to the
teaching of French to young learners in England:**

Impacts on teachers and pupils

Elaine Minett

PhD

University of York

Education

December 2025

Abstract

Learning a foreign language (FL) is statutory in the English National Curriculum (DfE, 2013b) for pupils aged 7-11, yet little guidance is provided on appropriate pedagogical approaches to facilitate progress. In this limited-exposure context, teaching may be by non-specialist teachers with low levels of FL competence and/or confidence, who may find ensuring progression challenging.

The mixed methods design addressed the need for experimental studies evaluating holistic approaches to the teaching of FL to young learners ($n = 149$). It evaluated the impact of the 'Novel', language-driven approach (systematically embedding vocabulary, grammar, phonics) on outcomes and motivation of 7-8-year-olds in their first year of French and on non-specialist teachers' confidence and enthusiasm, alongside a 'Standard' approach representing typical topic-/phrase-driven pedagogy. Pupils were pair-matched on ranked scores based on a language-analytic ability pre-test (Kasprowicz et al., 2019), then randomly allocated to two groups. Teachers alternated between approaches. Pupils' self-efficacy and enjoyment were surveyed pre-, mid- and end-project, supplemented by group interviews. Post- and delayed post-tests measured outcomes in vocabulary, symbol-sound correspondences, grammar, and speaking. Teacher interviews and questionnaires elicited confidence and perceived impact on subject knowledge and enjoyment.

Mixed-effects models indicated learning in both approaches, with descriptive trends for generally higher scores in Novel than Standard in vocabulary measures, phonics Read Aloud, spoken grammar recall, Elicited Imitation grammar, and both open-ended speaking measures; these reached statistical significance in written vocabulary recognition and spoken grammar recall. In both approaches, pupils' self-efficacy increased, whereas enjoyment decreased (with individual variations). Teachers' confidence grew, suggesting challenges delivering novelty but positive perceptions about learning and impact on their subject knowledge. Baseline language-analytic ability appeared to indicate language learning potential overall, with some descriptive trends suggesting that Novel compensated for individual learning aptitude differences. Pupils did not appear disadvantaged when taught explicitly.

List of Contents

Abstract.....	2
List of Contents	3
List of Abbreviations	18
List of Tables	20
List of Figures.....	23
Acknowledgements	27
Author's declaration	28
Chapter 1: Introduction	29
1.1 Foreign languages in English primary schools: background.....	29
1.2 Non-specialist provision	29
1.3 Young learners and foreign language learning.....	32
1.4 Factors influencing outcomes	34
1.5 Summary.....	39
Chapter 2: Literature Review I. Practice in foreign languages: vocabulary, phonics, grammar	41
2.1 Foreign language learning practice.....	41
2.2 Vocabulary.....	47
2.3 Phonics.....	57
2.4 Grammar	62
2.5 The two approaches: similarities and differences.....	70
Chapter 3: Literature Review II. Pupil and teacher motivation and related variables, focusing on research conducted with young learners	76
3.1 Pupil motivation	76
3.2 Teacher emotions.....	83
Chapter 4: Methodology and methods	88
4.1 Aims, rationale and research questions	88
4.2 Ethics	95

4.3 Methods	99
4.4 Quantitative analysis.....	129
4.5 Qualitative analysis.....	137
4.6 Summary of methods.....	142
Chapter 5: Results and Discussion I. Pupils' learning outcomes	143
5.1 General information on results	143
5.2 Using the pre-test outcomes to compare starting points in Novel and Standard groups	145
5.3 Vocabulary recognition in the oral modality: results	146
5.4 Vocabulary recognition in the written modality: results	148
5.5 Vocabulary recall in the written modality: results.....	153
5.6 Vocabulary recall in the oral modality: results.....	156
5.7 Pupils' learning outcomes in vocabulary: discussion.....	160
5.8 Phonics transcription: results.....	170
5.9 Phonics Read Aloud (RA): results.....	173
5.10 Pupils' learning outcomes in phonics: discussion	176
5.11 Written grammar recognition: results.....	180
5.12 Spoken grammar recall: results	183
5.13 Elicited imitation grammar test	187
5.14 Pupils' learning outcomes in grammar: discussion	190
5.15 Open-ended speaking Picture Response: results	198
5.16 Elicited imitation speaking: results.....	201
5.17 Pupils' learning outcomes in speaking tasks: discussion	206
5.18 Summary of learning outcomes	213
Chapter 6: Results and Discussion II. Pupils' self-efficacy and enjoyment	214
6.1 General information on results	214
6.2 Pupil self-efficacy: results from quantified data analysis.....	215
6.3 Pupil self-efficacy: results from qualitative data analysis	220
6.4 Pupil self-efficacy: discussion of quantitative and qualitative data analysis.....	231
6.5 Pupil enjoyment: results from quantified data analysis.....	236

6.6 Pupil enjoyment: results from qualitative data analysis	242
6.7 Pupil enjoyment: discussion of quantified and qualitative data analysis	253
6.8 Pupil perceptions of teacher enjoyment: results from quantified and qualitative data analysis	260
6.9 Summary of results and discussion of pupil affective variables.....	263
Chapter 7: Results and discussion III. Teacher findings	266
7.1 Comparing fidelity to treatment across the two approaches.....	266
7.2 Teacher starting points: results	269
7.3 Teacher starting points: summary and brief discussion.....	277
7.4 Teacher responses to each approach: results from post-lesson evaluations	278
7.5 Teacher responses to each approach: discussion of results from post-lesson evaluations	287
7.6 Teacher responses to each approach: results from teacher questionnaires	291
7.7 Teacher responses to each approach: discussion of teacher questionnaire results	295
7.8 Teacher responses to each approach: results from teacher interviews (Interview2/Interview3).....	296
7.9 Teacher responses to each approach: discussion of results from interviews	325
7.10 Comparing teacher reports/perceptions with pupil data: discussion drawing together key findings from Results I, II and III.....	337
7.11 Teacher findings: summary	340
Chapter 8: Conclusions and recommendations	341
8.1 Study context	341
8.2 Summary of study.....	341
8.3 Summary of findings in relation to research questions	342
8.4 Wider contribution of findings and implications for the primary foreign languages classroom and policy	348
8.5 Limitations and future research	352
8.6 Final concluding remarks	353
Reference List.....	354
Appendix A.1 <i>Detailed Discussion of Evaluation as a Research Method</i>	395

Appendix A.2 <i>Individual Vocabulary Items: Discussion</i>	398
Appendix A.3 <i>Individual Phonics Items: Discussion</i>	400
Appendix A.4 <i>Elicited Imitation Grammar (Rescored Model)</i>	402
Appendix B.1 <i>Pre-test: Language-analytic Ability Test (from Kasprawicz et al., 2019)</i> ...	407
Appendix B.2 <i>Piloting Information - The Language-analytic Ability Test</i>	414
Appendix B.3 <i>Language-analytic Ability Test Instructions</i>	415
Appendix C.1 <i>Teacher Information and Consent Form</i>	417
Appendix C.2 <i>Child Assent Form</i>	424
Appendix D.1 <i>Data Management Plan</i>	425
Appendix D.2 <i>Data Collection Logs</i>	427
Appendix E.1 <i>Mapping of the National Curriculum and Testing Content</i>	428
Appendix E.2 <i>Piloting the Vocabulary Tests</i>	429
Appendix E.3 <i>Vocabulary Content Allocated to Test Type</i>	431
Appendix E.4 <i>Distribution of Corpus-based Frequency Ratings According to Test Type</i> .	432
Appendix E.5 <i>Word Length in Letters According to Test Type</i>	433
Appendix E.6 <i>Recency by Lesson When Last Encountered and Number of Times Encountered for Vocabulary in Each Test Type</i>	434
Appendix E.7 <i>Distribution of Word Classes According to Test Type</i>	435
Appendix E.8 <i>Concreteness of Words According to Test Type</i>	436
Appendix E.9 <i>Cognate Status of Words According to Test Type</i>	437
Appendix E.10 <i>Differences Between English and French Symbol-Sound Correspondences According to Test Type</i>	438
Appendix E.11 <i>Syllable Count of Vocabulary According to Test Type</i>	439
Appendix E.12 <i>Sample Vocabulary Listening Test (Recognition, Oral Modality)</i>	440
Appendix E.13 <i>Sample Vocabulary Reading Test (Recognition, Written Modality)</i>	441
Appendix E.14 <i>Sample Vocabulary Writing Test (Recall, Written Modality)</i>	442
Appendix E.15 <i>Sample Vocabulary Speaking Test (Recall, Oral Modality)</i>	443
Appendix E.16 <i>Teacher Test Instructions</i>	444

Appendix E.17 <i>Language-analytic Aptitude Scores of Pupil Subsample for Oral Production Tests: Post-Tests</i>	447
Appendix E.18 <i>Host words for Symbol-Sound Correspondences in the Phonics Transcription Tests and Position in Post-tests and Delayed Post-tests</i>	448
Appendix E.19 <i>Host words for Symbol-Sound Correspondences in the Phonics Read Aloud Tests and Position in Post-tests and Delayed Post-tests</i>	449
Appendix E.20 <i>Piloting the Phonics Tests</i>	450
Appendix E.21 <i>Sample Phonics Transcription Test</i>	453
Appendix E.22 <i>Sample Phonics Read Aloud Test</i>	454
Appendix E.23 <i>Items for Written Grammar Recognition and Position in Post-tests and Delayed Post-tests</i>	455
Appendix E.24 <i>Items for Spoken Grammar Recall and Position in Post-tests and Delayed Post-tests</i>	457
Appendix E.25 <i>Items for Elicited Imitation Grammar and Elicited Imitation Speaking Tests and Position in Post-tests and Delayed Post-tests</i>	459
Appendix E.26 <i>Piloting the Grammar Tests</i>	460
Appendix E.27 <i>Sample Written Grammar Recognition Test</i>	463
Appendix E.28 <i>Sample Spoken Grammar Recall Test</i>	464
Appendix E.29 <i>Sample Slides from Elicited Imitation Grammar Test</i>	465
Appendix E.30 <i>Piloting the Speaking Tests</i>	466
Appendix E.31 <i>Picture Response: Pictures and Position in Post-tests and Delayed Post-tests</i>	467
Appendix E.32 <i>Picture Response Scoring Guidelines</i>	468
Appendix E.33 <i>Scoring Guidelines for Elicited Imitation Speaking Task</i>	469
Appendix E.34 <i>Vocabulary Items for Vocabulary Recognition in the Oral Modality (Listening) Tests and Position in Post-tests and Delayed Post-tests</i>	473
Appendix E.35 <i>Vocabulary Items for Vocabulary Recognition in the Written Modality Tests and Position in Post-tests and Delayed Post-tests</i>	474
Appendix E.36 <i>Vocabulary Items for Vocabulary Recall in the Oral Modality Tests and Position in Post-tests and Delayed Post-tests</i>	475

Appendix E.37 <i>Vocabulary Items for Vocabulary Recall Tests in the Written Modality and Position in Post-tests and Delayed Post-tests</i>	476
Appendix F.1 <i>Teachers' Post-lesson Evaluation</i>	477
Appendix F.2 <i>Piloting Teachers' Post-lesson Evaluations</i>	478
Appendix F.3 <i>Primary French Teacher Questionnaire 1 Final</i>	479
Appendix F.4 <i>Teacher Questionnaire 2 & 3</i>	484
Appendix F.5 <i>Number of Post-Lesson Evaluations Completed by Teachers by Approach</i>	485
Appendix G.1 <i>Primary French Questionnaire: Pupils 1</i>	486
Appendix G.2 <i>Pupil Questionnaire 2/3</i>	493
Appendix G.3 <i>Pupil Questionnaire 3 – Open-ended Questions</i>	498
Appendix G.4 <i>Piloting Pupil Questionnaires</i>	499
Appendix G.5 <i>Original Questions from Carreira (2006) Used as Measure of Pupil Enjoyment</i>	500
Appendix G.6 <i>Piloting Pupil Group Interviews</i>	501
Appendix G.7 <i>Pupils' Group Interview Schedule</i>	502
Appendix G.8 <i>Feelings Thermometer Activity</i>	504
Appendix G.9 <i>Self-efficacy Questions from Pupil Questionnaires</i>	505
Appendix G.10 <i>Response Options and Scoring System for Self-efficacy Questions</i>	506
Appendix G.11 <i>Response Options for Enjoyment Questions in Pupil Questionnaire 2 (PQ2) and Pupil Questionnaire 3 (PQ3)</i>	507
Appendix H.1 <i>Piloting Teacher Interviews</i>	508
Appendix H.2 <i>Teacher Interview Schedule 1</i>	509
Appendix H.3 <i>Teacher Interview Schedule 2 & 3</i>	512
Appendix I.1 <i>Box Plots Showing Data Distribution for Language-analytic Ability Test Scores According to Approach in Schools A, B, C, E, F and All Schools</i>	514
Appendix I.2 <i>Histograms Showing Data Distribution for Language-analytic Ability Test Scores According to Approach in Schools A, B, C, E, F and All Schools</i>	515
Appendix I.3 <i>Skew, Kurtosis and Shapiro-Wilk Scores for Language-analytic Ability Test Scores by Approach</i>	517

Appendix I.4 <i>Comparing Corpus-Based Frequency Ratings in Vocabulary Test Content in Receptive and Productive Modes, and Oral and Written Modalities – Normality of Distribution</i>	518
Appendix I.5 <i>Comparing Word Length in Vocabulary Test Content in Receptive and Productive Modes, and Oral and Written Modalities – Normality of Distribution</i>	520
Appendix I.6 <i>Comparing Symbol-Sound Correspondence Distance Between the English and the French in Vocabulary Test Content in Receptive and Productive Modes, and Oral and Written Modalities – Normality of Distribution</i>	522
Appendix I.7 <i>Comparing Syllable Count in Vocabulary Test Content in Receptive and Productive Modes, and Oral and Written Modalities – Normality of Distribution</i>	524
Appendix I.8 <i>Comparing Recency of Encounter in Vocabulary Test Content in Receptive and Productive Modes, and Oral and Written Modalities – Normality of Distribution</i>	526
Appendix I.9 <i>Comparing Frequency of Encounter in Vocabulary Test Content in Receptive and Productive Modes, and Oral and Written Modalities – Normality of Distribution</i>	529
Appendix J.1 <i>Checking Assumptions for Mixed-effects Models</i>	532
Appendix K.1 <i>Example Lesson Plan: Novel</i>	533
Appendix K.2 <i>Example Lesson Plan: Standard</i>	537
Appendix K.3 <i>Observed Lessons and Monitored Audio-recorded Lessons per Teacher</i> ...	542
Appendix K.4 <i>Observation Chart</i>	543
Appendix K.5 <i>Calculations for Lessons Delivered as Intended</i>	544
Appendix K.6 <i>Fidelity to Treatment: Descriptive Statistics</i>	545
Appendix K.7 <i>Descriptive Deviations Chart</i>	546
Appendix L.1 <i>A First-Draft Conceptual Framework Showing Identified Influences on Enjoyment and Anxiety of Non-specialist Teachers of Primary French</i>	548
Appendix L.2 <i>Sample Interview1 Emotions Grid</i>	549
Appendix L.3 <i>Example of First-stage Coding Taken from Pupil Questionnaire 1 Open-Ended Question</i>	551
Appendix L.4 <i>Example of Code Merging Taken from Pupil Questionnaire 1 Open-Ended Question</i>	553
Appendix L.5 <i>Examples of Coding Types Taken from Teacher Questionnaire 2 (Mid-project)</i>	561

Appendix L.6 <i>Interview Analysis – Taken from the Literature Review on Specialist Teaching and Used to Inform Coding</i>	563
Appendix L.7 <i>Example of Advanced Coding Taken from Teacher Interview 2 – Developing and Naming Themes</i>	565
Appendix L.8 <i>Thoughts on Teacher Interview 1 – Example of Researcher Memo</i>	566
Appendix L.9 <i>Example of Concept Table from Group Interview Data Showing Impact of Approach on Pupil Confidence at Mid-point: Novel</i>	568
Appendix L.10 <i>Analytic Memo (NVivo, 21.03.24)</i>	569
Appendix M.1 <i>Normality of Distribution for Vocabulary Recognition in the Oral Modality</i>	570
Appendix M.2 <i>Normality of Distribution Checks of Post-test Receptive Tests: Vocabulary Recognition in the Oral Modality (raw data)</i>	571
Appendix M.3 <i>Normality of Distribution Checks of Delayed Post-test Receptive Tests: Vocabulary Recognition in the Oral Modality (raw data)</i>	572
Appendix M.4 <i>Normality of Distribution for Vocabulary Recognition in the Written Modality</i>	573
Appendix M.5 <i>Normality of Distribution Checks of Post-test Vocabulary Recognition in the Written Modality Tests (raw data)</i>	574
Appendix M.6 <i>Normality of Distribution Checks of Delayed Post-test Vocabulary Recognition in the Written Modality (raw data)</i>	575
Appendix M.7 <i>Normality of Distribution for Vocabulary Recall in the Written Modality</i> .	576
Appendix M.8 <i>Normality of Distribution Checks of Post-test Receptive Tests: Vocabulary Recall in the Written Modality (raw data)</i>	577
Appendix M.9 <i>Normality of Distribution Checks of Delayed Post-test Receptive Tests: Vocabulary Recall in the Written Modality (raw data)</i>	578
Appendix M.10 <i>Normality of Distribution for Vocabulary Recall in the Oral Modality</i>	579
Appendix M.11 <i>Normality of Distribution Checks of Post-test Receptive Tests: Vocabulary Recall in the Oral Modality (raw data)</i>	580
Appendix M.12 <i>Normality of Distribution Checks of Delayed Post-test Receptive Tests: Vocabulary Recall in the Oral Modality (raw data)</i>	581
Appendix M.13 <i>Normality of Distribution for Phonics Transcription Tests</i>	582

Appendix M.14 <i>Normality of Distribution Checks of Post-test Phonics Transcription Tests</i>	583
Appendix M.15 <i>Normality of Distribution Checks of Delayed Post-test Phonics Transcription Tests</i>	584
Appendix M.16 <i>Normality of Distribution for Phonics Read Aloud Tests</i>	585
Appendix M.17 <i>Normality of Distribution Checks of Post-test Phonics Read Aloud Tests</i>	586
Appendix M.18 <i>Normality of Distribution Checks of Delayed Post-test Phonics Read Aloud Tests</i>	587
Appendix M.19 <i>Normality of Distribution Checks for Written Grammar Recognition Tests</i>	588
Appendix M.20 <i>Normality of Distribution Checks of Post-test: Written Grammar Recognition (raw data)</i>	589
Appendix M.21 <i>Normality of Distribution Checks of Delayed Post-test: Written Grammar Recognition (raw data)</i>	590
Appendix M.22 <i>Normality of Distribution for Spoken Grammar Recall Tests</i>	591
Appendix M.23 <i>Normality of Distribution Checks of Post-test: Spoken Grammar Recall (raw data)</i>	592
Appendix M.24 <i>Normality of Distribution Checks of Delayed Post-test: Spoken Grammar Recall (raw data)</i>	593
Appendix M.25 <i>Normality of Distribution for Elicited Imitation Grammar Tests</i>	594
Appendix M.26 <i>Normality of Distribution Checks of Post-test: Elicited Imitation Grammar (raw data)</i>	595
Appendix M.27 <i>Normality of Distribution Checks of Delayed Post-test: Elicited Imitation Grammar (raw data)</i>	596
Appendix M.28 <i>Normality of Distribution for Picture Response Test</i>	597
Appendix M.29 <i>Normality of Distribution Checks of Post-test: Picture Response (raw data)</i>	598
Appendix M.30 <i>Normality of Distribution Checks of Delayed Post-test: Picture Response (raw data)</i>	599
Appendix M.31 <i>Normality of Distribution for Elicited Imitation Speaking Tests</i>	600

Appendix M.32 <i>Normality of Distribution Checks of Post-test: Elicited Imitation Speaking (raw data)</i>	601
Appendix M.33 <i>Normality of Distribution Checks of Delayed Post-test: Elicited Imitation Speaking (raw data)</i>	602
Appendix M.34 <i>Confirmatory Factor Analysis Results for Elicited Imitation Grammar...</i>	603
Appendix N.1 <i>Model Code: Vocabulary Recognition in the Speaking and Writing Modalities</i>	604
Appendix N.2 <i>Linearity Check: Vocabulary Recognition in the Speaking and Writing Modalities</i>	605
Appendix N.3 <i>Model Summary: Vocabulary Recognition in the Speaking and Writing Modalities</i>	606
Appendix N.4 <i>Pairwise Comparisons Using Emmeans for Vocabulary Recognition in the Spoken and Written Modalities: Scores by Time-point and Approach</i>	608
Appendix N.5 <i>Vocabulary Recall in the Written Modality Model Code</i>	609
Appendix N.6 <i>Vocabulary Recall in the Oral Modality Model Code</i>	610
Appendix N.7 <i>Brant Test Results for Vocabulary Recall in the Written Modality</i>	611
Appendix N.8 <i>Summary of Model (Writing Vocabulary)</i>	612
Appendix N.9 <i>Pairwise Comparisons Using Emmeans for Vocabulary Recall in the Written Modality Scores by Time-point and Approach</i>	613
Appendix N.10 <i>Brant Test Results for Vocabulary Recall in the Oral Modality</i>	614
Appendix N.11 <i>Summary of Model (Speaking Vocabulary)</i>	615
Appendix N.12 <i>Pairwise Comparisons Using Emmeans for Vocabulary Recall in the Oral Modality by Time-point and Approach</i>	616
Appendix N.13 <i>Pairwise Comparisons Using Emmeans for Vocabulary Recall in the Oral Modality by Approach and Time-point</i>	617
Appendix N.14 <i>Phonics Read Aloud Model Code</i>	618
Appendix N.15 <i>Brant Test Results for Phonics Read Aloud Tests</i>	619
Appendix N.16 <i>Summary of Model (Phonics Read Aloud)</i>	620
Appendix N.17 <i>Pairwise Comparisons Using Emmeans for Phonics Speaking Read Aloud Scores by Time-point and Approach</i>	621
Appendix N.18 <i>Written Grammar Recognition Model Code</i>	622

Appendix N.19 <i>Linearity Check for Written Grammar Recognition Test</i>	623
Appendix N.20 <i>Summary of Written Grammar Recognition Model</i>	624
Appendix N.21 <i>Pairwise Comparisons for Written Grammar Recognition by Time-point</i>	625
Appendix N.22 <i>Pairwise Comparisons for Written Grammar Recognition by Approach..</i>	626
Appendix N.23 <i>Spoken Grammar Recognition Model Code</i>	627
Appendix N.24 <i>Brant Test Results for Spoken Grammar Recall Test</i>	628
Appendix N.25 <i>Summary of Model (Spoken Grammar Recall)</i>	629
Appendix N.26 <i>Pairwise Comparisons for Speaking Grammar Recall by Time-point and Approach</i>	631
Appendix N.27 <i>Elicited Imitation Grammar Model Code</i>	633
Appendix N.28 <i>Brant Test Results for Elicited Imitation Grammar Test</i>	634
Appendix N.29 <i>Summary of Model (Elicited Imitation Grammar)</i>	635
Appendix N.30 <i>Pairwise Comparisons for Elicited Imitation Grammar by Time-point and Approach</i>	636
Appendix N.31 <i>Elicited Imitation Speaking Model Code</i>	637
Appendix N.32 <i>Brant Test Results for Elicited Imitation Speaking Test</i>	638
Appendix N.33 <i>Summary of Model (Elicited Imitation Speaking)</i>	639
Appendix N.34 <i>Pairwise Comparisons for Elicited Imitation Speaking by Time-point and Approach</i>	640
Appendix O.1 <i>Sample Vocabulary Grid Activity</i>	642
Appendix O.2 <i>Sample Materials from Novel Approach on Menton's Fête du Citron</i>	643
Appendix P.1 <i>Normality of Distribution (Self-efficacy Data)</i>	644
Appendix P.2 <i>Normality of Distribution for PQ1 Self-Efficacy Data</i>	645
Appendix P.3 <i>Normality of Distribution for PQ2 Self-Efficacy Data</i>	646
Appendix P.4 <i>Normality of Distribution for PQ3 Self-Efficacy Data</i>	647
Appendix P.5 <i>Normality of Distribution (Enjoyment Data)</i>	648
Appendix P.6 <i>Normality of Distribution for PQ2 Enjoyment Data (raw ratings)</i>	649
Appendix P.7 <i>Normality of Distribution for PQ3 Enjoyment Data (raw ratings)</i>	650
Appendix Q.1 <i>Self-efficacy Model Code</i>	651

Appendix Q.2 <i>Brant Test Results for Self-efficacy Ratings</i>	652
Appendix Q.3 <i>Summary of Self-efficacy Model</i>	653
Appendix Q.4 <i>Pairwise Comparisons for Self-efficacy Ratings by Time-point</i>	655
Appendix Q.5 <i>Pairwise Comparisons for Self-Efficacy Ratings by Approach</i>	659
Appendix Q.6 <i>Enjoyment Model Code</i>	660
Appendix Q.7 <i>Brant Test Results for Enjoyment Ratings</i>	661
Appendix Q.8 <i>Summary of Enjoyment Model</i>	662
Appendix Q.9 <i>Pairwise Comparisons for Enjoyment Ratings of Pupils Not Looking Forward to French by Approach and Time-point</i>	664
Appendix Q.10 <i>Pairwise Comparisons for Enjoyment Ratings of Unsure Pupils by Approach and Time-point</i>	665
Appendix Q.11 <i>Pairwise Comparisons for Enjoyment Ratings of Pupils with a Positive Baseline by Approach and Time-point</i>	666
Appendix R.1 <i>Additional Pupil Quotations</i>	667
Confidence:.....	667
Self-efficacy.....	667
Enjoyment:.....	668
Dislikes:	668
Appendix R.2 <i>Examples from Audio Recordings Supporting Pupils' Enjoyment of Vocabulary Grid Activities</i>	669
Appendix R.3 <i>Full Quotation from T1 (Interview1)</i>	670
Appendix S.1 <i>Normality of Distribution for Pupil Perceptions of Teacher Enjoyment: Q1 "Do Your Teachers Enjoy Teaching You French?"</i>	671
Appendix S.2 <i>Normality of Distribution for PQ2 Pupil Perceptions of Teacher Enjoyment: Q1 "Do Your Teachers Enjoy Teaching You French?"</i>	672
Appendix S.3 <i>Normality of Distribution for PQ3 Pupil Perceptions of Teacher Enjoyment: Q1 "Do Your Teachers Enjoy Teaching You French?"</i>	673
Appendix S.4 <i>Normality of Distribution for Pupil Perceptions of Teacher Enjoyment: Q2 "Do Your Teachers Try to Make You Excited to Learn French?"</i>	674

Appendix S.5 <i>Normality of Distribution for PQ2 Pupil Perceptions of Teacher Enjoyment: Q2 “Do Your Teachers Try to Make You Excited to Learn French?”</i>	675
Appendix S.6 <i>Normality of Distribution for PQ3 Pupil Perceptions of Teacher Enjoyment: Q2 “Do Your Teachers Try to Make You Excited to Learn French?”</i>	676
Appendix T.1 <i>Normality of Distribution Post-lesson Evaluation: Item 1 “I enjoyed this lesson”</i>	677
Appendix T.2 <i>Normality of Distribution Post-lesson Evaluation: Item 2 “I was tense during this lesson”</i>	678
Appendix T.3 <i>Normality of Distribution Post-lesson Evaluation: Item 3 “Students understood the material during this lesson”</i>	679
Appendix T.4 <i>Normality of Distribution Post-lesson Evaluation: Item 4 “Students were motivated in this lesson”</i>	680
Appendix T.5 <i>Normality of Distribution Post-lesson Evaluation: Item 5 “Students were focused in this lesson”</i>	681
Appendix T.6 <i>Normality of Distribution Post-lesson Evaluation: Item 6 “I was satisfied with what the students knew by the end of the lesson”</i>	682
Appendix T.7 <i>Normality of Distribution Post-lesson Evaluation: Item 7 “I was satisfied that the French I taught was useful”</i>	683
Appendix T.8 <i>Normality of Distribution Post-lesson Evaluation: Item 8 “The activities gave me sufficient evidence about students’ actual learning”</i>	684
Appendix T.9 <i>Post-lesson Evaluation: Normality of Distribution for Teachers’ Perceptions of Pupil Knowledge and Understanding, and of Usefulness of French Taught</i>	685
Appendix T.10 <i>Post-lesson Evaluation: Normality of Distribution for Classroom Enjoyment</i>	686
Appendix T.11 <i>Normality of Distribution Post-lesson Evaluation: I2 “I was tense during this lesson”</i>	687
Appendix U.1 <i>Individual Teacher Data for Teachers’ Perceptions of Pupil Knowledge and Understanding, and Usefulness of the Language Taught</i>	688
Appendix U.2 <i>Individual Teacher Data for Enjoyment</i>	691
Appendix V.1 <i>Normality of Distribution Checks for Teacher 1’s Ratings for Perceived Pupil Outcomes and Linguistic Usefulness</i>	694

Appendix V.2 Normality of Distribution Checks for Teacher 2's Ratings for Perceived Pupil Outcomes and Linguistic Usefulness	695
Appendix V.3 Normality of Distribution Checks for Teacher 3's (SchA) Ratings for Perceived Pupil Outcomes and Linguistic Usefulness	696
Appendix V.4 Normality of Distribution Checks for Teacher 3's (SchC) Ratings for Perceived Pupil Outcomes and Linguistic Usefulness	697
Appendix V.5 Normality of Distribution Checks for Teacher 3's (SchE) Ratings for Perceived Pupil Outcomes and Linguistic Usefulness	698
Appendix V.6 Normality of Distribution Checks for Teacher 7's Ratings for Perceived Pupil Outcomes and Linguistic Usefulness	699
Appendix V.7 Normality of Distribution Checks for Teacher 8's Ratings for Perceived Pupil Outcomes and Linguistic Usefulness	700
Appendix V.8 Normality of Distribution Checks for Teacher 9's Ratings for Perceived Pupil Outcomes and Linguistic Usefulness	701
Appendix V.9 Normality of Distribution Checks for Teacher 1's Ratings for Enjoyment Construct	702
Appendix V.10 Normality of Distribution Checks for Teacher 2's Ratings for Enjoyment Construct	703
Appendix V.11 Normality of Distribution Checks for Teacher 3's Ratings for Enjoyment Construct (School A)	704
Appendix V.12 Normality of Distribution Checks for Teacher 3's Ratings for Enjoyment Construct (School C)	705
Appendix V.13 Normality of Distribution Checks for Teacher 3's Ratings for Enjoyment Construct (School E)	706
Appendix V.14 Normality of Distribution Checks for Teacher 4's Ratings for Enjoyment Construct	707
Appendix V.15 Normality of Distribution Checks for Teacher 7's Ratings for Enjoyment Construct	708
Appendix V.16 Normality of Distribution Checks for Teacher 8's Ratings for Enjoyment Construct	709
Appendix V.17 Normality of Distribution Checks for Teacher 9's Ratings for Enjoyment Construct	710

Appendix V.18 <i>Normality of Distribution Checks for Teacher 1's Ratings for Tension</i>	711
Appendix V.19 <i>Normality of Distribution Checks for Teacher 2's Ratings for Tension</i>	712
Appendix V.20 <i>Normality of Distribution Checks for Teacher 3's (SchA) Ratings for Tension</i>	713
Appendix V.21 <i>Normality of Distribution Checks for Teacher 3's (SchC) Ratings for Tension</i>	714
Appendix V.22 <i>Normality of Distribution Checks for Teacher 3's (SchE) Ratings for Tension</i>	715
Appendix V.23 <i>Normality of Distribution Checks for Teacher 4's Ratings for Tension</i>	716
Appendix V.24 <i>Normality of Distribution Checks for Teacher 7's Ratings for Tension</i>	717
Appendix V.25 <i>Normality of Distribution Checks for Teacher 8's Ratings for Tension</i>	718
Appendix V.26 <i>Normality of Distribution Checks for Teacher 9's Ratings for Tension</i>	719
Appendix W.1 <i>Interview2 Teacher Comments on Pupils' Progress.....</i>	720
Appendix W.2 <i>Interview3 Teacher Comments on Pupils' Progress.....</i>	723
Appendix W.3 <i>Additional Quotations Supporting Growth in Teacher Confidence (Interview2)</i>	726

List of Abbreviations

BERA	British Educational Research Association
CBI	Comprehension-based instruction
CEFR	Common European Framework of Reference for Languages
CFA	Confirmatory factor analyses
CI(s)	Confidence interval(s)
CPD	Continuing professional development
DCSF	Department for Children, Schools and Families
DfE	Department for Education
EFA	Exploratory factor analyses
EFL	English as a FL
EI	Elicited imitation
EIG	Elicited imitation grammar
EIS	Elicited imitation speaking
ELLiE	Early language learning in Europe
FL	Foreign language
FonF	Focus on Form
GCSE	General Certificate of Secondary Education
HLAA	High language-analytic ability
L1	First language
L2	Second language
LAA	Language-analytic ability
LLAA	Low language-analytic ability
ITE	Initial teacher education
MLAA	Medium language-analytic ability
MLPS	Modern Languages in Primary School
NCELP	National Centre for Excellence in Language Pedagogy
PBI	Production-based instruction
PI	Processing instruction
PQ	Pupil questionnaire
PR	Picture response
Q	Question
RA	Read Aloud
RCT	Randomised controlled trial

RQ	Research question
RTM	Regression to the mean
Sch	School (used only in tables/figures/quotations)
SFC	Silent final consonant
SSCs	Symbol-Sound correspondences
TQ	Teacher questionnaire
TSC	Teaching Schools Council
UNICEF	United Nations Children's Fund

List of Tables

Table 1. <i>Comparison of Explicit and Implicit Learning</i>	42
Table 2. <i>Commonalities and Differences of the Schemes of Work</i>	70
Table 3. <i>Research Design for Investigating French Learning Progress through Different Treatments</i>	93
Table 4. <i>Random Allocation of Pupils to Teaching Approaches</i>	93
Table 5. <i>Alternating Classes in Single-form Entry Schools</i>	94
Table 6. <i>Alternating Classes in Schools with Parallel Classes</i>	94
Table 7. <i>Overview of Data Collection Methods Showing Impact on Teacher Participant Time</i>	97
Table 8. <i>Information on Participating Schools 2022-23</i>	99
Table 9. <i>School Information: Average Participant Starting Age, Gender and Mean Language-analytic Ability Scores</i>	101
Table 10. <i>Relationship Between Methods and Research Questions</i>	102
Table 11. <i>Project Timeline Showing Data Collection</i>	103
Table 12. <i>Types of Knowledge and National Curriculum Objectives Addressed in Vocabulary Tests</i>	106
Table 13. <i>Checking for Difference Between Content of Four Test Types on a Range of Factors</i>	109
Table 14. <i>Checking for Difference Between Approaches on Recency and Frequency of Encounter</i>	109
Table 15. <i>Types of Knowledge and National Curriculum Objectives Addressed in Phonics Tests</i>	111
Table 16. <i>Types of Knowledge and National Curriculum Objectives Addressed in Grammar Tests</i>	114
Table 17. <i>Types of Knowledge and National Curriculum Objectives Addressed in Speaking Tests</i>	117
Table 18. <i>Fligner-Killeen and Fligner-Policello Results by School</i>	145
Table 19. <i>Test Quality Information for All Vocabulary Tests</i>	146
Table 20. <i>Descriptive Statistics: Vocabulary Recognition in the Oral Modality</i>	147
Table 21. <i>Descriptive Statistics for Vocabulary Recognition in the Written Modality Tests</i> ..	149
Table 22. <i>Descriptive Statistics for Vocabulary Recall in the Written Modality Tests</i>	153
Table 23. <i>Descriptive Statistics for Vocabulary Recall Tests in the Oral Modality</i>	157
Table 24. <i>Overview of Vocabulary Outcomes</i>	169
Table 25. <i>Test Quality Information for Both Phonics Tests</i>	170

Table 26. <i>Descriptive Statistics for Phonics Transcription Tests</i>	171
Table 27. <i>Descriptive Statistics for Phonics Read Aloud Tests</i>	173
Table 28. <i>Overview of Phonics Outcomes</i>	180
Table 29. <i>Test Quality Information for All Grammar Tests</i>	180
Table 30. <i>Descriptive Statistics for Written Grammar Recognition Tests</i>	181
Table 31. <i>Descriptive Statistics for Spoken Grammar Recall Tests</i>	184
Table 32. <i>Descriptive Statistics for Elicited Imitation Grammar Tests</i>	187
Table 33. <i>Overview of Grammar Outcomes</i>	198
Table 34. <i>Descriptive Statistics for the Picture Response Task</i>	199
Table 35. <i>Test Quality Information for Elicited Imitation Speaking</i>	202
Table 36. <i>Descriptive Statistics for Elicited Imitation Speaking Tests</i>	202
Table 37. <i>Total Number of Utterances Containing Recognisable Verbs or Verbs in the Infinitive in Picture Response</i>	209
Table 38. <i>Overview of Speaking Outcomes</i>	213
Table 39. <i>Quality of Measure Information for Self-efficacy Surveys</i>	215
Table 40. <i>Descriptive Statistics and Effect Sizes for Self-efficacy Questions</i>	216
Table 41. <i>Quality of Measure Information for Enjoyment Surveys</i>	236
Table 42. <i>Pupils' Positive Anticipation Baseline by Approach</i>	237
Table 43. <i>Descriptive Statistics and Effect Sizes for Enjoyment Ratings (Pupil Questionnaires)</i>	238
Table 44. <i>Descriptive Statistics for Pupil Perceptions of Teacher Enjoyment Question1 and Question2</i>	260
Table 45. <i>Comparison of Pupil Perceptions of Teacher Enjoyment at Pupil Questionnaire 2 and Pupil Questionnaire 3</i>	261
Table 46. <i>Between-Group Overview of Trends in Self-efficacy and Enjoyment</i>	264
Table 47. <i>Teacher Starting Points (Teacher Questionnaire 1)</i>	270
Table 48. <i>Descriptive Statistics for Individual Items in Teachers' Post-lesson Evaluations</i> ..	279
Table 49. <i>Exploratory Factor Analysis of the Teacher Post-lesson Evaluation Items</i>	280
Table 50. <i>Teacher Perceptions of Pupil Outcomes and Linguistic Usefulness (Post-lesson Evaluations)</i>	282
Table 51. <i>Classroom Enjoyment: Teacher Enjoyment and Perceptions of Pupil Motivation (Post-lesson Evaluations)</i>	283
Table 52. <i>Individual Teachers' Tension Ratings in Post-lesson Evaluations</i>	285
Table 53. <i>Overarching Summary of Post-lesson Evaluations: Teacher Perceptions of Pupil Outcomes and Linguistic Usefulness</i>	287
Table 54. <i>Overarching Summary of Post-lesson Evaluations: Classroom Enjoyment</i>	288

Table 55. <i>Overarching Summary of Post-lesson Evaluations: Teacher Tension</i>	290
Table 56. <i>All Teachers' Enjoyment: Mean and Median Ratings at Teacher Questionnaire 1, Teacher Questionnaire 2 and Teacher Questionnaire 3</i>	292
Table 57. <i>All Teachers' Tension: Mean and Median Ratings at Teacher Questionnaire 1, Teacher Questionnaire 2 and Teacher Questionnaire 3</i>	294
Table 58. <i>Summary of the Descriptive Impact of Approach on the Achievement of Pupils with Higher and Lower Language-analytic Ability Scores in Different Test Types</i>	349
Table 59. <i>Elicited Imitation Grammar (Rescored Model) Test Quality</i>	402
Table 60. <i>Descriptive Statistics for Rescored Elicited Imitation Grammar Task</i>	402
Table 61. <i>Assumptions for the Regression Model</i>	404
Table 62. <i>Individual Teachers' Perceptions of Pupil Outcomes and Linguistic Usefulness Ratings in Post-lesson Evaluations</i>	690
Table 63. <i>Individual Teachers' Perceptions of Classroom Enjoyment Ratings in Post-lesson Evaluations</i>	693

List of Figures

Figure 1. <i>Reciprocal Effects Linking Student and Teacher Enjoyment</i>	86
Figure 2. <i>Notched Box Plot Showing Scores in Vocabulary Recognition in the Oral Modality (Listening) and Written Modality (Reading) Tests at Post-test and Delayed Post-test by Approach</i>	147
Figure 3. <i>Scatter Plot Showing Interactions Between Language-analytic Ability, Approach and Time-point for Vocabulary Recognition in the Oral Modality</i>	151
Figure 4. <i>Scatter Plot Showing Interactions Between Language-analytic Ability, Approach and Time-point for Vocabulary Recognition in the Written Modality</i>	151
Figure 5. <i>Predictions by Language-analytic Ability Score for Each Modality of Vocabulary Recognition</i>	152
Figure 6. <i>Notched Box Plot Showing Vocabulary Recall Scores in the Written Modality by Time-point</i>	154
Figure 7. <i>Vocabulary Recall in the Written Modality Predictions by Language-analytic Ability Score, Approach and Time-point</i>	156
Figure 8. <i>Notched Box Plot Showing Vocabulary Recall in the Oral Modality Scores by Time-point</i>	158
Figure 9. <i>Vocabulary Recall in the Oral Modality Predictions by Language-analytic Ability Score, Approach and Time-point</i>	159
Figure 10. <i>Individual Item Scores in Vocabulary Recognition in the Oral Modality (Post-test and Delayed Post-test)</i>	164
Figure 11. <i>Individual Item Scores in Vocabulary Recognition in the Written Modality (Post-test and Delayed Post-test)</i>	165
Figure 12. <i>Individual Item Scores in Vocabulary Recall in the Oral Modality Tests (Post-test and Delayed Post-test)</i>	167
Figure 13. <i>Individual Item Scores in Vocabulary Recall in the Written Modality Tests (Post-test and Delayed Post-test)</i>	168
Figure 14. <i>Notched Box Plot Showing Phonics Transcription Scores by Time-point</i>	172
Figure 15. <i>Scatter Plot Showing Phonics Transcription Scores According to Language-analytic Ability Score by Approach and Time-point</i>	172
Figure 16. <i>Notched Box Plot Showing Phonics Read Aloud Scores by Time-point</i>	174
Figure 17. <i>Phonics Read Aloud Predictions by Language-analytic Ability Score, Approach and Time-point</i>	176
Figure 18. <i>Scores in Individual Symbol-Sound Correspondences in Transcription Tests (Post-test and Delayed Post-test)</i>	178

Figure 19. <i>Scores in Individual Symbol-Sound Correspondences in Read Aloud Tests (Post-test and Delayed Post-test)</i>	179
Figure 20. <i>Written Grammar Predictions by Language-analytic Ability Score, Approach and Time-point</i>	183
Figure 21. <i>Notched Box Plot Showing Spoken Grammar Recall Scores by Time-point</i>	185
Figure 22. <i>Spoken Grammar Recall Predictions by Language-analytic Ability Score, Approach and Time-point</i>	186
Figure 23. <i>Notched Box Plot Showing Elicited Imitation Grammar Scores by Time-point</i>	188
Figure 24. <i>Elicited Imitation Grammar Predictions by Language-analytic Ability Score, Approach and Time-Point</i>	190
Figure 25. <i>Notched Box Plot Showing Written Grammar Recognition Scores by Category by Time-point</i>	194
Figure 26. <i>Participants' Scores in Individual Spoken Grammar Recall Categories</i>	195
Figure 27. <i>Participant Scores in Individual Categories in Elicited Imitation Grammar</i>	195
Figure 28. <i>Notched Box plot Showing Open-ended Picture Response Scores by Time-point</i> .	200
Figure 29. <i>Notched Box Plot Showing Elicited Imitation Speaking Scores by Time-point</i>	203
Figure 30. <i>Scatter Plot Showing Elicited Imitation Speaking Scores According to Language-analytic Ability Score by Time-point</i>	205
Figure 31. <i>Elicited Imitation Speaking Predictions by Language-analytic Ability Score, Approach and Time-point</i>	206
Figure 32. <i>Participant Scores in Elicited Imitation Speaking Tests at Post-test and Delayed Post-test</i>	211
Figure 33. <i>Box Plot Showing Responses to Individual Items in Self-efficacy Questions by Approach</i>	217
Figure 34. <i>Impact of Approach on Confidence at Mid-point from Pupil Group Interview Data: Novel</i>	221
Figure 35. <i>Impact of Approach on Confidence at Mid-point from Pupil Group Interview Data: Standard</i>	222
Figure 36. <i>Impact of Approach on Confidence at End-point from Pupil Group Interview Data: Novel</i>	223
Figure 37. <i>Impact of Approach on Confidence at End-point from Pupil Group Interview Data: Standard</i>	224
Figure 38. <i>Pupils' Perceptions of What They Will Be Able to Do in French in Year 4 (Mid-point Pupil Group Interviews)</i>	225
Figure 39. <i>Pupils' Perceptions of What They Will Be Able to Do in French in Year 4 (End-point Pupil Group Interviews)</i>	226

Figure 40. <i>Pupil Perception of Progress: Mid-point Pupil Group Interviews</i>	227
Figure 41. <i>Pupil Perception of Progress: End-point Pupil Group Interviews</i>	228
Figure 42. <i>Range of Different Words and Phrases Volunteered (Unprompted) in Each Approach: Mid-point Pupil Group Interviews</i>	229
Figure 43. <i>Range of Different Words and Phrases Volunteered (Unprompted) in Each Approach: End-point Pupil Group Interviews</i>	230
Figure 44. <i>Notched Box Plot Showing Pupil Enjoyment Ratings from Pupil Questionnaire Data by Approach and Time-point</i>	238
Figure 45. <i>Individual Enjoyment Question Analysis by Approach and Time-point</i>	239
Figure 46. <i>Point Range Plot Showing Predicted Probability of Enjoyment Ratings on Any Item from Pupil Questionnaires by Baseline Positive Anticipation Rating, Approach and Time-point</i>	241
Figure 47. <i>Pupils' Reasons for Liking French: (Mid-point Pupil Group Interview)</i>	244
Figure 48. <i>Pupils' Reasons for Liking French: (End-point Pupil Group Interview)</i>	247
Figure 49. <i>Negative Impacts on Pupil Motivation: (Mid-point – Pupil Group Interview 1)...</i>	249
Figure 50. <i>Negative Impacts on Pupil Motivation: (End-Point – Pupil Group Interview 2) ..</i>	251
Figure 51. <i>Negative Impacts on Pupil Motivation: (End-Point – Pupil Questionnaire 3)</i>	252
Figure 52. <i>Self-efficacy and Enjoyment: Overview of Mean Ratings over Time</i>	263
Figure 53. <i>Teacher-identified Issues Arising from Teaching Schemes (Interview2)</i>	267
Figure 54. <i>Teacher-identified Issues Arising from Teaching Schemes (Interview3)</i>	269
Figure 55. <i>Teachers' Pre-project Positive Emotions and Their Causes (Interview1)</i>	273
Figure 56. <i>Teachers' Pre-project Negative Emotions and Their Causes (Interview1)</i>	274
Figure 57. <i>Factor Analysis of Teachers' Post-lesson Evaluations</i>	281
Figure 58. <i>Individual Teacher Ratings for Tension Construct by Lesson (Post-lesson Evaluations)</i>	284
Figure 59. <i>Individual Teacher Enjoyment: Mean Ratings at Teacher Questionnaire 2 and Teacher Questionnaire 3</i>	293
Figure 60. <i>Teacher Tension: Mean Ratings at Teacher Questionnaire 2 and Teacher Questionnaire 3</i>	294
Figure 61. <i>Infographic Summarising Teachers' Positive (+) and Negative (-) Comments about Pupil Progress in Each Scheme (Interview2)</i>	298
Figure 62. <i>Infographic Summarising Teachers' Positive (+) and Negative (-) Comments about Pupil Progress in Each Scheme (Interview3)</i>	299
Figure 63. <i>Teacher Perceptions of Usefulness of French Taught (Interview1/2/3)</i>	300
Figure 64. <i>Time-ordered Meta-matrix Summarising Individual Teacher Perceptions of the Usefulness of French Taught (Interview1, Interview2, Interview3)</i>	301

Figure 65. <i>Main Sources of Teacher Enjoyment (Interview2)</i>	304
Figure 66. <i>Main Sources of Teacher Enjoyment (Interview3)</i>	305
Figure 67. <i>Teachers' Perceptions of Sources of Pupil Enjoyment (Interview2)</i>	307
Figure 68. <i>Teachers' Perceptions of Sources of Pupil Enjoyment (Interview3)</i>	307
Figure 69. <i>Teachers' Preferences for Schemes (Interview2)</i>	308
Figure 70. <i>Teachers' Preferences for Schemes (Interview3)</i>	309
Figure 71. <i>Teachers' Perceptions of the Positive Impact of Each Approach on Subject Knowledge (Interview2)</i>	310
Figure 72. <i>Teachers' Perceptions of the Positive Impact of Each Approach on Subject Knowledge (Interview3)</i>	311
Figure 73. <i>Causes of Teachers' Feelings of Worry (Interview2)</i>	312
Figure 74. <i>Causes of Teachers' Feelings of Worry (Interview3)</i>	314
Figure 75. <i>Teachers' Confidence According to Approach (Interview2)</i>	315
Figure 76. <i>Teachers' Feelings of Confidence Completing Certain Tasks in Each Approach (Interview2)</i>	316
Figure 77. <i>Teachers' Confidence According to Approach (Interview3)</i>	318
Figure 78. <i>Teachers' Feelings of Confidence Completing Certain Tasks in Each Approach (Interview3)</i>	320
Figure 79. <i>Teachers' (Non-approach-specific) Negative Emotions Associated with Pronunciation (Interview2)</i>	322
Figure 80. <i>Teachers' (Non-approach-specific) Negative Emotions Associated with Pronunciation (Interview3)</i>	323
Figure 81. <i>Teachers Perceptions of Support for Pronunciation and Symbol-Sound Correspondences in Each Approach (Interview2)</i>	323
Figure 82. <i>Teachers' Perceptions of Support for Pronunciation and Symbol-Sound Correspondences in Each Approach (Interview3)</i>	324
Figure 83. <i>Notched Box Plot Showing Elicited Imitation Grammar Scores by Time-point</i>	403
Figure 84. <i>Summary of Model (Elicited Imitation Grammar Rescored Model)</i>	404
Figure 85. <i>Elicited Imitation Grammar (Rescored Model) Predictions by Language-analytic Ability Score, Approach and Time-point</i>	406
Figure 86. <i>Individual Teacher Ratings for Perceptions of Pupil Outcomes and Usefulness by Lesson (Post-lesson Evaluations)</i>	688
Figure 87. <i>Individual Teacher Ratings for Enjoyment Construct by Lesson (Post-lesson Evaluations)</i>	691

Acknowledgements

I would like to thank my supervisor, Professor Emma Marsden, for your unwavering support and enthusiasm for this project from its inception. Emma, you are a model of integrity and professionalism and your support and knowledge are second to none. I could not have asked for a better supervisor: you knew exactly when to encourage and when to challenge, and were adept at offering support when needed. A very heartfelt thank you indeed.

My thanks too to Professor Rob Klassen for your input into pupil self-efficacy and enjoyment measures, as well as invaluable suggestions for monitoring teachers' fidelity to treatment. Thank you, Dr Zoe Handley, for stepping in and taking over the role of Thesis Advisory Panel supervisor in the latter stages of this PhD.

I wish to thank Dr Amber Dudley: you so generously offered your time, expertise and support; your work on the mixed-effects models made a more sophisticated piece of statistical analysis possible. Dr Jess Brown, I hope you know how much I have appreciated your patience and support which were crucial in helping me develop my statistical knowledge and understanding – thank you. I also wish to thank Professor Bev Hale for your early support and encouragement in my statistics journey.

The pupils and teachers who agreed to work with me have been central to this project – thank you for welcoming me into your schools so warmly and for making me remember how good it is to be in the school environment again, and how wonderfully insightful and refreshing young learners' comments can be. Thank you too to all the schools and colleagues who supported me through taking part in the pilot, enabling this project to get off to a smooth start.

Thanks to my colleagues, family and friends who have supported me throughout this project: every expression of interest has been welcome and appreciated. A special thank you to Gill Newsom and Graham, for your support with interrater reliability. And to Graham, Leah, and Alex, thank you for encouraging me every step of the way, accepting my absences and my preoccupations over the last few years. Thank you for believing in me.

Author's declaration

This thesis is a presentation of original work. All data were collected and analysed by the author, Elaine Minett. The regression models were initially created by Dr Amber Dudley, who also offered support in preparing part of the learner data for the regression analyses. Additional support in checking code and developing statistical understanding was provided by Dr Jess Brown, the departmental research support assistant. A single proof-read was provided by Graham Minett to check for, but not correct, typographical errors. This work has not previously been presented for a degree or other qualification at this University or elsewhere. All sources are acknowledged as references.

Chapter 1: Introduction

This chapter contextualises current foreign language (FL) provision in English primary schools, providing a broad rationale for research of the kind adopted in this study. Explaining non-specialist delivery and curriculum expectations, it addresses key challenges in ensuring quality provision and relevant factors influencing pupil outcomes.

1.1 Foreign languages in English primary schools: background

As The Nuffield Foundation (2000) called for action to build on public support for early language learning in England, European countries were encouraged to introduce FL teaching from an early age (European Commission, 2018; European Council Barcelona, 2002). The resulting decade of government backing led to FLs becoming a statutory part of the primary National Curriculum in 2014 (Department for Education - DfE, 2013b). By 2018, 75% of schools offered a FL in the Key Stage 2 (7-11) curriculum (Tinsley & Doležal, 2018), rising to 99% in 2023-24 (Collen & Duff, 2024), although schools most likely to respond to such national survey requests are those with better provision (response bias). Indeed, over half of responding schools in 2024 were from less deprived areas of England (Collen & Duff, 2024). Furthermore, ‘provision’ varies widely (Kaspruwicz, 2025) without guidance ensuring minimum teaching time and quality of input. Statutory status alone cannot guarantee successful provision (Ayres-Bennett & Carruthers, 2019). Thus, FL provision, once “patchy and uncoordinated”, “haphazard” and “piecemeal” (The Nuffield Foundation, 2000:39), is still characterised by inconsistency and disparities in teaching quality and time allocation (Ayres-Bennett & Carruthers, 2019; Myles & Mitchell, 2012; Tinsley & Board, 2017).

1.2 Non-specialist provision

National surveys in England evidenced the dearth of suitably qualified teachers, jeopardising successful delivery of FL initiatives (Board & Tinsley, 2014; Driscoll, Jones & Macrory, 2004; Tinsley & Doležal, 2018) for which teacher supply and quality are crucial (McLachlan, 2009; Murphy, 2014; Nikolov & Mihaljević Djigunović, 2006). Class teachers, usually non-native speakers with FL levels below native-like proficiency (Nikolov, 2009), are qualified to teach all curriculum subjects regardless of training (Enever, 2011). Responsible for pupil progress across all areas, they usually see the same children every day; they may be learning the FL themselves (noted by Myles, 2017). This model encompasses generalists (with limited linguistic competence and confidence) and those with existing FL experience or training to teach FL (Department for Children, Schools and Families - DCSF, 2005; OFSTED, 2005; Wade et al., 2009).

Macrory and McLachlan (2009) investigated undergraduate trainee teachers' responses to university-based primary FL training: 87% (non-specialists) preferred generalist over specialist delivery, regarding class teachers as better placed to teach FL. The generalist model has grown in popularity (Tinsley, 2019), recently accounting for provision in two-thirds of schools (Collen & Duff, 2024) or just under (Collen & Duff, 2025).

1.2.1 Class teacher foreign language qualifications

In a recent national survey (Collen, 2020), the highest FL qualification for 38% of respondents was General Certificate of Secondary Education (GCSE). Given possible response bias, the number of underqualified staff is potentially higher. Minimum requirements for primary FL teaching are not stipulated, although studies here and abroad (Butler, 2025; Nikolov & Mihaljević Djigunović, 2006; Unsworth et al., 2015; Woodgate-Jones, 2008) seem to favour A-level, or B2 (Common European Framework of Reference for Languages, CEFR, Council of Europe, 2001). However, Enever (2011) concluded from the *Early language learning in Europe* (ELLiE) project that higher fluency levels were needed, preferably C2. Insufficient linguistic competence remains concerning: 47% of national survey respondents in 2021 identified language proficiency as a major challenge in meeting National Curriculum (DfE, 2013b) requirements, rising to 52% in 2022 (ranked the second biggest factor, after insufficient curriculum time) (Collen, 2022).

Policy initiatives may be implemented too quickly to equip staff with necessary skills and capabilities (Nikolov & Mihaljević Djigunović, 2011); teachers may lack time to learn the language (Cable et al., 2010; Driscoll, Jones & Macrory et al., 2004). Ongoing calls for continuing professional development (CPD) suggest that pedagogical and linguistic training has been inadequate (Holmes & Myles, 2019; Johnstone, 2002; Tinsley, 2019; Tinsley & Board, 2012, 2017; Wade et al., 2009).

1.2.2 Subject knowledge for teaching primary foreign languages

Shulman (1986) argued the importance of teachers developing content *and* pedagogical knowledge, but in the field of primary FL in England, opinion is mixed. Research in Scotland (Low et al., 1995, cited in Cable et al., 2010; Driscoll, Jones, Martin, et al., 2004) recommended basic competence in key areas rather than high degrees of linguistic proficiency. However, subject knowledge can affect teacher confidence (reported by Barnes, 2006; Finch et al., 2018; Woodgate-Jones, 2008 - see [1.4.4.3](#)).

The extent of knowledge necessary to enable pupil progress in FLs remains unclear. Given that English provision prioritises language competence (DfE, 2013b), teachers may need considerable support. Martin (2000) suggested that defining the body of language to be covered

in Key Stage 2 could facilitate transition into secondary school (see also Marsden et al., 2023); it could also inform CPD and initial teacher education (ITE). Indeed, respondents to recent national surveys would appreciate such guidance (80% in 2024 - Collen & Duff, 2024; 78% in 2025 – Collen & Duff, 2025), and this has emerged as a recommendation from the recent *Curriculum and Assessment Review* (DfE, 2025).

1.2.3 The advantages of class teacher provision

This section addresses advantages; disadvantages are discussed in [1.4.4](#).

1.2.3.1 Links with first language literacy

Muñoz (2014a) recommended training FL teachers to develop pupils’ grapho-morphological awareness. In the policy document *Modern Foreign Languages Pedagogy Review*, Teaching Schools Council (TSC) (2016) also advised teaching symbol-sound correspondences (SSCs), with which primary teachers are familiar. As regular teachers of literacy, they can make effective links with pupils’ first language (henceforth, L1) (noted by Driscoll, Jones, Martin, et al., 2004; OFSTED, 2005, 2011) and potentially apply this existing pedagogical understanding to FL teaching. However, McLachlan (2009) suggested that teachers often fail to make such connections.

1.2.3.2 The teacher’s importance to primary-aged children

Evaluating the ELLiE project, Enever (2011:25) highlighted the “central role” of teachers for younger pupils, facilitating their adjustment to a more formal learning environment. In a Croatian experiment introducing English as a FL (EFL) to primary-aged learners in the 1990s, Mihaljević Djigunović and Vilke (2000) speculated that the strong emotional attachment that 6-7-year-olds form with teachers can influence their FL attitudes (highlighting the importance of teacher positivity). Nikolov and Mihaljević Djigunović (2006) reviewed primary FL research, concluding that, whereas older learners were more instrumentally motivated, younger pupils were influenced by teacher-pupil relationships.

In Muñoz’s (2014a) questionnaire-/interview-study, probing young pupils’ reflections on FL learning, 8-9-year-old learners mentioned teacher support or influence more regularly than 11-12-year-olds. Teachers appeared to affect how young learners perceived their success, whereas older learners drew motivation from the outside world more frequently or were more aware of the importance of learning English (see also Tragant, 2006).

1.3 Young learners and foreign language learning

Nikolov and Mihaljević Djigunović (2011) defined young learners as typically aged 7-12. In the English primary school context, they are 7-11. It is widely accepted that children learn languages more easily than adolescents or adults (The Nuffield Foundation, 2000), although evidence fails to support this assertion in classroom contexts (Baïdak et al., 2017; Baumert et al., 2020; Enever, 2011; Marinova-Todd et al., 2000); Nikolov and Mihaljević Djigunović (2006:243) dismissed it as “folk wisdom”, neglecting important aspects of provision such as quality or quantity of input.

Many researchers (e.g., Enever, 2011; Larson-Hall, 2008; Muñoz, 2006; Myles, 2017) clarify that this ‘younger is better’ assumption stems from experiments in immersion or naturalistic contexts offering vast amounts of FL exposure. This contrasts limited-time classroom input, the instructed context (Cenoz, 2003; Muñoz, 2003; Szpotowicz & Lindgren, 2011), possibly characterised by “less social urgency” than immersion settings (Hilton, 2017:148).

In research reviews, Edelenbos et al. (2006) and Tinsley and Comfort (2012) acknowledged benefits of early FL learning but cautioned that these depend on high-quality teaching and secure, year-on-year continuity and progression.

1.3.1 National Curriculum expectations

The Programme of Study (DfE, 2013a) promotes a language competence model: teaching a single language (Driscoll, Jones & Macrory, 2004) with ambitious linguistic outcomes related to practical communication and language manipulation. The National Curriculum (DfE, 2013b) states that primary languages should underpin later language study, but engendering a passion for FLs or enthusing pupils to want to continue studying is not mentioned, contrasting the regular references to excitement, enjoyment and enrichment in the *Key Stage 2 Framework for Languages* (DCSF, 2005) and the aims promoted by teachers involved in the longitudinal study of FL provision conducted by Cable et al. (2010). For many of these teachers, enthusing learners was considered more important than linguistic development or intercultural awareness.

Content (e.g., vocabulary, grammar, specific language domains of use, or functions/notions to be learned) is not prescribed. The aims promote pupils’ ability to ‘understand and respond to spoken and written language’, ‘speak with increasing confidence, fluency and spontaneity’ and ‘write at varying length’ (DfE, 2013b:193). Linguistic components are unsurprising, but the ambitious expectations in receptive *and* productive skills requiring “substantial progress” in a single language (DfE, 2013b:194) are challenging with limited time allocated. The document has been criticised for insufficient detail and support in assessment, content and expectations (Ayres-Bennett & Carruthers, 2019; Holmes & Myles, 2019; Tinsley

& Board, 2017), making effectiveness difficult to define. Additionally, evidence of primary-aged learners' likely FL outcomes is scant (Mitchell & Myles, 2019).

Some expectations may be more challenging than others. Earlier national surveys and studies indicated (positive) teacher perceptions of young learners' progress in listening skills (Bolster et al., 2004; Cable et al., 2010; Powell et al., 2001), whereas later survey findings suggested that other outcomes were likely unattainable for most pupils finishing Year 6 (Tinsley & Board, 2016). In 2016-17, some schools had only taught FL since 2014, but as noted earlier, responding schools likely had most established provision: only 58% (Tinsley & Board, 2017) indicated that most/all pupils would 'Understand basic grammar, key features and patterns of the language and how to apply these' (DfE, 2013b:194). Similarly, Mitchell and Myles's (2019) study confirmed that, after 38 hours' instruction, pupils had developed their receptive vocabulary and used high-frequency formulaic language but could not produce verbs and verb inflections independently. This picture echoes findings from OFSTED (2008a, 2011) and Cable et al. (2010): poor reading and writing outcomes; limited understanding of how languages work.

Lower outcomes do not render primary FL worthless: children *do* make progress, even with limited time and exposure (Baumert et al., 2020; Kasprowicz, 2025; Unsworth et al., 2015), but expectations must be realistic. Further research could determine realistic pupil outcomes in such contexts (Mitchell & Myles, 2019).

1.3.2 Age-appropriate approaches

Although the National Curriculum (DfE, 2013b) offers no guidance on pedagogical approaches, its aims and objectives are not inconsistent with those presented by DCSF (2005); this did not recommend a single approach but suggested "broad agreement" based on a "distinctly primary approach" (p.25, Part 2). Tinsley and Comfort's (2012) research review supported this, prioritising language for interactive, real-life purposes (see also Dearing & King, 2007).

Class teachers, trained in primary pedagogy (see [1.4.4.2](#)), may adopt age-appropriate approaches in FL lessons, an advantage noted in inspections and surveys of primary provision (Cable et al., 2010; Martin, 2000; Muijs et al., 2005; OFSTED, 2011), and considered a strength of the non-specialist model in Scotland's primary languages initiative (Crichton & Templeton, 2010, an opinion piece on the Modern Languages in Primary School project - MLPS). Nguyen (2018), researching age-appropriate pedagogy for young Vietnamese EFL learners, advocated story-telling, songs and role-plays, given pupils' social, emotional, physical and cognitive developmental stage.

Successful provision involves exposing motivated learners to rich FL input (Dearing & King, 2007). Holmes & Myles (2019) and Butler (2025) reiterated the need identified by Muijs et al. (2005) and Martin (2012) for more challenging activities which move pupils beyond fun

and games, engendering long-term motivation and engagement. [Chapter 2](#) discusses pedagogical approaches.

1.4 Factors influencing outcomes

Understanding factors influencing FL learning processes, particularly significant in instructed contexts (Muñoz, 2014b), should enable quality provision (Driscoll, Jones, Martin, et al., 2004). Aims aside, pupils' *actual* progress will always be constrained by learning contexts and conditions (Holmes & Myles, 2019). Varied provision results in varied outcomes (Muijs et al., 2005; Tinsley & Board, 2012), with less satisfactory provision more likely in schools in challenging circumstances (i.e., with higher levels of Free School Meals and lower overall attainment, Board & Tinsley, 2015; Tinsley & Board, 2017). This section considers relevant macro- and micro-level factors (Mihaljević Djigunović, 2017) influencing outcomes.

1.4.1 Out-of-school exposure

Frequent out-of-school FL exposure, inextricably linked to its societal status, is deemed influential in early language learning success (Araújo & Dinis da Costa, 2013; Edelenbos & Kubanek, 2009; Jones et al., 2012). It merits close attention as, in time-limited instructional settings, any additional FL contact could be beneficial. Although Unsworth et al. (2015) did not find that out-of-school exposure significantly predicted pupil outcomes, Szpotowicz and Lindgren (2011) noted positive correlation between high listening scores and contexts where FLs were used frequently in society. In England, FLs are encountered relatively infrequently outside school (Enever, 2009). In the ELLiE project (Lindgren & Muñoz, 2013; Muñoz & Lindgren, 2011), pupils' listening and reading outcomes in Croatia, Sweden, the Netherlands and Poland were positively influenced by, on average, 5 or more hours' weekly out-of-school FL exposure, through television, films, video games, music, books and other reading types (reported by parents). Pupils in England experienced under 3 hours per week of exposure to French or Spanish outside lessons. Out-of-school exposure to English also explained Belgian learners' higher vocabulary knowledge scores in English (FL2) than French (FL1), despite learning French for 3 years longer (E. Peters et al., 2019); similarly, Puimège and Peters (2019) found that watching television and playing videogames predicted vocabulary knowledge.

Such rich input may support younger pupils to develop implicit learning capabilities (Graham et al., 2017), providing immersive opportunities unavailable in typical English primary classrooms (Holmes & Myles, 2019). Such contexts are unlikely to provide the multiple exposures to linguistic features required for pupils to use implicit learning mechanisms (Muñoz, 2006), particularly compared to the consistent exposure and interaction characterising L1

acquisition (Jaekel et al., 2017). This highlights the advisability of prioritising effective use of available time and tempering expectations in terms of potential outcomes.

1.4.2 Quantity of input

Teaching time matters, as illustrated by Graham et al. (2017), which showed that amount of instruction predicted scores on grammar and vocabulary, and Kasprowicz (2025), reporting preliminary findings from the Progress in Primary Languages project, which found that 7-11-year-old learners made significantly more progress in linguistic outcomes when taught for 50-55 minutes per week than pupils receiving shorter inputs. Similarly, Unsworth et al. (2015) found that learning English for more than 60 minutes per week conferred advantages in receptive grammar and vocabulary tests. Pupils receiving under 60 minutes did not show any advantage in vocabulary over the control group (no input), although pupils' out-of-school FL exposure in the Netherlands is significantly higher than in England (1.4.1). Less than 60 minutes possibly conveyed no observable benefit above and beyond that of out-of-school exposure alone. Johnstone (2002), considering implications for language policy, concluded that sufficient time matters in all contexts, but particularly where out-of-school exposure is limited.

There is no recommended amount of time for FLs in the National Curriculum (DfE, 2013b). Provision in 81% of responding schools reporting a set time for language teaching (likely those with more established provision) was 30-60 minutes (Tinsley, 2019). A year later, 50% of respondents reported provision of under 45 minutes; 30% reported 45-60 minutes weekly (Collen, 2020). In an average 24-hour school week, this represents 2-4% of available curriculum time, below the average of 6% and one of the lowest allocations of FL learning in Europe (OECD, 2018). In a recent national survey of FL provision in England (Collen, 2022), respondents identified insufficient curriculum time as a challenge to implementing National Curriculum requirements (51% in 2021; 61% in 2022, the highest-ranking barrier) – see also Collen and Duff (2025). A weekly minimum of 60 minutes, with a non-statutory minimum of 140 hours across Key Stage 2 is recommended (Ayres-Bennett & Carruthers, 2019; Holmes & Myles, 2019).

Time pressure affects content and outcomes, necessitating decisions about what can realistically be covered. Supporting the importance of frequency of contact with the FL, OFSTED (2011) noted that follow-up input additional to discrete lessons improved pupils' progress. However, there has been little research into how approaches affect outcomes in time-limited settings.

1.4.3 Individual pupil factors

This section briefly acknowledges that learner characteristics and individual responses to FLs, discussed thoroughly in [Chapters 2](#) and [3](#), can affect outcomes. Motivation and attitudes clearly contribute (Edelenbos et al., 2006; Muñoz, 2014b), with individual differences apparent from the onset of language learning (Enever, 2011). Language aptitude is also instrumental: Unsworth et al. (2015) found correlations between phonological short-term memory and receptive vocabulary outcomes.

1.4.4 The drawbacks of class teacher provision

As the non-specialist class teacher model is integral to provision in England, awareness of its potential pitfalls is crucial. Quality of input may impact on pupil outcomes. For many researchers (e.g., Jaekel et al., 2017; Lopriore & Krikhaar, 2011; Marinova-Todd et al., 2000; Nikolov & Mihaljević Djigunović, 2006; Šamo, 2009), (lack of) teacher proficiency is one of the most important factors in pupils' learning success. The impact is apparent in many ways (see [1.4.4.4](#)). Unsworth et al. (2015) concluded that teachers influenced pupil outcomes more than quantity of time, with those pupils taught only by non-native speakers achieving the lowest results and making least progress in vocabulary knowledge.

Shulman (1987), considering how teachers promote comprehension, defined categories of knowledge: most relevant here are content knowledge, general pedagogical knowledge and pedagogical content knowledge. These are now explored with reference to non-specialist teachers of primary FL.

1.4.4.1 Linguistic competence: content knowledge

Various surveys and studies document potential issues if teachers possess insufficient linguistic proficiency, particularly when language competence models are promoted (found by Muijs et al., 2005). Issues in the following areas are summarised here:

- oral fluency, pronunciation and intonation (reported in headteacher interviews, Cable et al., 2010, DCSF, 2005; observed by Moon, 2009, Muijs et al., 2005, OFSTED, 2011);
- grammatical accuracy (Cable et al., 2010; Muijs et al., 2005);
- promoting and sustaining pupil FL use (Moon, 2009; OFSTED, 2005);
- difficulty ensuring progress beyond basic language; lack of longer-term vision (reported by Driscoll, 1999, from ethnographic study data);
- teaching language chunks rather than grammatical and syntactical understanding (Nguyen, 2018, from classroom observations and interviews; also reported by Cable et al., 2010:6, who observed "an emphasis on developing children's ability to produce

memorised language items and formulaic phrases, rather than creating their own independent sentences”, albeit with some evidence of Year 6 pupils beginning to use language more creatively);

- appropriate pupil support and challenge (Cable et al., 2010; Driscoll, 1999; Muijs et al., 2005);
- decreased error correction (Cable et al., 2010; Muijs et al., 2005).

Encouragingly, Rowe et al. (2012) found that non-specialist post-graduate trainee teachers felt capable of providing effective FL models for pupils, after guidance.

1.4.4.2 Subject-specific teaching approaches: pedagogical content knowledge

Class teachers may understand child development (part of general pedagogical knowledge), but lack appropriate FL-specific pedagogical knowledge, considering these approaches different from other subject areas, even “daunting” (Finch et al., 2018:10). Documented as a general concern by Cable et al. (2010), specific examples when compared with specialist teacher provision include:

- decreased assessment (Driscoll, 1999; Muijs et al., 2005);
- focusing on oracy rather than literacy (found by Cable et al., 2010, and Roothoof's 2017 study of teacher beliefs; observed by Driscoll, 1999 [ethnographic study]; acknowledged by Porter & Graham, 2022);
- overuse of games; insufficient pupil participation (Muijs et al., 2005).

Teachers should be able to communicate the essence of a subject, focus on key knowledge and understanding, and determine what can be disregarded (Shulman, 1987), potentially challenging for non-specialists.

1.4.4.3 Teacher confidence

Teacher motivation and confidence are addressed in [Chapter 3](#), but it is noted here that lack of confidence is repeatedly identified as significantly concerning in non-specialist provision (Board & Tinsley, 2014; Cable et al, 2010; Collen, 2023; Collen & Duff, 2024, 2025; Tinsley, 2019; Tinsley & Board, 2012, 2016, 2017; Wade et al., 2009). It may derive from insufficient linguistic proficiency, affecting lesson pace and motivation of staff (all found by Muijs et al., 2005) and pupils (Barton et al., 2009; Crichton & Templeton, 2010), and sometimes engendering reluctance in teachers (Tinsley & Board, 2016).

1.4.4.4 Negative impacts on pupil outcomes

This section outlines possible negative impacts on pupil outcomes arising from the challenges summarised above. It presents empirical evidence focusing on input quality in four areas: pupils' achievements in writing, speaking, and vocabulary, and lesson challenge.

(i) Pupils' performance in writing

In national surveys (Board & Tinsley, 2014, 2015; Tinsley & Board, 2012), the development of pupils' grammatical understanding and the teaching of reading and writing were repeatedly identified as concerning by teachers. In support of this (although not specifically related to the impact of non-specialist teaching), Bolster et al. (2004:38) reported secondary teachers' comments that pupils who had learnt FL at primary school "could be over-confident and occasionally lack written accuracy". Pupil performance in writing fell short of national expectations (Cable et al., 2010); where written outcomes *were* strong, teachers' linguistic proficiency was good. Lower levels of teacher confidence in FL literacy appeared to result in decreased reading and writing provision in lessons.

(ii) Pupils' performance in speaking

Non-specialists' insufficient linguistic skills may limit pupils' outcomes through poor pronunciation models, found by Muñoz (2009), who concluded from the Barcelona Age Factor project that teachers' accented input may adversely affect pupils' phonological ability. Larson-Hall (2008) similarly noted improved phonological outcomes for Japanese EFL learners when taught by native speakers, although statistical evidence was unavailable due to small participant numbers.

(iii) Pupils' performance in vocabulary

In Poland, Szpotowicz (2009) investigated factors affecting 67 EFL pupils' vocabulary recall and recognition. Although the four teachers had differing qualifications and experience, the experimental technique for vocabulary teaching was held constant. Szpotowicz (2009) found the lowest scores on all vocabulary tests among pupils taught by the teacher with least experience and no English qualifications. An English longitudinal study (Graham et al., 2014), measuring pupils' progress in vocabulary knowledge and grammatical understanding from Year 5 to Year 7, found similar results. Two different teaching approaches aimed to develop language competence: one prioritising oracy, one integrating oracy and literacy. The most significant factors in pupil outcomes on sentence-repetition and photo description tasks (after learners' L1 literacy levels) were teachers' levels of French and the amount of training received. Confidence

and linguistic proficiency in French, combined with weekly provision of 60 minutes' FL input, produced the highest outcomes, regardless of teaching approach.

Similarly, Unsworth et al. (2015) considered how classroom exposure and teacher linguistic proficiency affected pupils' progress in receptive vocabulary knowledge and grammatical understanding in the first two years of learning EFL. Pupils taught by a non-native speaker of English with a level lower than B1 on the CEFR (Council of Europe, 2001) achieved significantly lower results than any other group of pupils.

(iv) Reduced challenge

Insufficient linguistic competence may lead to reduced challenge in lessons (documented by Cable et al., 2010); Muijs et al. (2005) observed more complex content in lessons taught by specialist teachers. Graham et al. (2017) also found that Year 6 pupils taught by teachers with lower levels of proficiency in French made less progress, attributing this tentatively to inadequate subject knowledge which problematised planning for progression beyond beginner levels.

1.5 Summary

The 'who', 'when', 'what' and 'how' issues (DCSF, 2005) of primary languages are only partially resolved, currently determined more by necessity than research-informed measures of efficacy. The largely non-specialist workforce may lack confidence and linguistic competence; time is dictated by availability once other higher profile subjects have been prioritised. 'What' and 'how' are, however, areas which merit further exploration, to assess how the nature of input affects pupil outcomes (recommended by Graham et al., 2017), and whether different approaches and content yield improved results in time-limited provision. Reports have argued that effective teaching can be achieved in different ways (Tragant Mestres & Lundberg, 2011; TSC, 2016), with scope for 'innovative pedagogical approaches' (Baïdak et al., 2017:20). Such research could provide realistic guidance on what constitutes "substantial progress" (DfE, 2013b:194) in early FL learning.

Clear guidance should outline content needed by non-specialists and advise on effective teaching approaches to maximise pupil outcomes. Mihaljević Djigunović (2009a) speculated that Croatia's language-rich context, where English holds high status, could compensate for less favourable learning conditions, including non-specialist teachers with lower levels of proficiency or confidence. However, as Graham et al. (2017) suggested, in England it would seem that teacher proficiency *grows* in importance as out-of-school exposure is minimal.

Attempting to provide the guidance and support needed by non-specialist teachers of FL, this study considers a vast body of research in vocabulary, phonics and grammar teaching, as

well as discussing teacher and pupil response. It does this in as much depth as possible, whilst necessarily retaining a broad overview, providing much-needed investigation of age-appropriate pedagogy. Because of the breadth of this study, results are first presented, then discussed, in each results/discussion chapter ([Chapter 5](#): pupils' learning outcomes; [Chapter 6](#): pupils' self-efficacy and enjoyment; and [Chapter 7](#): teacher findings) to minimise repetition.

Chapter 2: Literature Review I. Practice in foreign languages: vocabulary, phonics, grammar

2.1 Foreign language learning practice

This chapter contextualises key elements of the approach adopted in the work of the National Centre for Excellence in Language Pedagogy (NCELP - 2018) – and thus characterising the Novel scheme of work. Addressing the challenge of ensuring success in FL provision in English secondary schools, the *Modern Foreign Languages Pedagogy Review* (TSC, 2016) made a number of recommendations relating to key areas of language learning such as vocabulary, phonics and grammar which were developed and operationalised by NCELP. It is the commonly occurring content of these areas which could usefully constitute the minimum core content of the new primary curriculum, as acknowledged in the *Curriculum and Assessment Review* (DfE, 2025). These are explored in detail in this chapter. The chapter concludes by drawing together the extensive review and illustrating its relevance to the current study ([section 2.5](#)).

Opinions differ concerning the form that practice essential to second language (L2) learning should take (DeKeyser, 2017; Lightbown, 2019; Spada & Tomita, 2010). This section examines different types of practice, which should suit context (acknowledged by Suzuki et al., 2019a), be transferable to long-term availability for use (R. A. Schmidt & Bjork, 1992), and be meaningful, involving learners understanding rather than simply reproducing (Lightbown, 2019). The following are considered:

- intentional and incidental learning.
- explicit and implicit learning and instruction.
- types of knowledge.
- modes and modalities of input.
- noticing.
- desirable difficulty.
- distribution of learning.

2.1.1 Intentional and incidental learning

2.1.1.1 Incidental learning

Most children generally acquire their L1 through incidental learning (Miralpeix, 2020), exposed to vast amounts of input (File & Adams, 2010; Pellicer-Sánchez, 2020; Uchihara et al., 2019; S. Webb, 2020; S. Webb et al., 2020). Such exposure is rare in FL learning contexts, with learners unlikely to have repeated encounters with new items (Miralpeix, 2020; van den Broek et al.,

2018) and possibly not according them sufficient attention to support memorisation (Kormos, 2020). Attention *may* be present in incidental learning but not intent (Kormos, 2020; Laufer, 2020).

2.1.1.2 Intentional learning

Intentional learning, by contrast, is deliberate, explicit, involving effort to learn and attention to items (Barcroft, 2020; Pellicer-Sánchez, 2020; S. Webb et al., 2020). Learners may be aware that learning will be tested (Laufer & Hulstijn, 2001; Uchihara et al., 2019). Its benefits are acknowledged (e.g., in vocabulary knowledge, Nation, 2007, 2020; Schmitt, 2008; S. Webb et al., 2020), particularly short-term (S. Webb & Nation, 2017).

2.1.2 Explicit and implicit learning; explicit and implicit instruction

2.1.2.1 Explicit and implicit learning

Table 1.

Comparison of Explicit and Implicit Learning

Explicit learning	Implicit learning
closely related to intentional learning.	can happen without formal instruction (Laufer & Hulstijn, 2001) – e.g., children learning L1 (Roehr-Brackin & Tellier, 2019).
features consciousness (Ellis, 1994; Suzuki et al., 2019a); awareness (Suzuki & DeKeyser, 2017); attention to form (Erlam, 2003; Suzuki & DeKeyser, 2017);	intuitive (Erlam, 2003).
generally considered efficient.	lack of intention/awareness (DeKeyser, 2003; Suzuki & DeKeyser, 2017).
requires cognitive maturity.	only incidental (Laufer & Hulstijn, 2001).
often considered less suited to young learners (Roehr-Brackin & Tellier, 2019).	usually requires time/substantial exposure to supply implicit learning mechanisms with sufficient data.
possibly affected by language learning aptitude/metalinguistic awareness	

.(Kasprowicz et al., 2019; Roehr-Brackin & Tellier, 2019). rapid gains vis-à-vis time invested – suits L2 classroom contexts (Muñoz, 2006; Roehr-Brackin & Tellier, 2019).	

2.1.2.2 Explicit and implicit instruction

The presence of rule explanation or focus on grammatical forms defines explicit instruction (Norris & Ortega, 2001); absence of these components indicates implicit instruction. Explicit instruction produces larger effect sizes, found in meta-analyses conducted by Norris and Ortega (2001) of 49 studies and Spada and Tomita (2010) of 41 studies, which investigated the impact of instruction type on simple and complex grammatical concepts. Advantages of explicit instruction were maintained in delayed post-tests. However, explicit instruction may benefit from study designs and outcome measures lacking ecological validity (DeKeyser, 2017), for example short duration (DeKeyser, 2003) or fewer measures of spontaneous language production that may better capture implicit learning (Norris & Ortega, 2001).

2.1.3 Knowledge types

Both implicit and explicit learning contribute to L2 acquisition (Lichtman, 2016), because of the type of knowledge they make available to learners. Declarative knowledge ('knowledge that') describes knowledge of facts and events, usually consciously available (positively associated with explicit knowledge, established during explicit learning). With practice and feedback, declarative knowledge changes (e.g., can be 'chunked' into more accessible representations in the mind) to become procedural knowledge ('knowledge how') ('skill acquisition theory', DeKeyser, 2017). This is more 'instinctive', governing actions taken without conscious awareness (likely positively associated with implicit knowledge, established during implicit learning). Proceduralised knowledge is characterised by instinctive, often involuntary, rapid and largely correct retrieval, not subject to memory limitations (R. W. Schmidt, 1990). In language (and general skill) acquisition, proceduralisation signifies fluency and ready availability (DeKeyser, 2017). With further practice, this proceduralised knowledge can become automatised, usually more stable, less prone to errorful access (relative to declarative or proceduralised knowledge), enabling fast, unconscious access less prone to decay over time.

The relationship between explicit and implicit knowledge (the interface issue) has been widely debated in L2 acquisition. The non-interface argument holds that explicitly acquired declarative knowledge cannot be proceduralised and automatised, whereas Suzuki and DeKeyser

(2017) maintained that implicit knowledge could be developed in different (conscious) ways. Initial declarative knowledge may develop from explicit instruction, progressing into proceduralised and automatised knowledge through systematic, repeated practice (DeKeyser, 2017), with regular use improving fluency and accuracy (DeKeyser, 2003; Suzuki et al., 2019a), developing learners' ability to apply language to different contexts (DeKeyser, 2017). Repeated engagement in practice can produce habit-formation and highly automatised language, in turn potentially creating implicit knowledge (DeKeyser, 2017).

University students receiving intensive grammatical instruction prior to their FL degree (Macaro & Masterman, 2006) did not increase their ability to make grammaticality judgements when compared to a comparison group (although participants could correct incorrect sentences once identified), suggesting that explicitly taught grammar was unavailable as instinctive, implicit knowledge. In contrast, although further research is needed, Suzuki and DeKeyser (2017) concluded from their study of 100 L2 Japanese speakers that explicit knowledge may enable implicit knowledge development. The issue could depend on distinctions (DeKeyser, 2003): explicit knowledge cannot easily contribute to implicit, instinctive behaviour if procedural knowledge is considered unconscious, as learners are aware of language rules; however, if procedural knowledge implies automaticity, then language automatised through practice could be thus defined. Practice should ideally replicate real-world engagement, as transfer-appropriate processing suggests that engaging in similar cognitive processes to practice tasks facilitates language retrieval (DeKeyser, 2017) – see [2.1.4](#), [2.2.4.2](#).

2.1.4 Modes and modalities of input

Experiencing linguistic features in comprehension *and* productive modes is important. In the written modality, less time-pressured input offers greater opportunities for revisiting and noticing, whereas in the oral modality input is “inherently fleeting and timed” (K. Kim & Godfroid, 2019:651), burdening working memory capacity (Kormos, 2020). DeKeyser's (1997) study, involving 61 participants learning vocabulary and morphosyntactic rules in either comprehension or productive modes, found that ultimate performance related to the mode in which language features were practised, with outcomes limited when reversed. When participants used newly acquired linguistic features in skills not previously practised, their reaction times and errors increased (see also DeKeyser & Prieto Botana, 2015; Li & DeKeyser, 2017).

2.1.5 Noticing

The process of engagement also enables learning and retention (e.g., with reference to vocabulary, see Huang et al., 2012; Laufer & Nation, 2012; Nation, 2020; Pellicer-Sánchez,

2020; Schmitt, 2008). Noticing requires consciousness; conscious processing is essential to learning (R. W. Schmidt, 1990). Attending to, or noticing, key features occurs in incidental and intentional learning, with the latter more effective (Nation, 2020). However, noticing from enhanced input does not necessarily entail learning (Mackey, 2006): rather, Issa and Morgan-Short (2019) found that input orienting learners to the function of the form increased learning and attention on the target form more than input enhancement in the form of bolding. Learners possibly have to connect the form they notice to a meaning or function (Kasprowicz & Marsden, 2018; Marsden & Chen, 2011; VanPatten, 2002) or actually do something with the input.

2.1.6 Desirable difficulty

Intentional learning should be effortful, occurring “outside one’s comfort zone” (Suzuki et al., 2019b:713). In desirable difficulty, originated by R. Bjork (1994), challenging practice conditions are considered most likely to increase durable learning gains, by forcing engagement in “the very encoding and retrieval processes that support long-term retention and transfer” (R. Bjork, 2017:147). Challenge should be optimal: if too demanding, learners experience failure; difficulty is no longer desirable. Learners may mistakenly assume that short-term achievements enabled by less desirably difficult tasks represent lasting learning, whereas these conditions are unlikely to lead to permanent changes in knowledge and understanding. Difficulty during acquisition may degrade short-term performance (Bjork & Bjork, 2011; R. Bjork, 2017; Suzuki et al., 2019b), highlighting the importance of focusing on the after-effects of practice (R. A. Schmidt & Bjork, 1992).

2.1.7 Distribution of learning

Congruent with desirable difficulty is the concept of spaced learning. In *massed* learning, the same items are repeatedly studied sequentially (Nakata, 2020), whereas in *spaced* practice, multiple repetitions of items are spread over intervals (e.g., occurring once every 20 minutes or once a week). Spaced learning conditions are generally considered more effective for long-term gains (Nakata & Suzuki, 2019a). Although R. A. Schmidt & Bjork (1992) found that massed learning produced better short-term performance, S. Kim and Webb’s meta-analysis (2022) of 37 studies (48 experiments, $N = 3,411$) found that spaced practice significantly improved both immediate and longer-term L2 learning outcomes compared with massed practice, with effect sizes of $g = .58$ in immediate post-tests, and $g = .80$ at delayed post-tests. Although S. Kim and Webb found that adult learners’ retention (of grammar and L2 vocabulary) was greater with longer spacing than young learners’, the study sample size was very small ($k = 4$). When Nakata and Suzuki (2019a) investigated the impact of practice on word retention both in the learning

phase and immediate post-tests/delayed post-tests, the spacing condition was significantly more effective than the massed at securing long-term vocabulary gains.

The optimal lag effect, or interval between occurrences in spaced (or distributed) practice (Kasprowicz et al., 2019) remains unclear, and may depend on the timing of the post-instruction test (Nakata & Suzuki, 2019b). Nakata (2020) recommended weekly intervals over daily recycling, but longer delays introduce desirable difficulty; delays of at least two months were deemed most beneficial by Bahrack et al. (2016). Intervals can be equal (e.g., every two weeks) or expanding. Kasprowicz et al. (2019) examined the outcomes of 113 L1 English learners of French aged 8-11 when learning French verb inflections through practice spaced at 3.5 or 7 days. Outcomes in sentence-picture matching and acceptability judgement tests (demonstrating little group-level learning over the period of the study) did not reflect the accuracy demonstrated during the practice. High levels of individual variation were noted within each condition, suggesting that language-analytic ability (LAA) was the key predictor of learning outcomes (more than distribution of practice) (see [2.4.3.3](#)).

Instruction can involve practice of a single concept (blocked) or different concepts at the same time (interleaving). Blocked practice enables learners to develop familiarity with commonalities of a particular concept; interleaving provides practice in distinguishing between concepts (Nakata & Suzuki, 2019b). They can be combined, with blocking followed by interleaving (increasing practice). In Nakata and Suzuki's (2019b) study, 115 EFL Japanese learners learnt grammatical structures (simple past/present perfect; first/second/third conditional) after random assignment to three conditions: blocking, interleaving or increasing. Differential outcomes were not significant immediately, but in delayed post-tests, the interleaving group significantly outperformed other learners. Nakata and Suzuki had hypothesised that blocking practice might better suit beginners, yet participants with lower pre-test scores benefited more from interleaving practice than those with higher pre-test scores. However, Hwang (2025) found that interleaving alone introduced undesirable difficulty in a study conducted with 15-16-year-old low-achieving LI Korean students learning EFL, whereas a hybrid approach combining blocked and interleaving practice better promoted longer-term learning gains in vocabulary meaning recognition (than blocked or interleaving practice alone). Gains made in all three approaches in vocabulary recall knowledge were not sustained at delayed post-test.

Learners themselves may not realise the benefits of spaced or interleaving practice, precisely because the increased challenge will likely degrade performance in the learning phase (Nakata, 2020; Nakata & Suzuki, 2019a). Participants in the massed condition may overestimate likely final outcomes on vocabulary tests because of higher performances in the learning phase (Nakata & Suzuki, 2019a).

Having discussed the nature of *practice*, the nature of *knowledge* is now considered - particularly as different components of language (e.g., vocabulary, grammatical, and sound-symbol relations knowledge) may react differently to different practice types. All language competence models include core components of lexical, morphosyntactic, phonological and orthographical knowledge (among other components such as sociolinguistic competence and pragmatic competence; see Bachman, 1990; Canale & Swain, 1980). Indeed, more recently and in the context of a policy-document, the *Modern Foreign Languages Pedagogy Review* (TSC, 2016:5) reflected this basic conceptualisation of language competence, stating that learners in English schools should gain “systematic knowledge of the vocabulary, grammar, and sound and spelling systems (phonics) of their new language”. These components are now explored with reference to the roles they play in promoting effective language learning.

2.2 Vocabulary

Vocabulary knowledge is essential for successful language learning (Nakata et al., 2021; Orosz, 2009; Qian & Lin, 2020; Vermeer, 2001; Zhao & Murphy, 2017). The National Curriculum for FL (DfE, 2013a, 2013b) requires learners aged 7-11 to use vocabulary knowledge to: communicate verbally and in writing; comprehend spoken and written text, engaging with stories, poems and songs. However, TSC (2016) questioned whether, among secondary school learners, unsatisfactory vocabulary knowledge and poor vocabulary teaching may partly explain low achievement levels.

This section explores aspects of vocabulary knowledge and size, considering utility value, and how this understanding might influence taught lexical content and word list compilation. It reflects on factors affecting learnability, retention and recall, before investigating approaches to vocabulary teaching.

2.2.1 Word knowledge

2.2.1.1 Terms and definitions

The linguistic unit ‘word’ is actually a token, or ‘lemma’ (Miralpeix, 2020), which includes the base, or headword, and its inflected forms. Lemmas belong to the same word class (e.g., all verbs, or all adjectives), whereas a ‘flemma’ allows forms from two different word classes with the same spelling (Cobb & Laufer, 2021; Dang & Webb, 2016; Dang et al., 2020) – for example, a verb, its inflections, and a noun (e.g., ‘to bank’ and ‘bank’). Word families, typically larger, contain headword, inflected forms and derivations, possibly covering different word classes (Dang et al., 2020). Sometimes it is (erroneously) assumed that learners, knowing a headword and inflected forms, will understand closely related derivations (Schmitt, 2008; Vilkaitė-Lozdienė & Schmitt, 2020). However, this can challenge younger beginner learners with lower

levels of morphological understanding. When deciding how many words learners should know, counting lemmas may be better than counting word families for productive modes; ultimately, the reason for counting should determine the measure adopted (Nation, 2016).

Word lists may contain multi-word units (lexical bundles, formulaic sequences or collocations, [2.2.4.5](#)), usually learnt and recalled as complete units and thus supporting fluency development (Boers, 2020; Nation, 2020). Function words (high frequency, usually a closed class, including, for example, conjunctions, prepositions, pronouns, qualifiers, and auxiliary verbs) may also be included in word lists alongside content words, as they often convey essential meaning (Nation, 2016).

2.2.1.2 Receptive and productive knowledge

Receptive (‘passive’) knowledge involves recalling meaning of written (reading) or spoken text (listening). Productive (‘active’) knowledge is required to recall word forms to express meaning in speaking or writing (Nation, 2016; Schmitt, 2014). Receptive knowledge is generally considered larger than productive (Laufer & Nation, 2012; Schmitt, 2008) and easier to develop (Nation, 2016; Qian & Lin, 2020; Schmitt, 2014; Szpotowicz, 2000; S. Webb & Nation, 2017). In four experiments (two testing recall, two receptive knowledge), De Groot and Keijzer (2000) found the highest scores and shortest retrieval times of 40 experienced FL learners on receptive tests. Recognition may be easier, but translation into the FL involving “effortful retrieval practice” (Kemp & McDonald, 2021:2) may produce *more effective* learning. Huang et al.’s (2012) meta-analysis findings suggested that learners’ awareness of language inadequacy when trying to convey meaning refocused attention on key forms.

Debate continues over whether receptive and productive knowledge are distinctive constructs or part of a continuum “starting from superficial familiarity with the word and ending with an ability to use the word correctly in free production” (Qian & Lin, 2020:67).

2.2.1.3 Breadth, depth and fluency

(i) Breadth

Breadth refers to the number of words known, and can predict learner performance in the four skills (as reviewed by Milton, 2013): in a meta-analysis (Zhang & Zhang, 2022), the highest predictor of success in reading comprehension was learners’ ability to recall meaning. Knowing 95-98% of words in a written text enables accurate comprehension (Schmitt, 2008; Stæhr, 2008), requiring a word size of 5000-8000 to access any text (Laufer & Nation, 2012). However, 95% coverage of any text not prepared for specific educational purposes may require dictionary use (Laufer & Aviad-Levitsky, 2017), possibly still adversely affecting some learners’ comprehension (Qian & Lin, 2020).

In listening, 2000-3000 word families would need to be known for 95% coverage, although this is unconfirmed by empirical evidence (Giordano, 2021). For simple conversation, 2000 word families may suffice (Laufer & Nation; 2012; Milton, 2013). Speaking, perhaps more repetitious or incomplete (Nation, 2016), requires fewer words than more sophisticated written communication (Milton, 2013).

Stæhr's (2008) Scandinavian study found that, after learning EFL for seven years (400-700 hours), 68/88 15-16-year-old Dutch learners did not know the 2000 most common English words. Their vocabulary levels were compared with their performances in listening, reading, and writing: 62% and 72% of those who did not know the 2000 most common words scored below average on reading and writing tests, respectively. Listening results were less clearly related to vocabulary size, suggesting that participants perhaps employed strategies to compensate for insufficient knowledge. Knowledge of the most frequent 2000 words appears necessary for successful FL learning (Milton, 2013).

Learners in instructed EFL contexts are unlikely to learn > 1000 words in two years (Dang & Webb, 2016), offering 60-75% coverage levels of texts. Referring to secondary education in Britain, Milton (2006) explained that previous research suggested that pupils learn on average 170 words per year aged 11-16, with five years' input leaving learners far short of even 1000 words, yet according to A. David (2008), 1000 words constituted the core vocabulary in the GCSE. Learners with word knowledge < 2000 words are unlikely to be able to function independently in the language. The first 1500 words is a "bare minimum for simple oral conversations"; the second 1500 takes learners to "a level where more independent learning becomes possible" (Tschichold, 2012:10).

(ii) Depth

In addition to size, vocabulary knowledge includes how much is known about words (Stæhr, 2008). Learners need to understand and produce orthographical and phonological forms, and how word meanings are related (Zhao & Murphy, 2017). Depth of vocabulary knowledge explains linguistic performance in all four skills (Milton, 2013) and is highly relevant to curriculum design and pedagogy.

(iii) Fluency

In-depth vocabulary knowledge may ensure greater speed and automaticity (stability, reliability) in word retrieval (Stæhr, 2008). Fluency refers not to word knowledge itself but to the degree of control over it (Laufer & Nation, 2012).

2.2.2 Word choice

FL learners have limited exposure to the language and instruction, so word selection must be made carefully (Cobb & Laufer, 2021; Laufer & Nation, 2012), possibly taking account of frequency ([2.2.2.1](#)) and usefulness ([2.2.2.3](#)), considered below.

2.2.2.1 Frequency

In time-poor contexts, word selection should offer the best return for the effort invested (Dang et al., 2020): the “cost/benefit principle” (Nation, 2016:5). Some words occur more frequently than others; the most frequent words represent “the core meaning system” (Gu, 2020:274), although less frequent words still obviously carry meaning (Nation, 2016). Frequency varies according to modality: spoken words, less formal, feature greater use of very high-frequency items such as ‘I’ or ‘you’ (Nation, 2008, 2016). If we accept (and apply to FLs) Nation’s (2016) suggestion that 50% of text is made up of the most frequent 100 words in English, and that knowing the most frequent 1000 words enables learners to understand 70-90% of a text (with percentages affected by modality and context), focusing on the highest frequency words appears to offer the best return (Dang et al., 2020).

Nation (2016) classed high-frequency words as the 2000 most frequent, although Schmitt and Schmitt (2014) advised a high-frequency cut-off point of 3000 word families. The advice for FL course design is therefore to prioritise developing learners’ knowledge of high-frequency vocabulary (Nation, 2016). Vilkaitė-Lozdienė & Schmitt (2020) suggested that, while learning the most frequent 2000 words is ideal, in England a more appropriate initial L2 learning goal is 1000.

Focusing on high-frequency words increases the likelihood of regular re-encounters, as these words constitute most texts (Hashimoto & Egbert, 2019; Nation, 2016) encountered in naturalistic acquisition, although instructed contexts may require the creation of opportunities for deliberate learning (Gu, 2020). Frequent occurrence may facilitate word learning (suggested by Milton, 2007; demonstrated by E. Peters, 2014), and increase the probability of retention (Vermeer, 2001; Vilkaitė-Lozdienė & Schmitt, 2020). In contrast, focusing on low(est)-frequency words risks hampering learners’ ability to function effectively in the FL (Sakata, 2019); these may be best acquired incidentally (Gu, 2020).

2.2.2.2 Course books and frequency

When learners rarely encounter FLs outside school (see [1.4.1](#)), teachers must create opportunities for engagement with new vocabulary (S. Webb & Nation, 2017). Course books may therefore significantly influence pupils’ developing vocabulary knowledge (Orosz, 2009). Their content may not be high-frequency and may not represent the most useful words (Sakata,

2019); learners thus develop knowledge of low(er)-frequency vocabulary (E. Peters, 2020). Lower-frequency language features because course books are often organised thematically (A. David, 2008; Häcker, 2008; Milton, 2007; Wingate, 2018), focusing on language unrepresentative of that learnt by native speakers (A. David, 2008; Northbrook & Conklin, 2018) and not easily transferable to other topic areas (Häcker, 2008). Examining a then popular course book for teaching French in England (*Encore Tricolore*), Tschichold (2012:16) found that “several dozen words” from the most frequent 1000 did not appear in the text-book corpus; it included a high number of anglicisms and fewer verbs than expected. Tschichold (2012) speculated that this may explain 16-year-old British learners’ lower FL achievement levels compared with counterparts abroad. More recently, analysis of vocabulary content of the GCSE (in French, German and Spanish) found low(er)-frequency, non-recurring items (Dudley & Marsden, 2024). Vocabulary lists created by awarding bodies for this examination support teachers and textbook creators (thus influencing taught content), yet approximately half the lexical items were not used in examinations conducted in 2018-2020 and a sample examination (Marsden et al., 2023). This suggested non-optimal use of the limited curriculum time available: potential time for learning higher-frequency items likely to occur in examinations (and real life) was spent covering word list content.

In summary, FL learners develop vocabulary knowledge from three sources: course books; teacher target language; or out-of-school exposure (Häcker, 2008). English learners cannot easily draw on large amounts of out-of-school FL exposure (unlike EFL learners who may be exposed to English in their free time – see, e.g., Puimège & Peters, 2019, [1.4.1](#)). This emphasises the importance of lesson content. Non-native FL teachers are likely to use text-books (Tschichold, 2012) or provided schemes; course design and content therefore become increasingly significant.

2.2.2.3 Subjective usefulness

Vocabulary selection should be guided not just by objective frequency (i.e., based on data from corpora) but also by ‘perceived usefulness’ (He & Godfroid, 2019; Laufer & Nation, 2012). Of course, high-frequency words may be most useful (as suggested by Newton, 2020) but may also be influenced by learners’ age, needs (Vilkaitė-Lozdienė & Schmitt, 2020) and context (Cobb & Laufer, 2021): “young native speakers of the language speak and read about different things than adults do” (Nation, 2016:9; see also Nation & Sorrell, 2016). For example, vocabulary in children’s fairy tales may be low-frequency; young learners are often taught animal- or environment-related words - teachers should not exclude these, even if lower-frequency (S. Webb & Nation, 2017). Szpotowicz (2000) found higher levels of pupil recall and recognition on animal words (their favourite topic), then fairy tale items, then household words. However,

this was a very small-scale study ($N = 8$). Lists may include a greater proportion of words common in spoken language, reflecting language encountered by younger learners (Nation & Sorrell, 2016).

Deciding on the most useful words for learners depends largely on their purpose, and should be supported by a clear, strong rationale (Nation & Sorrell, 2016). It is not clear how claims of ‘relevance to young learners’ have been or could be ascertained, and learners themselves may not be best placed to know what the most useful words are likely to be. Vilkaitė-Lozdienė & Schmitt (2020) suggested that, for general learning purposes, objective (corpora-based) frequency is helpful.

2.2.3 Factors affecting learnability

As discussed, learning contexts can affect learnability (Häcker, 2008; Newton, 2020; E. Peters, 2020; S. Webb, 2020). “Learner-related factors” (E. Peters, 2020:125) also affect retention, including: aptitude differences (Häcker, 2008; Orosz, 2009); motivation (Laufer & Hulstijn, 2001; Nation, 2020; Newton, 2020; S. Webb, 2020); working memory (E. Peters, 2020; Uchihara et al., 2019); L1 knowledge (particularly important among young learners) (S. Webb & Nation, 2017); and concentration levels (Szpotowicz, 2000). Additionally, learnability may relate to the words themselves. This section explores this higher “learning burden” (Nation, 2020:15).

2.2.3.1 L1-L2 relationship

Cognates, FL words resembling words in our L1, have “high form and meaning overlap” (E. Peters, 2020:125). Cognates are generally considered easier to learn than non-cognates (Laufer & Nation, 2012; Milton, 2007; Tschichold, 2012; S. Webb & Nation, 2017). Indeed, De Groot and Keijzer (2000) tested 40 experienced FL learners receptively and productively on 20 pseudo-words and found higher receptive and recall scores (and faster response times) for cognates than non-cognates. They, along with Tonzar et al. (2009), hypothesised that learners build on existing memory entries when L1 resembles L2, decreasing processing times (see also Conklin, 2020). Tschichold (2012) speculated that cognates facilitate comprehension, creating a sense of early success in learners. However, word (form) similarity can be problematical (S. Webb & Nation, 2017): interlexical interference occurs when learners expect L1 words to be translated in the same way in the FL (Boers, 2020). Indeed, in Otwinowska et al.’s (2020) study, Polish EFL learners ($N = 83$) able to spot cognates were more likely to try to guess the meaning of orthographically similar words (which had different L1-L2 meanings). Learners’ knowledge of cognates was not greater than of non-cognates, even after intensive awareness raising.

2.2.3.2 Word length; pronounceability; word class; abstract versus concrete nouns; input modality

Hashimoto and Egbert (2019) suggested that word length affects learnability, although Szpotowicz (2000), in a small-scale study ($N = 8$) investigating learnability and retention of words from different categories (animals, fairy tales, everyday objects) among 7-year-old Polish EFL learners, found recall and recognition unaffected by the number of syllables in words. Regular spelling patterns may facilitate learning, whereas words which are difficult to pronounce may be harder to learn and remember (noted by S. Webb & Nation, 2017); E. Peters (2020) suggested that FL-L1 phonemic differences complicate learning.

Difficulty may depend on word class. Tschichold (2012) suggested that nouns are easier to learn than verbs; adverbs are more difficult still (Milton, 2007). E. Peters (2020) summarised available research as inconclusive but explained that difficulties may arise because of potential variations inherent in verbs, also suggesting that nouns are often more concrete or imageable, with abstract words harder to visualise. Concrete words are generally considered easier to learn (Boers, 2020; Milton, 2007; E. Peters, 2020); indeed, in the Dutch study cited ([2.2.1.2/2.2.3.1](#)), learners demonstrated higher scores and faster recall response times for concrete words than for abstract (De Groot & Keijzer, 2000); the easiest words to learn were concrete cognates. The authors suggested that dual-coded storage (in image and verbal systems) possibly supported learners. Abstract words can also apply to a range of concepts or ideas in various situations, whereas concrete relate more reliably to one construct (the object).

Regardless of words' inherent features, Kormos (2020) posited that learnability is affected by the different modalities in which words occur. Written texts afford learners time to engage and revisit, whereas spoken words are fleeting and heavily dependent on memory capacity.

2.2.4 Teaching vocabulary

Learning activities should develop breadth, depth and fluency, or automaticity (see Barcroft, 2020; Stæhr, 2008). Schmitt (2008) summarised effective ways to promote learning, including: maximising interaction with, and attention to, words; deliberate learning; a requirement or need to learn; and opportunities to manipulate words. Schmitt (2008:353) cited key principles from Hunt and Beglar (1998) as ever relevant in promoting effective learning: incidental and intentional learning opportunities; identifying the most useful words to learn; developing depth and fluency of vocabulary knowledge; and strategies such as dictionary use or contextual guessing. Schmitt also foregrounded the following: setting ambitious vocabulary targets; adopting flexible learning approaches; establishing form-meaning links quickly through

intentional learning (see Dang et al., 2020), and supporting this process through repeated encounters.

2.2.4.1 Semantic relatedness

FL words are often taught in semantically related sets (e.g., animal nouns, or related groups of adjectives, such as colours). Semantically related items risk interference (S. Webb & Nation, 2017), potentially limiting learners' knowledge to the category to which words belong, rather than individual meaning (Tschichold, 2012). Tinkham's (1997) study, involving 48 university students, investigated the learnability of four sets of pseudo-words with English equivalents, tested in the listening and reading skills after being learnt in four conditions: semantically related/unrelated from the same word class (e.g., all nouns); thematically related/unrelated from different word classes (e.g., noun, adjective, verb). Scores in thematic clustering were higher than in thematically unrelated groupings; semantically unrelated words were easier to learn than those that were related; most participants considered thematically clustered words easier to learn.

However, empirical evidence supporting vocabulary teaching in semantically unrelated groupings is inconclusive (Nakata, 2020). In Nakata and Suzuki's (2019a) study, 133 Japanese university students learnt 48 low-frequency English words (half semantically related, half unrelated) with Japanese equivalents; variables affecting word difficulty were controlled for in both languages. Post-test and delayed post-test scores were not significantly different, although Nakata and Suzuki concluded that semantic clustering should be avoided, as within-set errors were more frequent in semantic sets during the learning period and in tests (e.g., participants produced the word 'teacher' instead of 'student').

2.2.4.2 Frequency of encounters

Frequency of encounters with words is a key factor in promoting learning (Barcroft, 2020; Nation, 2020; Schmitt, 2008; S. Webb & Nation, 2017), likely to affect how effectively they are processed (Conklin, 2020) and remembered (Pellicer-Sánchez, 2020; Schmitt, 2008). Recycling enables repeated reinforcement of form-meaning relationship, speeding up the retrieval process (Conklin, 2020; Zhao & Murphy, 2017).

The ideal number of exposures appears indeterminate, possibly affected by learner engagement and testing approach (Schmitt, 2008), but S. Webb and Nation (2017) recommended seven intentional encounters, rising to twelve incidental. E. Peters (2020) suggested that eight to 10 repetitions should facilitate learning, with additional encounters supporting form-meaning associations (E. Peters, 2014). Greater depth may necessitate more encounters (E. Peters, 2020; S. Webb & Nation, 2017), as may word difficulty (Tschichold, 2012).

The advisability of encounters in different modalities (S. Webb & Nation, 2017) and different forms is highlighted (S. Webb et al., 2020; Tschichold, 2012). Brown (2011) suggested that, after initially focusing on meaning, subsequent exposures could target words in spoken form or in context. Each encounter reinforces knowledge, minimising forgetting (Schmitt, 2008). According to Nation (2020), teachers should prioritise frequency of contact, particularly with most useful vocabulary (Nation, 2007), over the form that initial teaching takes. Opportunities for recycling key language may be scarcer in topic-based syllabuses, with content-specific, less transferable words (S. Webb, 2020).

2.2.4.3 Retrieval practice and forewarning

Retrieval practice requires active recall of previously learnt information (Lindstromberg, 2020), impacting positively on future retention. Roediger and Karpicke's (2006) study, involving 120 undergraduate students, demonstrated that, although repeated *studying* offered short-term benefits, repeated *testing* was more beneficial one week later. Both approaches repeat encounters with new words, but the testing condition involves higher levels of challenge characteristic of desirable difficulty (2.1.6). A later study (Karpicke & Roediger, 2008) confirmed that retrieval practice increased learning gains beyond repeated studying. Additionally, forewarning learners that new vocabulary will be tested may promote greater attention to items and increase retention (Roediger & Karpicke, 2006; Uchihara et al., 2019; S. Webb et al., 2020).

R. A. Schmidt and Bjork (1992) stressed the importance of retrieval from memory, with this process strengthening memory traces (Barcroft, 2007; Lindstromberg, 2020). In van den Broek et al.'s (2018) study based on three experiments involving adult Dutch speakers ($N = 45$), participants learnt and practised nouns in an unknown language in two conditions. One forced recall from memory; in the other, contextual information supported word retrieval. From higher accuracy scores in receptive and productive knowledge delayed post-tests in the uninformative context, van den Broek et al. concluded that memory retrieval reproduced the beneficial testing effect. Contextual information, supportive in practice, reduced long-term retention.

Adopting a cumulative approach, involving "incrementally incorporating both new and previously learned materials in daily or weekly quizzes" (Maie et al., 2025:1), can enhance the benefits of testing for vocabulary learning. Nakata et al. (2021) found that Japanese university students ($N = 72$) in the cumulative group tested on both recently and previously studied low-frequency English words achieved significantly higher scores on receptive and productive delayed post-tests. However, learners should be made aware of the potential for short-term degraded performance: Nakata et al. (2021) found that participants in the cumulative approach group achieved lower scores than those in the non-cumulative group during the 9-week *learning*

phase, consistent with the desirable difficulty framework, with only long-term benefits apparent. The lack of successful outcomes in the learning phase risks negatively affecting motivation.

Additionally, Nakata (2020) suggested that productive retrieval (i.e., recalling L2 form) probably engenders more effective learning than receptive retrieval (recalling L1 meaning).

2.2.4.4 Repetition and imitation

Repetition is sometimes viewed negatively, with learners passively reproducing (Kolb, 2007). Imitation alone may involve superficial processing and thus result in poorer retention (Sagarra & Alba, 2006), promoting learning less effectively. Kang et al. (2013) compared outcomes of retrieval practice with imitation ($N = 100$). In imitation, English-speaking undergraduate students heard and repeated Hebrew words; in retrieval practice, learners supplied Hebrew words from memory before hearing the correct pronunciation. Retrieval, containing desirable difficulty, resulted in better final performance in both immediate tests and post-tests. Nation (2016) also alluded to the advantages of combining repetition with retrieval or using it in different contexts, with the importance of elaboration, expansion and reformulation in varying contexts supported by Tschichold (2012). Where repetition is used, form-meaning associations must be foregrounded (Barcroft, 2020); then, additional exposure to target items can strengthen retention (E. Peters, 2014).

2.2.4.5 Additional approaches

This section summarises key points relating to other appropriate classroom approaches.

- Presenting vocabulary through stories widens the range of lexical items; because language is not restricted to a particular topic, opportunities for revisiting and recycling are easily created (Häcker, 2008).
- Combining new vocabulary with explicit grammar instruction may support learners' understanding of individual sentence components (Häcker, 2008).
- Including verbs early in the learning phase, recommended in the National Curriculum for Key Stage 2 (DfE, 2013a, 2013b) and supported by TSC (2016), could accelerate learning ([2.4.3.4](#)).
- Opportunities to use new vocabulary in authentic communicative tasks may increase learning beyond what can be derived from input alone (Laufer & Nation, 2012).
- Focus on Forms ("decontextualised vocabulary learning", E. Peters, 2014:74) involves deeper processing and increased engagement, potentially resulting in higher gains than, for example, reading for general meaning alone (Laufer & Nation, 2012; E. Peters, 2014).

- ‘Chunks’ of language, or multi-word units have been considered “‘starting points’ for generalisation and abstraction processes” (Schulz et al., 2023:4) for beginner learners, set phrases which facilitate communication and from which learners eventually develop rule-based understanding (Myles et al., 1998). However, Schulz et al.’s (2023) systematic review (two studies, 5-12-year-olds learning L2 in instructed contexts) failed to find trustworthy evidence of the value of teaching multi-word units to young learners.
- Sharing the written form with young learners can be beneficial, with Hu (2008) finding that orthographic forms facilitated learning. Two groups of 37 Mandarin Chinese-speaking beginner learners of English (with higher/lower levels of phonological awareness) learnt new lexical items through two conditions: (a) accompanied by visuals; (b) with orthographic forms. The vocabulary knowledge of these young learners (aged 8-10), particularly those with higher phonological awareness, seemed to benefit from the addition of orthographic forms while learning. Although Graham et al. (2017) did not find a strong advantage for young learners when taught following an oracy- or literacy-based approach, they tentatively suggested that access to written forms in the more literacy-based approach supported learners with lower L1 literacy levels to retain vocabulary (and grammar) after transfer to secondary school.
- Glossing (providing L1 translations for unknown words alongside written FL text) enables access to more challenging texts with minimal interruption (Schmitt, 2008; S. Webb & Nation, 2017). It draws attention to key items, possibly supporting learning (Jung, 2016; Schmitt, 2008), although the number of glossed words should be limited (S. Webb & Nation, 2017).
- High expectations regarding FL production rather than choral repetition can maintain motivation (Wingate, 2018).

2.3 Phonics

Phonological decoding, or recoding (Ehri, 2005), describes readers’ transformation of the written form of words or letter combinations (graphemes) into their corresponding sounds (phonemes) (Woore, 2018). This knowledge of grapheme-phoneme correspondences, or symbol-sound correspondences (SSCs) (Woore, 2007, 2010), enables learners to pronounce unknown words by mapping visual representations to spoken forms (Ehri, 2005). Knowledge of SSCs also benefits the ‘other direction’: the ability to hear unfamiliar words and write them down with increased likelihood of plausibly correct orthography. This section considers approaches to phonological decoding in English FL classrooms, particularly in French teaching.

2.3.1 Foreign languages phonics instruction: the English context

Despite well-established provision in L1 phonics, FL phonics instruction lacks a strong tradition in English schools (Erler, 2004; TSC, 2016; Woore, 2007, 2010). It appears to be assumed that pupils develop competence through ongoing contact with the language (Erler, 2004; Woore, 2010), making intuitive connections between new written forms and words previously encountered only orally (Woore, 2007), perhaps drawing on knowledge of L1 SSCs. However, given the limited time allocated to FLs in English schools, contact alone is unlikely to achieve successful outcomes in phonological decoding (Woore, 2009, 2010).

Prior to primary FLs becoming compulsory in 2014, OFSTED's (2008a) national survey, conducted 2004-2007, revealed that pupils were not taught SSCs and struggled with FL pronunciation. Cable et al. (2010) found that 10-11-year-old pupils demonstrated some SSC understanding, also reporting teacher perceptions that pupils' exposure to sounds in different languages benefited their pronunciation and spelling. In Tinsley and Board's (2012) national survey, 35% of responding primary schools considered the teaching of intonation and pronunciation a strength. These developments perhaps reflected the inclusion in the *Key Stage 2 Framework for Languages* (DCSF, 2005) of the need for learners to respond to sound patterns, identify specific phonemes, sounds and words, and make links between spoken and written forms. However, intonation and pronunciation were identified as a major challenge or quite challenging by half of responding schools three years later (Board & Tinsley, 2015); additionally, Graham et al. (2014) reported little evidence of SSC-focused teaching when comparing oracy- and literacy-based teaching approaches to primary FL.

Phonological decoding skills have been associated with vocabulary acquisition (Hamada & Koda, 2008; Woore, 2009, 2022) and are considered important by learners (Erler, 2004). Studies in English secondary schools have shown that learners struggle to decode accurately in L2 French, even after some years of learning. For example, Erler and Macaro's findings (2011) ($N = 1,735$ pupils; 25 different schools) aligned with Erler (2004), indicating that 11-12-year-old pupils ($N = 359$) showed little understanding of key pronunciation issues after one year of learning French. If not taught explicitly, learners may apply their own (incorrect) understanding of decoding patterns, as found by Woore (2010), reflecting L1 influence (2.3.3). Woore (2018) found that, among 31 lower-achieving secondary-aged learners of French, decoding was inaccurate in 60% of cases, even when knowledge of word meaning was secure.

2.3.2 Recommendations

Calls for systematic, explicit approaches to teaching phonological decoding have emerged in primary and secondary phases (Holmes & Myles, 2019; TSC, 2016). The National Curriculum for Key Stage 2 (7-11) (DfE, 2013b) requires pupils to explore FL word patterns and sounds,

and develop intonation and pronunciation to be able to read aloud intelligibly; this continues in the secondary phase (DfE, 2014).

Systematicity is important. Topic-based approaches introduce words (and therefore spelling patterns) randomly, with item selection governed by meaning rather than SSCs (Siddons, 2001), possibly complicating the process of making connections between letter strings and sounds. Woore (2007) showed systematic, explicit instruction to be effective. Twenty-eight 11-12-year-old beginner learners of German were taught short poems containing key SSCs. After 18 sessions of input alongside their usual instruction, learners demonstrated improved pronunciation when reading aloud unknown words compared with a comparison group; results reached statistical significance, suggesting that an explicit, systematic, sustained approach could improve learners' phonological decoding ability further.

To teach SSCs effectively, and understand how to “introduce their patterns clearly, directly and in a sensible sequence from the beginning of the course” (TSC, 2016:12), teachers themselves must understand SSCs - potentially challenging for non-specialists. Young learners may reasonably be expected to develop good SSC understanding (suggested by Edelenbos & Kubanek, 2009; Johnstone, 2002), but clear modelling is required. However, primary teachers are generally skilled in teaching English phonics and may transfer this competence to L2 teaching (TSC, 2016).

2.3.3 L1 influence on L2 phonological decoding

TSC (2016) cautioned that, without systematic, explicit phonics instruction, learners may resort to their familiar L1 phonological decoding system (found by Woore, 2014), provoking communication and comprehension issues as they produce non-existent FL sounds (noted by Bassetti et al., 2020). Indeed, 9-11-year-old Croatian beginner learners of French ($N = 42$) struggled when reading aloud to pronounce certain vowels accurately when these sounds were non-existent in their L1 (Marijanović et al., 2009). E. Peters (2020) suggested that such phonemic issues affect word pronounceability because the contrasting phonological structures are unfamiliar. Additionally, Woore (2009) explained that L1-L2 SSC differences could adversely affect vocabulary learning and motivation ([2.3.4](#)).

When learners mispronounced even familiar FL words incorrectly, Woore (2018) speculated that either incorrect pronunciation had been memorised, or the sight of the orthographic form engaged knowledge of stored L1 SSCs stronger than stored L2 pronunciation. This was not found by Hamada and Koda (2008), in which Korean and Chinese EFL learners ($N = 35$) drew on their L2 SSC understanding, ignoring L1 influence. However, these learners were adults resident in an English-speaking environment, with greater potential for L2 SSCs to be

strengthened than in FL classroom contexts. Regular opportunities to enforce FL SSCs support learners to suppress L1 SSC influence.

L1 SSC knowledge could positively influence learners' L2 SSC understanding, for example, if the languages have congruent orthographic systems (Hamada & Koda, 2008). The influence may also be positive if consonants and vowels share the same L1-L2 pronunciation (Woore, 2014, 2018), obviating the need for explicit instruction (Porter, 2020). Facilitative effects may benefit some languages more than others, depending on similarities. Much research in English classrooms has focused on French as a FL. Both languages have alphabetic writing systems and use Roman script (Woore, 2014), although these similarities could potentially increase confusion (Woore, 2018). However, they often differ in vowel pronunciation (Woore, 2014). In terms of orthographic depth, English is deeper (Katz & Frost, 1992, cited in Woore, 2014), more opaque (Erler & Macaro, 2011). Essentially this refers to how consistently graphemes correspond to their spoken representation (Hamada & Koda, 2008), or vice versa (Murphy et al., 2015). Woore (2014) cited the example of the sound [ou] in both languages, decoded more consistently in French than English.

In some languages (e.g., Italian, German or Spanish), early readers make greater use of individual graphemic units ('synthetic phonics') (Goswami, 2005). In shallow orthographies, individual letters or letter clusters correspond more consistently to single phonemes (Ehri, 2005; Goswami, 2005), whereas in English consistency only appears in larger units or letter strings (e.g., 'ight'). An analytic approach to phonics, involving sight reading whole words, may then be appropriate (Goswami, 2005). Although less transparent than Spanish or Italian, French has 'clear and finite' (Erler & Macaro, 2011:497) pronunciation rules; however, learners may struggle to master them given the volume of patterns, and because some French phonemes are non-existent in English (Porter, 2020). Transcribing French may be challenging for English learners as graphemes do not always represent the expected sounds, given their L1 (e.g., [oi] sounds to English learners as if it should be represented by [w], Erler & Macaro, 2011).

2.3.4 Motivation

Wanting to sound 'nativelike' may incentivise learners, as found by Bassetti et al. (2020), yet motivation may be adversely affected if they cannot decode words effectively. Learners struggling to decode words probably experience negative responses to FL, including low self-confidence (Erler, 2004, among others); this highlights their lack of independence as learners (Woore, 2022) and "constantly reinforces their own perception of their 'beginner status'" (Woore, 2007:177). Erler (2004) asked 11-12-year-old L1 English learners of L2 French their feelings about aspects of French after one year of learning: just 13.6% described themselves as

‘happy’ speaking French; 9.6% were happy with their pronunciation; even fewer were happy reading aloud.

A randomised controlled trial (RCT) evaluating the impact of phonics and strategy instruction on 900 11-12-year-old learners of French in English secondary schools (Woore et al., 2018) highlighted no change in pupils’ positivity regarding FL learning in treatment or control groups. However, duration (16 weeks) was possibly too short for lasting impact. Learners apparently enjoy phonics instruction: Porter (2020) noted positive responses to phonics-related activities, and Woore (2007) documented the motivational influence of poems targeting SSC understanding.

Erler and Macaro’s (2011) cross-sectional study, conducted in 25 secondary schools with 11-14-year-old learners ($N = 1735$), investigated the relationship between participants’ phonological decoding ability and aspects of motivation, including intention to continue FL beyond the compulsory age of 14. Erler and Macaro found that learners’ confidence to read aloud decreased with age; pupils frequently commented on the perceived ‘foreignness’ of French. The positive correlations between (a) self-efficacy experienced by learners on decoding tasks and their positivity; and (b) learners’ performance in rhyme/word segmentation tests and their intention to continue French, suggested that these constructs are related. Interestingly, learners’ *perceptions* of their decoding ability, rather than *actual* performance, appeared the key predictor of positivity.

It therefore seems reasonable to conclude that increasing learners’ confidence and achievement through explicit phonics instruction should contribute to their enjoyment of FL learning.

2.3.5 Suggested pedagogical approaches

Schemes could incorporate the following opportunities for young learners to benefit from phonics instruction, as illustrated in the Novel intervention in the current study:

- explicit approaches (advocated by Erler & Macaro, 2011; Sturm, 2013; Woore, 2010; among others) supporting learners to suppress instinctive L1 decoding mechanisms (Woore, 2014).
- provision of metalinguistic information on key aspects of FL pronunciation, enabling learners to differentiate between new sounds and highlighting L1-L2 similarities and differences (suggested by Saito & Plonsky, 2019).
- opportunities for age-appropriate fluency training (recommended by Saito, 2021).

- inclusion of appropriately challenging texts containing unfamiliar L2 words, so that learners utilise L2 phonological decoding skills instead of previously memorised pronunciation patterns (Woore, 2018).
- use of poems illustrating focal SSCs (see Woore's observation-study, 2007).
- a systematic, long-term approach (Woore et al., 2018) featuring regular input (Woore, 2007), part of a carefully planned sequence of learning (TSC, 2016) supporting written expression and reading (Woore, 2022).

2.4 Grammar

This section explains relevant approaches to grammar teaching, establishing the English context for primary FL grammar teaching, then considers findings concerning young learners.

2.4.1 Approaches to teaching grammar

Focused L2 instruction increases learning gains (see Norris & Ortega's meta-analysis, 2001). This section addresses whether instruction should: focus on form, meaning or both; be deductive or inductive; be input-based or output-based.

2.4.1.1 Focus on form (FonF)

Norris and Ortega (2001) adopted Long's (1991) definition of three types of grammatical instruction: meaning-based (likely to promote incidental L2 acquisition); focus on forms, highlighting key grammatical features and teaching them sequentially in isolation; and FonF, combining grammatical features and meaning, reactive (relying on error correction and other unplanned focus on form, as opposed to the intentional, planned focus on forms). FonF activities include: opportunities to encounter meaning before addressing grammatical form; L2 forms essential to task completion; noticing of grammatical features; attempts to make instruction unobtrusive. Norris and Ortega (2001) concluded that both FonF and focus on forms activities benefit learning, although FonF has larger effect sizes.

2.4.1.2 Deductive versus inductive teaching

Deductive and inductive approaches, both types of explicit instruction ([2.1.2.2](#)), vary in degree of explicitness (Erlam, 2003), with inductive learning involving identification of patterns and form-meaning relationships by learners (Ellis, 1994). For Haight et al. (2007), the difference related to the point at which target grammatical structures are presented: the initial grammatical focus can *precede* (deductive) or *follow* (inductive) practice. Varied operationalisation of

inductive approaches in particular and diverse research contexts have produced inconclusive evidence regarding effectiveness (Erlam, 2003; Haight et al., 2007; Norris & Ortega, 2001).

Haight et al. (2007) operationalised the deductive approach as FonF before meaning and practice, and the inductive approach as grammatical analysis following oral practice of the key structure. University-aged learners ($N = 47$) in the inductive condition demonstrated largest short-/long-term gains. However, this study did not control for potential teacher effect, with conditions taught by different instructors.

Contrasting findings emerged from Erlam's (2003) New Zealand study: 69 14-year-olds in their second year of French received either deductive or inductive instruction on direct object pronouns, to investigate the impact on FL comprehension and production. Pupils in the deductive condition made statistically significant gains in both comprehension modes compared with the inductive condition learners and the control group, although advantages disappeared in listening delayed post-tests. Only results from the deductive condition were consistently higher than the control group. However, Erlam (2003) advised caution, given the low test reliability rates in comprehension tasks. Learners receiving deductive instruction also outperformed inductive and control groups in productive tests (in post-tests/delayed post-tests in writing; and in one immediate speaking test). These gains strengthen the importance of a deductive approach, although it should be noted that the medium and large effect sizes related to these gains were unsustained in delayed post-tests, whereas gains associated with the inductive condition decreased less over time. This finding aligns with Cerezo et al. (2016), in which 70 English-speaking learners of Spanish were taught the "*gustar*" structure via videogame (guided induction) or in a traditional classroom setting (deductive instruction); those taught deductively retained less over time in oral and written production tasks. Cerezo et al. speculated that this resulted from the greater cognitive processing involved in guided induction.

DeKeyser and Prieto Botana's (2015) narrative review examined the importance of explicit instruction, concluding that learners can occasionally work out rules for themselves (inductive approaches). However, in time-restricted teaching contexts (typifying FL classrooms in England), learners benefit from focused explicit instruction establishing form-meaning connections quickly. A further advantage is that *all* learners gain, as Erlam (2005) concluded that deductive teaching combined with language production may compensate for individual aptitude differences ([2.4.1.5](#)). The importance of output is examined after consideration of input.

2.4.1.3 Input processing

This describes how learners use the input to which they are exposed, the form-meaning connections made (VanPatten, 2017). This intake is held in working memory (VanPatten, 2002), but the amount that can be processed is restricted because capacity is limited. Consequently,

learners only attend to the most salient grammatical features in inputs, resulting in diminished communicative value of certain grammatical forms redundant in the overall sentence meaning; lexical items therefore assume greater importance for learners than grammatical form (VanPatten, 2002). To exemplify, in the sentence “*mes parents parlent*” (‘my parents speak’), the plural noun indicates that more than one person is speaking, and learners therefore pay less attention to the plural verb ending “*parlent*” (and in the oral modality, this inflection is inaudible). Because of this tendency to process meaning before form, providing opportunities to focus on form within examples that do not compete for learners’ attention is important (e.g., by stripping away temporal adverbs which would indicate tense).

2.4.1.4 Processing instruction (PI)

PI is a form of grammatical instruction, or pedagogical intervention (VanPatten, 2017) developed from input processing, which encourages learners to override natural inclinations to focus on lexical rather than grammatical features (VanPatten, 2002). PI focuses on comprehension activities rather than production, although VanPatten (2002) reiterated the importance of output in developing fluency and communication. PI usually includes (VanPatten, 2002):

- explicit information about target features (overtly addressing the relationship between form and meaning - DeKeyser & Prieto Botana, 2015).
- strategic advice on avoiding common processing errors.
- input practice focusing attention on key grammatical forms.

Structured input practice comprises referential and affective activities (VanPatten, 2002). In the former, learners derive meaning through their understanding of form, for example, selecting the correct picture to match written or spoken input (Henry et al., 2009). Answers are correct or incorrect. In affective structured input activities, feedback does not target the grammatical feature (Marsden & Chen, 2011); answers are personalised, as learners express opinions, beliefs, or affective responses to statements embedding the feature related to the real world (e.g., agreeing/disagreeing). Affective activities contextualise target forms (VanPatten, 2017), assuming that learners will process them better than prior to the referential activities. In VanPatten and Cadierno’s study (1993, cited in VanPatten, 2002), 80 learners were assigned to one of three groups: PI; traditional instruction; or control. Learning gains on word order and object pronouns in Spanish were identified for the PI group in interpretation tests; durable gains were also found for both treatment conditions in productive tests, indicating that PI (referential plus affective activities, preceded by explicit information about the feature) helps production as well as comprehension.

More recently, Marsden (2006) tested 13-14-year-old L1 English learners of L2 French in interpretation and verb production, comparing a PI (explicit information plus referential plus affective) approach with enriched input (explicit information plus an affective input task), comprising identical grammatical input as PI but not requiring attention to form or meaning. In the first experiment, learners in the PI condition outperformed those in the enriched input group on all measures, with durability of gains indicated by statistically significant results on delayed post-tests. Even though learners in the enriched input were exposed to the same number of examples of the target grammatical feature and the initial explicit information, learners did not appear to process key forms as well as the PI group. A second experiment, identical but with a different school, produced similar outcomes. As both groups had received explicit information at the start of the study, Marsden (2006) concluded that explicit information could not explain the PI group's greater success, otherwise the enriched input group would have shown similar improvement.

The benefits of explicit information were also noted by Henry et al. (2009) as learners in a PI group 'with explicit information' condition processed sentences more quickly than those in PI 'without explicit information' ($N = 38$). Additionally, when explaining strategy use after tests, these learners made specific reference to their understanding of relevant grammatical rules.

2.4.1.5 Input-based versus output-based

In comprehension-based instruction (CBI), learners do not produce language but encounter key forms embedded in input; production-based instruction (PBI) *may* involve comprehension but also requires learners to produce key features, on which they receive feedback (Shintani et al., 2013). Shintani et al.'s meta-analysis (2013) concluded that CBI appears to have an initial advantage over PBI, as gains were apparent (although not durable) in immediate tests for the comprehension mode. Notably, CBI involving structured input in PI studies provided advantage for learners in receptive modes. In the productive mode, PBI led to greater, more durable gains than CBI, indicating that it may consolidate newly acquired features. Importantly, both approaches led to higher gains than those achieved by control groups.

DeKeyser and Prieto Botana's (2015) narrative review also compared studies focusing on the advantages of PI and PBI conditions. Compared with PBI conditions, PI consistently led to greater learning gains in comprehension tasks and was more beneficial or equal in the productive mode. Shintani et al.'s (2013) meta-analysis produced similar findings in the comprehension mode. For production, PBI and PI were equally effective, although adding explicit instruction to PBI increased learning gains. This appears to support Erlam's (2005) assertion that all learners benefit from a combination of productive tasks and explicit instruction.

Considering production, DeKeyser and Prieto Botana (2015) advocated situating production-based activities within communicative tasks, rather than mechanical drills, to maximise benefit. Hopman and MacDonald (2018) recommended measurements requiring production of utterances involving forced recall, including the ordering and planning of words into utterances, rather than simple verbal repetition. DeKeyser and Prieto Botana (2015) warned that outcomes are partly determined by the nature of the measurement, recalling that few studies have been completed with young learners. This matters, as Shintani (2015) speculated that the voluntary attention to grammatical features necessary in PI alone might better motivate adult learners and may be difficult to monitor in classroom environments. However, form-focus in input-based instruction could be easier to achieve in classroom contexts compared with making the grammatical focus an essential element of an output task (Marsden & Chen, 2011). Shintani et al.'s (2013) conclusion seems sensible: both PBI and CBI have roles to play, and combining both approaches is advisable.

2.4.2 Grammar teaching in the English primary FL context

The current National Curriculum for FL (learners aged 7-11, DfE, 2013a) refers explicitly to grammatical features such as gender, linguistic patterns, sentence-building and high-frequency verbs, and the use of grammatical structures to support comprehension and written/oral expression.

Expectations in this National Curriculum – focusing on skills and outcomes rather than specifying content - may lack clarity (Tinsley & Board, 2017), but they recall content from the *Key Stage 2 Framework for Languages* (DCSF, 2005) which, although non-statutory, provided detailed guidance on expectations and pedagogical approaches for young learners prior to FLs becoming compulsory. National surveys and inspections after this document was introduced revealed issues in primary FL grammar teaching. For example, only in schools with better provision were pupils beginning to spot language patterns and demonstrate understanding of basic grammar such as plurality and gender (OFSTED, 2008a); writing opportunities were limited, often only offered in upper primary (OFSTED, 2011). Prioritisation of oracy over literacy was also reported by Graham et al. (2014) and Cable et al. (2010), with verb inflections causing particular problems. Despite some helpful comparisons with (L1) English supporting pupils' developing grammatical understanding (Cable et al., 2010; OFSTED, 2011) and provision enabling *some* learners to create sentences demonstrating syntactic knowledge, most production was rooted in formulaic utterances, with weak verb knowledge “a limiting factor” (Cable et al., 2010:7).

National surveys have repeatedly identified issues in FL grammar teaching. The teaching of writing including grammatical understanding was considered: a strength by only 5% of

responding schools (Tinsley & Board, 2012); potentially problematic for non-specialist teachers (Board & Tinsley, 2014); a major challenge/quite challenging for 80% of responding schools in 2014-15 (Board & Tinsley, 2015). Cable et al.'s (2010) finding that primary FL provision was characterised by memorised language and formulaic phrases has been echoed by others (e.g., Mitchell & Myles, 2019). In a transition study, Courtney (2017) identified a marked contrast between provision in the final year of primary (with 10-11-year-old pupils often unable to produce utterances containing verbs) and the first year of secondary school, evidencing a focus on grammar explanations – see [2.4.3.4](#).

Such tendencies apparently contradict key messages from published research (summarised by Edelenbos et al., 2006; Edelenbos & Kubanek, 2009; Johnstone, 2009) highlighting the value of early exposure to reading and writing and benefits of focusing on form *and* meaning. The next section explores whether this apparent reluctance to teach grammar to young learners is related to assumptions about how they learn.

2.4.3 Young learners and grammar

2.4.3.1 Implicit versus explicit learning

According to Roehr-Brackin and Tellier (2019), the teaching of grammar to young learners may be less prevalent because of the widespread belief that children learn implicitly, rather than explicitly (see, e.g., the summary by Nikolov & Mihaljević Djigunović, 2006). However, implicit learning requires considerable time and exposure, neither characteristic of English primary FL classrooms (Tinsley, 2019). Explicit teaching suits classroom contexts (Roehr-Brackin & Tellier, 2019), yet some may argue that young learners aged 7-11 lack the cognitive maturity necessary for such methods.

Evaluating the potential development of young learners' metalinguistic awareness, a Canadian experiment by Bouffard and Sarkar (2008) involved 8-year-old L1 English learners of French ($N = 43$) in reflection on their free oral production. Young learners noticed mistakes and suggested corrections, demonstrating an apparent capability in grammatical analysis at odds with the ability to learn only implicitly. This skill developed from guided talk with the researcher, demonstrating that, when instructed explicitly, young learners can make progress in this area.

2.4.3.2 Explicit teaching of young learners

Although children are assumed to learn implicitly, explicit instruction in grammar, spelling and punctuation constitutes a major part of their English (as L1) school education (DfE, 2013b). Non-statutory guidance for primary FL teaching (DCSF, 2005:27) expected pupils to discuss how language works, referring to their "relative cognitive maturity". Indeed, unlike youngest learners, 7-11-year-olds are developing greater metalinguistic awareness and logicity in their

thinking, drawing on a (usually) well-developed L1 in terms of breadth and complexity (Philp et al., 2008). Lichtman's (2016) study, involving 40 5-7-year-olds, investigated whether children learn implicitly or explicitly, speculating that the assumption of implicit learning arises from a predominance of research with adults and children in immersion settings. Lichtman based her study on two hypotheses: 'maturational', where children learn implicitly because of their young age and corresponding cognitive developmental stage; and 'instructional' - children learn the way they are taught. If teachers adopt "whole-language" instruction (2016:710), basing FL learning on songs, stories and games, rather than focusing on grammatical form and rules, learners' outcomes will reflect progress made through these approaches.

Adults and children, assigned to either implicit or explicit conditions, were taught an artificial mini-language (Lichtman, 2016). If the maturational hypothesis were responsible for how children learn, these learners should not make progress when taught explicitly. However, although adults learnt more quickly, children in both conditions could create sentences and recognise grammatically correct utterances. Nikolov and Mihaljević Djigunović (2006) suggested replicating naturalistic L2 acquisition opportunities in FL classrooms, but in Lichtman's (2016) study, children's grammatical understanding increased when instructed explicitly to focus on key forms. This supports the instructional hypothesis, indicating that, when taught explicitly, young learners develop and apply grammatical knowledge. Additionally, form-focused teaching may *develop* LAA in young learners (Roehr-Brackin, 2020, [2.4.3.3](#)).

Relevant to this study, Kasprowicz and Marsden (2018) compared input-based approaches teaching the Accusative case definite article to 9-11-year-old learners of L2 German in England. Pupils completed receptive and productive tests in written and spoken modalities. Differential gains were not apparent between the group which received training in establishing task-essential form-meaning connections (referential activities – [2.4.1.4](#)) and the group assigned to task-essential form-spotting (which necessitated noticing but omitted the need to focus on function, though did focus the learners' attention on the meaning of the word 'the'). However, both groups outperformed the control group (who only did the tests), demonstrating that young learners can develop metalinguistic awareness when taught explicitly.

2.4.3.3 Language-analytic ability

As discussed ([2.1.7](#)), high levels of individual variation in Kasprowicz et al.'s (2019) study indicated the influence of LAA on performance. In Roehr-Brackin and Tellier (2019:1111), pupils' pre-test language learning aptitude scores included LAA ("the ability to treat language as an object of analysis and arrive at linguistic generalizations"). These scores correlated significantly with pupils' final L2 proficiency scores, indicating the potential of aptitude to predict L2 achievement (see Roehr-Brackin et al., 2024; Tellier & Roehr-Brackin, 2013). Roehr-

Brackin (2020:5) suggested that this indicated more ‘adult’ patterns of learning, rather than implicit processes or memory capacity.

Erlam’s (2005) study observed that a deductive approach combining explicit rule explanation (of direct object pronouns in French) with opportunities for learners to produce language, benefited *all* learners, apparently compensating for individual levels of language learning aptitude (of which LAA is a component). Tellier and Roehr-Brackin (2013) concluded that the focus on form, nature and function of the language inherent in explicit teaching could stimulate children’s explicit learning processes (evidenced by learning outcomes of 8-9-year-old pupils after instruction in Esperanto). A structured, analytical approach to instruction could benefit lower-achieving students by contributing to “‘equalizing out’ the IDs [individual differences] which produce differences in achievement in less helpful circumstances” (Skehan, 1989:129).

2.4.3.4 Teaching verbs to young learners

Verb morphology is considered late to develop. In Graham et al. (2014), pupils struggled to produce present tense verbs in sentence repetition and photo description tasks. In Courtney’s (2017) study, focusing on primary-secondary transition, 46.6% of Year 6 learners’ utterances included a verb, increasing to 79.9% in Year 7. Knowledge of verbs characterised 17-18-year-old learners’ performance compared with noun-dominated language used by 13-14-year-olds in a study by Marsden and David (2008). In Rule and Marsden (2006), 80.7% of utterances produced by 11-12-year-old learners after approximately 40 hours of instruction at secondary school were verbless (out of 332 tokens of contexts when verbs were needed). Verbless utterances were also the norm when Mitchell and Myles (2019) tracked 7-8-year-olds taught through an oracy-led approach. After 38 hours of instruction, in the mid-test, pupils produced “*c’est*” (‘it is’) inconsistently in the Elicited Imitation Task and omitted it during role-play tasks requiring spontaneous production; other inflected verb forms were rare outside formulaic utterances.

2.4.4 Grammar: summary

Further research should ideally explore how different approaches to FL grammar teaching affect young learners. Grammar has often been considered off-putting for young learners (observed by Cenoz, 2003; noted by Tierney & Gallastegi, 2005), but in Muñoz’s (2014a) qualitative study into young learners’ awareness of FL learning, children completing the primary phase valued FonF activities as benefitting learning (as well as activities promoting oral production). The call for a pedagogy which enables young learners to grasp basic grammatical concepts (Graham et al., 2014) is not at odds with recommendations from TSC (2016) foregrounding the following:

- concise, explicit grammatical rule explanation, linked to L1 where appropriate.
- forced noticing of grammatical features in the receptive mode, linked to meaning.
- supported progression to production of grammatical features.

Non-specialist teachers lacking confidence or grammatical understanding may need support to achieve this (Driscoll, Jones & Macrory, 2004). However, their additional knowledge of learners (Low et al., 1995) and their understanding of L1 grammar teaching in the primary curriculum should be beneficial.

2.5 The two approaches: similarities and differences

This chapter concludes by presenting how research-informed principles outlined above in vocabulary, phonics and grammar were operationalised in the Novel approach, and how this compares with the Standard scheme. The evaluation will address research question (RQ) 1: for each French teaching approach, what are the effects on Year 3 pupils' performance in vocabulary learning, SSCs, and simple sentence manipulation in one academic year?

Table 2.

Commonalities and Differences of the Schemes of Work

Areas of language/practice	Topic-driven 'Standard' approach	Language-driven 'Novel' approach
Vocabulary	Vocabulary taught in 'chunks' as well as in the form of individual items (e.g., 'Quel âge as-tu?')	Vocabulary taught mainly as individual items; focus on manipulation
	Pupils supported at sentence-and word-level	
	Question forms & answers	
	Semantic clustering	Some semantic clustering (Marsden et al., 2023). Greater focus on range of word classes and thematic clustering; core sets of lexical items included (e.g., classroom items, numbers, days, Hawkes & Guillou, 2022)
	Larger number of items in each lesson (e.g., lesson 9, "j'ai" + nine nouns,	Fewer items in each lesson for active use, ca. four per lesson (Hawkes & Guillou, 2022)

	singular/plural forms, gender)	
	Written	word shown
	More nouns than other word classes	Balance of nouns, adjectives, pronouns, verbs
	No consideration of corpus-based frequency ranking	Lexicon draws on “corpora-informed frequency information” (Marsden & Hawkes, 2023:96); words mainly from the most common 2000 words (Hawkes & Guillou, 2022)
	Age-appropriate	vocabulary
		Use of “patterns in cognates and word families” (Hawkes et al., 2019:3)
	Reference materials available for learners	Some recall/recognition vocabulary activities promote working from memory and active retrieval (Marsden & Hawkes, 2023:107)
		Cumulative approach adopted (2.2.4.3)
		Some timed recall/recognition activities “to promote the development of automatised explicit knowledge” (Marsden & Hawkes, 2023:108)
Phonics		Structured, systematic approach to selecting/teaching SSCs: “informed by principles of difficulty and frequency and by practitioner knowledge” (Hawkes et al., 2019:2)
	SSCs presented in isolation/embedded in whole words	SSCs initially presented in isolation; practised when

		embedded in source words (Hawkes et al., 2019)
	Whole lessons initially dedicated to phonics (term 1); explicit teaching	“Short, discrete segments” (Woore, 2022:182) dedicated to phonics in (most) lessons; explicit teaching
	Initial focus on vowels	
	Practised through rhyme	
	Large number of SSCs covered quickly	Regular focus on smaller number of SSCs
		Minimal pairs activities in listening, reading aloud and transcription used for discrimination practice of phonemically similar SSCs most confusing for L1-English learners (Hawkes et al., 2019)
	Familiar/unknown words used to exemplify patterns	
		High-frequency vocabulary exemplifies SSCs
	Some opportunity for reading aloud	Regular opportunity for reading aloud, including unfamiliar words (including timed activities, with increase in speed one indicator of successful proceduralisation), (Marsden & Hawkes, 2023:106)
	Fewer opportunities for knowledge application	Greater application of phonics knowledge
Grammar	Some explicit explanation (gender, plural nouns, adjectival agreement of colours)	Regular, explicit grammar focus
		Initial brief explicit description of grammatical point prior to practice, to

		potentially “speed[s] up learning”, possibly “compensat[ing] for learners who tend not to analyse or break down language themselves (learners with lower ‘analytic ability’)” (Hawkes et al., 2019:5)
		Regular minimal pairs activities force learners’ attention onto key grammar points difficult for learners (e.g., features which have “low perceptual salience (i.e. ... difficult to hear in speech or notice in writing) or low communicative value (i.e. they often co-occur with other language features that reduce the chance of learners’ attending to them))” (Hawkes et al., 2019:5)
		Explicit focus on question formation
		Grammatical points included in input tasks before production: “an additional, intermediate stage of input language processing tasks, in which the noticing of grammatical forms and the recognition of their associated meanings (or functions) are essential” (Hawkes et al., 2019:4)
	Verbs taught/practised within chunks, set phrases	Verb manipulation
	Some sentence-level work (without explicit grammar focus)	Opportunities for sentence creation
		Grammatical features practised and revisited in new lexical contexts (Marsden & Hawkes, 2023): “noticing how the system works across the lexicon (not just in specific, routinised phrases)” (Hawkes et al., 2019:4)

	Comparisons made	with L1 English
Practice	Regular but unsystematic recaps of prior learning	Systematic recaps of prior learning in planned cycles (Marsden & Hawkes, 2023) including recent/less recent items: an “expanding practice schedule” (p.99). Vocabulary, after introduction, is revisited after 3 and 9 weeks; additional, incidental recycling (Hawkes & Guillou, 2022). Interleaving “used at various different levels of language representation (phonemic, morphological, lexical, and syntactic), modes (receptive, productive), modalities (oral, written), and at different levels of distribution (within a single activity, within a lesson, and across weeks, months, and even years), leveraging both spacing and interleaving” (Marsden & Hawkes, 2023:107)
Communicative activities	Yes	Yes. Tasks require “both the speaker(s) and the listener(s) to use the target feature to understand or produce meaning from the lexicon and/or from morphosyntax (e.g., the meaning of question words; information about tense, number, person, gender ...)” (Marsden & Hawkes, 2023:105-6)
Feedback	Yes, via teachers	Yes, via teachers; also embedded in PowerPoints, aiming to “establish[ing] reliable declarative knowledge for all learners” (Marsden & Hawkes, 2023:106)
Songs	Yes	Yes

Classroom instructions	Yes, in detail, taught as chunks	Yes, key verbs
Poems	Yes	Yes
Games	Yes	Yes
Cultural references	Yes	Yes
General		“Connection between phonics, vocabulary and grammar through the focus on frequency” (Hawkes et al., 2019:4) Vocabulary and grammar “systematically established and practised across different semantic and linguistic contexts” (Marsden & Hawkes, in press:4) Conceptualisation of phonics, grammar and vocabulary as “‘strands’ of progression ... braided together to form the overall ‘rope’ of ‘essential language knowledge’” (Woore, 2022:174) Aims to increase “transparency for learners, enabling them to understand patterns, make links within and between languages, tackle head-on what is most difficult, and connect form and meaning securely and with increasing confidence” (Hawkes et al., 2019:5)

Note. Blank cells indicate that this is not evident in the scheme or there is no information available indicating that it featured in the design. Entries appearing across Novel and Standard columns apply to both approaches.

Chapter 3: Literature Review II. Pupil and teacher motivation and related variables, focusing on research conducted with young learners

Motivation is accepted as a principal influence in successful language learning (Dörnyei, 1994, 1998; Dörnyei & Csizér, 1998; The Nuffield Foundation, 2000), potentially enhancing learner involvement and achievement (Pintrich & Schrauben, 1992). Subject to internal and external influences (Dörnyei, 2013), motivation explains the impetus to engage with activities (Dörnyei, 1998, 2013; Ryan & Deci, 2000), willingness to sustain this (Dörnyei, 1998; Dörnyei & Csizér, 1998), and effort invested in the process. The decision to persevere with something can be influenced by perceived self-efficacy (Bandura, 1986). In this study, measures related to these constructs for both pupils and teachers were compared across approaches.

Dörnyei's framework (1994) presented three main components of language learners' motivation: the *language level* explains motivation towards the FL, culture, and communities; the *learner level* draws on self-efficacy, anxiety, linguistic confidence, need for achievement; the *learning situation level* encompasses classroom experience relating to the course, teacher, and group. Recommendations address each level, including contact with the FL and culture, promoting self-belief in linguistic competence, and ensuring the appeal and relevance of course content and materials. Justification for the use of this framework is presented in [4.5.3](#).

This chapter first considers pupil motivation (including self-efficacy) ([3.1](#)), and then teacher emotions ([3.2](#)).

3.1 Pupil motivation

This section explores variables affecting positive attitudes and potential attitudinal gains which may prolong commitment to FL learning and possibly improve achievement.

3.1.1 Links between motivation (including self-efficacy) and attainment

Summarising early language learning research, Edelenbos et al. (2006) suggested that motivation, together with language aptitude, predicts linguistic proficiency (see also Murphy, 2014). In the ELLiE project, young learners displaying more positive attitudes towards FL learning achieved higher scores in listening comprehension and oral lexical proficiency tests (Mihaljević Djigunović & Lopriore, 2011). In contrast, in the 1960s' national pilot scheme in England to introduce French from the age of 8, although learner attitudes were positively affected, more motivated pupils were not necessarily higher-achieving (Burstall, 1975). A similar lack of positive correlation between learning attitudes and achievement was reported by

Courtney et al. (2017) among 11-year-old pupils, and Fenyvesi et al. (2020), which suggested that EFL enjoyment did not affect outcomes in receptive grammar and vocabulary tests. Motivation may not ‘cause’ better achievement; higher-achieving pupils may just enjoy learning FLs more or it may constitute a complex achievement-motivation cycle. For example, Alamer and Lee’s (2024) cross-lagged panel analysis explored the L2 anxiety-achievement relationship in L1 Arabic university-aged learners of English ($N = 226$), as higher levels of anxiety may impact negatively on motivation to learn and use the FL (Mercer et al., 2012, cited in Alamer & Lee, 2024). Higher achievement scores early in the project were found to decrease later anxiety, whereas achievement was not predicted by anxiety. Similarly, Alamer et al. (2025) found that higher scores in vocabulary knowledge reduced anxiety levels over time.

Some studies investigating the achievement-motivation relationship have included self-efficacy (3.1.6). A meta-analysis of 39 studies ($N = 4998$), 60.6% of which involved elementary-aged learners, explored the relationship between self-efficacy beliefs and academic performance and persistence (Multon et al., 1991), finding that ca. 14% of variance in academic performance (and 12% of persistence) was explained by learners’ self-efficacy beliefs. Erler and Macaro (2011) also found that positive feelings towards French correlated positively with self-efficacy ratings on decoding tasks, and decoding tests, in a study conducted with 11-14-year-old learners of French in England ($N = 1,735$). They also reported a moderate, significant correlation between decoding test scores on rhyme and word segmentation tasks with pupils’ intention to continue learning French beyond the age of 14. Pupils’ test results accounted for 8.2% of the variance in their decision to continue; combining test results, self-efficacy measures and positive feelings towards French increased this to 23.1%.

3.1.2 Motivational differences in young learners

Research suggests that EFL learning is viewed positively by young learners (Cenoz, 2003; Heinzmann, 2013; Nikolov, 1999). In England, primary-aged learners responded positively to the introduction of FL learning (Cable et al., 2010; Muijs et al., 2005; Myles & Mitchell, 2012; OFSTED, 2008a, 2011); these observations were supported by teacher perceptions of pupils’ positive attitudes towards FL (Bolster et al., 2004; Cable et al., 2010; Driscoll, Jones & Macrory, 2004; Low et al., 1993; Powell et al., 2001). However, positivity is not guaranteed (Enever, 2009): Tragant (2006), investigating EFL learning in Catalan-Spanish-speaking Catalonia, reported higher motivational levels among pupils starting at 11 than pupils starting at 8.

Empirical studies have shown that early language learning can encourage longer-term positive attitudes (Courtney, 2017; Courtney et al., 2017; Hunt et al., 2008; among others), considered a principal aim in England (DCSF, 2005; Department for Education & Skills, 2002; Wade et al., 2009). The evaluation of the 1996-8 MLPS initiative in Scotland highlighted young

learners' enthusiasm and favourable attitudes, although 18% of boys expressed negativity in their first year learning FL (Tierney & Gallastegi, 2005), emphasising the importance of positive learning experiences (Dörnyei, 1994).

Indeed, young learners are not a homogenous group (Mihaljević Djigunović, 2017): the ELLiE project reported attitudinal and motivational differences from the outset (Enever, 2011): the majority of pupils were positive towards FL learning, a quarter ambivalent, and a small number negative (Mihaljević Djigunović & Lopriore, 2011). Similarly, in Courtney et al.'s (2017) study almost one fifth of 10-11-year-old pupils ($N = 243$) reported not enjoying learning French and a third demonstrated negative ability beliefs. However, as Courtney et al. (2017) noted, negative attitudes may develop *during* primary-aged FL learning. In Cable et al.'s (2010) longitudinal study, 81.6% of Year 3 learners ($n = 136$) in 2006-7 and 86.7% ($n = 144$) in 2007-8 reported enjoying FL, but this decreased to 58.5% and 69.25% ($n = 133$) in Year 6. Realising the time and effort required to speak FLs fluently may demotivate young learners (Muijs et al., 2005; Myles, 2017), although finding a language difficult appears not to impact negatively on enjoyment (Butler, 2017).

3.1.3 Motivational orientations and young learners

This section introduces types of motivation demonstrated by young learners: instrumental, intrinsic and integrative. Understanding different motivational orientation matters because resulting experiences and outcomes may differ.

3.1.3.1 Instrumental motivation

Instrumental, or extrinsic, motivation is practically- and goal-oriented (Gardner & Lambert, 1972; Williams & Burden, 1997), with students possibly aiming to please others (Pintrich & Schrauben, 1992) or gain qualifications. Exemplifying instrumental motivation, the utility value of FL for travel/holidays was reported in primary-aged pupils' interview/questionnaire responses in the 2005 Pathfinder evaluation (Martin, 2012 - see also Cable et al., 2010; Courtney, 2017; Graham et al., 2014, 2016).

3.1.3.2 Intrinsic motivation

Intrinsic motivation is likely to be geared towards learning and mastery (Pintrich & Schrauben, 1992) and involves acting freely for the reward of the activity (Ryan & Deci, 2000), rather than external recognition (Deci & Ryan, 1985). Intrinsic motivation is often considered powerful in sustaining learning (thereby affecting achievement), for example: "being intrinsically motivated to learn improves the quality of learning" (Deci & Ryan, 1985:256), potentially enhancing cognitive engagement (Pintrich & Schrauben, 1992).

Young learners generally demonstrate intrinsic motivation (found by Cable et al., 2010; Carreira, 2006; Mihaljević Djigunović, 2009b; Mihaljević Djigunović & Lopriore, 2011; Nikolov, 1999, among others), describing FL learning as fun. Intrinsic motivation can be stimulated by satisfaction from intellectual challenge (Bandura, 1986; Deci & Ryan, 1985; Edelenbos & Kubanek, 2009; Ryan & Deci, 2000; Wu, 2003), and experiencing success on appropriately challenging tasks is beneficial (Johnstone, 2009; Wingate, 2018; Wu, 2003). Indeed, teachers in Cable et al.'s (2010) longitudinal study perceived primary-aged learners to be intrinsically motivated by their ability to use known language creatively. However, pitching challenge appropriately can be problematic. Learners need to feel that they are making progress, as this may help sustain motivation (as argued by Holmes & Myles, 2019; [3.1.4.2](#) provides further discussion), but this awareness may be lacking among learners in English schools (noted by Hawkes et al., 2019). Perceptions of the FL as too difficult may impact on pupils' desire to continue language learning (found by F. Taylor & Marsden, 2014).

3.1.3.3 Integrative motivation

Motivation can also be integrative, developed by positive contact with the FL community (Gardner & Lambert, 1972). Alongside intrinsic motivation (demonstrated through FL enjoyment), 9-10-year-old pupils in the MLPS initiative displayed integrative motivation through their desire to communicate with others in the FL (Tierney & Gallastegi, 2011 – see also Cable et al., 2010; Chambers, 2019; Courtney, 2017; Courtney et al., 2017; Graham et al., 2014, 2017). Learning about other countries and cultures also interests young learners (Graham et al., 2014; Muijs et al., 2005; Myles, 2017).

3.1.4 Variables contributing to young learners' positive attitudes

Understanding factors promoting positive FL attitudes is important as these may explain behaviour (Ben-Eliyahu, 2019). Interest, a “general liking of or attitude towards the task” (Pintrich & Schrauben, 1992b:158), may stimulate intrinsic motivation, potentially creating enjoyment (Deci & Ryan, 1985).

3.1.4.1 Classroom activities

Songs, drama, stories and games, typifying interactive provision observed by, among others, Muijs et al. (2005), Graham et al. (2014), Courtney (2017) and Mitchell and Myles (2019), are thought to promote learning through engagement (DCSF, 2005). Indeed, this ‘fun’ approach appears on first appraisal to make FL learning more appealing than other subjects (Powell et al., 2001).

3.1.4.2 Sense of progress

Early in FL learning, learners appear motivated by learning new words (Cable et al., 2010; Mihaljević Djigunović & Lopriore, 2011), equating learning with vocabulary acquisition (Muñoz, 2014a). Whilst noting learner engagement, Mitchell and Myles (2019) argued that, if only minimal progress is secured, motivation may be jeopardised and pupils' self-efficacy negatively impacted. Fun by itself may be insufficient: "enjoyment is more than mere pleasure. It implies challenge, hard work, and a sense of fulfillment at completing something that is really difficult" (Dewaele, 2022:201). Indeed, learners in the Pathfinder evaluation (Martin, 2012) demonstrated frustration and decreased commitment at perceived lack of progress in the second year of learning, whereas teachers in Cable et al.'s study (2010) reported increased motivation and positivity among pupils who experienced a sense of pride in their achievement. Upper-primary-aged learners wanting greater linguistic challenge responded to games less positively (see also Butler, 2017; Chambers, 2019; Fenyvesi, 2020). Games risk signalling that FL learning is light-hearted, often lacking challenge beyond single words, representing "negative correlation between time investment and learning outcomes" (Wingate, 2018:451).

3.1.5 Variables contributing to young learners' negative attitudes

Young learners report few dislikes in FL learning (Courtney, 2017; Mihaljević Djigunović & Lopriore, 2011). This section discusses relevant aspects of FL learning reported as demotivating for learners: writing and grammar ([3.1.5.1](#)) and negative learning experiences ([3.1.5.2](#)).

3.1.5.1 Writing and grammar

FL writing is not always positively perceived (Mihaljević Djigunović & Lopriore, 2011; Nikolov, 1999), considered demanding (Courtney, 2017), unappealing (Chambers, 2019), and uninteresting when linked to worksheets (Graham et al., 2016). Form-focused activities were cited in interviews as less popular among 76 primary-aged Catalan-Spanish EFL learners (Muñoz, 2014a). However, in Cable et al.'s (2010) longitudinal study, some upper-primary-aged learners requested *more* writing. Planning decisions to make grammar subordinate to 'fun' (singing, games) approaches in the MLPS initiative aimed to avoid learner anxiety (Tierney & Gallastegi, 2005), but perhaps contributed to pupils' difficulties in extending and manipulating spoken language (noted by Low et al., 1993). As the National Curriculum (DfE, 2013a) includes oracy *and* literacy objectives, we should understand how oral and written activities (or combinations thereof) might impact on learners' engagement.

3.1.5.2 Negative learning experiences

Tragant (2006) found learners' (average age: 10.9) dislike of EFL usually related to negative learning experiences; positive classroom experience was also considered crucial in sustaining motivation in primary-secondary transition (Graham et al., 2016). Repetition of topic-related vocabulary (Cable et al., 2010), task type (Fenyvesi, 2020) and task and content (Graham et al., 2016) has been found to be demotivating, as has choral repetition (Courtney, 2017; Martin, 2012).

3.1.6 Pupil self-efficacy

An important component of social cognitive theory (Bandura, 1986) differing from *general* confidence, self-efficacy explains individuals' beliefs in their ability to succeed in *specific* tasks. Learners may over/underestimate ability but self-perception influences decisions regarding actions taken and duration of commitment (Bandura, 1986). For example, as argued by Phakiti (2025), self-perceived higher levels of linguistic self-confidence may increase willingness to engage in FL learning (see also Dörnyei & Csizér, 1998). Sources of self-efficacy originate in four principal areas: previous attainment; vicarious experiences (based on observations of others' success and failure); verbal persuasion and social influences; and physiological states (Bandura, 1986:399). School contexts may affect all these sources. Competence judgements arising from personal experiences are considered the main component of self-efficacy.

3.1.6.1 Why self-efficacy matters

Self-efficacy has been hypothesised to influence learner participation, effort, persistence and achievement (Wigfield & Eccles, 2002); belief in potential success in tasks may affect commitment (Heinzmann, 2013). In a longitudinal study tracking 254 9-12-year-old L1 English learners, self-efficacy explained variance (after L1 literacy levels and school context) in pupils' performance on a French sentence reproduction task (Courtney et al., 2017).

Higher self-efficacy may impact positively on enjoyment, with positive correlations between these constructs found in primary-aged learners (Courtney et al., 2017); this may engender lasting positive FL attitudes. Motivation among learners of English and French in German-speaking Switzerland increased when pupils felt confident about their current and future FL competence (Heinzmann, 2013). Lacking confidence and experiencing anxiety had a particularly negative impact when combined with low EFL competence beliefs: 9-10-year-old Danish learners made little gain in vocabulary when affected by both low competence beliefs and high anxiety, although younger learners were less affected (Fenyvesi et al., 2020).

3.1.6.2 Observed and perceived variables impacting on self-efficacy

Time spent learning FLs might be expected to increase self-efficacy (Mihaljević Djigunović, 2009a). Indeed, in Graham et al.'s (2014) study comparing pupil outcomes in oracy- and literacy-based approaches, learners' motivation and self-efficacy levels, already high in the final year of primary (Year 6), increased significantly at the start of Year 7. Learners' perceived progress led to further increases in self-efficacy by the end of Year 7, although motivation levels had decreased, influenced by changes in pedagogical approach. Higher levels of self-efficacy among learners experiencing oracy-based approaches were only apparent in Year 6. However, younger learners typically demonstrate greater self-efficacy than older learners (Eccles et al., 1993; Fenyvesi et al., 2020; Muenks et al., 2018; Wigfield & Eccles, 2000), with this difference perhaps explained by older students' more realistic views of self-competence and greater understanding of teachers' feedback.

Evaluations of the implementation of primary languages provision included data relating to teacher perceptions of pupil confidence, which showed an overall increase in studies conducted in England (Powell et al., 2001) and Scotland (Low et al., 1993). Cable et al. (2010) reported that teachers perceived a positive impact on pupils' confidence arising from a sense of achievement in their FL progress. Teachers also reported noting increased confidence when pupils were speaking French, as they benefited from the repetition of new content (Driscoll, Jones & Macrory, 2004); such confidence in speaking among primary learners with experience of FLs was also noted by secondary teachers in Bolster et al.'s transition study (2004).

3.1.7 The teacher's impact on pupil motivation

Young learners' opinions are not yet fixed (Mihaljević Djigunović, 2009b) and, given their considerable time in school (Eccles, 1999), it is unsurprising that teachers can influence their attitudes, as: noted by Ben-Eliyahu (2019) and Mihaljević Djigunović (2009b); reported by pupils in interviews (Chambers, 2019; Muñoz, 2014a) and questionnaires (Tierney & Gallastegi, 2011); perceived by teachers in a questionnaire-study (Mierzwa, 2019). The influence of classroom experiences and teachers was identified by Dörnyei (1994) as the learning situation component of L2 learners' motivation. In Muñoz's (2014a) comparative study of students' perspectives as FL learners, 11-12-year-olds demonstrated greater self-regulation and less teacher influence than younger learners. The teacher was also more important for the youngest Danish EFL learners (aged 8), indicating greater reliance on external authorities (Fenyvesi, 2020; Fenyvesi et al., 2020) and perhaps reflecting the strong emotional attachments which young pupils form with teachers (Mihaljević Djigunović & Vilke, 2000; Nikolov, 1999). Teachers' enthusiasm may also influence pupils' positivity (noted by Martin, 2012). When introducing L2

English to 6-year-old Croatian pupils, teachers were advised to ensure a “happy and relaxed” learning atmosphere (Mihaljević Djigunović & Vilke, 2000:68).

3.1.8 Pupil motivation: summary

Causal motivation-achievement relationships are difficult to establish: learners may display greater motivation *because* they are successful FL learners. Wigfield and Eccles (2002) advocated recognising the reciprocal impact of achievement and motivation, rather than focusing on causal relationships. The largely oral approach adopted in primary FL learning in England is assumed to motivate young learners (e.g., by Cable et al., 2010), yet pedagogical methods in primary and secondary need greater alignment to avoid negative impact on pupil motivation, achievement and self-efficacy (Courtney, 2017; Courtney et al., 2017).

Graham et al. (2016) did not find significant differences in impact on young learners’ motivation from literacy- or oracy-based approaches. A key rationale for primary FL teaching is promoting enjoyment to instil motivation; given that longer-term motivation remains low overall, an alternative approach seems worth exploring. Measuring the impact of a progress-promoting approach would be beneficial, to determine if satisfaction from optimal challenge (in terms of achievement in understanding and using the language), could engender equivalent or superior sustained motivation. Following negative responses to literacy work consisting of worksheets, copy writing and text-book activities (found by Graham et al., 2014), careful consideration of how best to integrate a carefully sequenced, knowledge-based curriculum with communicative speaking activities is needed. This aspect of the study addresses the following RQs: for each approach, what is the (perceived) effect on pupils’ self-efficacy and enjoyment? (2a); what are pupils’ and teachers’ perceptions of the effect on pupils’ progress in French? (2b).

3.2 Teacher emotions

Teacher emotions influence teachers, teaching and learners (Burić et al., 2018; Frenzel, Goetz, Stephens, et al. 2009), with enjoyment, anxiety and anger identified as common emotional teaching responses (Frenzel et al., 2015). Much L2 research has focused on negative emotions (Dewaele et al., 2018), with comparatively little emphasis on teacher enjoyment (Mierzwa, 2019) or the relationship between teacher and pupil enjoyment (Becker et al., 2014). As consideration of both positive and negative emotions is important (Dewaele et al., 2018), this section examines teachers’ linguistic and pedagogical insecurity ([3.2.1](#)), anxiety ([3.2.2](#)), enjoyment and enthusiasm ([3.2.3](#)).

3.2.1 Linguistic and pedagogical insecurity

3.2.1.1 Definitions and impact

Perceived competence in linguistic and pedagogical expertise may influence teachers' attitudes towards FL teaching. Linguistic insecurity may describe feelings experienced by non-native speakers lacking confidence in their knowledge of grammar, vocabulary and pronunciation required to teach effectively (Daftari & Tavit, 2017). Although recognised in learners (Horwitz et al., 1986), teachers' FL anxiety is less widely researched (Horwitz, 1996). The semantic field for anxiety is defined as "anxiety/anxious, tense, nervous, worry/worried, uneasy" (Frenzel et al. 2016:150). Anxiety may occur when teachers doubt their competence to achieve classroom goals, resulting in physical and emotional responses (Amengual Pizarro, 2019) and possibly affecting delivery (Frenzel, Goetz, Stephens, et al., 2009; Horwitz, 1996). In Amengual Pizarro's (2019) questionnaire-study of 75 prospective primary EFL teachers in Spain, language proficiency negatively predicted FL teaching anxiety, but emotional responses may differ according to the individual (Frenzel et al., 2015) and teaching context (Frenzel et al., 2015; Frenzel et al., 2016), including learners' FL level (T. Farrell & Richards, 2007). As language learners themselves, non-specialist primary teachers may also suffer *learner* anxiety.

Combined perceived linguistic and pedagogical competence is essential for effective teaching (Choi & Lee, 2016). If teacher knowledge influences instructional behaviour (T. Farrell & Richards, 2007; Rohaan et al., 2012), perceived or real insufficient linguistic and pedagogical competence may affect primary teachers' FL teaching in England (Woolhouse et al., 2013). In school-based case studies in New Zealand, the earlier introduction of FL to 11-year-old pupils involved teachers with lower levels of subject knowledge teaching FL for the first time. H. Richards et al. (2013) reported negative impacts on teaching practice, many identified by Horwitz (1996) as commonplace among anxious teachers, including: reduced ability to explain accurately and meaningfully, including new vocabulary (see also Tsui, 2003; T. Farrell & Richards, 2007); less accurate and spontaneous feedback (see Andrews, 2007; T. Farrell & Richards, 2007); and less confident correction of pronunciation. Insecurity experienced by teachers teaching outside their field of expertise can induce less positive student-teacher relationships (Du Plessis, 2015); teaching FL as a new subject may even challenge teachers' identities as competent practitioners (Woolhouse et al., 2013). Reduced levels of perceived competence may result in certain concepts being avoided in lessons (Du Plessis, 2015). Indeed, limited FL literacy input was observed in the primary FL longitudinal study in England, attributed partly to teachers' lack of confidence and expertise (Cable et al., 2010). Horwitz (1996) also warned that high levels of teacher FL anxiety may be perceived by students as negativity about language learning.

3.2.2 Aspects of linguistic competence causing particular concern

Research has repeatedly referenced teachers' concerns in linguistic proficiency (Board & Tinsley, 2014; Cable et al., 2010; Legg, 2013; Muijs et al., 2005), emphasising pronunciation (Kasprowicz, 2025; Tinsley & Board, 2016) and grammar issues (Macrory & McLachlan, 2009). Although Macrory and McLachlan (2009) highlighted grammar as concerning for prospective primary practitioners, in a study of self-efficacy beliefs among EFL teachers in Venezuela (Chacón, 2005) teachers' confidence to teach grammar was high (see also Daftari & Tavit, 2017). Grammar can be prepared, whereas spontaneous oral response makes different (public) demands on teachers' knowledge. Indeed, Horwitz (1996) and Amengual Pizarro (2019) suggested that, just as speaking can provoke high levels of anxiety in FL *learners* and affect communication (Dewaele, 2019; Horwitz et al., 1986), this skill may be most challenging for *teachers* (see also Machida, 2016). Anxiety in speaking may lead to limited interaction with learners and controlled activities. A survey of L1 Japanese EFL teachers revealed a causal relationship between greater confidence in practitioners' own pronunciation and positive attitudes towards teaching pronunciation to learners (Uchida & Sugimoto, 2019).

Greater experience of FL teaching may boost confidence levels (Cable et al., 2010; Cooke & Faez, 2018; Muijs et al., 2005; Rowe et al., 2012). Teachers with more general teaching experience may also feel less insecure (Daftari & Tavit, 2017), having perhaps developed strategies for teaching unfamiliar subjects (Machida, 2016). Fortunately, linguistic training can alter teachers' views of their language competence (Machida, 2016; Woolhouse et al., 2013). For example, although concerned about subject knowledge, 80% of prospective non-specialist primary teachers ($n = 150$) in England felt generally able to provide accurate linguistic models for young learners after training (Rowe et al., 2012). Possibly, *perception* matters more than *actual* competence (Dörnyei & Csizér, 1998; Horwitz, 1996; Pattison, 2014; Sharpe, 1995). Schemes of work and materials can also be supportive and model good pronunciation (Driscoll, Jones, Martin, et al., 2004; Muijs et al., 2005).

3.2.3 Teacher enjoyment and enthusiasm

Enjoyment is an experienced emotion, whereas enthusiasm explains its physical manifestation. Teacher enthusiasm is often discussed as two categories: behavioural/displayed; and affective/experienced (Keller et al., 2018; Keller et al., 2016). Displayed enthusiasm may be non-verbal, or may be observed in teachers' classroom practices conveying "energy and excitement" (Keller et al., 2016:747; Frenzel et al., 2018). Keller et al. (2016) recommended integrating both constructs in research.

Teachers' positive emotional responses, like anxiety, may be context-dependent (Keller et al., 2018), affected by groups or subjects (Frenzel et al., 2015), and positively impacted by

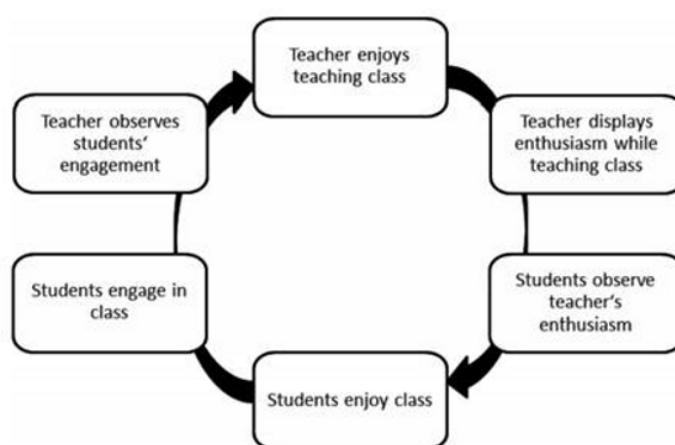
greater FL proficiency or knowledge of other languages, as found by Dewaele & MacIntyre (2014).

3.2.3.1 The importance of enjoyment

Teacher enthusiasm helps establish a positive learning environment (Fauth et al., 2019), contributing to student learning (OECD, 2005) and facilitating clearer, more comprehensible explanations (Becker et al., 2014). Developing the Teacher Emotions Scales, Frenzel et al. (2016) found positive correlations between teacher enjoyment and desirable classroom practices, such as instructional clarity and support. Teacher enthusiasm, while unrelated to pupils' achievement, stimulated primary-aged learners' interest in science in a correlational study (Fauth et al., 2019). Additionally, learners' FL enjoyment (to a greater extent than anxiety) has been linked to teachers' instructional behaviour (Dewaele et al., 2019; Keller et al., 2018) and to perceived teacher positivity (Frenzel et al., 2018:670, [Figure 1](#)).

Figure 1.

Reciprocal Effects Linking Student and Teacher Enjoyment



Learners' engagement may also increase teachers' positive emotional responses reciprocally (Frenzel et al., 2018), and learner enjoyment can mitigate negative emotions such as anxiety, and boost motivation (Mierzwa, 2019).

3.2.4 Teacher emotions: summary

Research has found generally high levels of enthusiasm among serving and prospective primary FL teachers in England (e.g., Cable et al., 2010; Muijs et al., 2005; OFSTED, 2008a, 2008b). However, insecurity persists (Pattison, 2014; Woolhouse et al., 2013). The challenge of ensuring substantial progress (DfE, 2013b) may affect teachers' enjoyment and anxiety, with many

lacking opportunities to develop linguistic proficiency, and schools repeatedly identifying confidence as an area for development (Board & Tinsley, 2014, 2015; Collen & Duff, 2024, 2025; Tinsley, 2019; Tinsley & Board, 2016, 2017; Tinsley & Doležal, 2018). As teacher emotions are related to learner emotions (Becker et al., 2014; Frenzel et al., 2018; Frenzel, Goetz, Lüdtke, et al., 2009), maximising teacher enjoyment and instructional enthusiasm may promote learners' positive emotional responses, especially as anxiety is likely to inhibit enthusiastic teaching behaviour (Frenzel, Goetz, Stephens, et al., 2009).

Graham et al. (2016:686) advocated exploring the impact of “contrasting teaching approaches experienced in the primary school” on *pupil* motivation, but it is also unclear how different pedagogical approaches may increase *teachers'* confidence, competence and enjoyment, and reduce anxiety. None of the literature reviewed here has addressed this question, for any subject area, including FLs at primary school. The current study aims to address this gap, responding to the following RQs: for each approach, what are teachers' perceptions of the effect on their motivation (3a) and their developing subject knowledge (3b)?

Chapter 4: Methodology and methods

Discussion has thus far focused on pedagogical approaches to teaching FLs in English primary schools, and issues pertaining to young learners' motivation and teachers' confidence and competence. This chapter now explores the rationale for a quasi-experimental study evaluating two different FL teaching approaches, encompassing investigation of learner outcomes and affective variables and impact on teachers. It presents methodological context, ethical issues, methods and analysis.

4.1 Aims, rationale and research questions

4.1.1 Aims

The overarching aim was to evaluate the impact, in time-poor contexts with non-specialist teachers, of a novel (language-driven) approach to teaching French on the learning outcomes, self-efficacy and enjoyment of 7-8-year-old learners, and on teachers' confidence, enthusiasm and subject knowledge, when compared with standard, topic-driven pedagogy. The language-driven ('Novel') approach, adopting some of the pedagogical choices recommended and operationalised by NCELP (2018), prioritised high-frequency vocabulary, systematic phonics, and sequenced and meaningful explicit grammar practice; the 'Standard' approach taught topic-based vocabulary and non-systematic phonics, with little explicit grammar explanation or practice (see [Table 2](#)).

4.1.2 Rationale

This section presents the rationale for this study linked to three focal areas: pupil outcomes, and pupil and teacher affective variables.

4.1.2.1 Impact on learning outcomes

Thus far, there has been little research into how different pedagogical approaches affect FL learning outcomes in time-limited settings (Mitchell & Myles, 2019). This study has potential to inform our understanding of what young learners can achieve in such instructional contexts, including what constitutes "substantial progress" (required by the DfE, 2013b:194). It will support practitioners in making best use of the time available for FL learning and feed into our understanding of how young learners learn FLs, comparing progress learners make through a standard approach with the impact of one which prioritises high-frequency vocabulary, and explicit, sequenced SSCs and sequential grammar.

This focus on young learners is important. In Andringa and Godfroid's (2020) examination of applied linguistics meta-analyses published in the preceding five years (863

studies), only 25% included learners from primary or secondary schools (see also Godfroid & Andringa, 2023). Graham et al. (2014, 2017) and the *White Paper: Primary languages policy in England - The way forward* (Holmes & Myles, 2019) encouraged research development in age-appropriate FL pedagogies. This is not new: Mitchell (2000:298) called for “process-product research ... both on the macro scale of programme evaluations and the micro scale of classroom experiments and quasi experiments” investigating the learning outcomes-instructional processes relationship and supporting teachers in deciding how best to teach particular groups of learners.

4.1.2.2 Impact on pupils’ affective variables

This element of the study addresses requests (see Martin, 2012; Muijs et al., 2005; among others) for challenging activities promoting pupil progress, comparing the impact of these on pupils’ enjoyment and engagement with that engendered by the existing teaching approach. Using LAA measures as a baseline indication of pupils’ language learning potential, this study adds to our knowledge of how this aspect of learning aptitude can predict FL learning success, and how it is affected by different teaching approaches.

The study has potential to develop our understanding of how young learners respond to learning high-frequency vocabulary as opposed to topic-driven language, and its impact on their enjoyment. Their response to explicit teaching of the sounds and patterns of the language will be evaluated, as well as to a combined oracy and literacy approach, involving explicitly taught grammar, an area which has produced mixed reactions in young learners (compare, e.g., Cenoz, 2003, and Tierney & Gallastegi, 2005, with Muñoz, 2014a).

4.1.2.3 Impact on teachers’ affective variables

Alongside inadequate teaching resources, Myles (2017) foregrounded as areas of concern non-specialist teachers and inadequate professional development. This study explores young learners’ outcomes when taught uniquely by non-specialist teachers, acknowledging this as an integral part of FL provision in the English context. It extends Rowe et al.’s (2012) work on the impact of FL teaching experience and guidance on non-specialists’ confidence and competence, considering whether teachers feel better supported by either approach. It presents the opportunity to gauge non-specialist teachers’ responses to a clearly defined body of language, including the perceived impact on their confidence when teaching French. Non-specialist teachers, as language learners themselves, may experience anxiety, and this study will evaluate whether this increases or decreases when teaching each approach.

4.1.3 Research questions

In summary, this study addresses for each French teaching approach the following RQs:

- RQ1: What are the effects on Year 3 pupils' performance in vocabulary learning, SSCs, and simple sentence manipulation in one academic year?
- RQ2a: What is the (perceived) effect on pupils' self-efficacy and enjoyment?
- RQ2b: What are pupils' and teachers' perceptions of the effect on pupils' progress?
- RQ3a: What are teachers' perceptions of the effect on their motivation?
- RQ3b: What are teachers' perceptions of the effect on their developing subject knowledge?

4.1.4 Research design: experimental methods evaluating educational interventions

Comparing teaching approaches required an empirical evaluative study, measuring learning outcomes and investigating teacher and pupil responses to, and experiences of, each approach. More detailed discussion is provided in [Appendix A.1](#).

4.1.4.1 Randomised controlled trials (RCTs)

Relying on at least two groups – experimental and control, and/or comparison - RCTs potentially establish causal relationships between intervention and outcomes (Burtless, 2002; Cohen et al., 2018; Torgerson & Torgerson, 2008). Influencing factors (independent variables) are evenly represented across both groups, affecting outcomes equally. Random allocation of participants enables outcomes to be reliably attributed to the intervention by ruling out alternative interpretations (Burtless, 2002; Connolly et al., 2017; Cook & Campbell, 1979; Mosteller & Boruch, 2002; Torgerson & Torgerson, 2008).

4.1.4.2 Threats to validity

Randomisation minimises threats posed by selection bias, temporal changes and regression to the mean (RTM) (Marsden & Torgerson, 2012; Torgerson & Torgerson, 2008) (see discussion in [Appendix A.1](#)). Attrition remains problematic: in this research, group equivalency could be affected by individuals withdrawing from data collection or leaving the school. Keeping data collection demands manageable minimises attrition through choice. Secondly, even for pupils often absent, studies can adopt intent-to-treat analysis (Shadish et al., 2002), with participants analysed as if having received the assigned intervention.

(i) Fidelity to treatment

This is discussed with reference to this study in [4.4.5](#).

(ii) Real-world contexts

Innovative approaches should be relevant (Norris, 2009) and recognisable to teachers delivering interventions, to “operationalise the learning theory ... as classroom practice” (Marsden, 2007:579). Nevertheless, they should be distinct from comparison/control treatments, for meaningful differential to be observable (Gueron, 2002).

(iii) The teacher variable

In educational interventions, eliminating the impact on outcomes of the teacher variable is difficult (Beretta, 1992), particularly if comparing two approaches to teaching (Ross, 1992) or testing theory (Marsden, 2007). Studies have found impact on learners’ outcomes from teachers’ age, enthusiasm, FL proficiency and ability to adapt language to pupils’ needs (Duff, 1991), and teacher FL qualifications and behaviour when teaching particular courses (Edelenbos & Suhre, 1994). The teacher variable could be minimised by two teachers delivering both treatments (see Marsden, 2007), although even when teaching both, teachers’ preferences for one could affect outcomes (Beretta, 1992). Teachers may also introduce bias through their reporting on programmes, perhaps supportive of innovations and anxious to prove their worth, or equally, loyal to existing approaches (Torgerson & Torgerson, 2008). Gaining perspectives from different sources is therefore important.

Despite potential issues, teacher involvement is crucial to educational contexts, increasing relevance (Marsden, 2007), our understanding of programme improvement (Kiely, 2009; Norris, 2009), and the potential for “meaningful, lasting change” (Norris & Watanabe, 2013:3). Involving teachers from the outset is recommended to ensure their support (Zoblotsky & Kaldon, 2018), clarify inputs, share terminology (Connolly et al., 2017; Karavas-Doukas, 1998), and develop collegial cooperation (Connolly et al., 2017). Without teachers embracing different approaches and materials, demonstrating willingness to change their beliefs, change is unlikely (Karavas-Doukas, 1998). Full involvement may also increase teachers’ fidelity to the treatment (Marsden, 2007). The practicalities of attempts to reduce teacher effect are discussed in [4.1.5.1, \(ii\)](#).

4.1.4.3 External validity

The “Achilles heel” (Connolly et al., 2017:11) of RCTs is external validity, the generalisability of findings to a wider population (Kiely & Rea-Dickins, 2005). Working with given samples with particular characteristics means that outcomes cannot necessarily be replicated in different contexts (Camburn et al., 2016; Torgerson & Torgerson, 2008). However, valid, trustworthy outcomes from smaller studies have been considered to trump invalid conclusions drawn from the wider population (Torgerson & Torgerson, 2008).

4.1.4.4 Summary of evaluation as a research approach

Greater rigour in language education research has been recommended for over two decades (Norris & Ortega, 2000). RCTs, with quantitative *and* qualitative datasets as determined by RQs, appeared most useful for the purposes of the current study.

4.1.5 Applying a mixed methods approach to this study

Mixed methods research draws on the strengths of each approach (Gorard & Taylor, 2004; K. Punch & Oancea, 2014; Riazi & Candlin, 2014); such “methodological pluralism” (Norris, 2016:7) regularly features in evaluations (Norris & Mills, 2016; Norris & Watanabe, 2013; Rea-Dickins, 1994). Drawing on varied sources is advised (Bennett, 2003; Chapelle, 2007; Rea-Dickins, 1994; Zohrabi, 2011), as triangulation improves validity (Lindgren & Enever, 2017; Norris & Watanabe, 2013; Rea-Dickins, 1999). It may increase understanding: Bennett et al. (2019) investigated the impact of a CPD intervention on primary school science teachers’ subject and pedagogical knowledge using a Randomised Controlled Trial (RCT), interviews and lesson observations. Whilst the RCT did not identify the intervention’s impact as statistically significant, teachers reported improved subject knowledge.

One mixed methods research approach is pragmatism (Riazi & Candlin, 2014), which guided this study. Rather than reflecting “philosophy and values” (Collins, 2017:290), pragmatism prioritises the most suitable method for the intended purpose (Teddle & Tashakkori, 2009). The combination of quantitative and qualitative approaches may vary, as may extent of integration and sequencing (Riazi & Candlin, 2014; Teddle & Tashakkori, 2009). Using Leech and Onwuegbuzie’s (2007) definition, this was mixed methods research with a partially mixed concurrent equal status (quantitative/qualitative) design. Each element is now explored.

4.1.5.1 Quantitative data collection

This research evaluated an educational intervention, with an independent measures (between-subject) (quasi-)experimental design: one group of participants received an intervention, and one received standard treatment (Gorard, 2017; Greig et al., 2013). Quantitative approaches measured whether pupils’ learning outcomes improved post-intervention (as suggested by Greig et al., 2013), and the impact on pupil and teacher affective variables. Manipulating the independent variable (the teaching approach, with beginners learning French either through a topic-driven or experimental language-driven approach for one academic year) enabled measurement of its impact on dependent variables (pupils’ performances in vocabulary learning, SSCs, and simple sentence manipulation). To add credibility, both schemes of work, freely available online (Novel: <https://www.rachelhawkes.com/Resources/KS2Lang.php>; Standard: https://www.rachelhawkes.com/Resources/Y3_French/Yr3French.php) and popular in English

schools (conservative estimates suggest ca. 150-200 schools using the Novel scheme in 2023, and at least this number using Standard), were created by a respected authority in FL pedagogy and NCELP co-director (Novel: Hawkes & Guillou, 2022; Standard: Hawkes, 2013; see also Marsden & Hawkes, in press).

The longitudinal research design had three pupil data collection points during the year ([Table 3](#)).

Table 3.

Research Design for Investigating French Learning Progress through Different Treatments

Group	Pre-test	Treatment	Post-test	Delayed post-test
Experimental 1	Yes	Treatment A – Novel (language-driven)	Yes	Yes
Experimental 2	Yes	Treatment B – Standard (topic-driven)	Yes	Yes

Testing addressed RQ1; quantitative data from pupil questionnaires (PQs) and teachers' post-lesson evaluations addressed RQs 2a, 2b and 3a ([4.1.3](#)).

(i) The pre-test

This was a LAA test designed for young learners (Kasprowicz et al., 2019; [Appendix B.1](#)). As addressed in [2.4.3.3](#), high levels of individual variation in Kasprowicz et al.'s study indicated the influence of LAA on performance.

Having been pair-matched on ranked pre-test scores, each pupil within the pair was randomly assigned without researcher involvement (recommended by Gorard, 2013; Oakley, 2000), using an online random generator, to the language-driven (Novel) or topic-driven (Standard) treatment ([Table 4](#)), and remained in this group for the academic year.

Table 4.

Random Allocation of Pupils to Teaching Approaches

Matched pair after LAA test	Novel	Standard
Pair 1	One pupil	One pupil
Pair 2	One pupil	One pupil
Pair 3, etc.	One pupil	One pupil

(ii) *Models of provision*

To maximise reliability of treatment implementation, because “the way a treatment is implemented may differ from one person to another if different persons are responsible for implementing the treatment” (Cook & Campbell, 1979:43), teachers taught both groups, alternating classes weekly and following scripted lesson plans to minimise error variance arising from lack of standardisation (Cook & Campbell, 1979). Pupils’ progress was thus less likely to be affected by the teacher variable (unlike in cluster RCTs, with schools allocated wholly to one approach). Two models were adopted, outlined below.

(a) Split-class design

In single-form entry schools, half the class was taught by the usual class teacher and half by an additional teacher within the school or the visiting teacher (Table 5), funded by the researcher’s workplace. Although smaller class sizes may impact positively on learning (see Johnstone, 2002, 2008), this affected both approaches.

Table 5.

Alternating Classes in Single-form Entry Schools

	Class teacher	Additional teacher
Week 1	Novel	Standard
Week 2	Standard	Novel
Week 3	Novel	Standard
Week 4, etc.	Standard	Novel

(ii) Parallel classes

In schools with two Year 3 classes, matched pairs were created from both groups based on pupils’ pre-test performances and subsequently randomly assigned to two new teaching groups, one following each approach. Class teachers taught simultaneously, alternating treatments weekly (Table 6).

Table 6.

Alternating Classes in Schools with Parallel Classes

	Class teacher A	Class teacher B
Week 1	Novel	Standard
Week 2	Standard	Novel

Week 3	Novel	Standard
Week 4, etc.	Standard	Novel

(iii) Summary of quantitative approaches

In summary, RCTs do not guarantee equality of groups but should randomly distribute differences (K. Punch & Oancea, 2014), with randomisation at the group formation stage enabling differential outcomes to be attributed to treatments, rather than other variables. With both groups taught for 45 minutes weekly by the same teachers, differential outcomes should only result from within-group differences or interventions. Section [4.2.1.1](#) discusses ethical issues of RCTs.

4.1.5.2 Qualitative data collection

This study constituted an impact and process evaluation within an RCT. *Impact* evaluations assess the effect of initiatives on learners’ educational outcomes, possibly considering participants’ perceptions (Higgins, 2017), and should always accompany field experiments “to describe and illuminate the context and conditions under which research is conducted” (Cook & Campbell, 1979:93). *Process* evaluations, documenting the process, can add meaningful qualitative data, enabling insight into quantitative outcomes and ensuring that RCTs “move on from the notion of ‘what works’ towards what works for whom, under what conditions and in what circumstances” (Connolly et al., 2018:290). Qualitative data were collected through pupil and teacher interviews and open-ended questions in PQ1 and PQ3 (section [4.3.13](#), [4.3.14](#)).

Qualitative trustworthiness is addressed in [4.3.14.6](#).

4.2 Ethics

Guidelines from the supervising institution’s ethics committee underpinned ethical decisions in this study (see Farrimond, 2017; Tracy, 2010), but also reflected the researcher’s personal values (Ethical research involving children, n.d.; Hammersley, 2017; King & Horrocks, 2010). This section discusses considerations arising early in, during and after the project.

4.2.1 Early considerations

4.2.1.1 Offering different interventions

One early research consideration should be worthiness (Hill, 2005), relating directly to the framework of harm/benefit. A clear rationale for this research was established in [4.1.2](#). However, within the comparison of innovative treatments with existing/different approaches lies an ethical dilemma: offering a novel intervention to certain learners denies potentially beneficial instruction to others (Burtless, 2002; Cohen et al., 2018; Connolly et al., 2017; Cook &

Campbell, 1979; Goldacre, 2013). Neither group can be denied statutory teaching (Gueron, 2002), and both treatments must hold educational value (Marsden, 2007). However, the superiority of the new approach was unestablished; untested innovations could be less effective (Connolly et al., 2017; Goldacre, 2013). Furthermore, teaching practices already vary, apparently causing no ethical concern (Goldacre, 2013; Marsden, 2007), thus objecting to innovation on these grounds makes little sense. Offering a different perspective, Connolly et al. (2017) suggested that teachers are morally obliged to test programmes to ensure that pupils benefit from the most effective teaching approaches.

The researcher considered that the Novel approach was likely to be effective, thus minimising potential harm to academic outcomes (Allen, 2005) and avoiding disadvantage (British Educational Research Association - BERA, 2018). This provided equipoise (Hammersley, 2017): no evidence should exist of superiority of one treatment over another. Both schemes were available in subsequent years (recommended by BERA, 2018; Gorard & Taylor, 2004), should teachers wish to use them, thus learners were not denied longer-term access to either approach.

4.2.1.2 Acquiring consent; challenges arising from involving child-participants

Permission was sought from primary school headteachers as gate-keepers (K. Punch & Oancea, 2014) and from teachers. In voluntary informed and ongoing consent (BERA, 2018:9), participants should: agree freely (P. Alderson & Morrow, 2004; United Nations Children's Fund - UNICEF, 2002); understand that they can refuse without repercussions (Greig et al., 2013); have autonomy to decide (Farrimond, 2017); and understand the entire process, including requirements and representation of their contribution (BERA, 2018; Hammersley, 2017; K. Punch & Oancea, 2014).

Adults gave active consent by signing consent forms ([Appendix C.1](#)) and were regularly reminded of their ongoing right to withdraw: consent is not a one-off event (P. Alderson, 2005; BERA, 2018; King & Horrocks, 2010; UNICEF, 2002). In all schools, pupils' participation in French instruction (as a statutory curriculum component) was mandatory. In one school, the headteacher did not wish to seek pupils' or parents' consent for data collection, as the entire study (interventions, PQs/group interviews, test data) was viewed as constituting normal practice. In this respect, the researcher had to be guided by school practice (Morrow, 2008). In the other four schools, parents/carers were asked for active consent for data collection. This risked some children being unable to participate because parental consent was withheld (Hammersley, 2017); conversely, some children may have participated unwillingly (Hill, 2005) – although there was no evidence of this.

The research was managed by adults, but children were consulted, informed, and their understanding of their experiences in each intervention valued (Hart, 1992:8, UNICEF, 2002). Acknowledging BERA's (2018:15) recommendation that children should express their views if mature and old enough, pupils in these four schools were asked for assent. Differing from consent (given by those able to make informed choices), assent represents agreement by parties lacking full understanding of the main issues (P. Alderson & Morrow, 2004). BERA's (2018) and Farrimond's (2017) guidance on securing assent was followed: pupils were introduced to the researcher, understood the nature of involvement and what they could refuse, and had time to consider their participation and ask questions.

Information sheets were issued to teachers and pupils (for the latter, simple, child-friendly leaflets, as recommended by P. Alderson & Morrow, 2004., and P. Alderson, 2005 - [Appendix C.2](#)), using comprehensible language (Greig et al., 2013; Mertens, 2015). Where the researcher could not present the study in person, a video-recording explained the project (recommended by Hill, 2005).

4.2.1.3 Ethical dilemmas arising from sampling decisions

With any hierarchical organisation, participants may feel coerced to participate (T. David et al., 2005; Hammersley, 2017). The researcher discussed the study fully with teachers to assess their willingness to participate and avoided research commitments becoming burdensome (recommended by BERA, 2018; Hobbs & Kubanyiova, 2008; King & Horrocks, 2010; K. Punch & Oancea, 2014). Table 7 provides assessment of the impact on teachers' time, showing a concise overview of the project. All teaching materials/resources including linguistic support were provided.

Table 7.

Overview of Data Collection Methods Showing Impact on Teacher Participant Time

<i>Data collection method & duration</i>	<i>Collection point 1 September 2022</i>	<i>Collection point 2 February 2023</i>	<i>Collection point 3 July 2023</i>	<i>Collection point 4 September 2023</i>
<i>Online teacher questionnaire</i>	1 x 10 minutes	1 x 10 minutes	1 x 10 minutes	
<i>Short online post-lesson evaluation</i>		1 x 1-minute weekly Sept-Feb (15 in total)	1 x 1-minute weekly Feb-July (15 in total)	
<i>Individual interviews</i>	1 x 30 minutes	1 x 30 minutes	1 x 30 minutes	
<i>Lesson observations to ensure fidelity to treatment; lesson audio recordings</i>		1-2 observations Sept-Feb Audio recording of lessons Sept-Feb	1-2 observations Mar-July Audio recording of lessons Feb-July	
<i>Administer pupil questionnaires</i>	1 x 15 minutes in French lesson	1 x 15 minutes in French lesson	1 x 15 minutes in French lesson	
<i>Administer LAA test</i>	2 x 20 minutes			
<i>Administer language tests</i>			2 x 50 minutes in French lessons	2 x 50 minutes in French lessons

The potential “hierarchy of power” (Hobbs & Kubanyiova, 2008:505; see also Greene & Hill, 2005) in researcher-teacher relationships unsurprisingly also features in research with children. Adults have undeniable authority, particularly in school environments (Farrimond, 2017; Greene & Hill, 2005), with pupils instructed to “comply with teachers’ requests from the moment they arrive” (T. David et al., 2005:130). Children may therefore struggle to disagree with adults (Hill, 2005; Kellett & Ding, 2004; Masson, 2004), leaving them open to suggestibility (Greig et al., 2013), especially if not normally asked for opinions (S. Punch, 2002).

The researcher avoided coercive language (e.g., appealing for pupils' help - Farrimond, 2017), and reminded learners that there was no right or wrong answer (S. Punch, 2002). As child participation in research engenders a need for child protection (A. Farrell, 2005; Greig et al., 2013), group interviews took place in school, with adults present, and the researcher followed standard practice regarding enhanced criminal records checks (P. Alderson, 2005; Farrimond, 2017; Hill, 2005).

4.2.2 During the project

Accepting BERA’s (2018) guidance to maintain confidentiality and anonymity, the researcher ensured that identities could not be inferred from reports and stored names by code number (Hammersley, 2017; Mertens, 2015; K. Punch & Oancea, 2014). A data management plan ([Appendix D.1](#)), reviewed regularly throughout the project, ensured compliance with General

Data Protection Regulations. Data were stored on a password-protected computer and in the researcher's institution's filestore (advocated by BERA, 2018).

4.2.3 After data collection

Farrimond (2017:85) cautioned against “cherry picking of data” in reporting, advocating acting with integrity and honesty, avoiding sensationalism, distortion and selective reporting (BERA, 2018). Post-project, findings were shared with participants (recommended by BERA, 2018; UNICEF, 2002).

Participants were told how anonymised data would be used (e.g., in publications and presentations) (see BERA, 2018; King & Horrocks, 2010), with consent (with opt-out options) sought for this and future analysis. Participants' (and parents' of child-participants) attention was drawn to the researcher's practice of Open Science, and the anonymisation and safe long-term storage of data.

4.3 Methods

This section explains the sampling approach adopted and outlines participants; it then explains methods used to collect quantified, then qualitative data ([Table 10](#)). Data analysis processes are considered in [4.4-4.5](#).

4.3.1 Participants and sampling

The sampling approach was purposive with selection based on key criteria (Collins, 2017): primary schools in England, within 25 miles of the researcher's workplace, with French taught in Year 3 by non-specialist class teachers (highest qualification in French - GCSE). Six classes across five schools participated ([Table 8](#)). These schools, clustered in a narrow geographical area across the south coast of England, may not have represented the total population, an issue inherent in school-based research (Lightbown, 2019). Although schools represented areas of different socio-economic status, this was not controlled by stratified sampling.

Table 8.

Information on Participating Schools 2022-23

School	Number of classes participating	Number on roll	School type	French provision	% of pupils eligible for Free School Meals*

A	1	191	Voluntary controlled Church of England Primary (4-11)	Class teacher/Visiting teacher**	19.90
B	2	344	Community Junior (7-11)	Two class teachers	32.80
C	1	210	Academy converter Church of England Primary (4-11)	Visiting teacher	10.50
E	1	192	Voluntary aided Catholic Primary (4-11)	Class teacher/Visiting teacher	5.70
F	1	213	Community Primary (4-11)	Two class teachers	21.10

Note.* For context, 23.80% of pupils in England were eligible for Free School Meals in October 2022 according to census data (DfE, n.d.). **The visiting teacher took over all teaching after T4 left the school after the first term.

Once identified, schools were invited to participate by the researcher. Headteachers agreed to provision of 45-minutes' weekly French instruction throughout the 30-week teaching period, and six hours additionally for testing and data collection. Non-specialist teachers became participants, as did pupils in Year 3 ([4.2](#) outlines ethical considerations; individual teachers are described in [Table 47](#)). Learners, mostly with L1 English, had little or no previous knowledge of French. Learners with other home languages were not excluded (practice advocated by P. Alderson, 2004); their inclusion was documented. The same participants constituted the sample for quantitative and qualitative data collection (identical sampling), a mixed methods research approach whereby group interview pupils were a nested sample (Collins, 2017; Mertens, 2015; Riazi & Candlin, 2014). The qualitative sample ($n = 58$), 12 Year 3 children with different LAA levels from each of four participating schools, and 10 from one school (evenly distributed across approaches), was formed from a subset of the quantitative after discussions with class teachers. This subset also completed individual speaking tests with the researcher (see [4.3.6.4, ii](#)).

4.3.1.1 Sample size

Participants ([Table 9](#)) were 149 pupils ($n = 71$ male, $n = 78$ female; mean age 7.50 years), six class teachers and one visiting teacher. Available time and resources constrained this study (general issues acknowledged by Gorard, 2013), which had to accommodate real-life practical issues (Collins, 2017). However, K. Punch and Oancea's (2014) statement that small-scale empirical research holds value was encouraging. Much as this study sought to inform on the differential impact of two teaching interventions, the qualitative focus provided interesting, relevant information for practitioners and decision-makers in this field. In this respect, sample size mattered less than narratives of experiences.

Table 9.

School Information: Average Participant Starting Age, Gender and Mean Language-analytic Ability Scores

School	Approach	Average age	Gender	LAA test scores (total /40)	
				Mean	SD
A	Novel ($n = 11$)	7.41	3 Male, 8 Female	22.91	10.60
	Standard ($n = 12$)	7.41	6 Male, 6 Female	21.00	9.85
B	Novel ($n = 29$)	7.61	11 Male, 18 Female	22.59	10.35
	Standard ($n = 29$)	7.47	12 Male, 17 Female	22.97	10.22
C	Novel ($n = 10$)	7.48	2 Male, 8 Female	17.70	10.56
	Standard ($n = 11$)	7.49	8 Male, 3 Female	14.82	9.22
E	Novel ($n = 14$)	7.48	10 Male, 4 Female	22.29	11.22
	Standard ($n = 12$)	7.45	7 Male, 5 Female	23.58	9.85
F	Novel ($n = 9$)	7.46	5 Male, 4 Female	11.56	9.71
	Standard ($n = 12$)	7.54	7 Male, 5 Female	9.92	7.48
Overall	Novel ($n = 73$)	7.52	31 Male, 42 Female	20.55	10.90
	Standard ($n = 76$)	7.47	40 Male, 36 Female	19.51	10.61

4.3.1.2 The link between sampling and ecological validity

Mirroring real-world settings promotes ecological validity. The multi-site contexts of the current research were helpful, as trialling interventions in different settings is beneficial (Goldacre, 2013). Such contextualisation may help practitioners evaluate the generalisability of findings to personal settings (J. Alderson & Beretta, 1992).

4.3.2 Schedule

[Table 10](#) outlines the methods used to collect quantified, then qualitative data and [Table 11](#) shows the longitudinal, comparative research design. Mid-study tests (February 2023), deemed worthwhile by Schmitt (2010), were planned but proved challenging to incorporate, for two reasons. In the different approaches, pupils encountered vocabulary at different times during the year: content of mid-point tests could have been very recently covered by learners in one approach but not yet covered by pupils in the other. Additionally, the practicalities of administering mainly online tests with young learners and the limited time available for French teaching/testing rendered mid-point tests unfeasible.

Table 10.

Relationship Between Methods and Research Questions

Methods	RQ1. What are the effects of each teaching approach on pupils' performance in vocabulary learning, SSCs, and simple sentence manipulation?	RQ2a. What is the (perceived) effect of each approach on pupils' self-efficacy and enjoyment?	RQ2b. What are pupils' and teachers' perceptions of the effect of each approach on pupils' progress in French?	RQ3a. What are teachers' perceptions of the effects of each approach on their motivation?	RQ3b. What are teachers' perceptions of the effect of each approach on their developing subject knowledge?
Pupil language tests (PT/DPT) in: vocabulary (recall/recognition, oral/written modality); grammar (recognition, written modality; oral production; elicited imitation); SSCs (transcription; Read Aloud); overall proficiency (elicited imitation); open-ended speaking (picture response)	✓				
Teacher questionnaires (3 times)				✓	✓
Teacher post-lesson evaluations (30 times)		✓	✓	✓	
Teacher interviews (3 times)		✓	✓	✓	✓
Pupil questionnaires (3 times)		✓	✓	✓	
Pupil group interviews (2 times)		✓	✓	✓	

DPTs were completed one year after starting the study, approximately eight weeks after last instruction (aligning with Mitchell & Myles, 2019; Porter, 2017; Roehr-Brackin & Tellier, 2019; conducted in similar contexts). While post-tests can show whether interventions had an effect (Kremmel, 2020), they risk overestimating lasting learning; delayed post-tests measure durability of effectiveness (Norris, 2006; Norris & Ortega, 2000, 2001). Any benefit from post-tests applied to both groups.

Table 11.

Project Timeline Showing Data Collection

Participants	Aug-Sept 2022	Sept 2022	Sept 2022	Sept 2022-Feb 2023	Feb 2023	Feb 2023-July 2023	July 2023	Sept 2023
Teachers	Teacher questionnaire 1; Teacher interview 1			Post-lesson evaluations (x15); lesson observations and audio-recordings	Teacher questionnaire 2; Teacher interview 2	Post-lesson evaluations (x15); lesson observations and audio-recordings	Teacher questionnaire 3; Teacher interview 3	
Novel pupils		Pupil questionnaire 1; Pre-tests (LAA)	R	X - Novel (15 lessons of 45 minutes)	Pupil questionnaire 2; Group interview 1	X - Novel (15 lessons of 45 minutes)	Post-tests; Pupil questionnaire 3; Group interview 2	Delayed post-tests
Standard pupils		Pupil questionnaire 1; Pre-tests (LAA)	R	X - Standard (15 lessons of 45 minutes)	Pupil questionnaire 2; Group interview 1	X - Standard (15 lessons of 45 minutes)	Post-tests ; Pupil questionnaire 3; Group interview 2	Delayed post-tests

Note. R denotes randomisation; X - Novel indicates a language-driven and X - Standard a topic-based teaching intervention (adapted from Gorard, 2017; Schmitt, 2010).

Data logs were kept of data from each site (see [Appendix D.2](#)).

4.3.3 Preparing the schemes

The schemes were created by the same well-respected figure in FL learning (Hawkes, 2013; Hawkes & Guillou, 2022), with the topic-driven approach already widely used (Marsden & Hawkes, in press). Slight adaptations were made to both schemes. In Novel (originally 60 minutes per week), the (unchanged) content of the first 22.50 lessons of the scheme was spread across the 30 weeks of the project. In Standard, some language (learning set phrases expressing likes/dislikes) and activities (describing the colour of a butterfly) were added (determined and designed by the researcher), to ensure sufficient material. These additions were entirely

consistent with the pedagogical approach adopted in Standard, with which the researcher was familiar. While Standard represented typical primary FL content/delivery in English classrooms, Novel was based on NCELP's (2018) "research-led and practice-informed" (Hawkes et al., 2019:1) pedagogical approach to FL teaching, suited to the time-limited context typifying provision. [Table 2](#) summarises the schemes' similarities and differences. Minor changes were made to lesson content timings in both schemes after the first 10 lessons, as non-specialist teachers needed longer to cover material.

4.3.4 Reporting reliability and validity

Reliability relates to consistency, over time and internally, helping us conclude more confidently that observed differences are real (K. Punch & Oancea, 2014). Reliability maximises the chances of obtaining similar outcomes if the same participants took tests in the same context but at different times. Measuring knowledge in different ways and pupils taking all tests supported reliability in this study. Addressing the issue of attitudinal differences towards tests, the term 'quizzes' was adopted with pupils. Instrument reliability measures are reported under each instrument where relevant.

4.3.4.1 Alpha and omega

As Cronbach's alpha (Cronbach, 1951) makes assumptions about data, a different measure of reliability may be more appropriate (G.-J. Peters, 2014). In this study, the categorical omega (McDonald, 1999) measure, a "more sensible index of internal consistency" (Dunn et al., 2014:405), was suitable for congeneric scales, where some items almost inevitably relate more strongly to the construct than others (Dunn et al., 2014; McNeish, 2018). Trizano-Hermosilla and Alvarado (2016) concluded that the alpha coefficient would only estimate reliability correctly under conditions of normality and tau-equivalence, with violations resulting in lower alpha scores, particularly in instruments with fewer items (e.g., below 10, McNeish, 2018). Additionally, alpha assumes that data are continuous, an assumption violated by Likert data often used in questionnaire scales, again resulting in underestimation of reliability (McNeish, 2018).

However, claims that alpha assumes continuity are contested (Raykov & Marcoulides, 2019), and alpha is still regularly cited (e.g., Courtney & Graham, 2019; Peng & MacIntyre, 2025), therefore this study reports both alpha and omega. As there appears to be little specific guidance on interpreting omega scores, the guidance generally applied to the interpretation of alpha has been applied: scores $\geq .70$ are acceptable; $\geq .80$ good; .90-95 excellent (Fein et al., 2022). Scores .60-.69 are questionable and below this poor/unacceptable.

4.3.4.2 Confirmatory and exploratory factor analyses

Measures of reliability are reported using a combination of confirmatory and exploratory factor analyses (CFA and EFA). CFA is recommended when the hypothesis predicting the relationship between the target constructs and observed item scores is well-defined, with EFA more appropriate in less well-established outcome measures (Flora, 2020). So, CFA would suit vocabulary tests (a thoroughly researched field), whereas for Elicited Imitation with young learners, an EFA may be more appropriate.

4.3.4.3 Construct validity

This refers to the accuracy of our interpretations of the measurement used (Hedges, 2017). As tests should measure what they claim to measure, validity was, in part, checked through piloting (clarity of instructions; time needed). Factors potentially affecting vocabulary measurement validity were considered: item selection, what was considered a word, and different aspects of word knowledge (Nation, 2007). Only content common to both interventions was tested ([4.3.6–4.3.9](#)), with account taken of potential recency effects, due to time of exposure to content in each intervention ([E.6](#)). Knowledge was tested in different ways (recall/recognition, oral/written modalities), addressing content validity. Some tests used multiple-choice responses, but avoided two-choice options, as participants can guess answers (Nation, 2007). Tests assessed language knowledge, independent of knowledge about the world, a particular issue when testing young FL learners (Johnstone, 2000).

4.3.5 Data collection: pre-test

The pre-test consisted of the LAA test designed for young learners (Kasprowicz et al., 2019) – see [4.1.5.1, \(i\)](#) and [Appendix B.1](#). As learners had not received French instruction before this project, LAA variations were interpreted as indicating learners’ potential language aptitude.

4.3.5.1 Piloting and finalising the pre-test

As the LAA test had originally been used with pupils aged from 8, minor changes were made prior to the pilot (adding colour, visuals, examples, increasing font size), rendering it more accessible to this age group ([B.1](#)). After the first pilot ([Appendix B.2](#)), it was decided that completing the test over two sessions was more appropriate for younger learners. Reliability scores were .84 (Year 3 pilot) and .81 (Year 2 pilot) using Cronbach’s alpha.

4.3.5.2 Administering the pre-test

The LAA test was completed on paper by pupils in two sessions, administered by teacher participants with guidance provided ([Appendix B.3](#)).

4.3.5.3 Scoring the pre-test

The test was scored by the researcher twice; each item was scored on a binary 0-1 basis.

4.3.5.4 Using the pre-test outcomes to compare starting points in Novel and Standard groups

Once pupils had been pair-matched on LAA scores and randomly assigned to condition, scores of Novel and Standard groups were compared. Results are presented in [5.2](#).

4.3.6 Data collection methods: vocabulary tests

In designing these tests, it was important not to judge performance on single measures (B. Richards & Malvern, 2007; Schmitt, 2010). Tests were completed by *all* pupils in the following forms: meaning recognition, oral modality (listening); meaning recognition, written modality (reading); and written production. A *subsample* of 58 pupils also completed oral production tests: six pupils with differing LAA scores were selected from each approach in Schools A, B, C, and E, and five from each approach in School F (see [4.3.4.4, ii](#)).

Interventions should be assessed against the aims of both experimental and comparison treatments (Haynes et al., 2012; Marsden, 2007). Vocabulary tests ([Table 12](#)) addressed different knowledge types and National Curriculum objectives (DfE, 2013a:2), helping establish ecological validity for both approaches. ([Appendix E.1](#) provides overall mapping of test content against the National Curriculum). Given that the two schemes taught different types of vocabulary, with Novel prioritising high-frequency words and Standard teaching topic-related vocabulary, it was challenging (but essential) to identify common content (see [Table 2](#) for a comparison of the schemes of work). Testing all lexical items from both schemes was impractical, as vocabulary, grammar and phonics tests were allocated just two 45-minute lessons on each testing occasion.

Table 12.

Types of Knowledge and National Curriculum Objectives Addressed in Vocabulary Tests

	<i>Recognition</i>	<i>Recall</i>	<i>Oral modality</i>	<i>Written modality</i>	<i>National Curriculum objectives</i>
<i>Listening</i>	✓		✓		listen attentively to spoken language and show understanding by joining in and responding

<i>Speaking</i>		✓	✓		speak in sentences*, using familiar vocabulary, phrases, and basic language structures
<i>Reading</i>	✓			✓	read carefully and show understanding of words, phrases, and simple writing
<i>Writing</i>		✓		✓	write phrases* from memory

Note. *This study also included individual words, appropriate to the first year of learning French.

4.3.6.1 Piloting vocabulary tests

In the interest of brevity, piloting details are provided in [Appendix E.2](#), explaining: the multiple rounds; the change from the original counterbalanced test design to words allocated to one test type; and the impact of participants' young age on what could practically be achieved in the available time.

4.3.6.2 Finalising design and content

After scrutinising content of both schemes and several iterations, a list of 48 words common to both schemes was finalised and allocated to test types ([Appendix E.3](#)): 12 words were only tested in meaning recognition, oral modality; 12 different words only in oral production; a different set of 12 words only in meaning recognition, written modality; and the final 12 words only in written production. Each test type was completed at both time-points, and words did not change across different test phases. This provided feedback on individual gains rather than group outcomes, and enabled understanding of whether certain knowledge types improve more than others when taught in a particular way.

This list featured nouns, (infinitive) verbs, adverbs, adjectives, pronouns, conjunctions and one preposition, consisting of individual lexical items rather than multi-word phrases. This decision reflected the main content of both approaches and was supported by findings from Schulz et al.'s (2023) meta-analysis, which showed that the efficacy of teaching multi-word units to young learners is not yet fully established. The predominance of nouns was difficult to avoid because of the focus in Standard on teaching this word class. The following factors were matched across test types (see [4.3.4.3](#)): corpus-based frequency rating ([Appendix E.4](#)); word length ([E.5](#)); recency and frequency of encounters in each approach (acknowledging that values would inevitably be lower in Standard because it featured less retrieval practice than Novel) ([E.6](#)); word class ([E.7](#)); concreteness ([E.8](#)); cognate status ([E.9](#)); syllable count (no determiners included) ([E.11](#)); and SSC distance between English and French ([E.10](#)). A 'distances removed' approach

was adopted in relation to phonics – i.e., the number of pronunciation differences between the English and the French (so, ‘cheval’ = 1).

The order of the three whole-class vocabulary tests was maintained at both time-points, although items within each test were randomised at delayed post-test (see [Appendices E.34-E.37](#)). Including all items only once in each battery eliminated the risk of receptive exposure influencing production, avoiding learning effects (S. Webb, 2007). Given the piloting results, writing was placed last of the three, as it was likely – as a productive skill - to be most challenging for these young learners.

(i) Meaning recognition, oral/written modality

Tests ([Appendices E.12, E.13](#)) were multiple choice (see Milton, 2009; Schmitt, 2010). In the oral modality, pupils listened to an audio recording of an individual French word; in the written, the French word was shown in bold blue text. In both, pupils selected their preferred answer from four English options, with answers balanced across response options (i.e., equal numbers of a, b, c, d). An inherent danger in multiple choice responses is acknowledged, as participants can guess answers (Nation, 2007): two-option answers were avoided. One question at a time appeared on screen. Pupils worked at their own pace in the written modality. They were able to omit any questions they could not answer but could not re-attempt answers in either modality. No feedback was provided on response accuracy.

(ii) Meaning recall, oral/written modality

In recall tests, learners said or wrote words in response to English stimuli ([Appendices E.14, E.15](#)). One item at a time appeared on screen. Pupils worked at their own pace but could not re-attempt words. In writing, letters with accents were provided within Qualtrics (<http://www.qualtrics.com>) if pupils wished to use them. No support from the teacher was offered in either test apart from explaining the task and reading aloud English words. No feedback was provided on response accuracy, although, in oral production, the researcher encouraged pupils.

4.3.6.3 Checking for difference across vocabulary tests

As pupils were tested on four vocabulary knowledge types, lists of items were compared to check that there was no statistically significant difference between them across various factors. As this involved comparing more than two test types, where data were not normally distributed ([Appendices I.4-I.9](#)), the non-parametric Kruskal-Wallis test was conducted in R version 4.1.3 (R Core Team, 2020). This revealed no statistically significant difference between the four test types when compared on the range of factors in [Table 13](#). (Recency and frequency of encounter

with each lexical item were calculated separately in Novel and Standard approaches for each test type).

Table 13.

Checking for Difference Between Content of Four Test Types on a Range of Factors

Factor	Kruskal-Wallis H	df	<i>p</i> -value
Corpus-based frequency rating	.07	3	.92
Word length	.20	3	.98
SSC distance between English and French	.33	3	.95
Syllable count	.19	3	.98
Recency of encounter (Novel)	.14	3	.99
Recency of encounter (Standard)	.54	3	.91
Frequency of encounter (Novel)	2.37	3	.50
Frequency of encounter (Standard)	.70	3	.87

Novel and Standard approaches were also compared with each other across each test type for recency and frequency of encounter ([Table 14](#)). As data were not normally distributed in any test type, Wilcoxon rank sum tests with continuity correction were conducted. Results showed no statistically significant differences in recency/frequency of encounter between approaches in any of the four test types.

Table 14.

Checking for Difference Between Approaches on Recency and Frequency of Encounter

Test type	Recency	Recency	Frequency	Frequency
	Wilcoxon W	<i>p</i> -value	Wilcoxon W	<i>p</i> -value
receptive mode, oral modality	86	.43	85	.46
productive mode, oral modality	93	.23	100.50	.10
receptive mode, written modality	88	.37	94	.21
productive mode, written modality	100.50	.10	85	.47

Distribution across test types of word classes, concreteness and cognate status appears in [Appendices E.7](#), [E.8](#), and [E.9](#).

4.3.6.4 Administering vocabulary tests

(i) Whole-class tests

Tests were created using Qualtrics and completed online in whole-class settings during scheduled French lessons, the post-test in July and the delayed post-test in September 2023, and took approximately 20-25 minutes. Learners accessed tests from individual devices. Clear instructions were issued to teachers on test conduct ([Appendix E.16](#)); pupils worked individually without conferring. To ensure consistency, tests were conducted with the support of the researcher (and visiting class teacher where applicable). Audio files were played aloud to the class as schools were unable to provide sufficient sets of individual headphones. French audio files were recorded twice with in-built 5-second pauses, controlled by the teacher from a separate version of the test, to avoid pupils clicking the extract at random on their own devices. These files were recorded by the researcher, who has a degree in French and lived in France. French test words were not spoken aloud by teachers, but the administrator read aloud English options as required.

(ii) Speaking test

The subsample of 58 pupils completed this test individually with the researcher in school as part of a longer test including oral production of SSCs and grammar, an open-ended speaking task and an elicited imitation task; this entire test battery lasted 12-19 minutes and responses were recorded (then transcribed). Pupils were asked to translate orally written English prompts from a PowerPoint presentation. Pupils (six from each approach in Schools A, B, C and E; five from School F) were selected based on their LAA test scores; selection also utilised teachers' knowledge of pupils to ensure that they would feel comfortable working one-to-one with the researcher (who was known to learners, having visited schools several times). Pupils' scores from Novel and Standard were as closely matched as possible within each school ([Appendix E.17](#)). The subsample included: two pupils each with high, mid-, and low LAA scores (*within each school*, the range of LAA scores was divided into three categories from high to low, so a pupil with high LAA in one school could have featured in the middle or lowest LAA category in another). At post-test, the mean LAA scores for pupils in this subsample were 10.67 (Novel) and 10.71 (Standard). At delayed post-test, two pupils in School B (one from each approach) replaced two original pupils (again, one in each approach) to maintain balanced scores as one pupil had left the school; the mean LAA scores were then 10.71 (Novel) and 10.72 (Standard).

4.3.6.5 Scoring vocabulary tests

(i) Recognition tests

Automatic scoring (using Qualtrics) was used in both oral/written modalities, with both tests scored /12 (1 point available per item).

(ii) Recall tests

Written production was scored /24: 2 points were available for each of 12 items, awarded for wholly accurate production; 1 point where one letter was incorrect/omitted or two letters transposed – deemed appropriate as this test prioritised pupils’ orthography, rather than their sound representation of the word. For example, in producing the French word for ‘cat’, “*chat*” scored 2 points; “*cha*” or “*chan*” scored 1. Omission of accents was not penalised; random, incorrect accents were ignored. In oral production, also /24 (12 items), 2 points were awarded for each correct, reasonably authentic-sounding item understandable to a sympathetic native speaker. If meaning was communicated but with some ambiguity, 1 point was given. Second attempts were permitted if pupils self-corrected immediately. In production tests, determiners were not required for nouns and were ignored if supplied (regardless of accuracy). Tests were scored twice by the researcher.

4.3.7 Data collection methods: phonics tests

As with vocabulary, phonics knowledge was tested in different formats addressing pupils’ recall ability: transcription (all pupils); and Read Aloud (RA), completed by the same subsample of 58 pupils who completed oral tests.

Both interventions taught SSCs explicitly, but in Novel fewer SSCs were covered, and instruction was more systematic and integrated than in Standard (see [Table 2](#)). The topic-driven scheme focused on SSCs more than most other standard approaches in current use; it taught several SSCs explicitly in the first term, but revisiting was not planned for. As with vocabulary, only SSCs common to both approaches (Appendices [E.18-E.19](#)) were tested (i.e., they were explicitly introduced or reinforced in lessons, or teachers were instructed to draw learners’ attention to them). Tests ([Table 15](#)) addressed knowledge in different modalities and National Curriculum objectives (DfE, 2013a).

Table 15.

Types of Knowledge and National Curriculum Objectives Addressed in Phonics Tests

<i>Testing Approach</i>	<i>Recognition</i>	<i>Recall</i>	<i>Oral modality</i>	<i>Written modality</i>	<i>National Curriculum objectives</i>

<i>Transcription (Productive knowledge of orthography – write French word in response to hearing word spoken)</i>		✓		✓	explore the patterns and sounds of language ... link the spelling, sound ... of words
<i>Read aloud (from printed French)</i>		✓	✓		develop accurate pronunciation and intonation so that others understand when they are reading aloud

Host words were authentic French words not encountered in either scheme during the project. Transcription and RA featured the same SSCs but in different host words ([E.18-E.19](#)); the transcription test was completed two weeks before RA, so any residual benefit for RA from hearing the phonemes in the transcription test would be unlikely.

4.3.7.1 Piloting phonics tests

Detailed accounts of piloting ([Appendix E.20](#)) explain: the multiple rounds; changes made to the original counterbalanced test design; challenges encountered in issuing such tests to young learners.

4.3.7.2 Finalising design and content

(i) Transcription

After piloting a multiple-choice option, which did not appear to confer distinct benefit, the original form recall format was maintained ([Appendix E.21](#)), acknowledging the challenge for pupils. Learners heard 10 recorded words twice and had to write the missing grapheme (including some digraphs) correctly to gain 1 point, with missing letters indicated. Words were: short; mono-/bi-syllabic; mainly low-frequency; not covered by either scheme; unfamiliar to pupils. Prompt letters were supplied; pupils provided the focal grapheme (e.g., l _ _ = *leu*). The diacritics bar in Qualtrics was restricted to [é], as other accents appeared distracting in the original test (used randomly and unnecessarily). No feedback was provided on response accuracy.

(ii) Read Aloud

This was completed by the subsample (58 pupils) at post-test and delayed post-test. Pupils were tested on 10 French words, all unknown, embedded in a PowerPoint presentation ([E.22](#)). Target graphemes were not highlighted.

4.3.7.3 Administering phonics tests

(i) Transcription

This was completed online using Qualtrics by all pupils in whole-class settings during scheduled French lessons. It took approximately 10-15 minutes, delivered immediately prior to the online grammar tests. The same procedure was followed as for vocabulary tests ([4.3.6.4](#)), with audio files played aloud to the class.

(ii) Read Aloud

This was completed by the subsample of 58 pupils who completed the oral tests individually with the researcher (see [4.3.6.4, ii](#)). No feedback was provided on accuracy, but encouragement was given.

4.3.7.4 Scoring phonics tests

(i) Transcription

Automatic scoring (in Qualtrics) was initially used, then checked by the researcher. The test was scored /10 (1 point per item). With the item “*jettes*”, where the stem “*jette _*” was provided, any permissible silent final consonant was accepted, even if it failed to produce a viable French word ([Appendix E.18](#)).

(ii) Read Aloud

Tests were recorded, transcribed and scored by the researcher, considering only production of targeted SSCs (e.g., [ch] in “*chauve*”). Items were scored on a 0-1-2 basis, with 1 for reasonable but not wholly accurate production.

4.3.8 Data collection methods: grammar tests

As with vocabulary and SSCs, grammatical content common to both approaches was identified and included: “*avoir/être*” verbs (‘to be’/‘to have’), regular ‘-er’ verb endings (first, second, third person), pronouns and adjectival agreement (Appendices [E.23](#), [E.24](#), [E.25](#)). The Novel scheme included more regular, explicit teaching of grammatical constructions (see [Table 2](#)) than Standard, particularly of verbs, making it difficult to create a test equitable for both approaches. Test content therefore included only items where teachers in both schemes were instructed to

draw pupils' attention to particular forms. In doing this, a key difference between the schemes, the more explicit grammar focus in Novel, was possibly minimised.

Tests (Table 16) included written grammar recognition, spoken grammar recall, and Elicited Imitation Grammar (EIG) and addressed National Curriculum objectives (DfE, 2013a:3).

Table 16.

Types of Knowledge and National Curriculum Objectives Addressed in Grammar Tests

<i>Testing Approach</i>	<i>Recognition</i>	<i>Recall</i>	<i>Oral modality</i>	<i>Written modality</i>	<i>National Curriculum objectives</i>
<i>Written grammar recognition</i>	✓			✓	understand basic grammar ...: feminine, masculine ... forms and the conjugation of high-frequency verbs
<i>Spoken grammar recall</i>		✓	✓		speak in sentences*, using familiar vocabulary, phrases and basic language structures; understand basic grammar ...: feminine, masculine ... forms and the conjugation of high-frequency verbs; key features and patterns of the language
<i>Elicited Imitation</i>	✓	✓	✓		understand basic grammar ...: feminine, masculine ... forms and the conjugation of high-frequency verbs; key features and patterns of the language; how to apply these, for instance, to build sentences; and how these differ from or are similar to English; speak in sentences, using familiar vocabulary, phrases and basic language structures

Note. *This study also included individual words, appropriate to the first year of learning French.

4.3.8.1 Piloting grammar tests

[Appendix E.26](#) outlines: the multiple piloting rounds; changes made to the original counterbalanced test design; the impact of participants' young age on what could practically be achieved in the short space of time available.

4.3.8.2 Finalising design and content

(i) Written grammar recognition

The final version contained 18 items, six items targeting each of “avoir”/“être”, ‘-er’ verb endings and adjectival agreement. Items from these three grammatical categories were randomly distributed and the order of test items was also randomised between post-test and delayed post-test ([Appendix E.23](#)). Pupils read short, written sentences in French, each with one or two words missing (indicated by a blank – [Appendix E.27](#)). They selected their answer from three options (with answers balanced across response options - i.e., an equal number of a, b, c). All options were semantically correct, and gaps could only be completed by pupils demonstrating their understanding of focal grammatical structures. Because selecting content common to both approaches was challenging, translations of individual lexical items were provided, as necessary. Pupils were able to work at their own pace but could not re-attempt responses once submitted.

(ii) Spoken grammar recall

The oral modality test avoided too heavy a literacy focus for these young learners. Fourteen items targeted four areas of grammatical production: “avoir”/“être” verbs (four items); pronouns (five items); present tense ‘-er’ verb endings (three items); and adjectival agreement (two items). Pupils were individually shown a series of short English sentences embedded in a PowerPoint presentation, with an incomplete French translation below. The word containing the target structure was represented by a blank ([Appendix E.28](#)) and pupils had to say this French word. Between slides, pupils were distracted by a picture remaining on screen for three seconds before automatically moving to the next item. This minimised any benefit from pupils occasionally encountering in the written modality items which they subsequently had to produce verbally (although translations were shown, attention was never drawn to target items). The order of test items was randomised between post-test and delayed post-test ([Appendix E.24](#)).

(iii) Elicited Imitation Grammar (EIG)

The final version targeted 25 items, 11 focusing on “avoir/être”, seven on regular ‘-er’ verb endings and seven on adjectival agreement. The order of utterances was randomised between post-test and delayed post-test ([Appendix E.25](#)). The EIG featured the following:

- utterances of six-eight syllables in length, recorded once (Graham et al., 2017, found floor/ceiling effects below/above this). Courtney et al. (2017) and Graham et al. (2017) used 18-19 items with slightly older learners, but creating sentences with limited vocabulary and core grammatical structures featuring in both schemes was challenging – this also made it difficult to control for imageability. Length of sentences and range of structures were controlled for.
- pictures supporting meaning of utterances, but visible only while the utterance was heard (so that pupils were not distracted when speaking).
- two examples.
- a three-second delay after hearing the utterance before being able to speak (reduced from five seconds after piloting).
- oral praise/encouragement regardless of success of utterance.
- no ungrammatical stimuli (Kostromitina & Plonsky, 2022, suggested that ungrammatical stimuli could increase confusion - this seemed pertinent given the age group of these learners).
- treasure hunt context (following an idea from Marinis and Armon-Lotem, 2015, with each utterance attached to a stepping stone taking pupils closer to hidden treasure – [E.29](#)).

Positioning the target grammatical structure initially/medially/finally was difficult as learners had not encountered many different sentence formats.

4.3.8.3 Administering grammar tests

(i) Written recognition

This was completed online by all pupils in whole-class settings, following the phonics transcription. The same procedure was adopted as for vocabulary tests ([4.3.6.4](#)). Pupils worked at their own pace, but could not re-attempt items once submitted.

(ii) Spoken recall

This was completed by the subsample ($n = 58$) with the researcher ([4.3.6.4, ii](#)).

(iii) Elicited Imitation Grammar

The same subsample individually listened to the 15 utterances once through headphones (repeated only if noisy in the room). Utterances were pre-recorded into a PowerPoint presentation to ensure parity of experience.

4.3.8.4 Scoring grammar tests

(i) Written recognition

Automatic scoring (in Qualtrics) was used. The test was scored /18 (1 point available for each item).

(ii) Spoken recall

Responses were recorded, transcribed and scored twice by the researcher. The 14 items were scored on a 0-1-2 basis, with 1 point for reasonable but not wholly accurate production. On two items, gr5:1 (___?___ is big – “___?___ *est grand*”) and gr5:2 (___?___ is brown – “___?___ *est brune*”), any appropriate noun was accepted, provided gender was correct. The determiner was unnecessary.

(iii) Elicited Imitation Grammar

Responses were recorded, transcribed and scored by the researcher. The 25 items were originally scored on a 0-1-2 basis (1 point for reasonable but not wholly accurate production). Items scoring 1 represented just 3% of all observations, so scores 1 and 2 were merged, resulting in a binomial approach (0-1). However, outcomes were very similar, so the ordinal model is presented in [5.13](#), with the binomial model provided in [Appendix A.4](#). If pupils spoke before the picture stopped spinning, this response was removed.

4.3.9 Data collection methods: speaking tests

Two measures of pupils’ speaking ability were included and addressed National Curriculum objectives (DfE, 2013a:2) ([Table 17](#)). The open-ended speaking was a Picture Response (PR) task, with images selected to enable pupils from both approaches to respond verbally to visual stimuli. The Elicited Imitation Speaking (EIS) was a re-scoring of pupils’ responses to the EIG, designed to capture general oral proficiency by considering global response rather than individual grammatical components.

Table 17.

Types of Knowledge and National Curriculum Objectives Addressed in Speaking Tests

<i>Testing Approach</i>	<i>Recognition</i>	<i>Recall</i>	<i>Oral modality</i>	<i>Written modality</i>	<i>National Curriculum objectives</i>
<i>Open-ended speaking – Picture Response</i>		✓	✓		speak in sentences, using familiar vocabulary, phrases and basic language structures;

<i>Elicited Imitation Speaking</i>	✓	✓	✓		understand basic grammar ... feminine, masculine ... forms and the conjugation of high-frequency verbs; key features and patterns of the language; how to apply these, for instance, to build sentences; and how these differ from or are similar to English; speak in sentences, using familiar vocabulary, phrases and basic language structures.

4.3.9.1 Piloting the speaking tests

For information on piloting, see [Appendix E.30](#).

4.3.9.2 Finalising design and content

(i) *Picture Response*

Pupils were shown unfamiliar pictures embedded in four PowerPoint slides, three of which offered opportunities for learners to compare two images in some way; these were introduced by the researcher ([Appendix E.31](#)). Pictures were chosen to enable pupils from both approaches to describe them according to the language they had covered (e.g., pupils in Standard could describe colour, and pupils in Novel size). Two pictures (1,4) targeted production of nouns and adjectives, and two (2,3) aimed to elicit verbs (eating, listening to/playing music), although any image-appropriate response was accepted.

(ii) *Elicited Imitation Speaking (EIS)*

The EIS test consisted of a rescoring of the EIG task ([E.29](#)); no changes were made.

4.3.9.3 Administering speaking tests

Tests were conducted with 58 pupils individually as part of oral testing ([4.3.6.4, ii](#)). Pupils did not receive feedback on accuracy, although were encouraged to respond and develop answers where appropriate.

4.3.9.4 Scoring speaking tests, including inter-rater reliability

(i) Picture Response

Scores were uncapped. Pupils were credited for as much relevant French as they could produce, although no pupil scored higher than 10.50 at either time-point. Responses were transcribed into an Excel document, one cell per word, with scoring underneath. Where a pupil gave no response, despite encouragement, it was assumed that they were unable to say anything; this scored zero. Scoring was based on ‘idea units’, each carrying 2 points (e.g., ‘the dog’ constituted one idea unit). Incorrect pronunciation was penalised by deducting .50. [Appendix E.32](#) provides detailed guidance.

As four pictures were provided for 58 pupils to describe, 232 utterances were possible at each time-point (assuming one utterance per image per pupil). Of the possible 232 post-test items, a second rater with a degree in French and several years’ French teaching experience at secondary school scored a subsample separately from the main researcher (60 utterances), including pupil responses from all schools for each picture. The scoring of 59 utterances showed complete agreement, representing 98.33% similarity. Cohen’s κ was run in R version 4.3.2 (R Core Team, 2023) using irr (Gamer & Lemon, 2019) and cadet (Kuhn, 2008) packages and showed almost perfect agreement between the two raters’ scores, $\kappa = .98$ ($p = 0$). The one difference was reviewed and agreed, establishing 100% similarity. At delayed post-test, 55 utterances were second-scored by the same rater. The scoring of 51 showed complete agreement (92.70% similarity). Cohen’s κ was run and showed almost perfect agreement between the two raters’ scores, $\kappa = .91$ ($p = 0$). After review and discussion, 100% agreement was reached.

(ii) Elicited Imitation Speaking

Pupils’ utterances were scored globally, with 4 points available for each. The scoring scheme was adapted from Tracy-Ventura et al., (2014), originally used with university students (see [Appendix E.33](#) with additions, modifications and illustrative utterances in red). There were occasions where scoring wording needed to be amended to account for learners’ age and context. Making minor changes enabled more learners to score 4/4 rather than 3/4. The principle adopted was to reward overall contributions for communication of meaning rather than penalise individual errors.

As 15 items were included in the EIS task and completed by 58 pupils, 870 items were possible at each time-point. At post-test, a subsample of 225 were scored separately as above ([4.3.9.4](#)) by the same rater, with schools again evenly represented across all items. This showed 100% similarity. At delayed post-test, the same inter-rater scored 225 utterances. There was initial difference of opinion on three items, demonstrating similarity of 98.67%. Cohen’s κ was

run and showed almost perfect agreement between the two raters' scores, $\kappa = .98$ ($p = 0$). After review and discussion, 100% agreement was reached.

4.3.10 Data collection: teacher post-lesson evaluations

In addition to learning outcomes, this study focused on non-specialist teacher-participants' emotional experiences of teaching French, specifically enjoyment and anxiety. Post-lesson evaluations provided one measure of this.

4.3.10.1 Design and content

Evaluations consisted of eight items to which teachers responded using a 6-point Likert scale ([Appendix F.1](#)). Some questions arose from previously tested surveys (e.g., Frenzel et al., 2016). [Appendix F.2](#) contains piloting information.

Item1 ('I enjoyed this lesson') and Item2 ('I was tense during this lesson') were one-item assessments of enjoyment and anxiety, slightly adapted from Frenzel, Goetz, Stephens, et al. (2009). Teachers' enjoyment can influence learners' emotional response, potentially increasing subject value for them (Frenzel, Goetz, Lüdtke, et al., 2009). Manifesting itself as enthusiasm, and supporting positive learning environments (Fauth et al., 2019), enjoyment may facilitate student learning (OECD, 2005). Anxiety may occur when teachers feel incapable of achieving classroom goals (Frenzel, Goetz, Stephen, et al., 2009). Identifying potential physical and emotional responses (Amengual Pizarro, 2019), Item2 targeted how each approach increased or decreased non-specialists' anxiety.

Items3-5 ('Students understood the material'; 'Students were motivated'; 'Students were focused') explored teachers' perceptions of the impact of each approach on learners' understanding, motivation and attention (from Frenzel, Goetz, Stephens, et al., 2009, with a minor change in Item5 from "disciplined" to "focused". Although Frenzel, Goetz, Stephens, et al. found discipline an important indicator for teacher emotions, "focused" seemed appropriate with young learners). Learners' positive engagement may increase teachers' positive emotional responses (Frenzel et al., 2018) and decrease anxiety (Mierzwa, 2019).

Teachers' perceptions of the usefulness of language taught and of impacts of each approach on learner knowledge and learning were addressed in Items6-8 ('I was satisfied with what the students knew by the end of the lesson'; 'I was satisfied that the French I taught was useful'; 'The activities gave me sufficient evidence about the students' actual learning').

4.3.10.2 Justification for this method

Despite the small number of teachers ($n = 7$), sufficient within-subject data points ($k = 276$) were collected across the year to enable tracking of individual case study trajectories. Immediate post-

lesson evaluations targeted *state-like* rather than *trait-like* emotions (Frenzel, Goetz, Stephens, et al., 2009), hopefully unaffected by memory loss. Burić et al. (2018) documented issues with self-reporting of emotions (memory, willingness, understanding), but nevertheless acknowledged its popularity. Adding assessment of trait-like emotions in questionnaires and interviews, in addition to post-lesson evaluations, facilitated more robust measures of experiences.

4.3.10.3 Scheduling and administering evaluations

Teachers received email links to evaluations immediately after each lesson ($k = 30$ per teacher, 15 for each approach). Piloting showed that they took an average of 46 seconds to complete. Frenzel, Goetz, Stephens, et al. (2009), using teacher-diaries to report emotions, recommended that teachers only complete evaluations directly after lessons; practically, this was sometimes impossible on busy teaching days. Evaluations completed more than two hours after taught lessons were removed from the dataset.

4.3.11 Data collection: teacher questionnaires

Teacher questionnaires (TQs) were primarily used to illustrate teachers' enjoyment and anxiety trajectories, and to inform interviews.

4.3.11.1 Questionnaire length and format

TQs, slightly adapted from the items used by Frenzel, Goetz, Stephens, et al. (2009), provided a snapshot of teachers' enjoyment and tension using a 6-point Likert scale, with 1 indicating very strong agreement, and 6 very strong disagreement ([Appendices F.3-F.4](#)). TQ1, the longest questionnaire, elicited background data, whereas TQ2 and TQ3 focused solely on teachers' responses to teaching each approach. Piloting with four primary teachers showed that TQs were completed in 10-11 minutes on average. TQs adopted a more fine-grained approach than PQs ([4.3.12.3](#)), with seven options rather than four, to capture smaller incremental changes in responses. Questions (Qs) were numbered, and scales assigned a score from 1-6 (1 representing 'Very strong agreement', 6 'Very strong disagreement'), giving a possible total /30 for *enjoyment* on five items (low ratings indicated strong agreement and therefore high enjoyment). Responses to the four *tension* items were reverse coded and summed to give a possible total /24, with low ratings indicating lower levels of tension.

4.3.11.2 Rationale for questions eliciting background data

TQ1 ([F.3](#)) collected background data as follows:

- age and primary teaching experience (Q3-4): greater experience has been found to correlate positively with reduced linguistic insecurity (Daftari & Tavit, 2017; Machida, 2016).
- experience and expertise in teaching English phonics (Q5-7): possibly transferable to teaching French SSCs.
- participants' highest qualification in French, and French teaching experience (Q17-19): both identified as potential sources of anxiety (Machida, 2016). Experience has been associated with increased confidence (Cable et al., 2012; Cooke & Faez, 2018; Muijs et al., 2005; Rowe et al., 2012).

4.3.11.3 Questionnaire content

Q8 and Q12 were adapted from Frenzel, Goetz, Lüdtke, et al. (2009), from which the measure of trait-like teacher emotions developed into the Teacher Emotions Scale (Frenzel et al., 2016). Baseline indications of teachers' attitudes towards general teaching were compared with responses for French teaching (see [Chapter 7, Results III](#)). The items within these questions constituted TQ2 and TQ3 ([Appendix F.4](#)).

No changes were made after initial piloting.

4.3.11.4 Administering teacher questionnaires

Questionnaires were completed online using Qualtrics approximately one week prior to teacher interviews, so that responses could also inform the interview schedule.

4.3.12 Data collection: pupil questionnaires (PQs)

Questionnaires collected data from all pupils at three points ([Table 11](#); [Appendices G.1](#) and [G.2](#) contain PQ1 and PQ2/PQ3 – PQ3 differed from PQ2 only in the addition of two open-ended questions, shown in [Appendix G.3](#)).

Questionnaires are often considered an easy, cost-effective way to collect large amounts of data (Anderson & Arsenault, 1998; Dörnyei & Taguchi, 2010; Gillham, 2008), particularly facts and opinions (Greig et al., 2013). As response rates can be poor (Dörnyei & Taguchi, 2010), issuing PQs during lessons ensured captive samples (Gillham, 2000; Tymms, 2017). However, potential issues identified by Dörnyei and Taguchi (2010), among others, are magnified with younger participants (Borgers et al., 2000); these are addressed here.

4.3.12.1 Item length and complexity

Questions were short, simple, and unambiguous (see Bell, 2007; Gillham, 2008; Greig et al., 2013). They can be misinterpreted even by adults (Dörnyei & Taguchi, 2010), and pupils' understanding may be hampered by literacy difficulties (Borgers et al., 2000; Gillham, 2000).

4.3.12.2 Questionnaire length and format

PQs, conducted online using Qualtrics, fitted within four to six screens (see Gillham, 2008) and took approximately 10-15 minutes, within Dörnyei and Taguchi's (2010) suggested limit of 30 minutes, avoiding fatigue effects. Brevity avoids respondents losing concentration (Borgers et al., 2000), and responding indiscriminately (Tymms, 2017). Appealing formats can increase motivation (Bielick, 2017; Dörnyei & Taguchi, 2010; Gillham, 2008). PQs included the following:

- clearly laid-out (online) format (Dörnyei & Taguchi, 2010; Anderson & Arsenault, 1998).
- introductory explanations (Bell, 2007).
- context-setting section titles (Bielick, 2017).
- examples (Assor & Connell, 1992), adapted from Courtney et al. (2017), conducted with similar-aged learners.
- questions according to theme/format (Anderson & Arsenault, 1998).

Instructional video guidance was also provided by the researcher.

4.3.12.3 Question format

PQs mainly used closed questions, facilitating analysis (Gillham, 2008; Greig et al., 2013), with rating scales easily coded (Dörnyei & Taguchi, 2010). Questions were numbered, and scales assigned a score from 1-4 (1 representing 'No', 4 'Yes, a lot' on smileys, or 'I definitely couldn't' and 'I'm sure I could' on self-efficacy questions). Four response options were offered (recommended for young children by Assor & Connell, 1992, and used by Courtney et al., 2017). Although limiting response options could affect reliability and sensitivity (Bandura, 2006), too many choices can make distinguishing between options harder for respondents (Bielick, 2017; Dörnyei & Taguchi, 2010). Some researchers recommend three for young children (Anderson & Arsenault, 1998; Dörnyei & Taguchi, 2010), but four or five options increase reliability (Borgers & Hox, 2001).

Open-ended questions, potentially challenging for young participants, were generally avoided and reserved for group interviews (see [4.3.13](#)). Exceptions to this were PQ1 (self-efficacy Q2), asking why pupils felt they would (not) be able to learn French well, and PQ3 (Q5-

6, [G.3](#)), requiring reflection on aspects of French pupils had particularly enjoyed or disliked. Minimal written responses were expected (see Tymms, 2017).

‘Smiley’ questionnaires are popular for collecting data on young learners’ attitudes and motivation (Mihaljević Djigunović, 2009a): the use of faces and visual stimuli is perhaps easy for children to interpret (Tymms, 2017), and adds interest (Assor & Connell, 1992; Borgers et al., 2000). However, Bandura (2006) cautioned that smiley faces risk misrepresenting confidence as happiness or sadness; self-efficacy responses were therefore illustrated by different-sized dots demonstrating degrees of importance (used with similar-aged learners in Fenyvesi et al., 2020, with similar visuals recommended by Eccles et al., 1993). These dots were easily understood in the pilot ([Appendix G.4](#)). All response options were labelled with explanatory text (Bandura, 2006; Bell, 2007; Bielick, 2017).

4.3.12.4 Questionnaire content

PQs measured pupils’ attitudes towards French and self-efficacy. Section 1, PQ1 ([G.1](#)), collected information on language learning history and FL exposure (Dörnyei & Taguchi, 2010). Section 1, PQ2/PQ3 ([G.2](#)), compared pupils’ attitudes in each approach, drawing on Carreira’s (2006) study with Japanese learners aged from 8 learning L2 English. The questions in section 1 were slightly modified from the Intrinsic Motivation section of Carreira’s “Motivation and Attitudes towards Learning English Scale for Children” ([Appendix G.5](#) provides the original questions); Cronbach’s alpha for the reliability of each subscale in Carreira’s study ranged from .65 to .86.

Section 2 of all PQs compared pupils’ self-efficacy ratings in each intervention, their belief that they would be able to accomplish certain things (Bandura, 2006) in French. Section 3 related to pupils’ perceptions of teachers’ enjoyment (‘Do you think your French teacher likes teaching you French?’; ‘Do you think your teacher tries to make you excited about learning French?’), modified from four used by Frenzel, Goetz, Lüdtke, et al. (2009), designed to provide feedback on the same teacher teaching both approaches. As established by various studies (Becker et al., 2014; Frenzel et al., 2018), teacher and learner emotions are related, with teacher enjoyment potentially affecting pupils’ emotional response and increasing or decreasing the perceived value of/interest in subjects for learners (Fauth et al., 2019). Of particular relevance to this study is the idea that learners’ perceptions of teachers as positive are likely to increase lesson enjoyment (Frenzel et al., 2018). Pupils were therefore asked in PQ2/PQ3 (and group interviews), whether they felt that teachers enjoyed teaching them French and tried to make them excited about learning the language. Pupils scored their perceptions on a scale of 1 (lowest) – 4 (highest).

4.3.12.5 Piloting, finalising content and reliability

Despite challenges, questionnaires can be successful with young learners if designed carefully and piloted (Bell, 2007; Borgers et al., 2000). Piloting (G.4) included cognitive interviewing, a “qualitative evaluation of respondents’ comprehension” (Bielick, 2017:654) involving pupils thinking aloud while answering questions, and “retrospective probing” (Bell, 2007:467) to check validity.

The following sections address reliability, calculated using Cronbach’s alpha.

(i) PQ Section 1: enjoyment

In the pilot of PQ2/PQ3 involving 25 pupils, using the original four questions (Q1 ‘Have you enjoyed learning French this year?’; Q2 ‘Do you look forward to learning French?’; Q3 ‘Do you like to try to use the French you have learned?’; and Q4 ‘Would you like to have more French lessons?’), reliability was .63. Removing Q3 improved reliability to .73; deemed acceptable, this iteration constituted the final version.

(ii) PQ Section 2: self-efficacy

Reliability was .81 (deemed good), calculated on 52 responses to PQ1 and PQ2/PQ3 using all seven common questions. No changes were made.

(iii) PQ Section 3: pupils’ perceptions of teachers’ enjoyment

These two questions were seen as measuring slightly different constructs; reliability was therefore not calculated.

4.3.12.6 Administering pupil questionnaires

PQs were conducted online using Qualtrics by all pupils, accessed on individual devices (except for PQ1 School B participants, where it was completed on paper, as pupils struggled to log on so early in the academic year to gain access). Using Qualtrics made it possible to force responses to multiple-choice questions (although open-ended questions were optional). Teachers supported pupils where necessary with question comprehension.

4.3.12.7 Validity

Different stages in questionnaire data collection can present issues. Researcher subjectivity may affect theme/question selection (Edwards & Talbot, 1994), although using questionnaires with tested construct scales mitigates this (Bielick, 2017). Poor design and lack of clarity can increase measurement error (Bielick, 2017). Additionally, respondents may not be representative

(Bielick, 2017; Tymms, 2017). In this study, all pupils were invited to participate in data collection.

Respondents may give answers they believe are required, to present themselves positively (social desirability bias), or because they are unsure of what to say (acquiescence bias) (Dörnyei & Taguchi, 2010). Veracity cannot be assured (Gillham, 2008). Therefore, pupils were reminded that: there were no right or wrong answers; honesty was important; answers were confidential (Assor & Connell, 1992). Young respondents' recall can produce less reliable responses (Bell, 2007), so PQ3 was in July, rather than September.

4.3.13 Data collection (qualitative): pupil group interviews

Group interviews took place on two occasions ([Table 11](#)). Group interviews enable primary-aged learners (Hennessy & Heary, 2005) to give feedback on preferences for classroom activities (Mihaljević Djigunović & Lopriore, 2011; Mihaljević Djigunović, 2012), although utterances may be short, dependent on the interviewer's ability to put them at ease (Chambers, 2019). Advantages and issues pertinent to group interviews with children are outlined here.

4.3.13.1 Group interviews

Group interviews seek to understand participants' experiences and attitudes. Some pupils benefit from group support (T. David et al., 2005; Greene & Hill, 2005; Hennessy & Heary, 2005), which can promote confidence and engagement (Greig et al., 2013; Hunleth, 2011); others may feel uncomfortable contributing in front of others, depending on group dynamics (Kellett & Ding, 2004; K. Punch & Oancea, 2014; Smithson, 2000). There is, however, less pressure to answer every question. Group interviews can be more interactive, like focus groups (King & Horrocks, 2010; Smithson, 2000), which potentially facilitate recall (Hennessy & Heary, 2005) and greater depth of inquiry (K. Punch & Oancea, 2014).

4.3.13.2 Piloting group interviews

[Appendix G.6](#) summarises changes in group interview duration and explains the use of manipulatives.

4.3.13.3 Conducting group interviews

From each intervention group in each school, the pupils with varying LAA scores who had participated in the individual speaking tests ([4.3.6.4, ii](#)) took part in group interviews ($n = 58$), with groups interviewed separately during French lessons by the researcher. Six is an optimal size (Greig et al., 2013), as large groups can struggle to remain focused (Hennessy & Heary, 2005). Interviews were audio-recorded. Seating ensured good eye contact (Hennessy & Heary,

2005; King & Horrocks, 2010), and interviewees were reminded to listen carefully to everyone's contributions. Pupils had French flags to signal a desire to answer, to avoid individuals dominating discussion (Greig et al., 2013). French names were used to maintain anonymity whilst facilitating transcription.

The researcher employed age-appropriate language and welcomed all contributions; key points were verbally summarised by the researcher. Adult interpretation of children's perspectives was unavoidable (see Greig et al., 2013; S. Punch, 2002), but this practice supported understanding of essential messages and gave pupils opportunities to revise details. It was agreed with pupils that contributions were confidential and discussions should not be shared outside group interviews (suggested by Hill, 2005; King & Horrocks, 2010).

Questions ([Appendix G.7](#)) were simple and open-ended (recommended by Greig et al., 2013). Q1-3 addressed pupils' attitudes towards learning French, whereas Q6-7 addressed self-efficacy (RQ2a/2b, [4.1.3](#)). Q4-5 related to perceptions of progress and approach-specific elements considered beneficial by pupils (RQ2b). Q8-9 revisited questionnaire content to gain greater understanding of pupils' perceptions of teachers' feelings about both approaches. A pre-planned list of questions was used, generally followed in a specific order, although free-flowing conversation was encouraged (see Fenyvesi, 2020).

4.3.13.4 Truthfulness and reliability

Young children may not always answer truthfully. They may be keen to please (Bell, 2007; Greene & Hill, 2005), unwilling to diverge from others' opinions (Hennessy & Heary, 2005), or conditioned to respond to adults in authority (Kellett & Ding, 2004). However, careful planning and design combatted these risks (see K. Punch & Oancea, 2014): for example, open-ended (e.g., 'wh' rather than 'yes'/'no') questions avoided children's tendency to answer 'no' if unsure (Greene & Hill, 2005). A "feelings thermometer" was used when discussing reactions to learning (see Greig et al., 2013:218; [Appendix G.8](#)).

4.3.14 Data collection (qualitative): teacher interviews

In addition to qualitative data pertaining to children's experiences, teacher interviews were also conducted. These did not attempt to obtain a representative, generalisable sample, prioritising instead individual experiences and enabling exploration of emerging findings from questionnaires and post-lesson evaluations.

4.3.14.1 Piloting teacher interviews

No issues arose from piloting ([Appendix H.1](#)).

4.3.14.2 Interview format

The semi-structured approach followed carefully thought-out schedules ([Appendices H.2-H.3](#)) to ensure consistency of experience across interviewees, potentially facilitating cross-case comparison and analysis (Miles et al., 2020). Interviews reminded participants of purpose, confidentiality, and voluntary participation (see Greig et al., 2013), and included opportunities for interviewees to add final comments.

4.3.14.3 Practical considerations

Because interviews are prone to interruptions in school (Hobbs & Kubanyiova, 2008), face-to-face and online options were offered. The latter facilitated recording, which can make interviewees uncomfortable if face-to-face (King & Horrocks, 2010), and note-taking, which risks reinforcing power relations (Hobbs & Kubanyiova, 2008), particularly if teachers have low FL levels (Valdera Gil & Crichton, 2018).

4.3.14.4 Content

The literature review guided questions. In Interview1 ([H.2](#)), Q1-3 focused on previous experience of teaching unfamiliar subjects or French and on CPD, as these may influence teachers' confidence (Machida, 2016; Woolhouse et al., 2013). Q4-6 targeted teachers' attitudes towards teaching French and perceived areas for development, to discover whether either approach allayed concerns (cf. Q8-9, Q11). Q7 addressed teachers' confidence to make links with English, given they were regular literacy teachers (see Driscoll, Jones, Martin, et al., 2004; OFSTED, 2005, 2011). Q10 tracked their reflections on vocabulary content in each approach.

In Interview2/Interview3 ([H.3](#)), Q1-8, Q11 and Q18 elicited experiences and the impact of each approach on teachers' emotions, relating to TQ items. Q7, Q12 and Q14 targeted teachers' knowledge/confidence in vocabulary, SSCs, and grammar in each approach. Q15-17 addressed RQ2a/RQ2b ([4.1.3](#)), evaluating the interventions' perceived impact on pupils' motivation and progress. Remaining questions targeted links made with English (Q9), perceived impact on subject knowledge (Q10) and thoughts on vocabulary (Q13).

4.3.14.5 Interviewee responses

Teachers were able to comment on interview transcripts, in case they disclosed something later regretted, as interviewees may say more than intended (King & Horrocks, 2010); participants were asked to flag comments they wished to change for discussion with the researcher, although no amendments were requested.

4.3.14.6 Reliability and trustworthiness

Including similarly phrased questions, to gauge response reliability, represents internal triangulation (Atkins & Wallace, 2012). This practice was incorporated where possible ([H.2-H.3](#)). TQs, an alternative source of data, served as a comparison. For example, teachers were asked to comment on anomalies in data from TQs and post-lesson evaluations (see [7.6.1](#), [7.6.2](#), [7.7.1](#)). As with questionnaires, interviewees may feel compelled to give answers the interviewer wants (King & Horrocks, 2010); they were therefore reminded to give honest, detailed responses.

4.4 Quantitative analysis

Having described the methods, this section now provides a rationale for descriptive and inferential statistics and outlines the analytical approach adopted for each data collection instrument, focusing first on pupil, then teacher, data.

4.4.1 The importance of reporting descriptive statistics

One risk in experimental research concerns the reliance placed on null hypothesis statistical testing. Several researchers (e.g., Gorard & Taylor, 2004; Larson-Hall, 2016; Plonsky, 2015; Shadish et al., 2002) highlight issues in the use of dichotomous *p*-values, measuring statistical significance. The arbitrariness is concerning: findings could be dismissed if narrowly above or below .05% (Plonsky, 2015), yet still be ‘nearly’ significant. Sample size also has an impact. Larger samples are less prone to variability; because variability affects our ability to detect significance (Mertens, 2015), larger sample sizes can lead to statistically significant findings even when differences between conditions are very small (Gorard & Taylor, 2004; Plonsky, 2015). Additionally, significance testing can only be used with randomly assigned samples, but assumes no attrition or nonresponse (Gorard & Taylor, 2004) (see [4.1.4.2](#) for discussion of impact). Generally speaking, inferential statistical tests are inappropriate for smaller sample sizes (Beretta, 1992), as an effect particular to that sample may be detected (Connolly et al., 2017). Analysis could conclude no statistical difference between groups when the opposite is true – a Type II error (Cohen et al., 2018; Torgerson & Torgerson, 2008).

Descriptive statistics are informative and respond effectively to the issue of insufficient sample size (Plonsky, 2015). The following were therefore reported:

- central tendency: standard deviation and sample size were also reported, adding meaning (Diamond & Jefferies, 2001).
- effect sizes, bounded by 95% confidence intervals (CIs), to indicate the extent of difference indicated by *p*-values (Plonsky, 2015).

- CIs, providing better indication of effect size than p -value (Larson-Hall, 2016), unaffected by sample size.

4.4.1.1 Interpreting effect sizes

Different benchmarks were considered for effect size calculations. Cohen's (1988) scale (small: $d = .20$; medium: $d = .50$; large: $d = .80$) may underestimate effects in L2 research; Plonsky and Oswald (2014) recommended that two field-specific scales should be used, between-groups (small: $d = .40$; medium: $d = .70$; large: $d = 1.00$) and within-group contrasts, where differences are usually larger (small: $d = .60$; medium: $d = 1.00$; large: $d = 1.40$). This was based on an investigation of L2 effects from 346 primary studies and 91 meta-analyses ($N > 604,000$) which showed similar results from two designs: experimental (median $d = .74$) and observational/non-experimental (median $d = .70$); and two settings: laboratory ($d = .70$) and classroom ($d = .74$). However, it is generally acknowledged that greater experimental control is facilitated by laboratory-based studies, so effect sizes in the current school-based context may be smaller. Also, many of the classroom-based samples in Plonsky & Oswald's meta-analysis are likely to have included older learners (e.g., from Higher Education settings) with potentially higher FL proficiency, and would not therefore replicate the messier context of an intact class of young L2 beginners where learning is likely to be slower (with smaller effect sizes), reflecting a sampling bias common in L2 research (Andringa & Godfroid, 2020; Godfroid & Andringa, 2023; see also Kraft, 2020).

Hattie's (2008) synthesis of over 800 meta-analyses of educational research concluded that $d = .40$ was the "hinge-point" (p.16), although smaller effects do not lack value. Explaining why effect sizes were small from 141 large-scale RCTs (Education Endowment Foundation and National Center for Education Evaluation research; 271 distinct effect sizes; 1,222,024 students), Lortie-Forgues and Inglis (2019) highlighted factors relevant to the current study: randomised design; and intent-to-treat analysis, aiming to assess the real-world impact of interventions. They concluded: "we would certainly expect effect sizes lower than the 0.4 reported by Hattie (2008) and probably lower than those associated with a year of maturation and instruction (e.g., 0.23 SDs from fifth to sixth grade; Bloom et al., 2008)" (Lortie-Forgues & Inglis, 2019:159). In their study, considering both researcher-designed and standardised tests, estimates of effect size ranged from $-.16$ to $.74$ (median = $.03$), with an unweighted mean of $.06$ (CI $.04$ -. $.08$) rising to a weighted mean of $.04$ (CI $.03$ -. $.05$) for random effects models.

Torgerson and Torgerson (2008) suggested that, with innovative educational treatments introduced alongside existing (and likely reasonably effective) approaches, effect sizes could be small, possibly $.20$. This is supported by Kraft (2020), indicating that smaller effect sizes are likely when control and treatment groups access similar resources. Indeed, both approaches in

this study were created by the same author (Hawkes, 2013; Hawkes & Guillou, 2022). Furthermore, Kraft (2020:244) pointed out that studies with outcomes based on a “cumulative set of decisions and sustained effort over time” were more likely to produce smaller outcomes. This applied to the current study, not just in length (involving sustained effort) but also in the ‘cumulative decisions over the long term’, given the varied decisions involved in NCELP’s (2018) work that were amalgamated into the Novel approach (see Marsden et al., 2023).

Kraft (2020) highlighted the potential danger of educational programmes with small effect sizes being disregarded when actually more effective than existing alternatives. Indeed, small effect sizes can still have substantial impact when scaled up to entire populations: the “‘small effect big impact’ phenomenon” (Carey et al., 2023:1646).

In the expectation that the current ecologically valid study was likely to return small effect sizes, a decision was made to use Kraft’s (2020) scale (small: $d < .05$; medium: $d = .05 - < .20$; large: $d > .20$), as it was proposed for causal studies investigating impact on student achievement. As a separate scale for within-group measures was not suggested, it should be acknowledged that these will likely be larger.

4.4.2 Inferential statistics

This study hypothesised that pupils’ learning outcomes would possibly increase more when taught through the Novel, language-driven approach compared with the Standard, topic-driven approach. Inferential statistics help “confirm or reject our predictions” (Field et al., 2012:50) and broaden our results to a wider population. For analysis not using mixed-effects models ([4.4.4.7](#)), normality of data distribution was checked first, using visual statistics (see Larson-Hall, 2016) and the Shapiro-Wilk test, which compares data scores with normally distributed scores (Field et al., 2012) and is particularly suited to small sample sizes (Lee, 2025). Parametric or non-parametric tests were then selected as appropriate (detailed below where relevant).

4.4.2.1 Handling outliers

Outliers were acknowledged but not removed, as they reflect natural variation given the wide ability range in this educational setting.

4.4.3 Quantified (pupil) data analysis: pre-test (language-analytic ability)

To be able to compare language test scores at post-test and delayed post-test across approaches, there should be no significant differences at the starting point. Once pupils had been pair-matched on LAA scores and randomly assigned to condition, the scores of Novel and Standard groups were compared. As data were not normally distributed ([Appendices I.1, I.2, I.3](#)), non-parametric tests were conducted to determine difference between Novel and Standard. In R

version 4.2.0 (R Core Team, 2020), the Fligner-Policello test was used for individual school scores instead of the Wilcoxon Rank sum, because it is more tolerant of ties (there were several tied values). It is also better suited to small sample sizes, relevant to individual schools. As it assumes symmetry of distribution between groups, this assumption was checked through the Fligner-Killeen test (see [5.2](#)).

4.4.4 Quantified (pupil) data analysis: mixed-effects models, including dealing with nonresponse and attrition

4.4.4.1 Comparison of knowledge outcomes

The comparison of pupils' L2 French knowledge outcomes across approaches was conducted through a pragmatic evaluation of the educational interventions, adopting an independent measures design (between-subject). These outcomes, together with measures of pupil self-efficacy and enjoyment, were mainly analysed using mixed-effects models.

4.4.4.2 Mixed-effects models: introduction and justification

Mixed-effects models include both *fixed* effects, representing independent variables (Linck & Cunnings, 2015), and *random* effects, which model the type of unpredictability possibly resulting from random population sampling (Cunnings, 2012) - "usually grouping factors for which we are trying to control" (Hajduk & Gallois, 2023: Should my variables be fixed or random effects? section, para. 2). There may be other influential factors unknown to the researcher (Winter, 2013).

For *learning outcome measures* in this study, fixed effects consisted of: teaching approach (Novel/Standard), time-point (post-test/delayed post-test), and LAA score (predicted to impact on learning outcomes – see Hajduk & Gallois, 2023); and random effects of school, individual pupil, and test item. For *self-efficacy measures*, fixed effects were teaching approach, time-point (mid-/end-test) and positive anticipation score (PQ1, Q1_SE: Are you looking forward to learning French this year?); the use of the latter assumed that any initial negativity may impact on pupils' self-efficacy. Indeed, pupil comments from PQ1's open-ended question ([Appendix G.1](#), Q2_SE) referred to concerns that they would not be able to learn French/would find it difficult. Random effects were school, individual pupil and question. Within the random effects structure, pupil_id was nested within school. Such approaches, with individual pupil and school as random effects, modelled unexplained variation (Hajduk & Gallois, 2023), enabling the potential variance and covariance between pupils in the same class/school to be considered (Cunnings, 2012; Cunnings & Finlayson, 2015). According to Hajduk and Gallois (2023), mixed-effects models particularly suit the messiness characterising ecological studies and can

be effective with small sample sizes. They also cope well with missing data (Cunnings & Finlayson, 2015).

Unmatched *t*-tests assume that individual observations or data points are statistically unrelated to each other, whereas they may share characteristics beyond those anticipated by the researcher. For example, in classroom settings, pupils may be impacted in similar ways by the same teacher or school context (Cunnings & Finlayson, 2015; Gries, 2021; Linck & Cunnings, 2015). Mixed-effects models allow consideration of a number of random effects at the same time (Gries, 2021), enabling examination of the relationship between multiple variables (Salkind, 2014), within and between subjects, without involving additional analyses (Cunnings, 2012). This increases potential generalisability of findings from studies conducted with one group to a larger population which may differ on numerous measures (Linck & Cunnings, 2015). Including item as a random effect also ensures that mixed-effects models account for variation arising from the materials tested, by acknowledging that linguistic items included in the test represent only a sample of potential language (Cunnings, 2012; Linck & Cunnings, 2015).

4.4.4.3 Types of models

Mixed-effects models were created for pupil enjoyment and self-efficacy data as well as for the following pupil tests: all vocabulary measures; phonics Read Aloud; all grammar measures; Elicited Imitation Speaking. Where data were binary (vocabulary recognition, written grammar recall), binomial logistic mixed-effects regression models were created (using the `glmer` function from the `lme4` package [Bates et al., 2015] in R version 4.2.3 [R Core Team, 2023]); for all other tests, ordinal cumulative link mixed-effects models were used (using the `clmm` function from the `ordinal` package [Christensen, 2023] in R version 4.2.3 [R Core Team, 2023]). The Elicited Imitation Grammar test originally used ordinal scoring (0-1-2), but as the items scored 1 represented just 3% of all observations, scores 1 and 2 were merged resulting in a binomial approach of 0-1. As outcomes were very similar, the ordinal model is presented in [5.13.3](#), with the binomial model provided in [Appendix A.4](#). For all outcome measures, model types and codes are explained within relevant sections ([Chapter 5, Results I](#)). Analyses (confidence level: 95%) present initial results in Rmarkdown before explaining post-hoc analyses conducted (using `ggpredict` [Lüdtke, 2018] and `emmeans` [Lenth, 2024]) to interpret regression output.

Models included random intercepts, which “account for overall mean differences between participants” (Linck & Cunnings, 2015:191). Although random slopes were initially considered, the models failed to converge and were thus simplified (recommended by Cunnings & Finlayson, 2015) through their removal. Categorical variables (e.g., teaching approach) were ANOVA-coded using the function `contr_code_anova()` from the `faux` package (DeBruine, 2023) and continuous variables (e.g., total LAA score) were centred and scaled.

4.4.4.4 Nonresponse and attrition

Nonresponse and attrition potentially threatened the benefits of randomisation. Missing data were handled as outlined below.

Language knowledge tests

(i) Reliability analyses

Missing item scores were replaced with 0 in cases where a participant had completed most (not all) of the test, assuming that the missing value was due to them not knowing the answer rather than any technical issue. Participants who did not complete the test at all were removed from the reliability analyses.

(ii) Regression analyses

Missing item scores were replaced with 0 in cases where a participant had completed most (not all) of the test (as above). In cases where a participant did not complete the test at all, missing item scores were replaced with an imputed median value for each item, approach, school, task, and time-point.

Self-efficacy and enjoyment measures

(i) Reliability analyses

Self-efficacy and enjoyment questions required a Likert scale response. In such cases, the missing values were imputed using the median, both for missing time-points (e.g., where no questions were answered for a specific time-point) and missing items (e.g., where some questions were answered, and some not).

(ii) Regression analyses

In self-efficacy and enjoyment analyses, missing values were imputed using the mode ('yes'/'no'/'I don't know' data), as the median could not be calculated for each approach, school, task, and time-point. This was particularly relevant in PQ1 in School B, where pupils completed hard copies rather than online ([4.3.12.6](#)), and a response could not be forced.

4.4.4.5 Checking assumptions for mixed-effects models

Normality of distribution is less important for regression models than for more traditional tests such as ANOVAs (Winter, 2013) – indeed, “the assumption of normality is barely important at all” (Gelman & Hill, 2007:46). [Appendix J.1](#) explains the process of checking assumptions.

4.4.4.6 Reporting R^2 values

Gries (2021) recommended reporting R^2 values alongside model outcomes. These explain the amount of variance in the dependent variable accounted for by the mixed effects. Fixed effects, indicated by the marginal R^2 value, preferably explain a greater proportion of the variance than random effects (Gries, 2021), indicated by the conditional R^2 value. Ideally R^2 values will be high, but vary according to field of study (Winter, 2013).

4.4.4.7 Tests where mixed-effects models were not appropriate

The phonics transcription test had very low reliability ([Table 25](#)), and only descriptive statistics were reported. In the PR test, assumptions were not met. As data were not normally distributed, a non-parametric Mann Whitney U test was conducted in R version 4.3.2 (R Core Team, 2023) to compare between-group scores.

4.4.5 Fidelity to treatment and its impact on statistical conclusion validity (teacher data analysis)

Statistical conclusion validity relates to the correlation between dependent and independent variables, and the strength of this relationship (Shadish et al., 2002). It was threatened in this study by potential for inconsistent treatment implementation, common in real-life settings (Shadish et al., 2002). Interventions are not always delivered as intended, with input shortened or amended (J. Alderson, 1992; Bennett, 2003; Camburn et al., 2016; Karavas-Doukas, 1998; Ross, 1992). Without fidelity to treatment, outcomes cannot be confidently attributed to interventions.

An online introduction to the schemes and ongoing support ensured clear guidance for teachers (recommended by Camburn et al., 2016; Karavas-Doukas, 1998), together with standardised, scripted lesson plans (advocated by Beretta, 1992; see [Appendices K.1-K.2](#)) and resources, including audio files to support non-specialist teachers. The following two sections explain how fidelity was monitored and calculated.

4.4.5.1 Monitoring fidelity to treatment and reliability of measure

To measure “delivery, receipt, and adherence” (Shadish et al., 2002:316), intervention delivery was monitored (recommended by Beretta, 1992; Connolly et al., 2017; Mitchell, 1994). The researcher listened to one in three audio-recorded lessons (83 in total: 42 Novel; 41 Standard) and observed in person each teacher in each school once per term (25 observations: 13 Novel, 12 Standard): 108/300 lessons checked in total. Teachers were thus each observed three times, except for the visiting teacher (T3), who taught in three schools and was observed nine times,

and T4, who was observed once before leaving the school in December 2022 ([Appendix K.3](#) contains details and distribution of monitoring).

During all monitoring, an observation chart was completed ([Appendix K.4](#)) and teachers were provided with feedback on fidelity to treatment in email/face-to-face communications. Lessons were scheduled to last 45 minutes: the researcher logged at nine five-minute intervals whether teaching was consistent with the lesson plan, recording: very minor deviations; minor deviations; and deviations ([K.4](#)). To address consistency, a colleague of the researcher (a French teacher with a French degree and several years' teaching experience in primary and secondary phases) acted as inter-rater and listened to 10 recordings, three whole and seven half-lessons. Of the 59 measures of fidelity to treatment available during these lessons, 56 showed complete agreement between inter-rater and researcher on first listening (94.92% similarity). Cohen's κ showed moderate agreement between the two raters' scores, $\kappa = .56$ ($p = .005$). The three differences were reviewed and agreed, establishing 100% similarity.

4.4.5.2 Calculating fidelity to treatment

The total possible number of measures was calculated per teacher, based on the number of monitored lessons. Some lessons legitimately produced fewer than nine five-minute measures (Novel: $M = 8.78$; Standard $M = 8.68$): recordings failed or were interrupted; lessons were occasionally truncated. The number of measures indicating appropriate delivery was compared to the total possible, and a percentage calculated. Considering all teachers' ratings, 93.34% of measures demonstrated fidelity to treatment in Novel lessons, and 90.87% in Standard. It could be extrapolated from this that 138.50/150 lessons were delivered as intended in Novel, and 136.30/150 in Standard ([Appendix K.5](#) provides calculations).

4.4.5.3 Comparing fidelity to treatment across the two approaches

This is addressed in [7.1](#).

4.4.6 Quantified (teacher) data analysis: post-lesson evaluations

Throughout the year, teachers completed evaluations after teaching each lesson, totalling 276 ([Appendix F.5](#)). These were analysed in two ways: item-by-item ([4.4.6.1](#)), and grouped according to the three factors which emerged from the factor analysis ([4.4.6.2](#)).

4.4.6.1 Item-by-item analysis

As data were not normally distributed, the means of each question in Novel and Standard were compared using the R equivalent of the Mann Whitney test (the non-parametric Wilcoxon rank sum with continuity correction, R Core Team, 2023), an appropriate analysis of variance

technique (University of St Andrews, n.d.). Data were first analysed item-by-item for all teachers (7.4.2), then according to the sub-clusters identified in the factor analysis (7.4.3).

4.4.6.2 Factor analysis of post-lesson evaluation data

An exploratory factor analysis using oblique rotation (oblimin) was conducted in R version 4.3.2 (R Core Team, 2023) to explore the extent to which the teacher post-lesson evaluation measured the different constructs and whether items could be clustered or separated. The Bartlett's test of sphericity was significant $\chi^2 = 1422.91$ ($p < .001$), indicating that the correlation structure was suitable for factor analysis; the accompanying Kaiser-Meyer-Olken result of .88 verified adequacy of sampling (Kaiser, 1974, cited in Field et al., 2012).

4.4.7 Quantified (teacher) data analysis: teacher questionnaires

Responses to TQ items were analysed on an individual teacher basis and globally across all teachers. Calculating the mean for Likert data is not meaningful for individual items, but averaging scores across all data points (five enjoyment and four tension items) converted the ordinal data into interval data (University of St. Andrews, n.d.). Only descriptive data were presented.

4.4.8 Quantified (pupil) data analysis: pupil perceptions of teacher enjoyment

PQ2/3 contained two questions asking pupils about teachers' enjoyment teaching French. Descriptive statistics compared responses in the two approaches, reporting the median rather than the mean for ordinal data. The two items were seen as separate rather than a scale. As data were not normally distributed at either time-point, non-parametric unmatched two-samples Wilcoxon rank sum tests with continuity correction were performed in R version 4.3.2 (R Core Team, 2023) for each item, comparing pupil perceptions of teachers' enjoyment of teaching each approach. Non-parametric tests are also more suitable for ordinal data.

4.5 Qualitative analysis

Thematic analysis, "a method for identifying, analysing and reporting patterns (themes) within data" (Braun & Clarke, 2006:79), helped understand participants' lived experiences of each approach. The researcher played an active role in constructing themes (Xu & Zammit, 2020) through an iterative process described below.

4.5.1 Introduction

[Table 10](#) shows the qualitative data instruments: teacher interviews, pupil group interviews and three open-ended questions in PQ1/PQ3. Qualitative analysis ran throughout the study, starting with a conceptual framework ([Appendix L.1](#)), which, although subject to change, acted as the “researcher’s “map” of qualitative territory being investigated” (Miles et al., 2020:15), highlighting relationships between key variables and concepts for non-specialist teachers.

4.5.2 Practical considerations

Teacher interviews and pupil group interviews were recorded and transcribed within two weeks of the interview. As non-verbal signals can be misinterpreted, increasing potential for researcher bias (Atkins & Wallace, 2012), transcripts adopted a consistent, accurate approach; they noted obvious hesitation or emphasis, laughter and re-starts, but disregarded body language.

4.5.3 Analytic framework

Qualitative data was analysed in an iterative process using NVIVO and on paper, using tables, pens, and highlighters to support interrogation and organisation of data. Pupils’ responses to group interviews and the open-ended questions in PQ3 were analysed within Dörnyei’s (1994) framework of L2 motivation, consisting of three levels: (a) language, (b) learner, and (c) learning situation (see [Chapter 3](#)). These suited the evaluative nature of the research, considering: (a) whether pupils in each approach differed in the type of motivation they manifested or expressed towards French and France; (b) whether either approach appeared to encourage greater self-confidence and self-efficacy than the other; (c) any approach-specific motivational component related to activities, materials, teaching method, teacher or classroom experience.

4.5.4 Analytic process

Pupil responses to open-ended questions in PQ1 ([Appendix G.1](#), Q2_SE) and PQ3 ([Appendix G.3](#)) were collated and analysed as per interview transcripts. Transcripts were initially read to gain general understanding of the entire narrative context (see Braun & Clarke, 2006; King & Horrocks, 2010). Pupil group interview transcripts were analysed first by time-point (mid-point, end-point) by approach, with codes re-used across time-points and conditions to check for longevity of potential themes. A similar approach was adopted with teacher interview data, with all teachers’ transcripts analysed first for each time-point (Interview1, Interview2, Interview3), coding cross-participant and re-using codes from Interview1 at Interview2 and Interview3, whilst remaining open to new coding as appropriate given teacher comments in later rounds of data collection. Individual teachers’ transcripts were then re-read to develop a view of within-case longitudinal progression in teacher attitudes. Then began a process of moving from the

concrete to the abstract, or the specific to the general (K. Punch & Oancea, 2014), described below.

4.5.4.1 Reliability and trustworthiness

It was acknowledged that the process of transcribing may have been influenced by the researcher's interpretation of intended meanings (as an adult interpreting children's voices, Alldred & Burman, 2005). This subjective filtering of others' experiences leaves qualitative approaches open to accusations of imprecision and lack of rigour (Greig et al., 2013; Riazi & Candlin, 2014). Recognising that researchers could draw different conclusions from the data (Atkins & Wallace, 2012), chronological code logs (or audit trails - King & Horrocks, 2010:164) were maintained, outlining coding decisions (creation, amendment, amalgamation and deletion - Gibson & Brown, 2011; see [Appendices L.3-L.4](#)).

Tables were used in analysis as a way of identifying patterns and tracking data, suggested by Cloutier and Ravasi (2021) to improve trustworthiness. For example, teachers' changing emotions regarding their subject knowledge were tracked across the study, with their comments paraphrased and contextualised, emotions identified and possible causes explored (see [Appendix L.2](#) from Interview1). Where teachers identified emotions themselves, their wording was used (i.e., "I was very nervous" = 'nervousness', [L.2](#)); if not, the researcher interpreted the emotion from the processes of paraphrasing and contextualising. Such coding schemes aimed to contribute to trustworthiness by demonstrating the researcher's "conscious, systematic engagement with their data to establish solid empirical support for their emerging interpretations" (Cloutier & Ravasi, 2021:121). A range of participant voices were represented, with detailed accounts supporting readers to evaluate potential transferability (see Miles et al., 2020). Longitudinal data collection also enabled checking of pattern codes at each data collection point, again recommended as good practice by Miles et al. (2020).

4.5.4.2 First cycle coding

In first cycle coding, labels representing meaning were assigned to data ([Appendix L.3](#)), facilitating storage and retrieval (Miles et al., 2020; K. Punch & Oancea, 2014). Coding identifies patterns or repeated occurrences in data (Gibson & Brown, 2011; King & Horrocks, 2010), possibly from different sources (Atkins & Wallace, 2012), building toward the construction of themes, capturing "something important about the data in relation to the research question, and represent[ing] some level of patterned response or meaning within the data set" (Braun & Clarke, 2006:82). At this stage the researcher refrained from engaging in interpretation (King & Horrocks, 2010). On re-reading transcripts, codes were merged (showing hierarchy at different

levels) where appropriate to sharpen the focus, representing data condensation (Miles et al., 2020, [Appendix L.4](#)).

A combination of coding types was used (Miles et al., 2020; [Appendix L.5](#) provides exemplification):

- descriptive coding (identifying basic topics), used as a starting point.
- concept coding, going beyond what is immediately apparent (particularly relevant if aiming for generalisability).
- sub-coding, amplifying original entries.
- simultaneous coding, using multiple codes for the same section of text (see Braun & Clarke, 2006).
- emotion coding, an affective method helpful when participants related experiences.
- In Vivo coding, using participants' own wording.

Initial coding was exploratory, partly based on provisional (researcher-generated) codes developed from the literature review ('deductive') and piloting ([Appendix L.6](#) presents exemplification of salient points pertaining to non-specialist teachers arising from the literature review and used to inform coding); these a priori codes were later combined with those constructed from interviews. Empirical (or inductive, Miles et al., 2020) codes arise from data collection, possibly relating to a priori codes or unexpected. The researcher remained open to new direction (recommended by Miles et al., 2020), avoiding the danger of rigid expectations developing into self-fulfilling prophecies (Atkins & Wallace, 2012). Thus, the coding mixed both theoretical and inductive approaches, coding for specific RQs (Braun & Clarke, 2006) and also to a certain extent allowing RQs to develop through the coding process, resulting in a "balanced, comprehensive view of the data" (Xu & Zammit, 2020:3).

4.5.4.3 Advanced coding

Revisiting coding formed an iterative process (Braun & Clarke, 2006; Gibson & Brown, 2011; Miles et al., 2020). Second cycle (or pattern codes, describing an advanced stage - K. Punch & Oancea, 2014), involved amalgamating early codes into fewer categories, themes, or concepts, delineation of main themes, and heightened researcher analysis, interpretation and inference (King & Horrocks, 2010; Miles et al., 2020; [Appendix L.7](#)). A miscellaneous heading was used to house codes which did not group easily with others (Braun & Clarke, 2006; Xu & Zammit, 2020). Some codes were discarded. Themes can be grouped naturally (Greig et al., 2013) or subdivided to avoid complexity (Gibson & Brown, 2011). Identifying themes enabled varied experiences and ideas to be amalgamated, which was important in this study with its opportunity to explore both cross-case and longitudinally within-case. For example, connections were made

between themes extracted from different teachers' pre-project comments (Interview1) and their thoughts on teaching each approach mid- and post-teaching based on experiences (see [Chapter 7, Results III](#)). Themes were then reviewed to ensure relevance to the RQs and representation of data (Xu & Zammit, 2020).

Rather than reporting every theme, those highlighting key differences between experiences in each scheme were prioritised (i.e., teachers' *general* comments on linguistic competence were more detailed than could be reported; the study therefore focused on the ways in which each approach affected teachers' subject knowledge and emotional response). Interrelationships between factors were analysed (Gibson & Brown, 2011), as well as contradictions and inconsistencies. This process marked a shift away from description to reflection and analysis, to interpretation and the production of creative insights about the data (Gibson & Brown, 2011; see [7.8.2.5](#) for exemplification).

4.5.4.4 Additional processes

Concurrent with coding cycles, the researcher used jottings, their thoughts on substantive and procedural analysis, and analytic memoing, "little conceptual epiphanies" (Miles et al., 2020:89) with scope for creativity and speculation (K. Punch & Oancea, 2014), which facilitated cross-case comparison ([Appendix L.8](#)). Regular summaries produced throughout the data collection process used meta-matrices, tables ([L.2](#)). Additionally, concept tables (Miles et al., 2020) showed, for example, the range of opinion towards each intervention ([Appendix L.9](#)). Whilst acknowledging that using tables may be a potentially reductionist approach, they are effective in organising and categorising data, affording broader perspectives, bringing "order into an otherwise vast and chaotic mass of data" (Cloutier & Ravasi, 2021:115).

Qualitative data were occasionally quantified by counting non-numeric data after extraction of themes, where questions had been asked consistently and structured in such a way as to be able to count reliably. Quantifying non-numeric data is an appropriate approach when mixed methods studies and field experiments explore differences between intervention groups (Saldaña, 2021), enabling examination of the "distribution of certain features among cases" (Cloutier & Ravasi, 2021:122). Where this occurred, it is explained in the relevant results section.

4.5.4.5 Findings and assertions

Findings were advanced at the final stage of analysis. Assertions reflected findings based on synthesis of evidence from the data, and regular comparisons and reference to existing research were made. Participants' voices were included in the reporting, alongside (but distinct from) the researcher's commentary, "to demonstrate the rigor of the themes" (Xu & Zammit, 2020:7).

Participant quotations were selected to exemplify a chosen theme, either as representative of majority teacher opinions or because they illustrated a different angle on the data (Xu & Zammit, 2020).

4.6 Summary of methods

The scope of this research was broad, involving: measures of pupil learning in vocabulary learning, SSCs and simple sentence manipulation; pupil and teacher affective variables; teachers' confidence in subject knowledge. Adopting mixed methods research enabled a balanced design based on quantified and qualitative data, providing a well-rounded representation of participants' experiences in each intervention.

Chapter 5: Results and Discussion I. Pupils' learning outcomes

This is the first of three results and discussion chapters, presenting pupils' learning outcomes, while the second and third address pupil self-efficacy and enjoyment, and teacher findings, respectively. Pupils' learning outcomes include their performance in: (a) vocabulary recognition and recall; (b) phonics transcription and RA tests; (c) written grammar recognition, spoken grammar recall and EIG tests; (d) an open-ended PR task and EIS test.

Within this chapter, each results section is followed by discussion addressing RQ1: what are the effects of each French teaching approach on Year 3 pupils' performance in vocabulary learning, symbol-sound correspondences (SSCs), and simple sentence manipulation in one academic year? The discussion situates the current findings within existing findings concerning young learners' FL learning outcomes.

The exception to this structure is the way in which results on *individual items* within tests are presented, as these are presented in the discussion sections. This is partly to achieve economy of words, but also because pupil outcomes on individual items in each approach were not a main research question and therefore not a designed manipulation in the study (indeed, items were a random effect within the regression models). For this reason, pupils' performance on individual items is presented in a non-approach-specific manner (with more detailed discussion available in appendices), with attention drawn to approach-specific differences only where particularly relevant.

5.1 General information on results

To avoid duplication, information pertaining to *all* pupil outcome measures is described here. In all figures, July refers to post-test scores and September to delayed post-test.

5.1.1 Test quality

Information on quality of all tests is presented at the start of each relevant outcomes section (vocabulary, phonics, grammar, speaking), with text in bold indicating satisfactory results.

5.1.1.1 Reliability

Measures of reliability are reported using a combination of CFA and EFA as discussed in [4.3.4.2](#).

5.1.1.2 Facility and discrimination indices

For facility indices, which indicate the difficulty of an item by measuring the number of correct responses, .50 is considered ideal (Fulcher, 2010). A discrimination index, considering how well

the tests discriminated between the top and bottom thirds of participants, is considered good if $> .30$ (Fulcher, 2010).

5.1.1.3 Handling missing data and outliers

Discussion pertaining to handling of outliers and missing data is provided in [4.4.2.1](#) and [4.4.4.4](#).

5.1.1.4 Normality of distribution

Normality of distribution and assumption checks were conducted ([4.4.2](#), [4.4.4.5](#), [Appendix J.1](#)); links to relevant appendices are provided in the reliability table for each outcome measure.

5.1.2 Effect sizes

Effect sizes were calculated using Cohen's d and interpreted using criteria suggested for education interventions by Kraft (2020) for mean differences between groups: small ($d < .05$), medium ($d > .05 < .20$), and large ($d > .20$) – see discussion in [4.4.1.1](#). Note that in all tables of effect sizes, bold text indicates the time-point or approach with the higher scores. CIs were set at 95%, unless indicated otherwise.

5.1.3 Visualising descriptive data

Notched box visuals, explained here, are presented for most findings. The rectangular boxes show the interquartile range, with the horizontal lines within these boxes representing the median. Notches within the boxes represent the CIs (95%) around the median. When comparing approaches, notches in the boxes which do not overlap suggest a statistically significant difference. Vertical lines indicate the range of data outside the interquartile boxes – the equivalent of whiskers on a traditional box plot. Outliers are shown by dots.

5.1.4 Predicted data: analyses using mixed-effects models

Details of mixed-effects models are provided in [4.4.4](#). For all outcome measures, the type of model is outlined within the relevant section, and model codes are included in appendices. In all analyses, the confidence level was .95%. Analyses provide an overview of initial results before explaining post-hoc analyses (conducted using `ggpredict` [Lüdtke, 2018] and `emmeans` [Lenth, 2024]) to interpret the regression output.

5.2 Using the pre-test outcomes to compare starting points in Novel and Standard groups

Once the pupils had been pair-matched on their LAA test scores and randomly assigned to condition, the scores of Novel and Standard groups were compared. The range of mean scores was 9.92 (School F, Standard) to 23.58 (School E, Standard) – [Appendices I.1- I.3](#). The high standard deviation showed scores spread out rather than clustered around the mean. Analysis of box plots and distribution for each school and for the overall data showed that there was overlap between the scores. Although the Shapiro-Wilk test showed the data to be apparently normal for individual schools, the samples were very small and making a group larger by one or two scores could make a considerable difference. Histograms showed that data were not normally distributed. The non-parametric Fligner-Policello test was used in R version 4.2.0 (R Core Team, 2020) to determine difference between Novel and Standard ([4.4.3](#)). The tests supported the impression from the boxplots and histograms: there were no significant differences between approaches in the individual schools ([Table 18](#)), and it was concluded that any differences were due to chance fluctuations between participants.

Table 18.

Fligner-Killeen and Fligner-Policello Results by School

School	Fligner-	Killeen	<i>p</i> -value	Fligner-	Policello
	med chi-squared	df		FP	<i>p</i> -value
A	.003	1	.96	-.44	.66
B	.05	1	.83	.11	.91
C	.009	1	.92	-.51	.61
E	.27	1	.60	.33	.75
F	< .001	1	.99	-.31	.76

A Wilcoxon rank sum test (also non-parametric) with continuity correction was used when checking for difference in the overall scores, because of the larger sample size. The results ($W = 29.5$, $p = .57$) also indicated no significant differences between Novel and Standard groups overall at the start of the project.

5.3 Vocabulary recognition in the oral modality: results

All pupil participants ($n = 69$ Novel, $n = 72$ Standard) completed the vocabulary recognition test in the oral modality (E.12) at both time-points. Information on test quality is provided in [Table 19](#) (for all vocabulary measures).

5.3.1 Test quality

Facility indices suggested that pupils coped similarly well at both time-points. Discrimination indices, $> .30$ in all tests, indicated sufficient test discriminability. However, reliability was low ([Table 19](#)).

Table 19.

Test Quality Information for All Vocabulary Tests

		Oral modality				Written Modality			
		recognition Post-test	Delayed Post-test	recall Post-test	Delayed Post-test	recognition Post-test	Delayed Post-test	recall Post-test	Delayed Post-test
% cases complete prior to imputation		97	97	97	97	97	97	97	97
Reliability (CFA)	alpha	.44	.55	.67	.67	.64	.64	.71	.76
	omega	.49	.58	.72	.71	.66	.66	.76	.79
Facility		.59 <i>SD</i> .22	.61 <i>SD</i> .21	.33 <i>SD</i> .25	.32 <i>SD</i> .25	.62 <i>SD</i> .22	.62 <i>SD</i> .22	.18 <i>SD</i> .16	.15 <i>SD</i> .14
Discrimination		.36 <i>SD</i> .15	.38 <i>SD</i> .13	.38 <i>SD</i> .20	.30 <i>SD</i> .15	.39 <i>SD</i> .11	.42 <i>SD</i> .15	.32 <i>SD</i> .24	.29 <i>SD</i> .22
Normality of distribution link		Appendix M.1		Appendix M.10		Appendix M.4		Appendix M.7	

Return to [5.3.1](#), [5.4.1](#), [5.5.1](#), [5.6.1](#).

5.3.2 Descriptive statistics

The 12 items were scored on a binary basis of 0-1.

5.3.2.1 Between-groups results

There appeared little difference between Novel and Standard approaches in post-test outcomes; at delayed post-test, pupils' scores were slightly higher in Novel ([Table 20](#)). The notched box plot ([Figure 2](#)) shows that the medians were similar across approaches and time-points, with

overlapping notches, suggesting no significant difference. Comparing Novel with Standard, effect sizes were small at post-test and large at delayed post-test in favour of Novel; the fact that the CIs crossed zero (though only just at delayed post-test) also suggested no statistically significant difference between the approaches, though with indications of a potential trend in favour of Novel at delayed post-test.

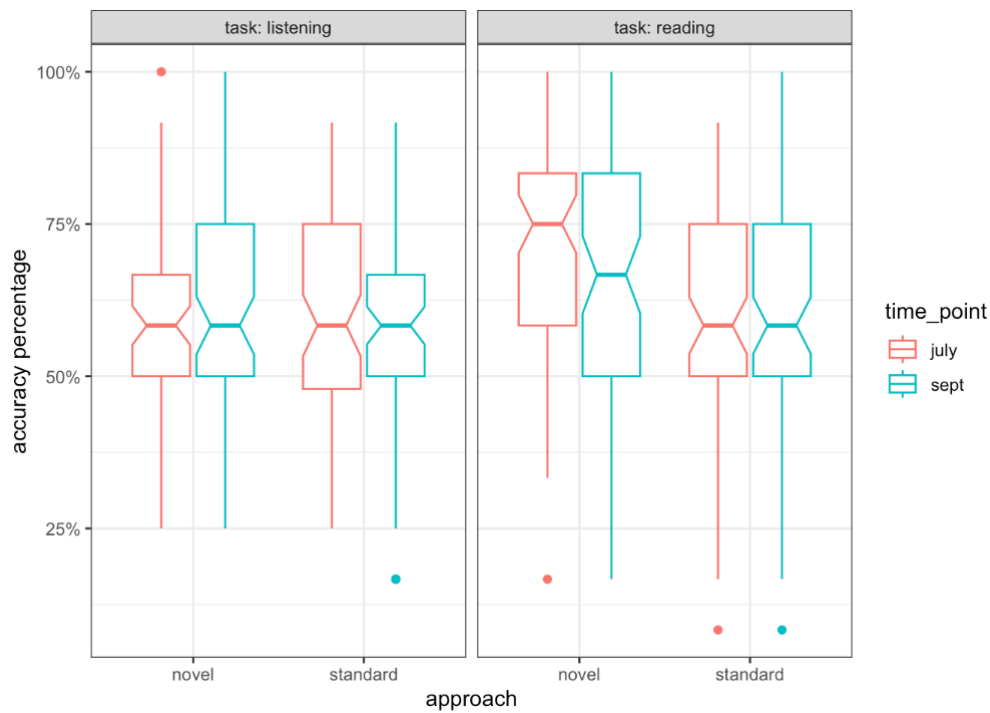
Table 20.

Descriptive Statistics: Vocabulary Recognition in the Oral Modality

Approach	Time-point ($k = 12$)	Descriptive statistics							
		n	Mean %	SD	CI %	Range /12		Effect size d	Effect size CI
Novel	Post-test	69	60	.17	56-64	3-12	Within-group: Novel Post-test – Delayed post-test	-.19	-.49 - .03
Standard	Post-test	72	59	.17	56-63	3-11	Within-group: Standard Post-test – Delayed post-test	.07	-.16 - .34
Novel	Delayed post-test	69	63	.18	59-67	3-12	Between-group: Novel – Standard Post-test	.02	-.32 - .34
Standard	Delayed post-test	72	58	.18	54-62	3-11	Between-group: Novel – Standard Delayed post-test	.26	-.06 - .59

Figure 2.

Notched Box Plot Showing Scores in Vocabulary Recognition in the Oral Modality (Listening) and Written Modality (Reading) Tests at Post-test and Delayed Post-test by Approach



5.3.2.2 Within-groups results

Scores in Novel increased slightly between time-points when comparing means, but not when comparing medians (Figure 2). The effect size between Novel scores at post-test and delayed post-test was medium, barely passing through zero, suggesting a meaningful increase at delayed post-test. In Standard, there was little difference between pupils' performance at Standard post-test and delayed post-test with a medium effect size.

5.3.3 Analyses using mixed-effects models

As vocabulary recognition in the oral and written modalities were included in the same statistical model, results are discussed in section 5.4.3 and summarised in 5.4.4.

5.4 Vocabulary recognition in the written modality: results

All pupil participants ($n = 69$ Novel, $n = 72$ Standard) completed this test (E.13) at both time-points.

5.4.1 Test quality

Reliability was more acceptable in vocabulary recognition tests in the written modality than in the oral (Table 19), with moderately reliable values at both post-test and delayed post-test. Facility scores indicated that pupils coped similarly well at both time-points, and tests showed good discriminability.

5.4.2 Descriptive statistics

The 12 items were scored on a binary 0-1 basis.

5.4.2.1 Between-groups results

At post-test, mean ([Table 21](#)) and median scores ([Figure 2](#)) were higher in Novel than in Standard; with the pattern for the mean replicated at delayed post-test. Effect sizes were large at both time-points ([Table 21](#)). Notches in the box plots ([Figure 2](#)) did not overlap, indicating a possibly significant difference between approaches at post-test.

Table 21.

Descriptive Statistics for Vocabulary Recognition in the Written Modality Tests

Approach	Time-point ($k = 12$)	Descriptive statistics							
		n	Mean %	SD	CI %	Range /12		Effect size d	Effect size CI
Novel	Post-test	69	68	.17	64-73	2-12	Within-group: Novel Post-test – Delayed post-test	.19	-.04 - .46
Standard	Post-test	72	60	.19	56-65	1-11	Within-group: Standard Post-test – Delayed post-test	.05	-.18 - .28
Novel	Delayed post-test	69	65	.20	61-70	2-12	Between-group: Novel – Standard Post-test	.45	.14 - .80
Standard	Delayed post-test	72	59	.19	55-74	1-12	Between-group: Novel – Standard Delayed post-test	.32	-.03 - .66

5.4.2.2 Within-groups results

In Novel, pupils performed better at post-test than delayed post-test, whereas in Standard there was little difference between time-points (see [Figure 2](#) for medians); both effect sizes were medium but passed through zero ([Table 21](#)).

5.4.3 Analyses using mixed-effects models

A binomial logistic mixed-effects regression model was constructed for vocabulary recognition in oral and written modalities ([Appendix N.1](#) provides code and details).

5.4.3.1 Assumptions for the regression models

Overdispersion was not detected in the model ($p = .59$). Additionally, the Box-Tidwell test ([Appendix N.2](#)) suggested that the linearity assumption between the predictors and the logit held.

5.4.3.2 Predictions based on mixed-effects models: vocabulary recognition (oral and written modalities combined)

The model summary ([Appendix N.3](#)) indicated a significant interaction between approaches, as well as between LAA and vocabulary recognition scores; these are explored below. Fixed effects (marginal R^2) accounted for 7% of variance, while fixed and random effects combined (conditional R^2) accounted for 37% of total variance.

Between-groups results

With approach (Novel versus Standard) and task (oral versus written modalities) as variables, pairwise comparisons were performed in emmeans (Lenth, 2024) using Bonferroni correction. This indicated no significant difference between approaches in oral recognition (odds.ratio = 1.11, SE = .14, $z = .81$, $p = .42$), but in the written modality, with results averaged across time-points, there was a significant difference between outcomes in Novel and Standard (odds.ratio = 1.44, SE = .18, $z = 2.88$, $p = .004$). The probability of a pupil in Novel scoring correctly in the written modality was 72%; in Standard it was 64% ([Appendix N.4](#)).

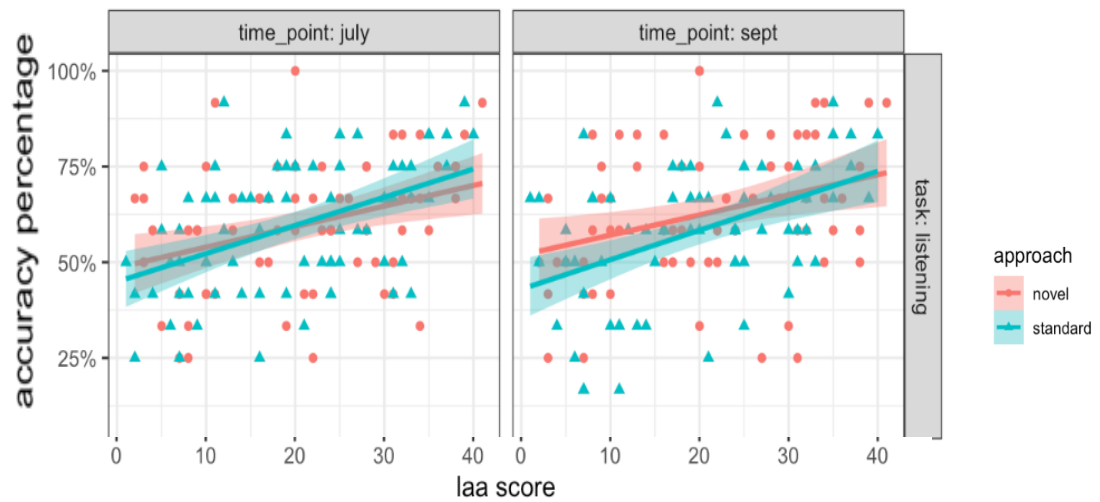
5.4.3.3 The influence of language-analytic ability on vocabulary recognition scores

(i) Oral modality

[Figure 3](#) presents the significant interaction ($< .001$) between recognition scores in the oral modality and LAA by approach and time-point. In both approaches, pupils appeared to perform better if they had higher baseline LAA levels. However, pupils with descriptively lower levels of LAA ($< \text{ca. } 15$ at post-test, $< \text{ca. } 35$ at delayed post-test) appeared to benefit from being in Novel, with this advantage moderately more pronounced at delayed post-test. At post-test, those with higher baseline LAA appeared to have a very slight advantage in Standard.

Figure 3.

Scatter Plot Showing Interactions Between Language-analytic Ability, Approach and Time-point for Vocabulary Recognition in the Oral Modality



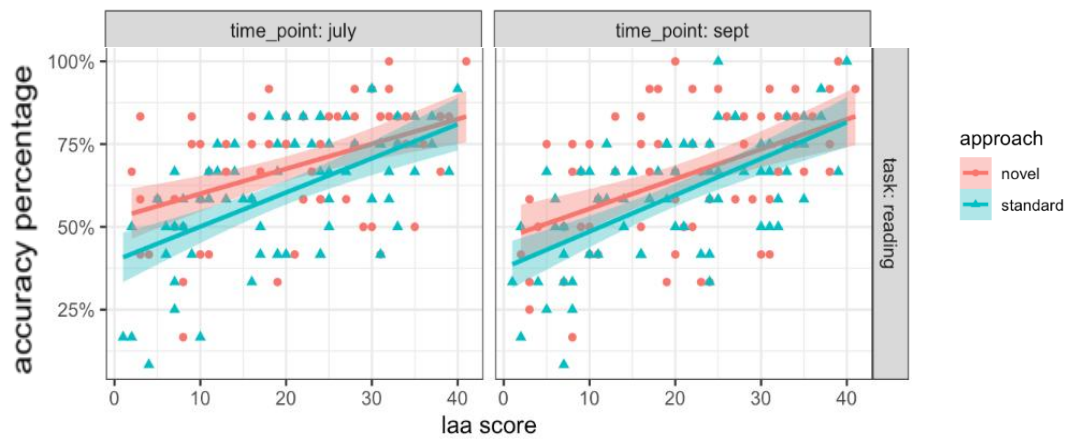
Note. ‘Listening’ refers to vocabulary recognition in the oral modality.

(ii) Written modality

[Figure 4](#) presents the relationship between recognition scores in the written modality (reading) and LAA by approach and time-point. In all cases, the higher the level of LAA, the higher the pupil outcomes. At post-test, nearly all pupils regardless of LAA performed better in Novel than Standard. At delayed post-test, the advantage for pupils with lower LAA scores in Novel was slightly less pronounced. There was no advantage from either approach for pupils with the highest LAA baseline levels at either time-point.

Figure 4.

Scatter Plot Showing Interactions Between Language-analytic Ability, Approach and Time-point for Vocabulary Recognition in the Written Modality



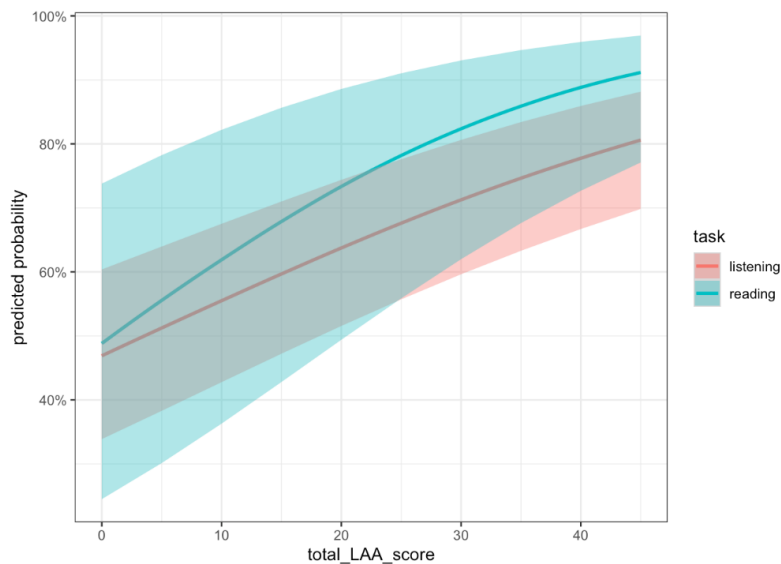
Note. ‘Reading’ refers to vocabulary recognition in the written modality.

(iii) Comparing the impact of language-analytic ability on each modality of vocabulary recognition

The ggpredict (Lüdtke, 2018) function indicated that the impact of pupils’ baseline LAA scores was greater on pupils’ recognition of words in the written than the oral modality (Figure 5). The wider range of results in the written than the oral modality suggested that performance in the written modality may be more affected by LAA.

Figure 5.

Predictions by Language-analytic Ability Score for Each Modality of Vocabulary Recognition



5.4.4 Vocabulary recognition: summary of results

Comparing approaches, there appeared little difference at either time-point in vocabulary recognition in the oral modality. However, in the written modality, there was a statistically significant difference across time-points with a large effect size in favour of Novel. In terms of change over time, there was very little difference between post-test and delayed post-test scores

in either approach in the oral modality. However, in the written modality, pupils in Novel appeared to perform slightly better at post-test than at delayed post-test, but this was not statistically significant.

Higher baseline LAA scores appeared to improve performance in both modalities at both time-points, in both groups, with some nuanced effects. In both modalities, lower LAA levels seemed to predict lower scores in Standard, whereas lower LAA levels in Novel did not seem to influence scores as much.

5.5 Vocabulary recall in the written modality: results

All pupil participants ($n = 69$ Novel, $n = 72$ Standard) completed the vocabulary recall test in the written modality (E.14) at both time-points.

5.5.1 Test quality

Reliability for this test was good (Table 19). Facility indices indicated that pupils found this writing test particularly challenging at both time-points – unsurprisingly, given their age and the productive nature of the tasks; discrimination indices were (just about) acceptable.

5.5.2 Descriptive statistics

The 12 writing items were scored 0-1-2 (total score of 24). Although higher scores *were* achieved (up to 19 in Novel and 15 in Standard, both at delayed post-test), the low means indicated that these were likely outliers (see Table 22).

Table 22.

Descriptive Statistics for Vocabulary Recall in the Written Modality Tests

Approach	Time-point ($k = 12$)	Descriptive statistics							
		n	Mean %	SD	CI %	Range /24		Effect size d	Effect size CI
Novel	Post-test	69	38	.33	30-46	0-18	Within-group: Novel Post-test – Delayed post- test	.34	.12-.59
Standard	Post-test	72	32	.28	25-39	0-13	Within-group: Standard Post-test – Delayed post- test	.30	.07-.51
Novel	Delayed post-test	69	31	.35	22-39	0-19	Between- group: Novel – Standard Post-test	.20	-.13-.53

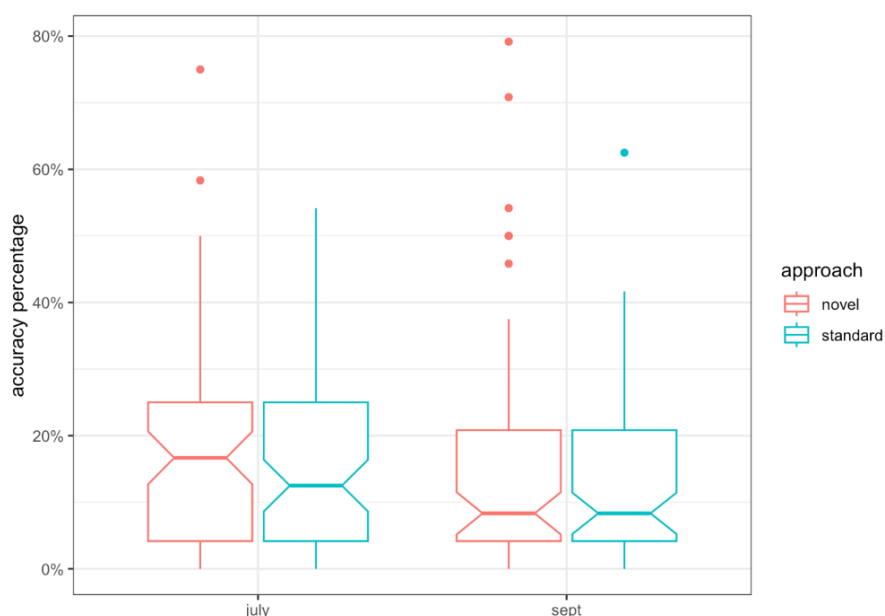
Standard	Delayed post-test	72	27	.25	21-33	0-15	Between-group: Novel – Standard Delayed post-test	.13	-.21-.46

5.5.2.1 Between-groups results

Novel's mean scores were slightly higher than Standard's at both time-points ([Table 22](#)). However, there was overlap in CIs around the mean; although effect sizes were medium to large, CIs crossed zero. [Figure 6](#) provides visual representation of these scores by time-point, showing overlapping notches.

Figure 6.

Notched Box Plot Showing Vocabulary Recall Scores in the Written Modality by Time-point



5.5.2.2 Within-groups results

Pupils scored slightly more highly in both approaches at post-test than at delayed post-test, with large effect sizes ([Table 22](#)), suggesting that learning deteriorated slightly over time. There were more outliers in Novel at both time-points ([Figure 6](#)).

5.5.3 Analyses using mixed-effects models

The model was an ordinal cumulative link mixed-effects model ([Appendix N.5](#) provides code).

5.5.3.1 Assumptions for the regression model

The Brant Test (Brant, 1990) checks whether observed deviations from ordinal logistic regression models are larger than what could be explained by chance. The Omnibus measure was $> .05$, but as p -values for approach were $\leq .05$ ([Appendix N.7](#)), results should be interpreted with caution.

5.5.3.2 Predictions based on mixed-effects models: vocabulary recall in the written modality

The model summary ([Appendix N.8](#)) indicated significant interaction between time-points, and also between LAA and writing vocabulary scores; these are explored below. Fixed effects (marginal R^2) explained 10% of variance and fixed and random effects combined (conditional R^2) accounted for 57% of total variance.

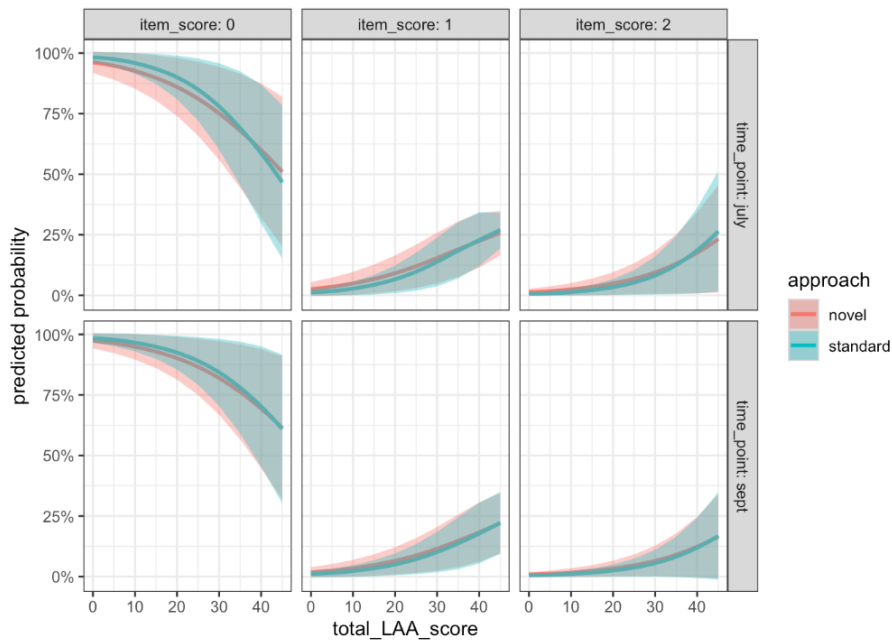
Pairwise comparisons, performed using emmeans (Lenth, 2024) with Bonferroni correction, used approach and time-point as variables. There were no statistically significant differences between approaches at either time-point, but comparisons suggested a statistically significant difference between time-points in each approach (Novel: estimate = .40, SE = .15, $z = 2.68$, $p = .007$; Standard: estimate = .32, SE = .15, $z = 2.12$, $p = .03$) ([Appendix N.9](#)).

5.5.3.3 The influence of language-analytic ability on written recall scores

As illustrated in [Figure 7](#), the ggpredict function (Lüdtke, 2018) indicated a significant interaction ($< .001$) between pupils' scores in written vocabulary recall and baseline LAA: pupils with higher levels of LAA were less likely to score 0 than those with lower LAA at both time-points, regardless of approach. Pupils in Novel with LAA scores of $< \text{ca. } 30\text{--}40$ were descriptively slightly more likely to score 1 and 2 than pupils with similar LAA scores in Standard at both time-points, whereas pupils with the highest LAA scores ($> \text{ca. } 35\text{--}40$) in Standard were slightly more likely than pupils in Novel to score 2. However, these differences did not reach statistical significance.

Figure 7.

Vocabulary Recall in the Written Modality Predictions by Language-analytic Ability Score, Approach and Time-point



5.5.4 Vocabulary recall in the written modality: summary of results

Reliability in these tests was good. Pupils scored higher at post-test than delayed post-test in both approaches, with this reaching statistical significance according to the models. As with both vocabulary recognition tests, pupils' performance in the written recall appeared influenced by their baseline LAA scores.

5.6 Vocabulary recall in the oral modality: results

A subset of 29 pupils from each approach ([4.3.6.4, ii](#)) completed the vocabulary speaking test ([E.15](#)) at both time-points.

5.6.1 Test quality

Reliability was acceptable at both time-points ([Table 19](#)). Although facility scores were higher than for written recall tests, participants nevertheless found these tests challenging. Discrimination indices were acceptable.

5.6.2 Descriptive statistics

As in written recall, the 12 speaking recall items were scored 0-2, out of a total of 24. The higher means indicated that pupils performed better in the speaking than in the written modality ([Table 23](#)).

Table 23.*Descriptive Statistics for Vocabulary Recall Tests in the Oral Modality*

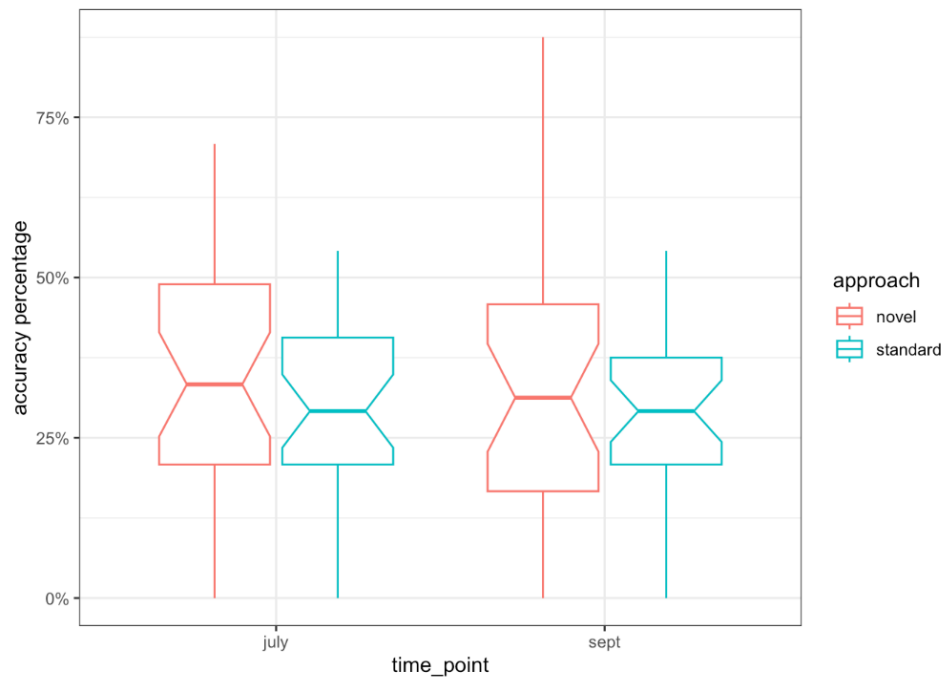
Approach	Time-point ($k = 12$)	Descriptive statistics							
		n	Mean %	SD	CI %	Range /24		Effect size d	Effect size CI
Novel	Post-test	29	66	.42	51-82	0-17	Within-group: Novel Post-test – Delayed post-test	-.10	-.46-.26
Standard	Post-test	29	61	.29	50-72	2-13	Within-group: Standard Post-test – Delayed post-test	.20	-.12-.73
Novel	Delayed post-test	29	68	.43	52-85	0-12	Between-group: Novel – Standard Post-test	.16	-.36-.67
Standard	Delayed post-test	29	57	.24	48-66	1-7	Between-group: Novel – Standard Delayed post-test	.32	-.21-.83

5.6.2.1 Between-groups results

As with the written modality, in the oral modality at post-test and delayed post-test, pupils' mean and median scores were higher in Novel than Standard ([Table 23](#), [Figure 8](#)). Effect sizes at post-test and delayed post-test were medium and large respectively, but CIs crossed zero.

Figure 8.

Notched Box Plot Showing Vocabulary Recall in the Oral Modality Scores by Time-point



5.6.2.2 Within-groups results

Novel scores increased from post-test to delayed post-test, while Standard decreased, both with medium effect sizes. This suggests potential durability of learning in Novel, although the standard deviations were quite high, suggesting individual variation ([Table 23](#)).

5.6.3 Analyses using mixed-effects models

As with the written modality, the model was an ordinal cumulative link mixed-effects model (see [Appendix N.6](#) for code).

5.6.3.1 Assumptions for the regression model

In the Brant test (Brant, 1990) the Omnibus measure was $> .05$, as were p -values for all other predictors ([Appendix N.10](#)); therefore, it was concluded that the proportional odds assumption had not been violated.

5.6.3.2 Predictions based on mixed-effects models: vocabulary recall in the oral modality

The model summary ([Appendix N.11](#)) indicated a possible significant relationship between LAA and vocabulary recall scores in the oral modality (explored below). Fixed effects (marginal R^2) accounted for 8% of variance, while fixed and random effects combined explained 56% of total variance.

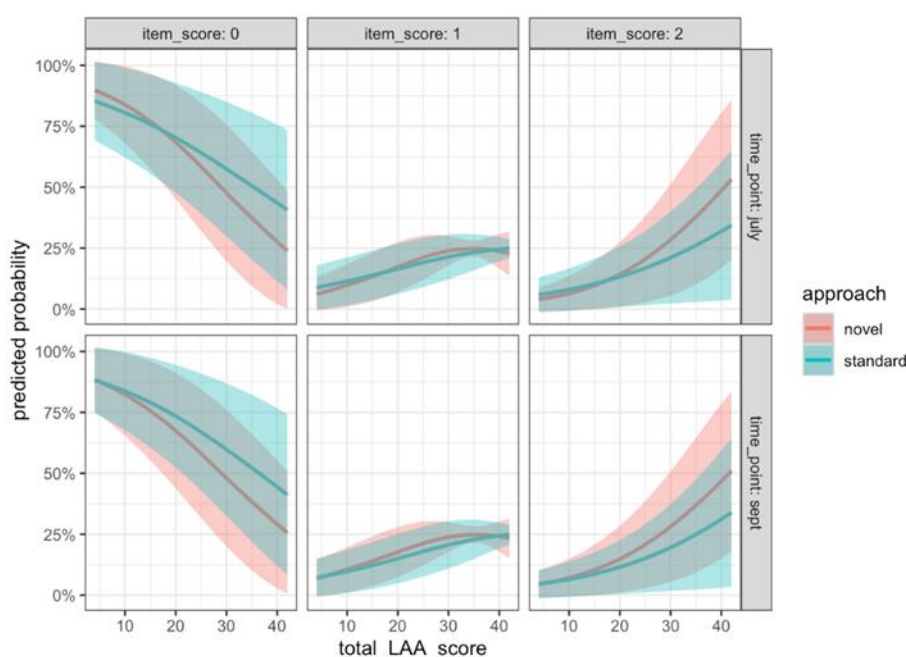
The pairwise comparisons performed in emmeans (Lenth, 2024) with Bonferroni correction (with approach and time-point as variables) confirmed that, while pupils were slightly more likely to score 1 or 2 at both time-points in Novel than Standard, there was no statistically significant difference between approaches in the oral production vocabulary scores at either time-point ([Appendix N.12](#)). The pairwise comparisons also indicated that there was little difference at post-test versus delayed post-test of a pupil in Novel approach scoring 0, 1 or 2, while pupils in Standard were slightly more likely to score 0 at delayed post-test than at post-test, although again these were not statistically significant ([Appendix N.13](#)).

5.6.3.3 The influence of language-analytic ability on oral recall scores

A significant interaction ($< .001$) between LAA scores and pupils' scores in oral vocabulary recall was apparent ([Figure 9](#)). For example, the lower the LAA score, the more likely pupils were to score 0, regardless of approach. Pupils with LAA scores $> \text{ca. } 15\text{-}20$ at post-test and $> \text{ca. } 10$ at delayed post-test were descriptively more likely to score 0 at both time-points if following Standard, whereas there was little difference between approaches among pupils with LAA scores below these. A similar number of pupils were likely to score 1 in both Novel and Standard regardless of LAA level. As LAA scores increased, so did the likelihood of scoring 2 points for pupils in Novel at both time-points. The difference between approaches did not reach statistical significance.

Figure 9.

Vocabulary Recall in the Oral Modality Predictions by Language-analytic Ability Score, Approach and Time-point



5.6.4 Vocabulary recall in the oral modality: summary of results

The reliability of this test was acceptable. As with the other vocabulary measures, there was a positive association between baseline LAA score and performance in this task. Pupils appeared to benefit slightly from being in Novel approach, regardless of baseline LAA, although this was not statistically significant. The effect size was medium, but CIs crossed zero.

5.7 Pupils' learning outcomes in vocabulary: discussion

Vocabulary outcomes are now discussed, comparing recognition and recall in both modalities (including pertinent supporting quotations from qualitative data as appropriate). Comparison sections ([5.7.1.3](#), [5.7.2.3](#), [5.7.3](#)) provide conclusions on the influence of LAA. Brief presentation of pupils' performance on individual lexical items consists of short summaries and illustrations of outcomes relating to word learnability in all tests ([5.7.4](#)); more detailed discussion is provided in [Appendix A.2](#).

5.7.1 Vocabulary recognition: quantitative trends

This section discusses quantitative trends emerging from spoken and written recognition tests.

5.7.1.1 Spoken recognition

At post-test there was little difference between outcomes in the two approaches ([Table 20](#)), suggesting that, at this young age and across a relatively short period of learning time (30 x 45-minute lessons), both approaches offered similar learning opportunities and short-term gains. The effect size between Novel and Standard increased from small (post-test) to large (delayed post-test). Given the greater emphasis in Novel on systematically revisiting vocabulary, generally seen as more effective for long-term gains (Nakata & Suzuki, 2019a; R. A. Schmidt & Bjork, 1992), it might have been hypothesised that a statistically significant difference between approaches would be found at delayed post-test, but this was not the case. The selection of vocabulary content for the test, in its attempt to offer an equitable experience for all learners, drew on items which had been encountered a similar number of times and similarly recently in both approaches ([Appendix E.6](#)). This reduced the number of items available for selection and minimised one of the main differences between the two schemes. Novel included systematic revision of all vocabulary at particular time intervals, whereas Standard taught more vocabulary in total, but not all of it was revisited. It was therefore impossible to test whether Novel was more effective in teaching 'more vocabulary well' relative to Standard.

Another explanation may be that scores were low-mid in spoken recognition with narrow standard deviations, suggesting quite narrow variance. As such, a kind of floor effect and/or the lack of variance meant that between-group differences did not emerge ([Table 20](#)). It should also be remembered that this test had low reliability scores.

5.7.1.2 Written recognition

The significant difference between outcomes in the written modality across time-points indicated an advantage for Novel over Standard ([Table 21](#)), possibly reflecting Novel's greater emphasis on systematic presentation and analysis of the written word, including explicit teaching of word patterns (Hawkes et al., 2019). Pupils' awareness of such patterns was raised when items of vocabulary were introduced, but also when revisited through SSCs. This supports the facilitative effect of orthographic forms on vocabulary learning found by Hu (2008) (see [2.2.4.5](#)).

Higher scores in Novel may also have resulted from the regular vocabulary grid activity, whereby pupils repeatedly revisited key vocabulary in the form of a self-challenge ([Appendix O.1](#)), providing cumulative testing (Maie et al., 2025). This appeared to reinforce form-meaning relationships, as explained by Conklin (2020). It entailed forced memory retrieval, as seen in van den Broek et al.'s (2018) study, which led to higher scores than the condition in which contextual information supported word retrieval in receptive and productive knowledge delayed post-tests. One pupil's comment from end-point Novel group interviews illustrated the benefit of these tasks: *"They helped, like, bring back words to my memory: 'I need to know this word! Oh, yes, now I remember'"* (High language-analytic ability - HLAA, SchE).

5.7.1.3 Comparison of oral and written recognition, with discussion of the influence of language-analytic ability

When disregarding approach, pupils' recognition scores in the written modality (post-test, $M = 7.12$; delayed post-test, $M = 7.26$) were similar to their performance in the oral modality (post-test, $M = 7.68$; delayed post-test, $M = 7.80$). Pupils in Novel performed better at post-test in the written modality than in the oral, but by delayed post-test, there was less difference. In Standard there was little difference between modalities at either time-point. The parity of performances across the two modalities is somewhat surprising: written input is generally assumed to be less time-pressured, thereby offering more opportunity for learners to reflect on meaning, while spoken input is by its very nature "fleeting" (Hu, 2008:824; K. Kim & Godfroid, 2019:651; see also Roehr-Brackin & Tellier, 2019), placing greater burdens on working memory (Kormos, 2020). However, learners were very young in the current study, with L1 literacy still in the early stages (and L2 literacy only just emerging and likely very fragile). As such, they may still be

(more) reliant on an oral modality of learning, meaning that the current findings do not reflect the differences between modalities observed in studies with older learners.

In both modalities in both approaches, pupils appeared to perform better if they had higher baseline levels of LAA. Roehr-Brackin and Tellier (2019) found that one of their tests, measuring grammatical sensitivity and LAA (“Matching Words”), correlated strongly with 8-9-year-old pupils’ performance in French grammar and reading, although less strongly in listening. In this current study, LAA impacted on both listening and reading, although slightly more so in the oral modality ([Figure 5](#)). Additionally, in both modalities, pupils with lower LAA levels appeared to benefit more from being in Novel than Standard. This suggests that the explicit, more analytic approach inherent in Novel minimised the effect of low LAA in vocabulary recognition generally. Although these tests target vocabulary learning (rather than grammar), the results align with Erlam (2005), which found that *all* learners benefited from a deductive approach combined with explicit rule explanation and opportunities for learners to produce the language, suggesting that such an approach may compensate for lower levels of language learning aptitude, perhaps stimulating explicit learning processes (Skehan, 1989; Tellier & Roehr-Brackin, 2013; see [2.4.3.3](#)).

In the oral modality at post-test, those with higher baseline LAA appeared to have an advantage in Standard, but this was not replicated in the written modality where there was no advantage from either approach for pupils with the highest LAA baseline levels. Skehan (1989) cited a study by Nation and McLoughlin (1986) which contrasted implicit and explicit learning conditions and found more effective learning of an artificial language for more novice learners in the explicit condition, while for multilingual learners (considered ‘experts’), instruction type had no impact. Skehan (1989) posited that the more expert learners did not require the explicit rule teaching which characterised the explicit approach and were able to make sense of the new language for themselves. It is possible that higher-achieving learners “are able to cope with less structured or even confusing material, and possibly even thrive on it” (1989:129). It suggests that pupils in this current study with the highest LAA levels possibly functioned more autonomously, relying less on the more analytical approach characterising Novel.

5.7.2 Vocabulary production: quantitative trends

This section discusses quantitative trends emerging from spoken and written recall tests.

5.7.2.1 Spoken production

The individual nature of the oral test necessitated a smaller sample size than the other vocabulary tests; a larger sample size may have led to statistically significant differences (Larson-Hall, 2016). Novel scores were higher at both time-points than Standard, with a large effect size at

delayed post-test ([Table 23](#)). Pupils in Novel appeared to retain their knowledge, with the delayed post-test mean 2% higher than post-test. This may reflect the greater emphasis in Novel on revisiting and recycling vocabulary at spaced intervals, supporting findings from Nakata & Suzuki's (2019a) study where the spacing condition secured long-term vocabulary gains significantly more effectively than the massed condition (whilst acknowledging that in the current study, the comparison was not a massed condition - though vocabulary was indeed less likely to be revisited).

5.7.2.2 Written production

Mean scores in Novel were higher, and the range wider, than in Standard at both time-points, with a more marked difference at post-test. Although smaller, a difference remained at delayed post-test ([Table 22](#)); neither difference was statistically significant. Similarly, Graham et al.'s (2017) study found no statistically significant difference between outcomes in vocabulary and grammar of pupils taught in a mainly oracy-based or literacy-based manner, when pupils' L1 literacy levels were not considered. That study has potential relevance to this one, in that Novel involved slightly more emphasis on the written modality relative to Standard.

5.7.2.3 Comparison of oral and written recall, with discussion of the influence of language-analytic ability

Mean scores for both approaches in oral recall were higher than written recall, with higher facility indices (although still $< .50$), and were similar to the means in both vocabulary recognition tests.

The interaction between baseline LAA and performance in oral and written modalities at both time-points appear to reflect findings from various researchers that learners' ability to recognise and recall vocabulary items may depend on a variety of "learner-related factors" (E. Peters, 2020:125) such as differences in aptitude (Häcker, 2008; Orosz, 2009) and L1 knowledge (S. Webb & Nation, 2017). These factors may have greater influence than the teaching approach, as seen in Hu's (2008) study, when the existing L1 phonological awareness levels of young Mandarin Chinese-speaking beginner learners of English affected the extent to which they benefited from the provision of orthographic forms alongside supporting visuals when learning new vocabulary.

However, in both oral and written recall, outcomes in Novel for those pupils with lower baseline LAA ($< \text{ca. } 30\text{-}35$) appeared to be impacted less than in Standard. This supports findings from oral recognition tests, suggesting that an explicit approach appears to benefit nearly all learners (see Erlam, 2005), but especially those with lower starting points, thus minimising the effect of individual LAA levels. It may also be that those pupils with lower levels of LAA (likely

to be lower achieving in FL generally) benefited from the cumulative testing approach provided by the regular vocabulary grid activity ([O.1](#)). In Maie et al.'s (2025) study with L1 Japanese university-level learners of L2 English, participants in the cumulative testing condition outperformed those in the non-cumulative group in both form and meaning recall tests. Outcomes were affected by number of practice trials but also L2 proficiency, with lower-achieving students appearing to benefit more from the approach than mid- or high-achievers.

5.7.3 Comparison of performance in receptive and productive (written modality) tests

When compared with recognition scores in both modalities, the low means in written recall (the highest being 38% in Novel post-test) support the assertion that productive knowledge is more challenging than receptive (De Groot & Keijzer, 2000; Nation, 2016; Qian & Lin, 2020; Schmitt, 2014; Szpotowicz, 2000; S. Webb & Nation, 2017) and that writing tasks are considered demanding for young learners (Courtney, 2017). Post-test gains in the present study were not reliably durable: scores decreased in both approaches by delayed post-test to around 27%. Nevertheless, even at delayed post-test, *some* written recall – albeit at low levels – was still evident, arguably a noteworthy finding, given these were young learners whose L1 writing skills are also in early stages with relatively little exposure to the written L2.

5.7.4 Individual vocabulary items: results

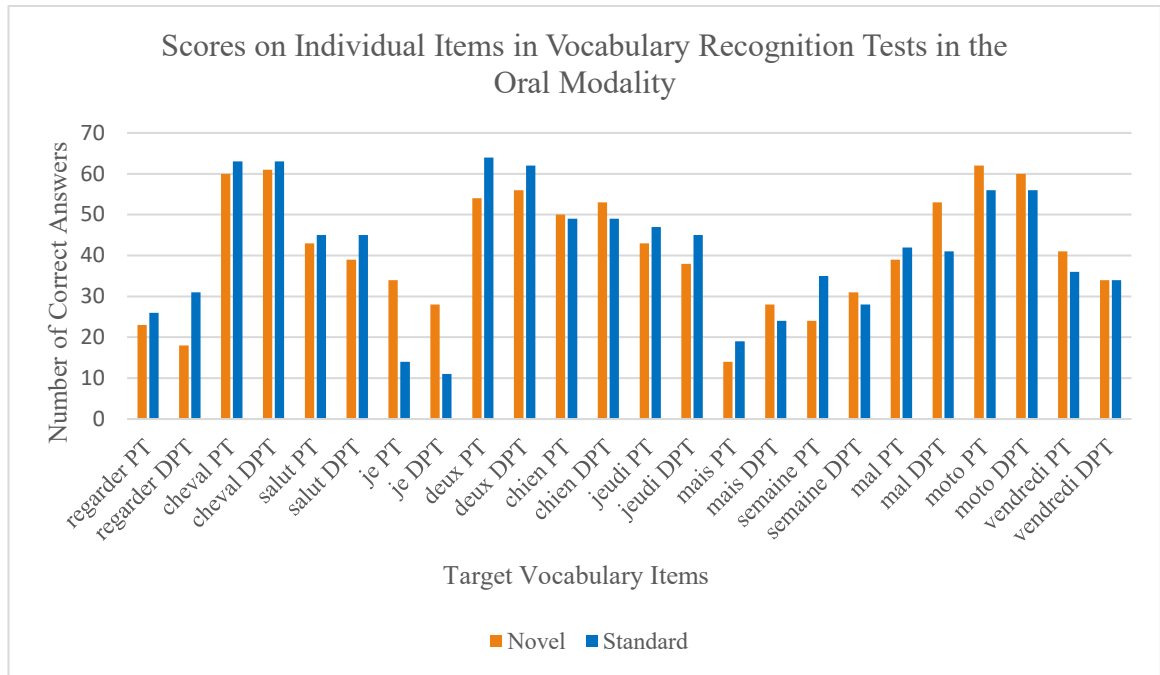
Pupils' performance in individual items in vocabulary recognition and recall is now presented. Discussion appears in [5.7.5](#).

5.7.4.1 Vocabulary recognition in the oral modality

Pupils struggled most with identifying the spoken forms of the verb “*regarder*” (‘to look’), the pronoun “*je*” (‘I’), and the conjunction “*mais*” (‘but’), at both time-points ([Figure 10](#)). (See [Appendix E.34](#) for translations and item order position at post-test and delayed post-test). The three highest-scoring words for both approaches were the nouns “*cheval*” (‘horse’), “*moto*” (‘motorbike’), and the number “*deux*” (‘two’).

Figure 10.

Individual Item Scores in Vocabulary Recognition in the Oral Modality (Post-test and Delayed Post-test)



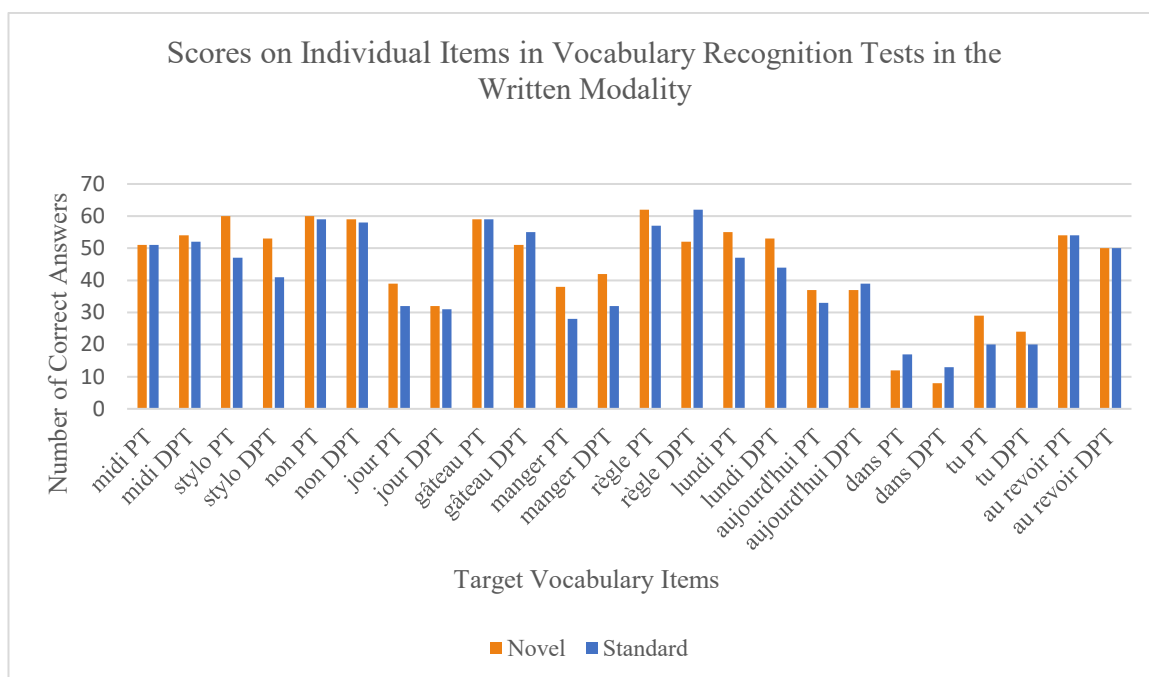
Note. In [Figure 10](#), PT stands for post-test and DPT for delayed post-test.

5.7.4.2 Vocabulary recognition in the written modality

The highest overall scores were on “*stylo*” (‘pen’), “*non*” (‘no’), “*gâteau*” (‘cake’) and “*règle*” (‘ruler’) ([Figure 11](#)), including three concrete nouns. (See [Appendix E.35](#) for translations and item test position). The two lowest-scoring items were the preposition “*dans*” and the personal pronoun “*tu*”.

Figure 11.

Individual Item Scores in Vocabulary Recognition in the Written Modality (Post-test and Delayed Post-test)



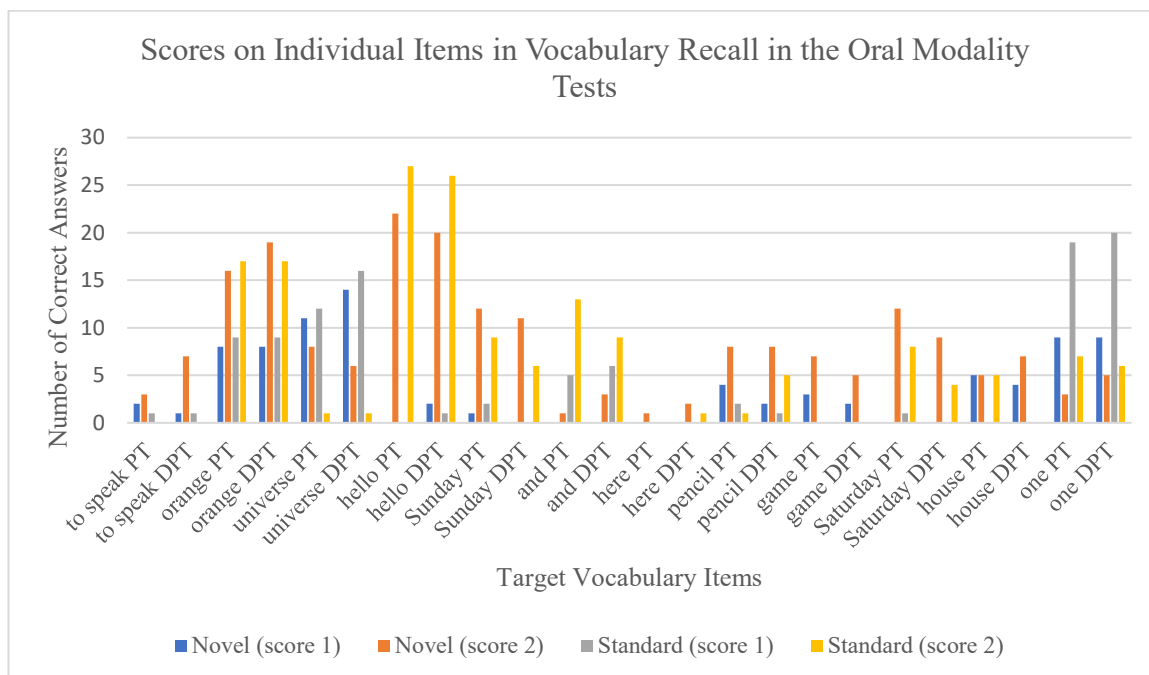
Note. In [Figure 11](#), PT stands for post-test and DPT for delayed post-test.

5.7.4.3 Vocabulary recall in the oral modality

Perhaps unsurprisingly, the most successfully recalled item was “*bonjour*” (‘hello’), likely used or heard fairly regularly by pupils incidentally with teachers ([Figure 12](#)). (See [Appendix E.36](#) for translations and item position at post-test and delayed post-test). The second-highest-scoring items were “*orange*” (‘orange’) for both groups and “*un*” (‘one’/‘a’) in Standard. Pupils recalled the adverb of place “*ici*” (‘here’), the noun “*jeu*” (‘game’) and the verb “*parler*” (‘to speak’) less successfully, especially in Standard.

Figure 12.

Individual Item Scores in Vocabulary Recall in the Oral Modality Tests (Post-test and Delayed Post-test)



Note. In [Figure 12](#), PT stands for post-test and DPT for delayed post-test.

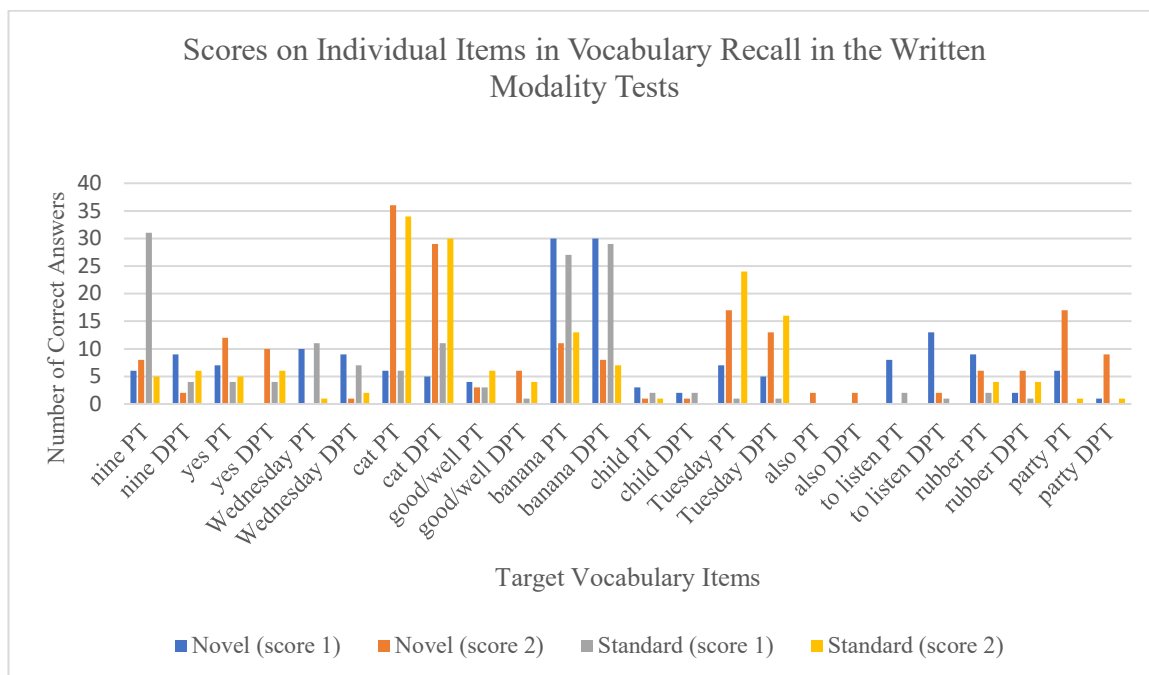
5.7.4.4 Vocabulary recall in the written modality

The item most consistently awarded two marks was “*chat*” (‘cat’), regardless of approach ([Figure 13](#)), then “*banane*” (‘banana’), although this more often scored 1 than 2 points. (See [Appendix E.37](#) for translations and item position at post-test and delayed post-test). Pupils in Standard coped well with “*mardi*” (‘Tuesday’). Some words proved challenging for pupils in both approaches: “*aussi*” (‘also’), “*bien*” (‘good/well’) and “*enfant*” (‘child’).

The only verb in this set of items (“*écouter*” - ‘to listen’) was more often partially successfully attempted by pupils in Novel than Standard. Pupils in Novel also performed better than those in Standard on “*oui*” (‘yes’) and the noun “*fête*” (‘party’/‘celebration’). Pupils in Standard more accurately produced “*neuf*” (‘nine’) than Novel.

Figure 13.

Individual Item Scores in Vocabulary Recall in the Written Modality Tests (Post-test and Delayed Post-test)



Note. In [Figure 13](#), PT stands for post-test and DPT for delayed post-test.

5.7.5 Discussion of individual items: the challenges of predicting word difficulty

Vocabulary content of the four tests was controlled for on a number of factors known to affect learnability (see [2.2.3](#)). Pupils' (non-approach-specific) performance on individual lexical items is discussed according to these factors in [Appendix A.2](#). ([Appendix E.3](#) provides a list of words allocated to each test type).

Often learnability appeared to be impacted by combinations of two or more factors from word class/length, frequency or recency in the experienced environment (to some extent positively correlated with frequency in the wider world), concreteness and cognate status. A possible additional explanation for learnability may arise from the approaches themselves, closely aligned with word class. For example, scores in Standard, which taught numbers as a lexical set, were higher across three tests on all three numerals than in Novel; Novel scores were higher on three of the four verbs and both personal pronouns. Novel's more explicit focus on sentence construction, including referential activities from PI ([2.4.1.4](#)), with the ultimate goal of enabling learners to "create their own sentences and increase[s] their ability to manipulate verbs" (Hawkes et al., 2019:4), possibly increased the salience of verbs for learners. This would support Häcker's (2008) assertion that explicit grammar instruction combined with vocabulary instruction increases learners' understanding of individual sentence components.

One final consideration is the relevance of vocabulary selection to these young learners. The high scores in both approaches on animal words aligned with findings from Szopotowicz’s (2000) small-scale study, in which the best-remembered words were from the semantic group of animals (the children’s favourite topic). In the current study, pupils cited animal vocabulary as reasons for liking French in both approaches, referring in group interviews to the appeal of animals generally: “[I like] naming the animals because I like animals a lot” (Low language-analytic ability - LLAA, Novel). While this is not suggesting that pupils should only be taught words for things of interest, it acknowledges the importance of word choices taking account of the learner’s age and needs (Vilkaitė-Lozdienė & Schmitt, 2020), particularly the potential appropriateness of animal-related words for younger learners (see S. Webb & Nation, 2017).

5.7.6 Overarching discussion of vocabulary outcome measures

Even when pupils found tests challenging, there was evidence that young learners *can* make noticeable progress in vocabulary. The overview (Table 24) shows that outcomes were more often durable in Novel than Standard, and pupils benefited slightly more from being in Novel approach in all measures except spoken recognition at post-test (where reliability was low at both time-points). At no point were pupils disadvantaged by being in Novel. Indeed, the significant interaction between LAA and learning outcomes appeared to be somewhat tempered by Novel, with its explicit approach possibly minimising the effects of low LAA levels. Pupils with the highest baseline LAA levels appeared to perform better in Standard only in spoken recognition and written recall tests.

Table 24.

Overview of Vocabulary Outcomes

Vocabulary measure	Over time		Between approaches	
	Novel	Standard	Post-test	Delayed post-test
Spoken recognition	Post-test < Delayed post-test (trend)	Post-test = Delayed post-test ($M = 59:58$)	$N = S$ ($M = 60:59$)	$N > S$ (trend)
Written recognition	Post-test > Delayed post-test (trend)	Post-test = Delayed post-test ($M = 60:59$)	$N > S$ (statistically significant)	$N > S$ (trend)
Spoken recall	Post-test < Delayed post-test (trend)	Post-test > Delayed post-test (trend)	$N > S$ (trend)	$N > S$ (trend)
Written recall	Post-test > Delayed post-test (trend)	Post-test > Delayed post-test (trend)	$N > S$ (trend)	$N > S$ (trend)

Note. Trend reflected in descriptives and (bordering on) medium effect size (.05 upwards).

5.8 Phonics transcription: results

The results from the two phonics tests will now be presented: transcription ([5.8](#)) and Read Aloud ([5.9](#)). [Table 25](#) provides information on test quality for both phonics measures. All pupil participants ($n = 69$ Novel, $n = 72$ Standard) completed a transcription test ([E.21](#)) at both time-points. [Appendix E.18](#) contains a list of host words for target SSCs in transcription and their position at post-test and delayed post-test.

5.8.1 Test quality

An EFA, used because of issues encountered in the CFA, showed unacceptable reliability ([Table 25](#)). For this reason, only descriptive statistics are presented and discussed, rather than the full model, as reliability may be questioned. The tests discriminated effectively, although low facility scores showed that pupils found transcription difficult.

Table 25.
Test Quality Information for Both Phonics Tests

		Transcription		Read	Aloud
		Post-test	Delayed post-test	Post-test	Delayed post-test
% cases complete prior to imputation		96	96	98	98
Reliability (EFA)	alpha	.21	.36	.57	.67
	omega	.40	.46	.69	.76
Facility		.28	.25	.29	.27
		<i>SD</i> .18	<i>SD</i> .18	<i>SD</i> .26	<i>SD</i> .22
Discrimination		.31	.30	.35	.42
		<i>SD</i> .18	<i>SD</i> .15	<i>SD</i> .24	<i>SD</i> .25
Normality of distribution link		Appendix M.13		Appendix M.16	Return to 5.9.1

5.8.2 Descriptive statistics for transcription

Transcription test data were binary (answers correct or incorrect), giving a possible total of 10 from 10 items. Scores ranged from 0-7 across all tests, but the mean was < 3 in both approaches

at both time-points ([Table 26](#)). Scores for both approaches were very slightly lower at delayed post-test.

Table 26.

Descriptive Statistics for Phonics Transcription Tests

Approach	Time-point ($k = 10$)	Descriptive statistics							
		n	Mean %	SD	CI %	Range /12		Effect size d	Effect size CI
Novel	Post-test	69	28	.15	24-32	0-7	Within-group: Novel Post-test – Delayed post-test	.18	-.05 - .44
Standard	Post-test	72	27	.14	24-31	0-6	Within-group: Standard Post-test – Delayed post-test	.22	.01 - .45
Novel	Delayed post-test	69	25	.15	21-28	0-6	Between-group: Novel – Standard Post-test	.04	-.32 - .36
Standard	Delayed post-test	72	24	.14	21-28	0-6	Between-group: Novel – Standard Delayed post-test	.02	-.32 - .36

5.8.2.1 Between-groups results

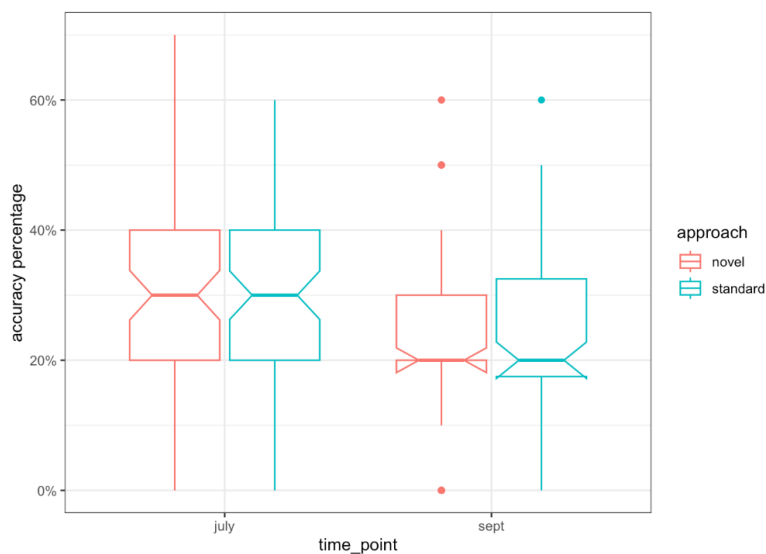
At both time-points, mean scores were similar for both approaches ([Table 26](#)) and effect sizes were small, with CI crossing zero, suggesting no differences between groups. Medians were similar across approaches ([Figure 14](#)).

5.8.2.2 Within-groups results

Effect sizes of the difference over time between the means in each approach were medium in Novel ([Table 26](#)) (crossing zero and suggesting the decrease at delayed post-test was not meaningful) and large in Standard (possibly more meaningful as the CIs did not cross zero).

Figure 14.

Notched Box Plot Showing Phonics Transcription Scores by Time-point

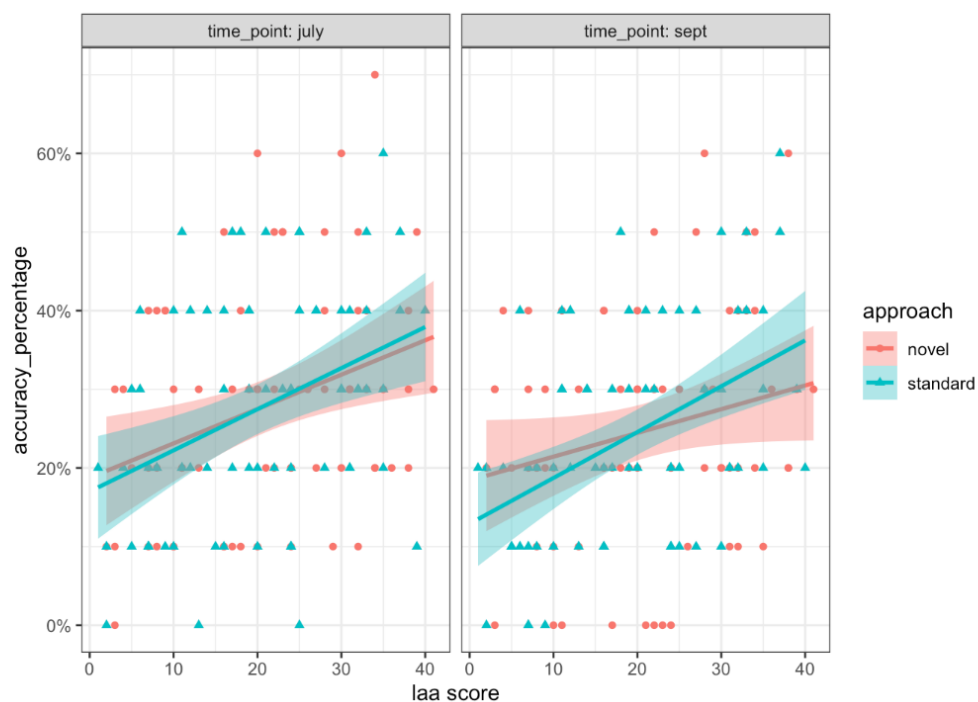


5.8.2.3 The influence of LAA on transcription scores

As can be seen in [Figure 15](#) – created from the inferential models which are not presented due to the very low reliability of the test itself - pupils with lower levels of LAA seemed to score lower in transcription than those with high LAA, regardless of approach.

Figure 15.

Scatter Plot Showing Phonics Transcription Scores According to Language-analytic Ability Score by Approach and Time-point



5.8.3 Phonics transcription: summary of results

Descriptive statistics showed similarity between Novel and Standard outcomes at both time-points, with scores very slightly lower at delayed post-test. They indicated the influence of LAA on pupils' scores in both approaches. However, reliability indices were low; statistical significance was not examined but seems likely to be absent.

5.9 Phonics Read Aloud (RA): results

A subset of 29 pupils from each approach ([4.3.6.4, ii](#)) completed RA tests ([E.22](#)).

5.9.1 Test quality

An EFA, used due to issues encountered in the CFA, showed moderately acceptable (post-test)/good (delayed post-test) reliability. As in transcription, RA tests were challenging for pupils, with facility scores slightly below the recommended benchmark. Discrimination scores were good ([Table 25](#)).

5.9.2 Descriptive statistics for Read Aloud

The RA test consisted of 10 items, targeting the same SSCs as transcription, but with different host words ([Appendix E.19](#)). Each item was scored on a 0-1-2 basis (giving a total of 20): 2 represented easily understandable, though not necessarily completely authentic, pronunciation.

5.9.2.1 Between-groups results

The range of scores was wider in Novel than Standard at post-test, although this difference decreased by delayed post-test ([Table 27](#)). [Figure 16](#) illustrates Novel's higher median at both time-points, although the overlapping notches indicate that this difference was unlikely to be statistically significant. The effect size of the post-test difference between approaches was large; at delayed post-test ([Table 27](#)) it was very small, in favour of Standard. However, both passed through zero.

Table 27.

Descriptive Statistics for Phonics Read Aloud Tests

Approach	Time-point (<i>k</i> = 10)	Descriptive statistics							
		<i>n</i>	<i>Mean</i> %	<i>SD</i>	<i>CI</i> %	Range /12		Effect size <i>d</i>	Effect size <i>CI</i>
Novel	Post-test	29	32	.18	25-39	0-15	Within-group: Novel	.34	.00-.66

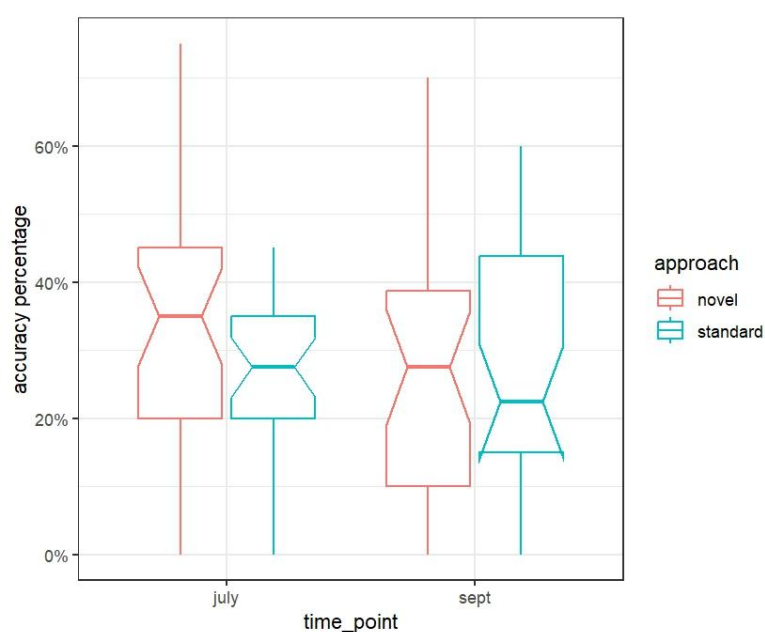
Standard	Post-test	29	26	.13	21-31	0-9	Post-test – Delayed post-test Within- group: Standard Post-test – Delayed post-test	-.03	.38-.42
Novel	Delayed post-test	29	27	.20	19-33	0-14	Between- group: Novel – Standard Post-test	.38	-.12 - .91
Standard	Delayed post-test	29	27	.18	20-33	0-12	Between- group: Novel – Standard Delayed post-test	-.01	-.54 - .52

5.9.2.2 Within-groups results

Pupils in Novel appeared to score less well at delayed post-test than post-test, with a large effect size, but they had more room to decrease than Standard, having made greater post-test gains (Table 27). Standard scores were very slightly higher at delayed post-test than post-test, with a small effect size.

Figure 16.

Notched Box Plot Showing Phonics Read Aloud Scores by Time-point



5.9.3 Analyses using mixed-effects models

An ordinal cumulative link mixed-effects model was constructed for the RA tests (see [Appendix N.14](#) for code).

5.9.3.1 Assumptions for the regression model

Assumptions were checked after running the model. The Omnibus measure was $> .05$, but as the p -value for approach was $< .05$ ([Appendix N.15](#)), results should be interpreted with caution.

5.9.3.2 Predictions based on mixed-effects models

The model summary ([Appendix N.16](#)) indicated significant relationships between LAA and RA scores ($p < .001$) and between LAA score and time-point ($p = .02$) (explored below). Fixed effects accounted for 5% of variance, while fixed and random effects combined accounted for 59% of total variance.

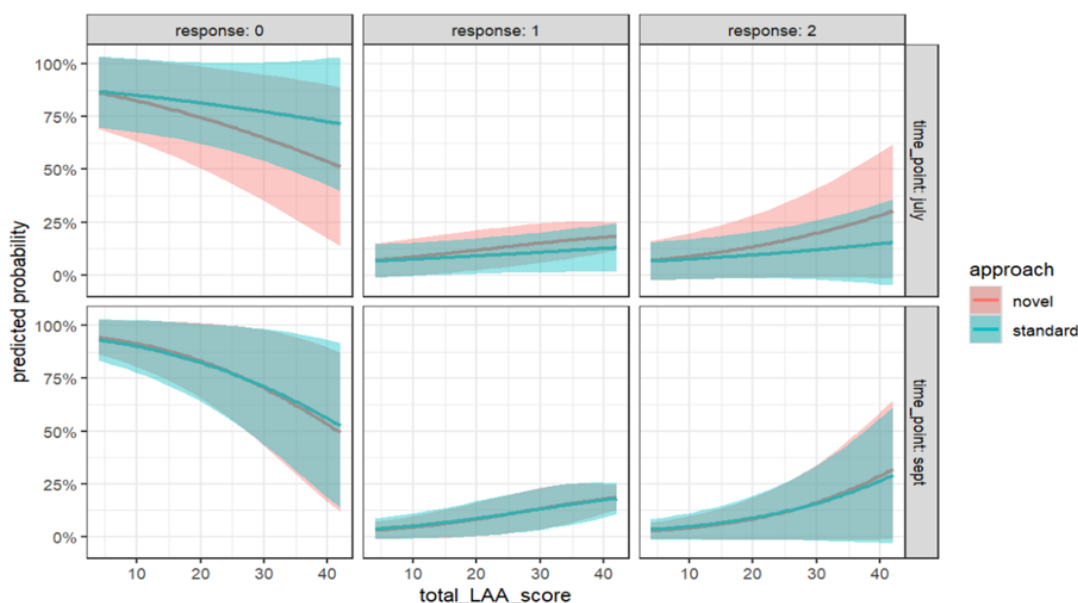
Pairwise comparisons using Bonferroni correction, with item score, approach and time-point as predictors, indicated a statistically significant difference between Novel scores over time (estimate = .50, SE = .22, $z = 2.31$, $p = .02$): the probability of pupils in Novel scoring 0 increased, and of scoring 1 or 2 decreased by delayed post-test ([Appendix N.17](#)). There was no statistically significant difference in Standard between the two time-points (estimate = .03, SE = .21, $z = .16$, $p = .87$), nor between approaches, at either post-test (estimate = .41, SE = .30, $z = 1.36$, $p = .17$) or delayed post-test (estimate = -.05, SE = .31, $z = -.17$, $p = .86$).

5.9.3.3 The influence of language-analytic ability on Read Aloud scores

As the model summary indicated a significant interaction between LAA and RA scores, predicted values were plotted ([Figure 17](#)). This showed that, at both time-points, the highest RA scores were achieved by pupils with the highest baseline levels of LAA. Nearly all pupils (i.e., with LAA scores of above ca. 5-15) were descriptively more likely at post-test to achieve a score of 2 in Novel than in Standard; this likelihood increased in accordance with their LAA baseline. Apart from those pupils with the lowest LAA scores (below ca. 5-10), participants in Novel were less likely to score 0 than those in Standard at post-test. Very little difference was apparent at delayed post-test, indicating that learning gains in Novel had disappeared.

Figure 17.

Phonics Read Aloud Predictions by Language-analytic Ability Score, Approach and Time-point



5.9.4 Phonics Read Aloud: summary of results

Scores in Novel were higher at post-test than delayed post-test, with a statistically significant difference found between these time-points, whereas Standard scores were similar over time. LAA appeared to interact with outcomes: pupils in both approaches with higher baseline LAA scores performed better in RA tests at both time-points.

5.10 Pupils' learning outcomes in phonics: discussion

Pupil outcomes in phonics tests are now discussed, including conclusions relating to the influence of LAA ([5.10.3](#)). Pupils' performance on individual SSCs is presented here, consisting of short summaries and illustrations of outcomes relating to SSC transcription and oral production. As this study did not compare the impact of approaches on performance in individual SSCs (indeed, item was a random effect in the mixed-effects models, as it was acknowledged that different SSCs would react differently to the treatments, and would equally change differently over time), this section therefore relates to general (rather than approach-specific) findings emerging from the phonics tests, linking it to existing literature in this field. A more detailed narrative discussion appears in [Appendix A.3](#).

5.10.1 Phonics transcription: quantitative trends

This section discusses quantitative trends emerging from transcription tests.

5.10.1.1 Between-groups

It is perhaps somewhat surprising that there was not a greater difference between outcomes in Novel and Standard in transcription. Novel's systematic and explicit focus on phonological decoding addressed a need identified by, among others, TSC (2016) and Holmes and Myles (2019). This was not emphasised to the same extent in Standard, although the latter did include an explicit focus on SSCs (only) at the start of the academic year, which very likely constituted a greater focus than would be found in other existing schemes of work currently in use in English primary schools.

5.10.1.2 Within-groups

The low mean scores (< 30%) of both groups at both time-points indicated how challenging pupils found these tests, a finding supported by low facility scores. Learning of SSCs did not appear secure, with performances in both approaches deteriorating over time ([Figure 14](#)). Pupils appeared to need longer exposure to key phonemic differences between the two languages in order to recognise FL equivalents, as identified by Woore et al. (2018), and to suppress the influence of their L1 understanding.

5.10.2 Phonics Read Aloud: quantitative trends

There was evidence of greater knowledge of SSCs at post-test, particularly among pupils in Novel, whereas the lower delayed post-test scores indicated that this learning was not durable. Novel scores showed greater deterioration between post-test and delayed post-test than Standard but, having increased at post-test more than the comparison group, they were more likely to decrease later (RTM). The more explicit SSC focus in Novel possibly had short-term benefits but pupils needed a longer period of instruction for this to become concretised (see Woore et al., 2018).

5.10.3 Comparison of transcription and Read Aloud, with discussion of the influence of language-analytic ability

Scrutiny of both tests indicates that these are areas which young learners find particularly challenging, at least after just 22.5 hours of instruction over 11 months. The similarity in pupils' means in both test types (transcription scores below 28%, RA scores below 32%) indicates that pupils generally found reading unknown French words aloud with approximately accurate pronunciation as difficult as they did transcribing SSCs. This appears to support findings from Erler (2004) and Erler and Macaro (2011), in which studies secondary-aged pupils struggled with accurate decoding in French.

The influence of LAA (even though the LAA test did not include a phonemic coding element) was apparent in both transcription and RA tests, with lower baseline levels appearing to impact negatively on performance. As with vocabulary outcomes, pupils in Novel appeared to have a slight advantage compared with Standard: lower LAA levels impacted less on these learners. Novel’s systematic, explicit approach to phonological decoding may have had short-term benefits but clearly needed longer input for durable impact.

5.10.4 Individual items: results

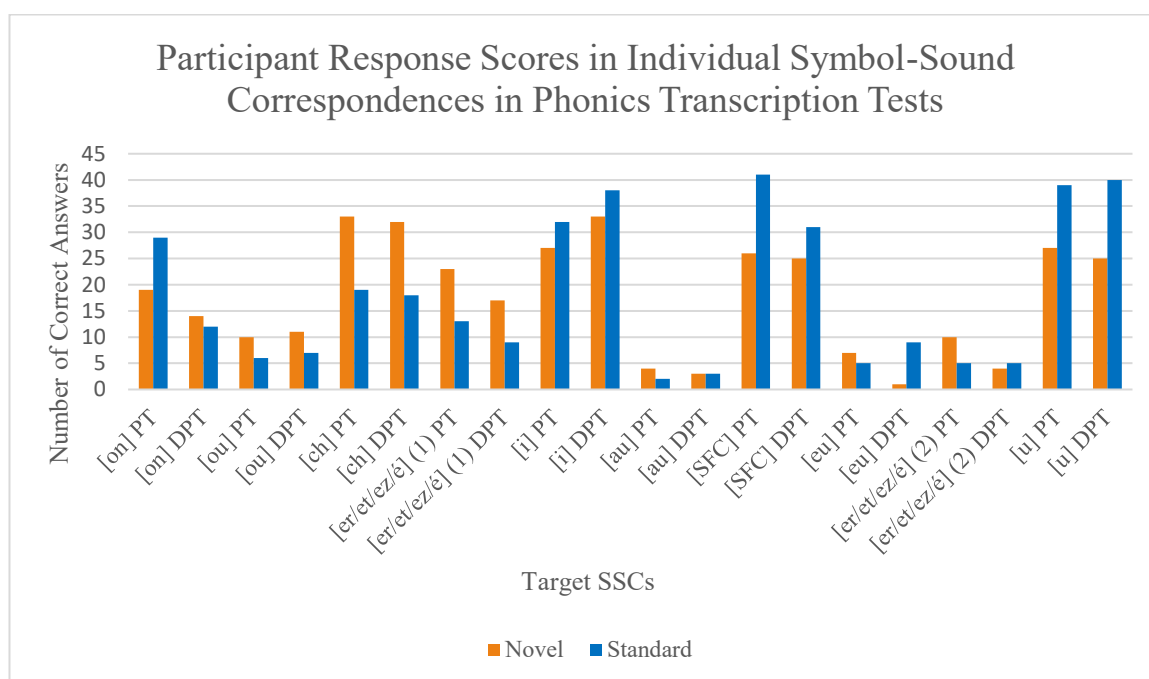
A brief overview of pupils’ performance on individual SSCs in transcription and RA tests is now presented.

5.10.4.1 Transcription

[Figure 18](#) shows the number of correct answers for each target SSC at post-test and delayed post-test, demonstrating that the following SSCs appeared easiest for all pupils to identify (albeit with approach-specific nuances): [ch]; [u]; [i]; and the silent final consonant (SFC). Some SSCs proved challenging at both time-points regardless of approach: [ou], [au] and [eu]. For all host words, see [Appendix E.18](#).

Figure 18.

Scores in Individual Symbol-Sound Correspondences in Transcription Tests (Post-test and Delayed Post-test)



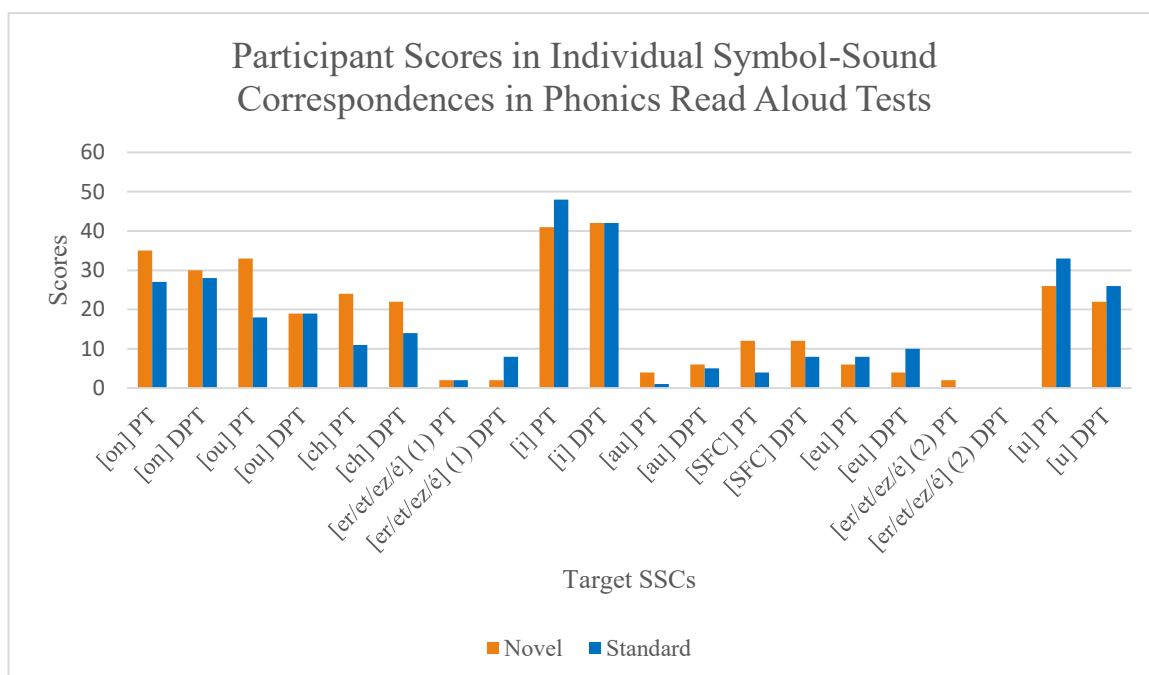
Note. In [Figure 18](#), PT stands for post-test and DPT for delayed post-test.

5.10.4.2 Read Aloud

As with transcription, pupils appeared to find some SSCs easier than others ([Figure 19](#)). For example, reading [i] aloud proved to be less demanding than other SSCs, as in transcription. Pupils coped reasonably well with [u], particularly in Standard, and [on] and [ch], particularly in Novel. Pupils found it challenging to orally produce [eu] and [au]. The SSCs [er], [et], [ez] and [é] appeared demanding, regardless of the host word.

Figure 19.

Scores in Individual Symbol-Sound Correspondences in Read Aloud Tests (Post-test and Delayed Post-test)



Note. In [Figure 19](#), PT stands for post-test and DPT for delayed post-test.

5.10.5 Overarching discussion of phonics outcome measures

[Appendix A.3](#) provides detailed discussion, situating outcomes in the field of existing research. Pupils clearly needed input of a longer duration for real progress to be made in phonics transcription and RA tasks. Scores decreased or remained unchanged at delayed post-test in both approaches. Although there was no difference between approaches in transcription scores (where reliability was unacceptable), there was a trend at post-test in RA for pupils in Novel to outperform those in Standard ([Table 28](#)); pupils were not disadvantaged by having experienced the systematic and explicit phonics teaching in Novel. As with tests of vocabulary knowledge, Novel's explicit approach appeared to counteract the effects of low LAA levels in RA at least.

Table 28.*Overview of Phonics Outcomes*

Phonics measure	Over time		Between approaches	
	Novel	Standard	Post-test	Delayed post-test
Transcription	Post-test > Delayed post-test (trend)	Post-test > Delayed post-test (trend)	N = S ($M = 28:27$)	N = S ($M = 25:24$)
Read Aloud	Post-test > Delayed post-test (trend)	Post-test = Delayed post-test ($M = 26:27$)	N > S (trend)	N = S ($M = 27:27$)

Note. Trend reflected in descriptives and (bordering on) medium effect size (.05 upwards).

Scores did not indicate statistically significant differences between approaches in either transcription or RA tests, but qualitative group interview data appeared to suggest greater awareness and understanding of SSCs in Novel. Pupils from both approaches demonstrated developing awareness that sounds differed between the two languages, and the enjoyment of these sounds among pupils in Novel was evident: “[*Learning French*] it’s really fun ... when we say ‘oui’ it means ‘yes’ but it sounds like you’re going flying” (Novel, LLAA, SchF). This impact on pupil attitudes (and particularly how they felt about writing words) is discussed in [6.7.2.2, ii.](#)

5.11 Written grammar recognition: results

This section presents written grammar recognition ([E.27](#)) results, the only grammar test completed by all participants (a spoken grammar recall test, [5.12](#), and an EIG test, [5.13](#), were completed by a subset of participants). [Appendix E.23](#) contains a list of items and their position at post-test and delayed post-test. Test quality for all grammar measures appears in [Table 29](#).

5.11.1 Test quality

The CFA results showed very low reliability ([Table 29](#)). An EFA showed higher reliability but, as values were still low, results from both descriptive and inferential statistics should be interpreted with caution. Facility scores, lower than ideal, suggested that pupils found written grammar recognition challenging at both time-points. Furthermore, tests did not discriminate well between the top and bottom thirds of participants.

Table 29.*Test Quality Information for All Grammar Tests*

Written recognition	Spoken recall	Elicited imitation

		Post-test	Delayed post-test	Post-test	Delayed post-test	Post-test	Delayed post-test
% cases complete prior to imputation		97	97	97	97	97	97
Reliability (CFA)	alpha	.33	.08	.73	.76		
	omega	.13	.19	.87	.88		
Reliability (EFA)	alpha	.41	.41			.79	.85
	omega	.48	.49			.82	.87
Facility		.39 <i>SD</i> .12	.38 <i>SD</i> .14	.16 <i>SD</i> .10	.16 <i>SD</i> .10	.37 <i>SD</i> .17	.40 <i>SD</i> .18
Discrimination		.24 <i>SD</i> .09	.20 <i>SD</i> .16	.33 <i>SD</i> .23	.39 <i>SD</i> .41	.39 <i>SD</i> .19	.47 <i>SD</i> .16
Normality of distribution link		Appendix M.19		Appendix M.22		Appendix M.25	

Return to [5.12.1](#), [5.13.1](#).

5.11.2 Descriptive statistics for the written grammar recognition task

Participants answered 18 multiple choice items, six targeting each of three areas: “avoir”/“être” verbs; present tense ‘-er’ verb endings; and adjectival agreement.

5.11.2.1 Between-groups results

At post-test, there was no meaningful difference between Novel and Standard mean test percentages ([Table 30](#)). At delayed post-test, Standard scores were very slightly higher than Novel with a medium effect size, but CIs crossed zero.

Table 30.

Descriptive Statistics for Written Grammar Recognition Tests

Approach	Time-point (<i>k</i> = 18)	Descriptive statistics							
		<i>n</i>	<i>Mean</i> %	<i>SD</i>	<i>CI</i> %	Range %		Effect size <i>d</i>	Effect size CI
Novel	Post-test	69	39	.14	35-42	5-72	Within-group: Novel Post-test – Delayed post-test	.08	-.05 - .31
Standard	Post-test	71	39	.13	36-42	17-61	Within-group: Standard	-.02	-.21 - .26

Novel	Delayed post-test	69	37	.09	35-40	11-72	Post-test – Delayed post-test Between-group: Novel – Standard	-.04	-.40 - .30
Standard	Delayed post-test	71	39	.13	36-42	11-78	Post-test Between-group: Novel – Standard delayed Post-test	-.12	-.46 - .18

5.11.2.2 Within-groups results

Standard scores remained constant across time-points, whereas those in Novel decreased slightly (with a medium effect size, but which crossed zero).

5.11.3 Analyses using mixed-effects models

A generalised linear mixed-effects binomial model was constructed for written grammar recognition (see [Appendix N.18](#) for code).

5.11.3.1 Assumptions for the regression model

Assumptions were checked after running the model. Overdispersion was not detected ($p = 1.00$), and the Box-Tidwell test results indicated that the linearity assumption held for the LAA score ([Appendix N.19](#)).

5.11.3.2 Predictions based on mixed-effects models

The model summary ([Appendix N.20](#)) indicated a possibly significant relationship between baseline LAA and written grammar recognition scores, and also between approaches, time-points and LAA score. Results may however be misleading due to involvement in interactions. Fixed effects (marginal R^2) explained 1% of variance, while fixed and random effects combined (conditional R^2) explained 10% of total variance.

In pairwise comparisons with Bonferroni correction ([Appendix N.21](#)), p -values $> .05$ at both time-points suggested no statistically significant difference between outcomes in each approach. This was also the case in within-groups comparisons ([Appendix N.22](#)).

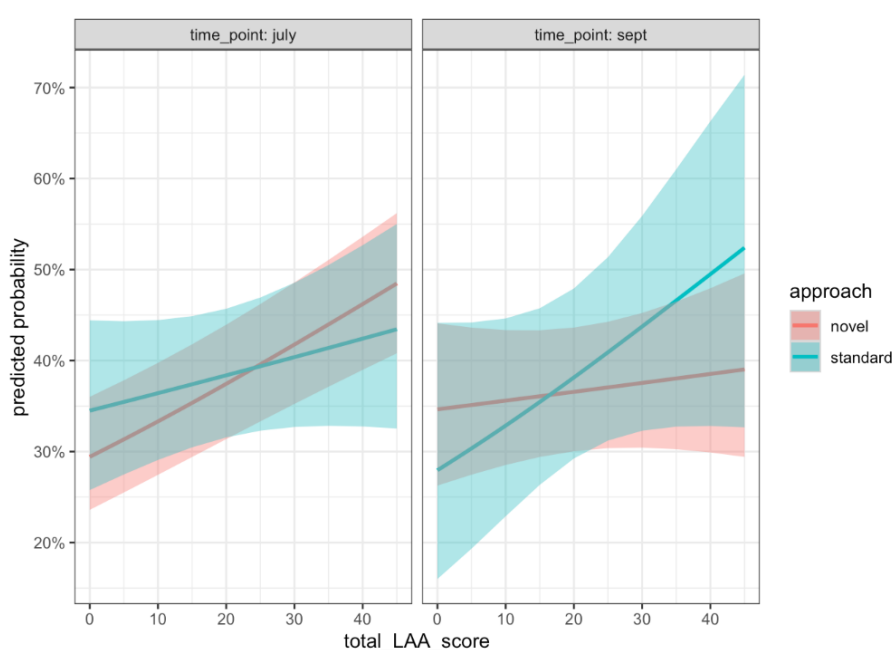
5.11.3.3 The influence of language-analytic ability on written grammar recognition scores

The significant interaction ($< .001$) between baseline LAA and pupils' scores varied by approach and time-point ([Figure 20](#)). At post-test, the relationship between LAA score and outcomes was evident: the higher the LAA score, the better the participant's performance in written grammar

recognition, particularly in Novel. By delayed post-test, the relationship between LAA scores and performance was less apparent in Novel, but more evident in Standard. At post-test, pupils with lower LAA scores (< ca. 20-30) appeared descriptively to perform better in Standard than Novel, whereas pupils with scores above this performed better in Novel than Standard. The reverse was apparent at delayed post-test, with pupils in Novel with LAA scores below ca. 12-17 achieving higher scores than those in Standard, while those with scores above this appeared to perform better in Standard.

Figure 20.

Written Grammar Predictions by Language-analytic Ability Score, Approach and Time-point



5.11.4 Written grammar recognition: summary of results

There was little difference in scores in each approach at either time-point. In Novel, the influence of LAA was more noticeable at post-test than at delayed post-test, whereas in Standard the opposite was the case. Reliability scores for these tests were however below satisfactory.

5.12 Spoken grammar recall: results

This test ([E.28](#)) was completed by a subset of 58 participants ([4.3.6.4, ii](#)). [Appendix E.24](#) contains a list of items and their position at post-test and delayed post-test.

5.12.1 Test quality

Reliability scores were good at both time-points. Although the tests discriminated effectively between the top and bottom thirds of participants, low facility scores indicated that pupils generally found this test very challenging ([Table 29](#)).

5.12.2 Descriptive statistics for the spoken grammar recall task

This consisted of 14 items targeting four areas of grammatical production: “*avoir*”/“*être*” verbs (four items); pronouns (five items); present tense ‘-er’ verb endings (three items); and adjectival agreement (two items). Items were scored 0-1-2, making a total score of 28 possible.

5.12.2.1 Between-groups results

Pupils in Novel appeared to perform better at both time-points than pupils in Standard ([Table 31](#), [Figure 21](#)). Standard deviations were slightly higher in Novel than Standard, indicating a wider spread of data with greater variance, although the raw mark range was similar. CIs were wider for Novel at both time-points; the lack of overlapping notches between Novel and Standard CIs at post-test ([Figure 21](#)) suggested a possibly statistically significant difference. Effect sizes between approaches were large at both time-points; CIs did not cross zero at post-test, but just crossed at delayed post-test.

Table 31.

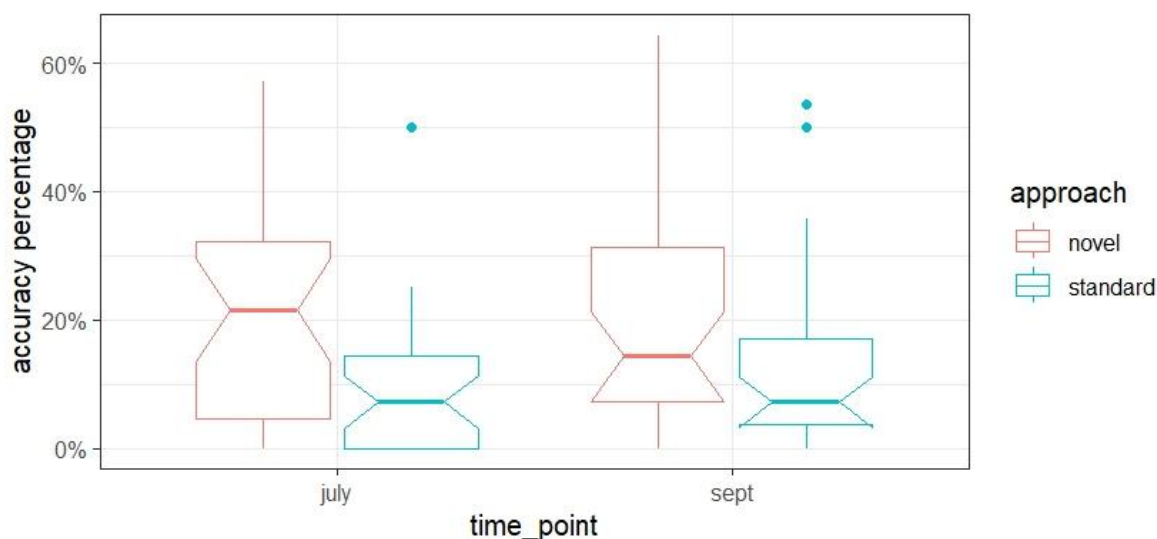
Descriptive Statistics for Spoken Grammar Recall Tests

Approach	Time-point (<i>k</i> = 14)	Descriptive statistics							
		<i>n</i>	<i>Mean</i> %	<i>SD</i>	<i>CI</i> %	Range %		Effect size <i>d</i>	Effect size <i>CI</i>
Novel	Post-test	29	42	.33	30-55	0-16	Within-group: Novel Post-test – Delayed post-test	.17	-.18-.63
Standard	Post-test	29	18	.22	10-26	0-14	Within-group: Standard Post-test – Delayed post-test	-.21	-.51-.18
Novel	Delayed post-test	29	38	.33	25-50	0-18	Between-group: Novel – Standard Post-test	.85	.35 - 1.47
Standard	Delayed post-test	29	24	.29	14-35	0-15	Between-group:	.43	-.07 - .99

Novel –
Standard
Delayed
post-test

Figure 21.

Notched Box Plot Showing Spoken Grammar Recall Scores by Time-point



5.12.2.2 Within-groups results

Standard scores increased between post-test and delayed post-test (with a large effect size), whereas pupils in Novel did not score as highly at delayed post-test as at post-test (with a medium effect size). However, Novel scores were higher at post-test than Standard and were therefore more likely to decrease in the delayed measure.

5.12.3 Analyses using mixed-effects models

The model constructed for spoken recall (see [Appendix N.23](#) for code) was a cumulative link mixed-effects ordinal model.

5.12.3.1 Assumptions for the regression model

Assumptions were checked after running the model using the Brant test (Brant, 1990). The Omnibus measure was $> .05$, as were all other p -values for assumption checks ([Appendix N.24](#)).

5.12.3.2 Predictions based on mixed-effects models

The model summary ([Appendix N.25](#)) showed significant interactions between outcomes and approach, and outcomes and baseline LAA scores. Fixed effects (marginal R^2) accounted for 10% of variance in the model, while 48% (conditional R^2) of total variance was explained by fixed and random effects.

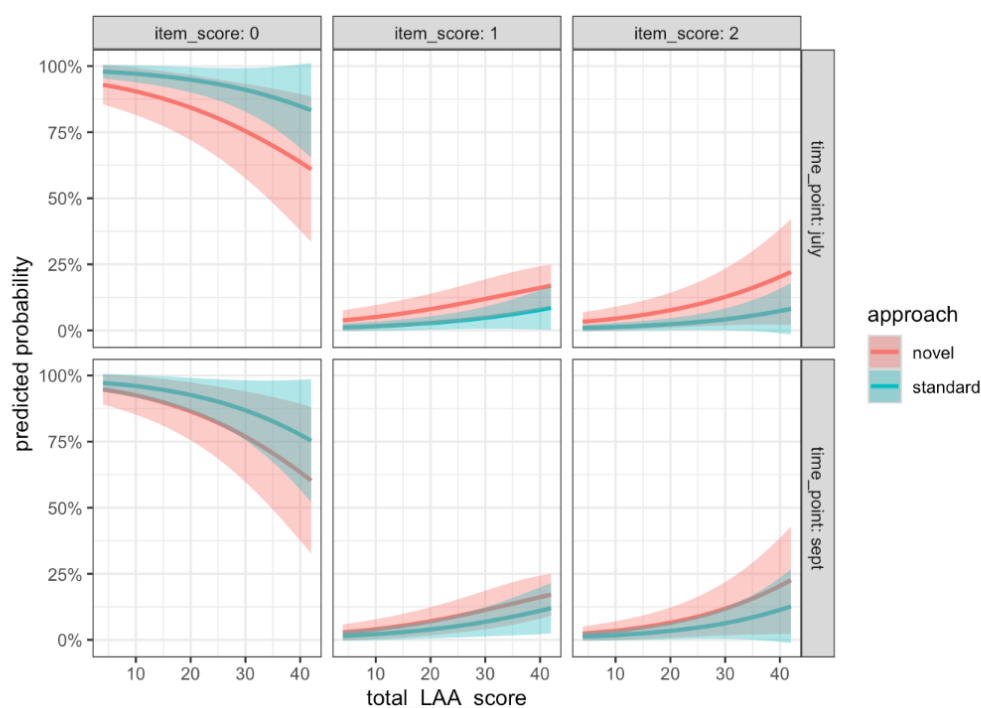
Pairwise comparisons did not show significant difference between time-points in either Novel (estimate = .17, SE = .19, $z = .89$, $p = .37$) or Standard (estimate = -.38, SE = .24, $z = -1.58$, $p = .11$). However, they did indicate a statistically significant difference between approaches at both time-points ([Appendix N.26](#)). At post-test, the probability of scoring anything at all was 16.36% in Novel, compared with 5.43% in Standard. While this probability decreased in Novel at delayed post-test, these participants were still more likely to score 1 or 2, rather than zero, than pupils in Standard (14.24% versus 7.77%).

5.12.3.3 The influence of language-analytic ability on spoken grammar recall scores

As with other tests, there was a significant interaction ($< .001$) between baseline LAA scores and outcomes ([Figure 22](#)): predicted values suggested that, in both approaches at both time-points, pupils with the lowest baseline LAA scores were more likely to score 0 in both approaches than those with higher baseline LAA; similarly, the higher the LAA score, the higher the spoken recall score. Pupils in Novel appeared descriptively to have an advantage over Standard, regardless of baseline LAA.

Figure 22.

Spoken Grammar Recall Predictions by Language-analytic Ability Score, Approach and Time-point



5.12.4 Spoken grammar recall: summary of results

Based on tests showing good reliability, there appeared to be a statistically significant difference between Novel and Standard outcomes, with pupils in Novel outperforming Standard at both

time-points, although the difference was greater at post-test than delayed post-test. Again, pupils with higher baseline LAA appeared to perform better than those with lower starting points.

5.13 Elicited imitation grammar test

This test (E.29) was completed by a subset of 58 participants (4.3.6.4, ii). [Appendix E.25](#) contains a list of items and their position at post-test and delayed post-test.

5.13.1 Test quality

An EFA was used to check reliability, as EIG is a less-researched construct ([Appendix M.34](#) contains CFA results); this showed good reliability. Facility scores indicated that pupils found the EIG challenging, but tests discriminated effectively ([Table 29](#)).

5.13.2 Descriptive statistics for the Elicited Imitation Grammar task

This consisted of 25 items, 11 targeting “avoir”/“être” verbs, seven present tense ‘-er’ verb endings and seven adjectival agreement. The test was originally scored using ordinal scoring (0-1-2), but as the items scored 1 represented just 3% of all observations, scores 1 and 2 were merged resulting in a binomial approach of 0-1. As outcomes were very similar, the ordinal model is presented here, with the binomial model provided in [Appendix A.4](#).

Table 32.

Descriptive Statistics for Elicited Imitation Grammar Tests

Approach	Time-point (<i>k</i> = 25)	Descriptive statistics							
		<i>n</i>	<i>Mean</i> %	<i>SD</i>	<i>CI</i> %	Range %		Effect size <i>d</i>	Effect size <i>CI</i>
Novel	Post-test	29	41	.19	33-48	8-84	Within-group: Novel Post-test – Delayed post-test	-.10	-.51 - .22
Standard	Post-test	29	33	.16	27-39	4-66	Within-group: Standard Post-test – Delayed post-test	-.51	-.97 - -.16
Novel	Delayed post-test	29	42	.22	34-50	0-88	Between-group: Novel – Standard Post-test	.45	-.07 - .97
Standard	Delayed post-test	29	39	.20	32-47	4-74	Between-group:	.13	-.43 - .64

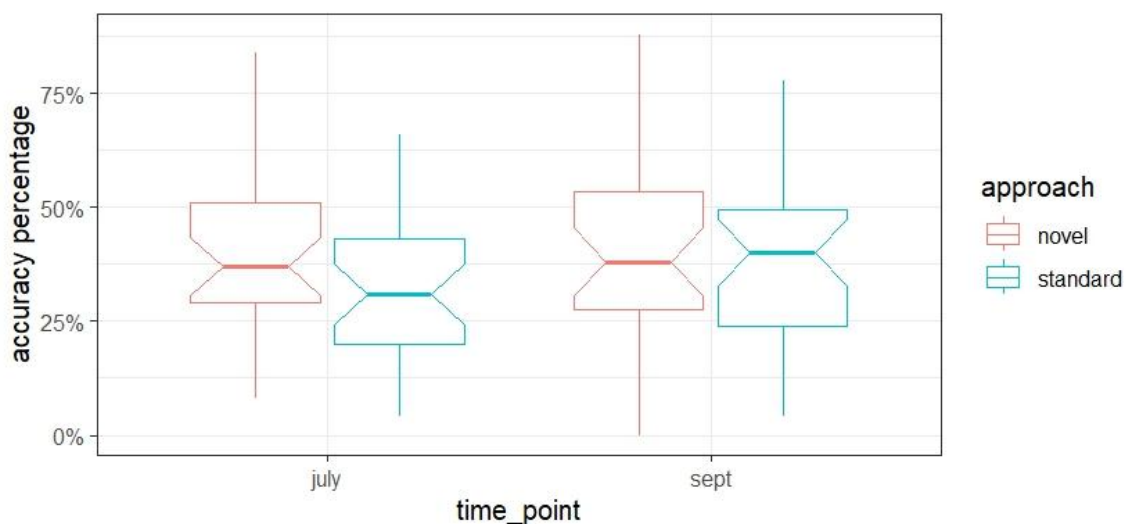
Novel –
Standard
Delayed
post-test

5.13.2.1 Between-groups results

Mean scores in Novel were higher than Standard at post-test and slightly higher at delayed post-test, with large and medium effect sizes respectively ([Table 32](#)). The CIs crossed zero, indicating that this may not be a significant difference, an observation supported by the notched box plot ([Figure 23](#)) where CI notches overlapped.

Figure 23.

Notched Box Plot Showing Elicited Imitation Grammar Scores by Time-point



5.13.2.2 Within-groups results

There was little difference between post-test and delayed post-test for Novel (medium effect size, with CIs crossing zero – see [Table 32](#)). In contrast, Standard scores increased ([Figure 23](#)), with a large effect size where CIs did not cross zero.

5.13.3 Analyses using mixed-effects models

A cumulative link mixed-effects ordinal model was created (see [Appendix N.27](#) for code).

5.13.3.1 Assumptions for the regression model

The Brant test (Brant, 1990), used to check assumptions after running the model, indicated that deviations from the model were larger than what could be explained by chance. The Omnibus measure was $< .05$, with p -values for time-point and LAA score also $< .05$ ([Appendix N.28](#)). Thus, results should be interpreted with caution.

5.13.3.2 Predictions based on mixed-effects models

Significant effects of time-point and LAA score ($< .001$) were revealed by the model ([Appendix N.29](#)). Fixed effects (marginal R^2) accounted for 5% of variance in the EIG model; a combination of fixed and random effects (conditional R^2) explained 34% of total variance.

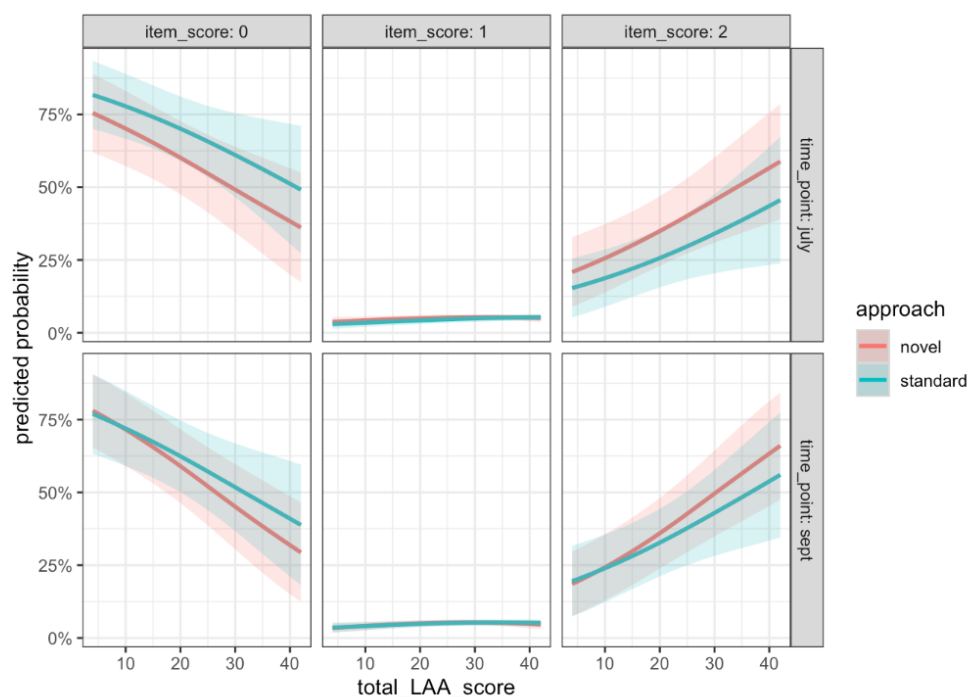
Pair-wise comparisons showed that pupils in Standard scored more highly in EIG at delayed post-test than at post-test (estimate = $-.35$, SE = $.12$, $z = -2.89$, $p = .004$), with the probability of a pupil in Standard scoring two points per item at delayed post-test increasing to 34% from 26% at post-test ([Appendix N.30](#)). In comparison, the probability of a pupil in Novel scoring two points per item remained stable at 36% at post-test and 37% at delayed post-test. There was no statistically significant difference between the two approaches.

5.13.3.3 The influence of language-analytic ability on Elicited Imitation Grammar scores

The significant interaction between baseline LAA score and EIG item score was explored ([Figure 24](#)). At both time-points, pupils with lower baseline LAA scores appeared more likely to score zero regardless of approach (a greater likelihood in Standard than Novel, but not statistically significant). Similarly, the higher pupils' baseline LAA score, the greater the likelihood of scoring 2. Again, pupils in Novel were descriptively more likely to score more highly than those in Standard at post-test; this appeared to apply at delayed post-test to those in Novel with raw baseline LAA scores $> ca. 12$. There seemed little interaction between pupils' LAA score and the probability of scoring 1 (possibly because of the low number of pupils scoring 1).

Figure 24.

Elicited Imitation Grammar Predictions by Language-analytic Ability Score, Approach and Time-Point



5.13.4 Elicited Imitation Grammar test: summary of results

Based on tests with good reliability indices, there was a large effect size between Novel and Standard outcomes at post-test, although not statistically significant. By delayed post-test, there was little difference between approaches. Baseline LAA scores appeared to impact on pupils' performances at both time-points, with lower starting points in LAA apparently limiting pupils' scores in EIG tests.

5.14 Pupils' learning outcomes in grammar: discussion

The Novel approach responded to Teaching School Council's (2016) recommendations to provide concise, explicit grammatical rule explanation, involving the noticing of grammatical features in the receptive mode, linked to meaning; it also addressed calls from Graham et al. (2014) to provide a pedagogy facilitating young learners' comprehension of basic grammatical concepts. This is not to say that Standard did not provide some explicit focus on grammatical patterns – there was, for example, explicit reference to adjectival agreement in singular and plural forms, and attention given to the impact of grammatical gender. Outcomes in the grammar tests are now discussed, including conclusions relating to the influence of LAA (5.14.4). Short summaries and illustrations of pupils' performance on individual grammatical categories in written recognition, oral production and EIG in this section are followed by more detailed

narrative discussion, included here rather than in the appendices as it is more approach-specific than discussion of individual items in vocabulary and phonics.

5.14.1 Written grammar recognition: quantitative trends

The low facility and discrimination indices indicated possible floor effects in written recognition, as pupils clearly found these particularly challenging. This illustrates the difficulty of testing knowledge at this level. Additionally, although two-choice answers were avoided, there is a risk in multiple-choice tasks that participants guess answers (Nation, 2007).

The similar outcomes across approaches suggest that these young learners - traditionally assumed to draw more on implicit-learning mechanisms (Nikolov & Mihaljević Djigunović, 2006), an assumption challenged by Kasprowicz and Marsden (2018), Lichtman (2016) and Roehr-Brackin and Tellier (2019) - were not disadvantaged by being taught explicitly in Novel. This may support Bouffard and Sarkar's findings (2008), in whose study 8-year-old learners of French demonstrated ability in grammatical analysis, possibly adding weight to Lichtman's instructional hypothesis (2016) that young learners learn in the way in which they are taught. It also supports findings from Kasprowicz and Marsden (2018) that explicit instruction can contribute to metalinguistic awareness and LAA.

5.14.2 Spoken grammar recall: quantitative trends

Although pupils in both approaches found these tests challenging, the tests appeared more reliable and to discriminate more effectively between high and low achievers than written recognition. This section discusses quantitative trends emerging from spoken recall tests.

5.14.2.1 Between-groups

Higher mean scores in Novel and the statistically significant difference between approaches at both time-points suggested by the mixed-effects model indicate that, in oral production, pupils benefited from the explicit grammatical focus. This may support findings from Graham et al. (2014), which tentatively suggested a benefit in production for pupils with lower L1 literacy as they moved into secondary school if they had followed an approach with a greater literacy focus in primary, although those results were not statistically significant.

5.14.2.2 Within-groups

Mean scores were more consistent across time-points in Standard than Novel, where they decreased at delayed post-test. This may support Erlam's (2003) findings, indicating that pupils following a deductive approach showed an advantage in speaking post-tests, but effect sizes related to these gains decreased at delayed post-test. In the inductive approach in Erlam's (2003)

study, learning gains decreased less by delayed post-test, which was also seen in this current research: indeed, Standard scores (less of a deductive approach than Novel) were consistently higher in all four spoken grammar recall categories at delayed post-test than at post-test, but still lower than Novel at both time-points. An alternative interpretation is that, having increased at post-test, Novel scores were more likely to decrease back to the mean at delayed post-test. These losses at delayed post-test relative to post-test suggest the importance of continuing the intervention to sustain the early gains.

5.14.3 Elicited Imitation Grammar: quantitative trends

5.14.3.1 Between-groups

A parallel can be drawn between outcomes in EIG and vocabulary recall in the oral modality (5.6), where Novel scores were also descriptively higher than Standard at both time-points, yet not statistically significantly: in both tests, the individual nature of the task limited the number of participants; a larger sample size may have altered findings (Larson-Hall, 2016). Nevertheless, the descriptively higher scores in Novel align with Graham et al.'s (2014, 2017) comparison of 10-12-year-olds' outcomes in L2 French after following a mainly oracy- or literacy-based approach in primary school. Scores were descriptively higher on a sentence repetition task targeting article-noun/adjective-noun agreement and simple present tense for pupils in the literacy-based approach at all three time-points (Year 5, Year 7, Year 8), but not statistically significant.

5.14.3.2 Within-groups

In line with spoken grammar recall, Standard EIG scores were higher in all four grammatical categories at delayed post-test than at post-test. This improved performance may have resulted from test or maturation effects, possibly explaining some of between-group difference at delayed post-test. The grammatical knowledge of pupils in Standard may have emerged more strongly only after time rather than immediately, whereas pupils in Novel showed larger immediate gains which were sustained. Novel scores, higher at post-test, had more room to decrease at delayed post-test.

5.14.4 Comparison of written recognition, spoken recall and Elicited Imitation Grammar tests, with discussion of the influence of language-analytic ability

All three tests were challenging for these young learners. There was little difference between outcomes in each approach in written recognition, but scores were higher in Novel than Standard in spoken recall and EIG, though not always statistically significant. Differences were less apparent at delayed post-test, suggesting that a longer period of instruction would have been

beneficial to maintain the advantage seen in Novel. In written recognition, less of the total variance in the model appeared to be explained by a combination of fixed effects (approach, time-point, LAA score) and random effects (school, individual pupil, test item) than in all other tests, suggesting that there may be other sources of variation to consider.

Regardless of approach, there appeared clear interaction between pupils' baseline LAA and their performance in all tests. This aligns with Kasprowicz et al.'s (2019) study, which highlighted the importance of LAA (rather than differences in spacing of practice) in predicting learning outcomes for young L1 English learners of French in sentence-picture matching and acceptability judgement. It supports findings from Roehr-Brackin and Tellier (2019) that LAA predicted learning outcomes in 8-9-year-old L1 English learners of French.

Novel scores in all tests appeared less influenced by baseline LAA scores than Standard, suggesting that its deductive approach ultimately benefited all learners when it came to written grammar recognition, not just those with higher baseline LAA. This benefit of deductive instruction was identified by Erlam (2005) and found by Roehr-Brackin et al. (2024), with both studies concluding that such an approach may compensate for individual aptitude differences when it came to grammatical production.

5.14.5 Individual grammar categories: results

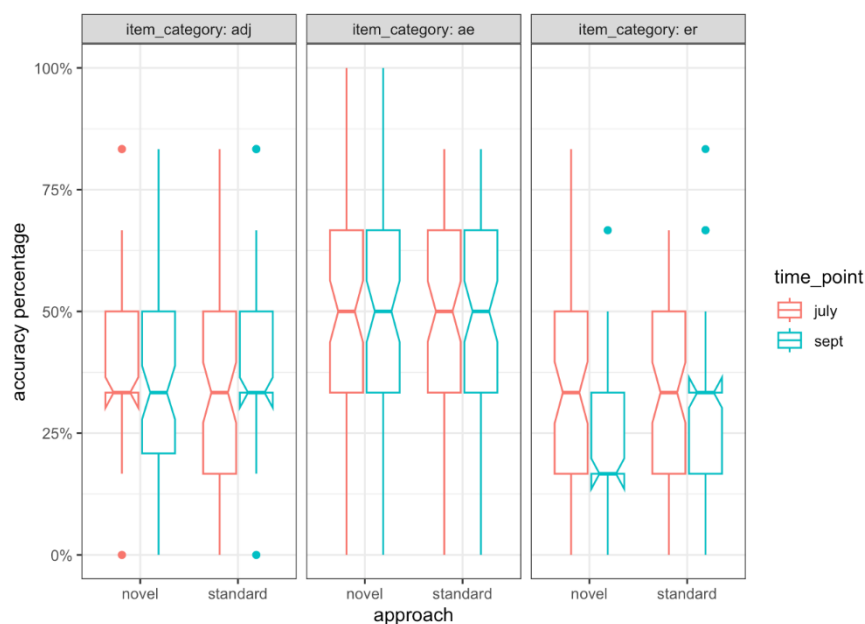
This study did not compare impact of approaches on pupils' performance in individual grammatical categories. However, an overview of pupils' performance in these areas in the three grammar tests is now presented and discussed, as there were some relevant approach-specific findings which also contextualise discussion in [5.17.4](#); these are linked to existing literature in this field.

5.14.5.1 Written grammar recognition

Each of the three categories ("*avoir*"/"*être*" verbs, '-er' verb endings, adjectival agreement) appeared similarly challenging for participants in both approaches, with all mean scores < 3 out of 6 points available. The highest mean scores for both approaches were on "*avoir*"/"*être*" verbs (Novel: $M = 2.80$, $SD = 1.35$; Standard: $M = 2.75$, $SD = 1.34$), followed by adjectival agreement (Novel: $M = 2.35$, $SD = 1.24$; Standard: $M = 2.42$, $SD = 1.23$) ([Figure 25](#)). The most challenging category for participants in both approaches was '-er' verb endings (Novel: $M = 1.85$, $SD = 1.16$; Standard: $M = 1.96$, $SD = 1.15$).

Figure 25.

Notched Box Plot Showing Written Grammar Recognition Scores by Category by Time-point

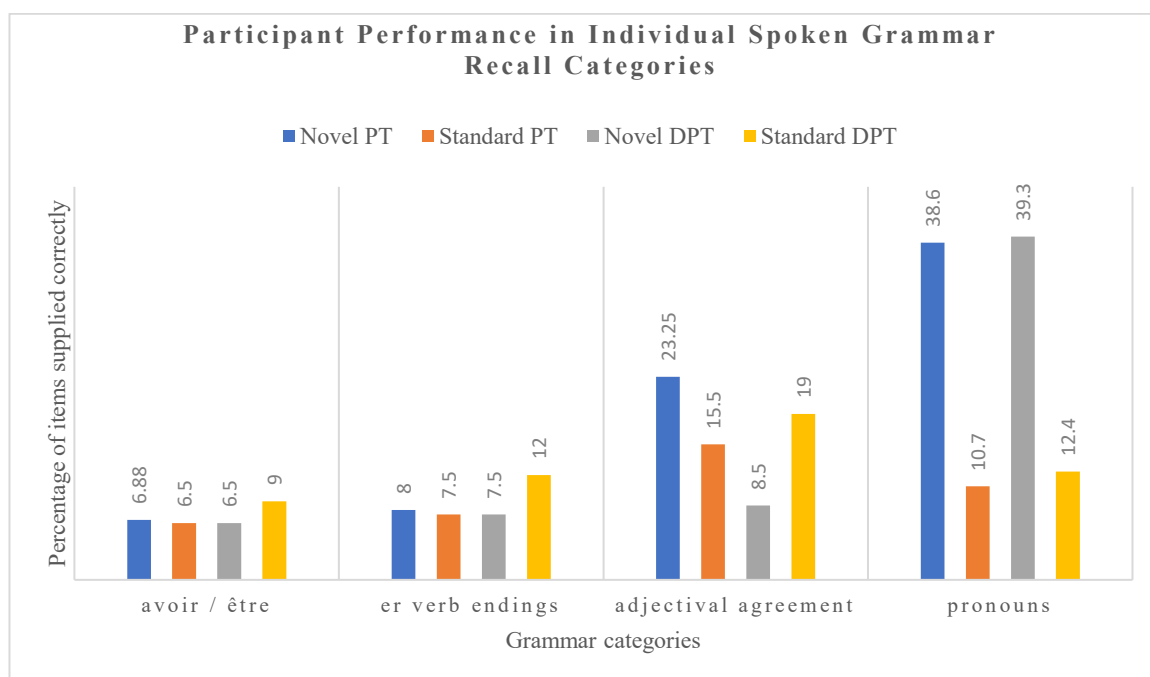


5.14.5.2 Spoken grammar recall

Pupils' scores on “avoir”/“être” were the lowest of all four categories ([Figure 26](#)), generally similar across approaches and time-point. Outcomes were similar in the ‘-er’ verb endings category. Participants generally performed better in adjectival agreement (with just two items) than in verbs. The highest scores were in pronoun knowledge, but only in Novel.

Figure 26.

Participants' Scores in Individual Spoken Grammar Recall Categories



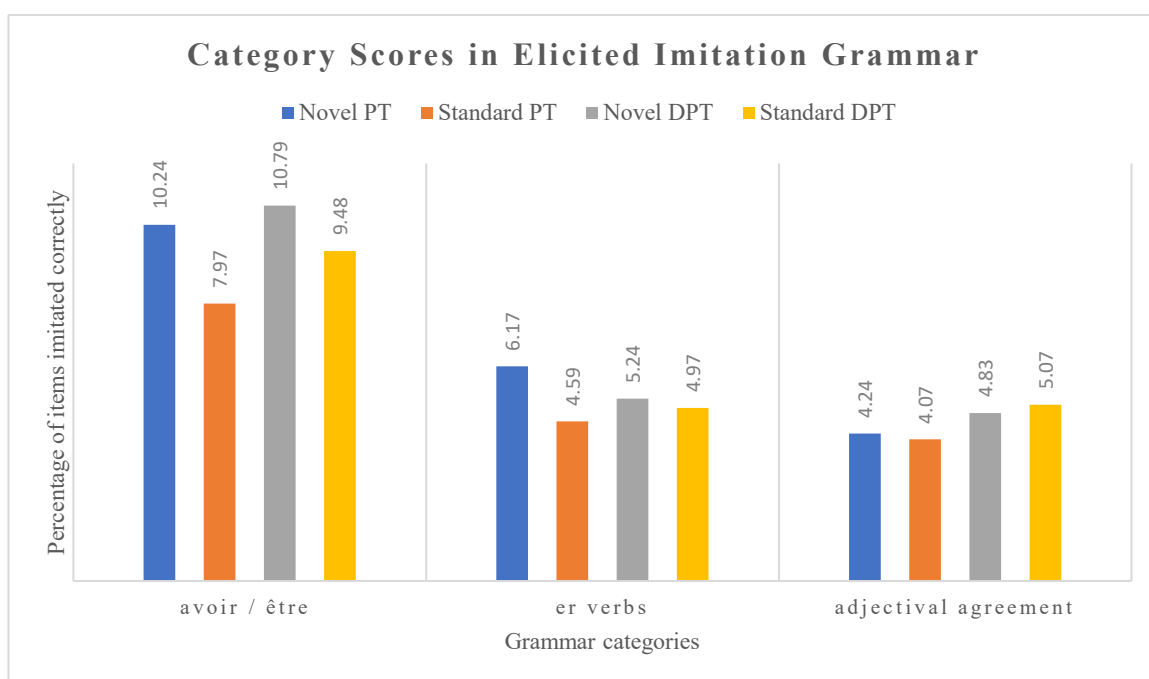
Note. In [Figure 26](#), PT stands for post-test, and DPT for delayed post-test.

5.14.5.3 Elicited Imitation Grammar

In line with pupils' performance in written recognition, pupils' performance at both time-points appeared better in the category of “avoir”/“être” verbs ([Figure 27](#)). Unlike spoken recall, pupils generally performed slightly better on ‘-er’ verb endings than on adjectival agreement (except for pupils in Standard at delayed post-test), where scores were lowest of all categories.

Figure 27.

Participant Scores in Individual Categories in Elicited Imitation Grammar



Note. In [Figure 27](#), PT stands for post-test, and DPT for delayed post-test.

5.14.6 Discussion of individual categories

In EIG, pupils scored highest in both approaches on “*avoir*”/“*être*” verbs ([Figure 25](#)), in line with written recognition outcomes, but not spoken recall, where this category was lowest-scoring. Although pupils had to verbally reproduce sentences they heard, the EIG task was perhaps less challenging than spoken recall, in which pupils relied entirely on personal knowledge, rather than drawing on the support of the utterance they heard (a possible syntactic priming effect – see McDonough & Chaikitmongkol, 2010). This suggests that young learners find it easier to recognise and re-produce, rather than independently produce, these verbs. Indeed, of all four grammar categories in spoken recall, pupils in both approaches appeared to find most demanding the requirement to supply verbs of any kind, whether “*avoir*”/“*être*” or ‘-er’ verb endings ([Figure 26](#)). This is consistent with findings from Courtney (2017), where L1 English pupils in the final year of primary school struggled to produce French utterances containing verbs, although did improve once they transferred to secondary school, where grammar explanations were more explicit. In the present study, pupils in Novel experienced this explicit focus from the outset yet still appeared to find this aspect of language production challenging in spoken recall, in both verb categories. These outcomes are in line with Cable et al. (2010) and Graham et al. (2014), in which pupils struggled particularly with the present tense of verbs in both sentence repetition and photo description tasks. This suggests that pupils in the present study needed longer than the 30 lessons experienced in Year 3 for their grammatical production of verbs to develop (regardless of whether they are the highly irregular, abstract and phonologically salient “*avoir*”/“*être*”, or regular -er verbs). Studies have also found verb

morphology late to develop (Marsden & David, 2008; Myles, 2005; Rule & Marsden, 2006), so *any* verb production is an achievement; indeed, Rule and Marsden's (2006) study suggested that, if findings from their study with 13-14-year-old L1 English learners of French (experiencing 195 hours' exposure and producing 20% verbless utterances) were applied to learners experiencing under 40 hours' instruction, this would likely rise to 81% verbless utterances. This is discussed further in [5.15](#) as these EIG findings contrast results in open-ended speaking tasks.

The lowest scores in EIG were in adjectival agreement, also found by Graham et al. (2014) in both sentence repetition and photo description tasks. In this EIG test, pupils: struggled to distinguish between masculine/feminine forms of adjectives; sometimes provided the English translation (e.g., 'brown' instead of "*brun*"); or confused adjectives ("*blanc*" for "*brun*"). Occasionally mispronunciation led to ambiguity ("*jeune*" instead of "*jaune*"). Adjectives tested in this EIG predominantly included adjectives of colour, as pupils in Standard had had few encounters with other descriptive language. Graham et al. (2014) cautioned that, although commonly taught in primary schools, French adjectives of colour can pose difficulties for learners as they are not always phonologically distinct, and learners can be confused by the fact that some are invariant. In contrast, Novel taught adjectives that were regular, to introduce regular patterns. As arose with the selection of vocabulary ([5.7.1.1](#)) for tests, this attempt to ensure equitability (i.e., choosing adjectives based on them being in both approaches) may have minimised some of the potential key benefits of Novel (i.e., its careful selection of the lexicon to support the grammar being introduced). The low scores generally in this category suggest that, even when reference to grammatical features such as gender (as required by the National Curriculum, DfE, 2013a, 2013b) is explicit, as in Novel, pupils take time to begin to notice language patterns such as gender (noted by OFSTED, 2008a).

Novel scores on pronoun knowledge were descriptively over three times as high as Standard at both time-points in spoken recall, suggesting that the greater number of opportunities to create and manipulate sentences characterising Novel had led to more secure understanding. Although the test format (with pupils provided with the English sentence and the French translation, with the target word missing - [Appendix E.24](#), section gr:2), could be said to constitute a vocabulary test rather than a requirement to demonstrate grammatical understanding, knowledge of pronouns - whether as lexical or morphosyntactic knowledge - is nevertheless critical for sentence creation and meaning making.

It is worth noting that, based on impressionistic observation of pupils' responses to all tests, EIG was the most popular. Pupils generally responded positively to the task, and most appeared excited about re-doing it at delayed post-test.

5.14.7 Overarching discussion of grammar outcome measures

In summary, pupils appeared to benefit from the Novel approach in spoken recall and EIG ([Table 33](#)), where tests showed good levels of reliability. Scores were the same across approaches or slightly higher in Standard than Novel in written recognition, but with low reliability. As with phonics and vocabulary tests, baseline LAA scores appeared to impact on all pupils, but less in Novel than Standard. A longer study may have produced more durable impact of the explicit focus on sequenced grammar on Novel scores.

Table 33.

Overview of Grammar Outcomes

Grammar measure	Over time		Between approaches	
	Novel	Standard	Post-test	Delayed post-test
Written recognition	Post-test > Delayed post-test (trend)	PT = Delayed post-test ($M = 39:39$)	$N = S$ ($M = 37:39$)	$N < S$ (trend)
Spoken recall	Post-test > Delayed post-test (trend)	PT < Delayed post-test (trend)	$N > S$ (statistically significant)	$N > S$ (statistically significant)
Elicited imitation	Post-test < Delayed post-test (trend)	PT < Delayed post-test (trend)	$N > S$ (trend)	$N > S$ (trend)

Note. Trend reflected in descriptives and (bordering on) medium effect size (.05 upwards).

5.15 Open-ended speaking Picture Response: results

This section presents the results of the first of two speaking tests completed by a subset of 29 pupils from each approach ([4.3.6.4, ii](#)): the open-ended PR task. It is followed by results from the EIS test.

Pictures were unfamiliar to pupils and are included in [Appendix E.31](#) with details of their positions at post-test and delayed post-test.

5.15.1 Test quality

The PR task consisted of four pictures, and did not have a maximum score, therefore reliability, facility and discrimination scores were not calculated. Normality of distribution checks ([Appendix M.28](#)) indicated that data were not normally distributed.

5.15.2 Descriptive statistics for the Picture Response task

As any appropriate verbal responses to the pictures were credited, there was no possible maximum score, although no pupil scored higher than 10.50 at either time-point (4.3.9.4). Although the range was quite wide on occasions (e.g., Picture 1, post-test, Table 34), the low means, standard deviation and CI scores show that pupils were generally scoring nearer 0. Scores were highest for Pictures 1 and 4 (Figure 28).

Table 34.

Descriptive Statistics for the Picture Response Task

Approach	Time-point ($k = 4$)	Descriptive statistics								
		n	Picture	Mean %	SD	CI %	Range %		Effect size d	Effect size CI
Novel	Post-test	29	1	2.23	2.34	1.34 - 3.13	0-9	Within-group: Novel Post-test – Delayed post-test	-.31	-.76 - .05
Standard	Post-test	29	1	2.18	1.79	1.52 - 2.85	0-6	Within-group: Standard Post-test – Delayed post-test	-.15	-.50 - .23
Novel	Post-test	29	2	.65	1.36	.14 - 1.16	0-6	Between-group: Novel – Standard Post-test	.19	-.38 - .70
Standard	Post-test	29	2	.25	1.01	-.13 - .63	0-5.50	Between-group: Novel – Standard Delayed post-test	.34	-.13 - .88
Novel	Post-test	29	3	.62	1.00	.24 - .99	0-4			
Standard	Post-test	29	3	.68	1.05	.29 - 1.08	0-4.50			
Novel	Post-test	29	4	2.15	2.32	1.28 - 3.02	0-10.50			
Standard	Post-test	29	4	1.58	2.06	.81 - 2.35	0-8			
Novel	Delayed post-test	29	1	2.32	1.97	1.58 - 3.05	0-7.50			
Standard	Delayed post-test	29	1	2.03	2.13	1.24 - 2.83	0-7.50			
Novel	Delayed post-test	29	2	.85	1.37	.34 - 1.36	0-4			
Standard	Delayed post-test	29	2	.12	.39	-.03 - .26	0-1.50			
Novel	Delayed post-test	29	3	1.50	2.19	.68 - 2.32	0-8.50			

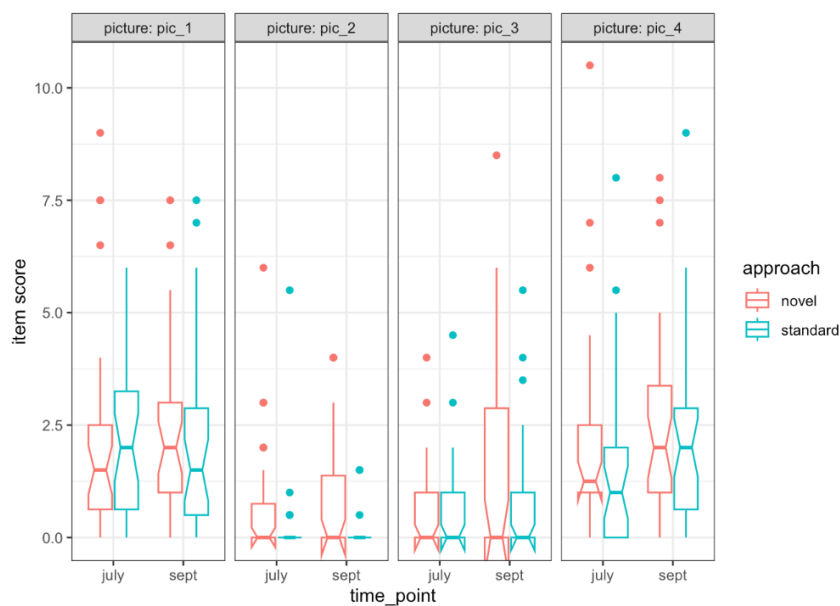
Standard	Delayed post-test	29	3	.95	1.40	.43 - 1.47	0-5.50
Novel	Delayed post-test	29	4	2.47	2.14	1.67 - 3.37	0-8
Standard	Delayed post-test	29	4	2.08	2.12	1.29 - 2.87	0-9

5.15.2.1 Between-groups results

Mean scores in Novel were higher than Standard in all picture responses at both time-points except for Picture 3 at post-test (Table 34). Similar medians (Figure 28) indicated that these higher mean scores may be due to a greater number of and higher outliers in Novel than Standard (e.g., Pictures 1 and 4). Effect sizes based on total scores between Novel and Standard approaches were medium (post-test) and large (delayed post-test) in favour of Novel, but crossed zero, suggesting no significant differences.

Figure 28.

Notched Box plot Showing Open-ended Picture Response Scores by Time-point



5.15.2.2 Within-groups results

Novel scores increased at delayed post-test (with a large effect size), indicating durable learning, combined with potential test and/or maturation effects, although CIs just passed through zero; Standard scores increased by delayed post-test for Pictures 3 and 4, but decreased for 1 and 2, with an overall medium effect size (again crossing zero).

5.15.3 Inferential statistics for the Picture Response task

The regression model was not run as assumptions were not met. As data were not normally distributed, a non-parametric Mann Whitney U test was conducted in R version 4.3.2 (R Core Team, 2023) to compare between-group scores. As data were re-used in within-groups comparisons, the level of permissible uncertainty for statistical significance was changed from .05 to .025.

5.15.3.1 Between-groups results

There was no statistically significant difference between approaches at either time-point. A further comparison was performed by subtracting post-test scores from delayed post-test scores to compare changes over time, but the Mann Whitney U test produced non-significant results ($W = 490.50, p = .28$), suggesting no significant differences between groups in terms of any change between the two test-points.

5.15.3.2 Within-groups results

A Wilcoxon signed rank test with continuity correction was performed in R version 4.3.2 (R Core Team, 2023) to compare within-groups data at the two time-points. Neither Novel scores ($V = 106.50, p = .08$) nor Standard scores ($V = 182, p = .64$) showed a statistically significant difference between post-test and delayed post-test.

5.15.4 Picture Response: summary of results

Scores ranged widely in both approaches. Pupils in Novel performed better at delayed post-test than at post-test, whereas this did not apply consistently to Standard. There was no statistically significant difference between performances in each approach at either time-point.

5.16 Elicited imitation speaking: results

For EIS, the EIG task ([E.29](#)) was rescored to provide a measure of pupils' general oral proficiency.

5.16.1 Test quality

[Table 35](#) outlines test quality. Reliability, checked through an EFA (used as EI is a less-researched construct) was very good ([Appendix M.34](#) shows CFA results). Low facility scores indicated that pupils found the EIS test challenging when scored for overall speaking proficiency.

Table 35.*Test Quality Information for Elicited Imitation Speaking*

		Elicited	Imitation
		Post-test	Delayed post-test
% cases complete prior to imputation		97	97
Reliability (EFA)	alpha	.87	.86
	omega	.89	.89
Facility		.22	.25
		<i>SD</i> .12	<i>SD</i> .12
Discrimination		.35	.36
		<i>SD</i> .14	<i>SD</i> .16
Normality of distribution link		Appendix M.31	

5.16.2 Descriptive statistics for the EIS task

Pupils were asked to listen to, then repeat 15 sentences; a maximum total of 4 points was awarded to each sentence ([4.3.9.4](#)), thus providing a possible total of 60. [Appendix E.25](#) provides a list of utterances and their position at post-test and delayed post-test.

5.16.2.1 Between-groups results

Mean scores ([Table 36](#)) in Novel were higher than Standard at both time-points, although the between-group difference was smaller at delayed post-test than post-test (see also median scores, [Figure 29](#)). Effect sizes were large at post-test and medium at delayed post-test, but both CIs crossed zero, suggesting that there may not be any between-group differences.

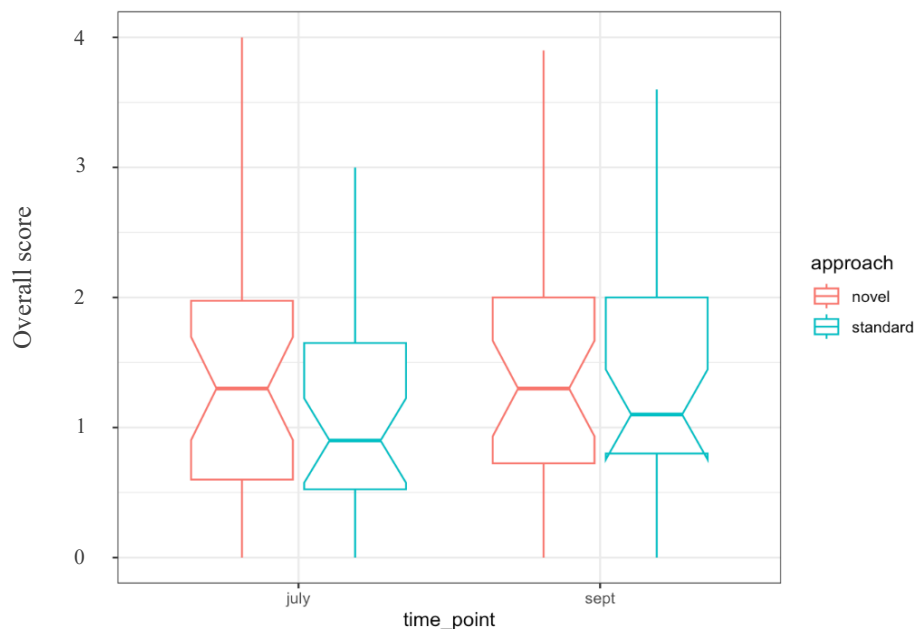
Table 36.*Descriptive Statistics for Elicited Imitation Speaking Tests*

Approach	Time-point (<i>k</i> = 15)	Descriptive statistics							
		<i>n</i>	<i>Mean</i> (total/60)	<i>SD</i>	CI	Range		Effect size <i>d</i>	Effect size CI
Novel	Post-test	29	14.60	10.79	10.50-18.60	0-40	Within-group: Novel Post-test – Delayed post-test	-.15	-.62 - .19
Standard	Post-test	29	11.10	8.29	8.00-14.20	0-30	Within-group: Standard	-.69	-1.15 - -.36

Novel	Delayed post-test	29	15.30	10.29	11.50-19.10	0-39	Post-test – Delayed post-test Between-group: Novel – Standard	.36	-.15 - .88
Standard	Delayed post-test	29	14.40	9.75	10.80-18.04	0-36	Post-test Between-group: Novel – Standard Delayed post-test	.09	-.45 - .64

Figure 29.

Notched Box Plot Showing Elicited Imitation Speaking Scores by Time-point



5.16.2.2 Within-groups results

The slightly higher scores for both approaches at delayed post-test versus post-test ([Table 36](#)) may indicate durable learning but may also reflect a test effect. The effect size was medium for Novel and large for Standard between post-test and delayed post-test (with delayed post-test higher). This passed through zero only in Novel, suggesting a more meaningful difference in Standard.

5.16.3 Analyses using mixed-effects models

A cumulative link mixed-effects model was constructed (see [Appendix N.31](#) for code).

5.16.3.1 Assumptions for the regression model

After running the model, assumptions were checked using the Brant test (Brant, 1990). Although the Omnibus measure was $> .05$, p -values for approach, time-point and LAA score were $< .05$ ([Appendix N.32](#)); for this reason, results should be interpreted with caution.

5.16.3.2 Predictions based on mixed-effects models

The model summary ([Appendix N.33](#)) showed a significant relationship between time-points, as well as between LAA and EIS scores ($< .001$). Fixed effects (marginal R^2) accounted for 11% of variance; a combination of fixed and random effects (conditional R^2) accounted for 53% of total variance.

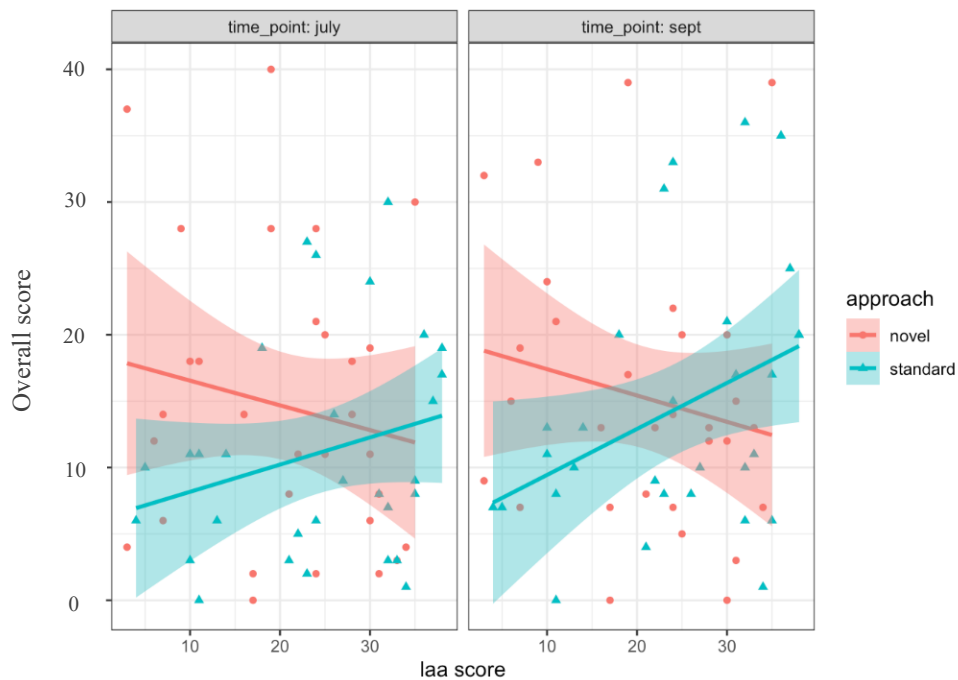
Pairwise comparisons were performed to explore these interactions further. The relationship between time-point and EIS scores revealed a statistically significant difference ($p = .002$) between time-points in Standard ([Appendix N.34](#)), but no difference in Novel. There was no statistically significant difference between approaches at either time-point.

5.16.3.4 The influence of language-analytic ability on Elicited Imitation Speaking scores

There only appeared to be an interaction between baseline LAA score and pupils' performance in Standard: the higher the LAA score, the higher pupils' post-test scores in EIS ([Figure 30](#)). In Novel, the opposite seemed to be the case, with pupils with lower LAA scores ($< 28-33$) descriptively likely to score higher than those in Standard. At delayed post-test, this pattern was repeated, but this time the point at which pupils with higher LAA scores in Standard outperformed those in Novel dropped to ca. 23-28.

Figure 30.

Scatter Plot Showing Elicited Imitation Speaking Scores According to Language-analytic Ability Score by Time-point

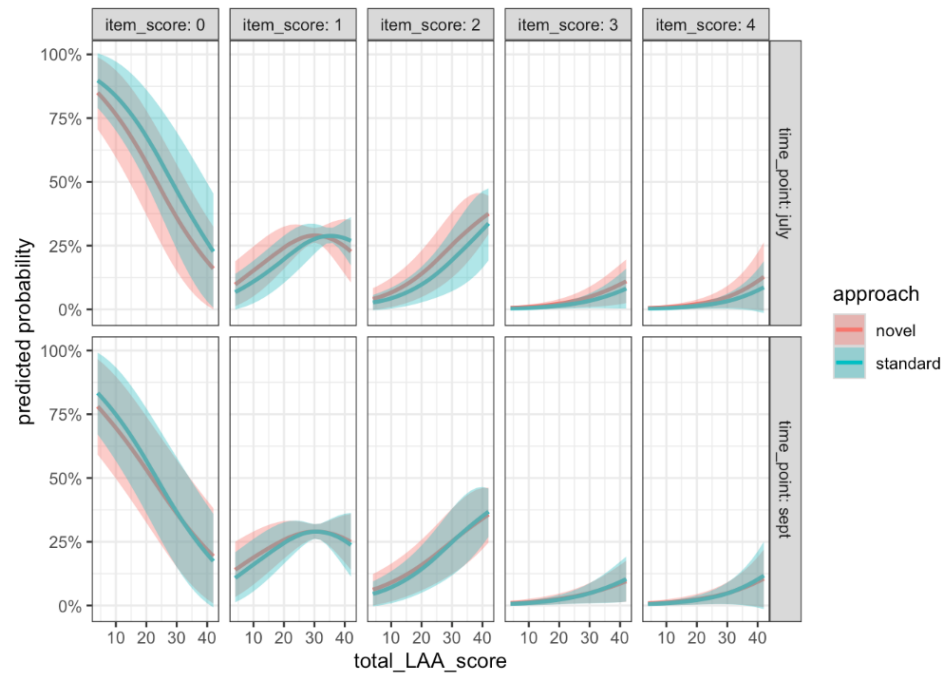


Predicted values ([Figure 31](#)) indicated the significant interaction between LAA and item score. Although there was no significant interaction between LAA score, approach, time-point and EIS (which could suggest potential ‘leveller’ effects), an exploration of the descriptive trends is considered useful, given the small sample size and the indications from visuals. Pupils in Novel were descriptively less likely at post-test to score 0 and more likely to score 1 and 2 than those in Standard, regardless of LAA score, except for those with the very highest LAA scores (> ca. 40), at which point Standard scores were higher. When it came to scoring 3 or 4 points, there was little difference between approaches at the lower end of LAA scores, but pupils with LAA scores of > ca. 23-27 appeared to have an advantage in Novel.

At delayed post-test, pupils in Novel were again descriptively less likely to score 0 than those in Standard, regardless of LAA score, apart from those with the very highest LAA scores (> 40). When it came to scoring 1 and 2, there was very little difference between approaches except for pupils with lowest LAA scores, when there appeared to be an advantage in Novel. There was little difference in the likelihood of pupils scoring 3 and 4 at delayed post-test, although there appeared to be a very slight advantage for pupils with the highest LAA scores (> ca. 40) in Standard.

Figure 31.

Elicited Imitation Speaking Predictions by Language-analytic Ability Score, Approach and Time-point



5.16.4 Elicited Imitation Speaking: summary of results

Based on tests which showed very good reliability, pupils in Novel scored higher than in Standard at both time-points, although there was no statistically significant difference. Standard scores increased at delayed post-test, with a large effect size. The influence of LAA only appeared evident in Standard.

5.17 Pupils' learning outcomes in speaking tasks: discussion

This section discusses pupil outcomes in both speaking tests, including conclusions relating to the influence of LAA ([5.17.3](#)). It includes short summaries and illustrations of pupils' performance on individual items in PR and EIS tasks, followed by narrative discussion.

5.17.1 Open-ended speaking Picture Response: quantitative trends

This measure of spontaneous language production offered a means of capturing general proficiency and some aspects of implicit knowledge (Norris & Ortega, 2001). In the study conducted by Graham et al. (2014, 2017), Year 5 learners (the closest age group to pupils in the present study) who had followed a more oracy-based teaching style ($M = 15.75$, $SD = 9.01$) outperformed in a photo description task those that had experienced a more literacy-based approach ($M = 13.85$, $SD = 8.14$), though the difference was small and not statistically

significant. The between-groups difference in the present study was also not statistically significant at either time-point, but mean scores in Novel (more literacy-based) were higher than in Standard in all picture responses at both time-points (except for Picture 3 at post-test), suggesting that pupils were not *disadvantaged*, and were potentially *advantaged*, in terms of this kind of knowledge by being taught in a more explicit manner.

5.17.2 Elicited Imitation Speaking: quantitative trends

Although Novel outcomes were higher than Standard at both time-points, high standard deviations and wide ranges in both approaches ([Table 36](#)) indicated considerable individual variation. This is in line with, for example, findings from Myles et al.'s (1998) study considering near-beginners' verbal production in structured information-gap tasks.

5.17.3 Comparison of Picture Response and Elicited Imitation Speaking, with discussion of the influence of language-analytic ability

Mean scores were relatively low in both PR and EIS tests, but with high variation, suggesting that individuals differed in their ability to perform on these tasks. Novel scores were higher than Standard in both measures, at both time-points, though not statistically significant. In EIS, Standard scores increased significantly from post-test to delayed post-test (though were still lower than Novel). This aligns with findings from Erlam's (2003) study, as discussed in [2.4.1.2](#).

Once again, baseline LAA scores seemed to influence outcomes more in Standard than Novel in the EIS measure of oral proficiency, appearing to support the advantages found in deductive instruction by both Erlam (2005) and Roehr-Brackin et al. (2024). The Novel approach seemed to provide an equalising effect for learners with lower baseline LAA, compensating for some individual differences in aptitude, as in this study their lower LAA seemed not to 'penalise' them in EIS. This aligns with findings from EIG ([5.14.4](#)).

5.17.4 Individual pictures/items: results and discussion

This section addresses performance on individual PR pictures and then individual items in the EIS task, presenting results and following this with approach-specific discussion ([5.17.4.2](#)). See [Appendix E.31](#) for pictures.

5.17.4.1 Picture Response: results

Pictures were chosen to prompt creative communication without being tied to any particular focus covered in taught lessons, avoiding a highly structured exchange. Regardless of approach, pupils had encountered appropriate lexical items to enable them to produce utterances on all pictures, given that any relevant response was accepted.

(i) *Picture 1*

Responses generally consisted of single words (nouns or adjectives). Colour adjectives were most commonly supplied (covered more in Standard than Novel, whereas pupils in Novel encountered adjectives of size more often than Standard). Pupils in Novel were beginning to form sentences, particularly at post-test, whereas pupils in Standard made greater use of single word answers, or linked these using “*et*” (‘and’). The sentences produced by pupils in Novel used the third person singular form of the verb “*être*” (errors in pronunciation are indicated by capital letters):

“*Le chat eSt petit, le chien eSt grande* [sic].” (Novel, HLAA, post-test)

“*Le chien is brune* [sic], *le chat is noir*.” (Standard, HLAA, post-test)

One pupil in Novel attempted to use the verb “*écouter*” (‘to listen’) to describe the dog’s action:

“*Le CHat eST brun*. The cat is brown, and the dog ... *le chien eST* ... the dog is ... *écouter*? Listening, it looks like.” (Novel, HLAA, delayed post-test)

(ii) *Picture 2*

The lowest scores across all pictures and time-points were on Picture 2. Most pupils struggled to produce the pronouns “*il*” (‘he’) and “*elle*” (‘she’): across both time-points, “*il*” was produced correctly only twice; “*elle*” was comprehensible on 12 occasions (produced almost exclusively in Novel), although this included variants such as “*elli*” and “*ellé*”. Picture 2 was the picture most likely to require verbs in responses, even if pupils only offered single words. Producing verbs verbally was challenging for most pupils, although some in Novel were able to produce the infinitive form and there were isolated instances of correct conjugation, rather than simple repetition of a pre-learnt chunk of language. There was evidence in Novel that pupils were beginning to use language to create their own sentences:

“That one’s *chanter* [pointing to image 1]. I can’t remember what ‘listening’ is. Oh, *écoute*. *Elle écoute* to *chanter*”. (Novel, HLAA, post-test)

(iii) *Picture 3*

Most pupils described the boy, the cake or the bottle, and “*bleu*” (‘blue’) was particularly common among Standard responses, whereas pupils in Novel drew on a wider range of adjectives: “*amusant*” (‘funny’), “*jeune*” (‘young’), “*petite*” (‘small’) [sic] (each used once at post-test). Mean scores were < 1 for pupils in both approaches, but there was evidence at both time-points that pupils were trying to form sentences:

“The boy is eating. *Le, le bébé ... mange*.” (Novel, LLAA, post-test)

“*Le ... I don’t know what ‘boy’ is ... I don’t know what ‘it’ ... dANs mange une madeleine*”.
(Standard, HLAA, post-test)

There were more examples of (partially) correct sentences at delayed post-test than post-test in both approaches:

“*Boutaille [sic] eSt bleue*” (Novel, HLAA);

“*Madeleine est bleue*” (Standard, Medium language-analytic ability, MLAA);

“*Il est bleu*” (Novel, HLAA);

“*Le bébé est mange bleu*” (Standard, HLAA, delayed post-test);

“*Le gâteau eST blue et il eST ... eating ... mange*” (Novel, HLAA).

(iv) Picture 4

Responses to Picture 4 were structurally similar to picture 1, typically consisting of nouns and adjectives, sometimes linked using “*est*” (‘is’). Adjectives were most often colours, particularly in Standard, but included adjectives of size in Novel. There were examples at post-test of pupils beginning to manipulate language for themselves to build sentences with verbs, particularly in Novel:

“That one uses ink [points to pen] ... *le stylo utiliser ink*.” (Novel, LLAA, post-test)

“*GommÉ* - cause that's a rubber. And that is ... I know it, er, begins with a 's'. I know ... *stylo!*
Le stylo - the pen - *le stylo est, eSt brun. La gommÉ, c'est rose et bleu.*” (Novel, HLAA, post-test)

There was greater evidence of pupils in Standard putting words together within a sentence by delayed post-test than post-test:

“*Une stylo [sic], une gomme. Une stylo est brune. Une, une gomme est bleue.*” (Standard, LLAA, delayed post-test)

Some pupils in Novel appeared to have maintained their ability at delayed post-test to produce whole sentences verbally, with similar accuracy:

“*Le stylo eST brun and la gommÉ eST bleue.*” (Novel, HLAA, delayed post-test)

(v) Verb production

The number of utterances including a verb in some form was higher in Novel than Standard, and these were produced by a higher number of pupils in Novel than Standard ([Table 37](#)).

Table 37.

Total Number of Utterances Containing Recognisable Verbs or Verbs in the Infinitive in Picture Response

Approach	Time-point	Descriptive statistics

		Number of utterances with recognisable verbs	Number of different pupils producing these utterances
Novel	Post-test	28	9
Standard	Post-test	5	3
Novel	Delayed post-test	37	12
Standard	Delayed post-test	9	4

5.17.4.2 Picture Response: discussion

The higher scores ([Figure 28](#)) on Pictures 1 and 4, where responses generally included nouns and adjectives with the verb “*être*” (‘to be’) to describe animals and stationary items, possibly suggest that nouns are easier to learn than verbs (Milton, 2007; E. Peters, 2020; Tschichold, 2012). Consistent with this, the fact that Picture 2 (showing people listening to/playing music) would have been most often associated with action verbs may explain the low scores.

As the picture prompts were unfamiliar to pupils, there was little opportunity for them to resort to the memorised language and formulaic utterances which characterised primary FL provision observed by Cable et al. (2010). Given the more explicit approach in Novel towards using and manipulating verbs, it is perhaps unsurprising that pupils’ scores were slightly higher in this approach. Indeed, almost one third of pupils in Novel (nine of 29) produced 28 utterances containing recognisable verbs within a syntactically correct sentence or the infinitive of an appropriate verb at post-test ([Table 37](#)), increasing to 12 pupils producing 37 utterances containing verbs at delayed post-test. This may support the suggestion that explicit grammatical explanation such as that provided in Novel can enable even the youngest FL learners in England to “describe people, places, things and actions orally ...” (DfE, 2013a:2) as required by the National Curriculum.

This evidence that pupils in Novel were beginning to use verbs to create their own sentences is interesting, given that, in a study conducted by Marsden and David (2008), a reliable knowledge of verbs was a stronger feature of older learners’ (aged 17-18) performance, whereas younger pupils aged 13-14 tended to use more noun-dominated language. It would also appear favourable in comparison with the findings of Rule and Marsden’s (2006) study, which found that verbless utterances were common among 13-14-year-old L1 English learners of French, as did Mitchell and Myles (2019), with inflected verb forms rarely produced outside formulaic utterances (see also Myles, 2005).

Novel’s approach taught pronouns as individual lexical items. Consequently, pupil responses in Novel included occasional accurate use of the correct third-person pronoun preceding a (sometimes conjugated) verb other than “*est*” (‘is’). This suggests that, when taught

explicitly how to construct sentences rather than learning set phrases (often using the first-person “je”), pupils in Novel were able to bypass the stage identified in Myles et al.’s (1998) study with 11-14-year-old L1 English learners of French, where pupils had to modify first-person set phrases they had learnt in order to use verbs and pronouns creatively.

Although there were instances of pupils producing utterances containing verbs, there was nevertheless a predominance of unlinked nouns and adjectives which supports Courtney’s (2017) findings, in which study 10-11-year-old pupils struggled to produce sentences with verbs. However, it should be remembered that learners in the current study were three years younger than the youngest pupils in Courtney’s transition study, having only received 22.5 hours of French instruction by this point; consistent provision in subsequent years would hopefully develop and consolidate these rudimentary attempts.

Assuming that greater use of utterances containing verbs is indicative of increased grammatical production (see Marsden & David, 2008; Myles, 2005; Rule & Marsden, 2006), these outcomes may support findings of Bouffard and Sarkar (2008) indicating that, when instructed explicitly, young learners can indeed make progress in grammatical analysis. They may reinforce the concept of Lichtman’s (2016) ‘instructional hypothesis’, which asserted that children learn the way in which they are taught (see also Kasprovicz & Marsden, 2018).

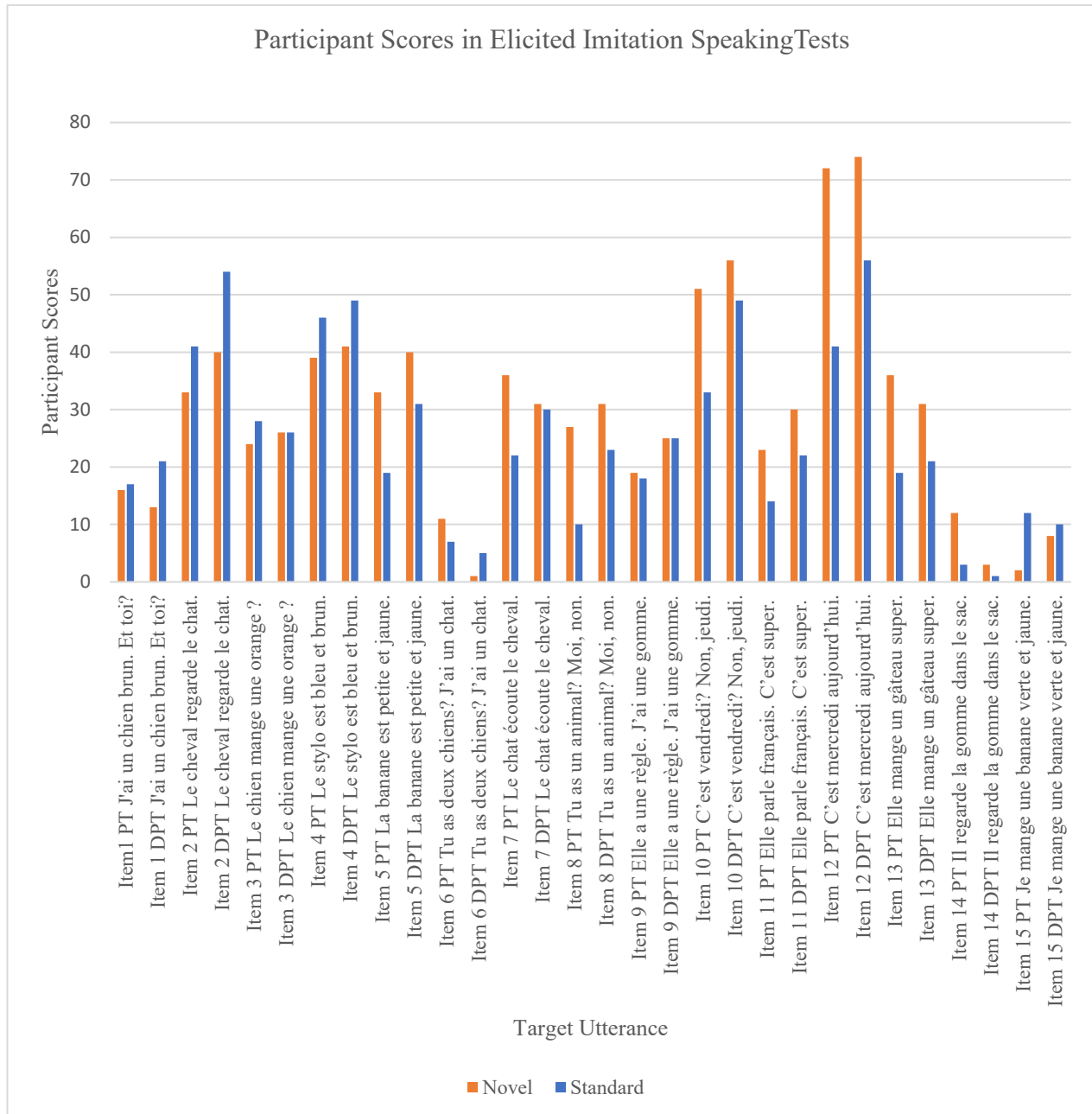
5.17.4.3 Elicited Imitation Speaking: results and discussion

Pupils were tested on 15 items ([Appendix E.25](#), which includes translations) with 4 points available for exact reproduction of meaning on each item ([Appendix E.33](#)). Pupils’ performance on individual items and comparisons of total scores on each utterance across approaches are shown in [Figure 32](#).

High-scoring items included “*C’est mercredi aujourd’hui*” (item 12) and item 10, where days of the week also featured. The two lowest scoring utterances for both approaches were item 6 (“*Tu as deux chiens? Moi, j’ai un chat*”) and item 14 (“*Il regarde la gomme dans le sac*”). Although utterances had been matched for syllable length ([4.3.8.2, iii](#)), these items were perhaps more challenging because of the number of ideas which pupils had to retain and reproduce (i.e., you have – two – dogs – question – me – I have – a cat), especially when compared with item 12 (it is - Wednesday – today). Interestingly, scores over time remained consistent or increased on several items, showing perhaps some durability of learning or, as discussed in [5.16.2.2](#), reflecting possible test and/or maturation effects.

Figure 32.

Participant Scores in Elicited Imitation Speaking Tests at Post-test and Delayed Post-test



Note. In [Figure 32](#), PT stands for post-test, and DPT for delayed post-test.

Pupils' superior performance in Novel (versus Standard) on personal pronouns other than the first person "je" (items 6, 8, 9, 11, 13 and 14 at post-test; [Appendix E.25](#)) aligns with outcomes in spoken grammar and the open-ended PR task, possibly reflecting the greater explicit emphasis placed on these in Novel. In contrast, pupils in Standard experienced greater success on items containing adjectives of colour (1, 4, 15), taught as a topic in Standard.

5.17.5 Overarching discussion of speaking outcomes

Although none of the between-group measures were statistically significant, there appeared to be an advantage for pupils in Novel versus Standard at both time-points in both measures ([Table](#)

38). Pupils performed better at delayed post-test than post-test in PR and EIS, again not statistically significantly. This possibly indicated durable learning, although may also be due to test and/or maturation effects. Baseline LAA scores seemed to influence outcomes more in Standard than Novel, with learners in Novel with lower LAA appearing to benefit particularly. There was a very slight advantage in Standard for learners with the highest baseline levels of LAA.

Pupils' ability to respond verbally to picture prompts appeared influenced by picture content; they seemed to cope better where nouns and adjectives could be supplied than where action verbs were needed. However, pupils in Novel produced a higher number of utterances containing verbs and also demonstrated greater knowledge of pronouns than Standard.

Table 38.

Overview of Speaking Outcomes

Grammar measure	Over time		Between approaches	
	Novel	Standard	Post-test	Delayed post-test
Open-ended PR	Post-test > Delayed post-test (trend)	Post-test = Delayed post-test	N > S (trend)	N > S (trend)
Elicited imitation	Post-test < Delayed post-test (trend)	Post-test < Delayed post-test (trend)	N > S (trend)	N > S (trend)

Note. Trend reflected in descriptives and (bordering on) medium effect size (.05 upwards).

5.18 Summary of learning outcomes

Overall, indicators of pupil progress in all learning outcomes were positive, regardless of approach: young learners *can* make progress in time-limited contexts when taught by non-specialist teachers. Although not always statistically significant, there were trends in favour of pupils in Novel outperforming those in Standard in all reliable tests. LAA appeared to influence outcomes, but slightly less so in Novel than Standard, with some nuances. Findings regarding pupils' ability to produce verbs on open-ended speaking tasks were encouraging, appearing to show a benefit of explicit instruction.

Chapter 6: Results and Discussion II. Pupils' self-efficacy and enjoyment

Whereas the previous chapter related to *learning* outcomes, this second results and discussion chapter presents and analyses the outcomes of two measures related to pupil *affective variables*: self-efficacy and enjoyment, with cross-reference where appropriate to learning outcomes. It also considers measures of pupil perceptions of teacher enjoyment.

Pupils (Novel: $n = 74$; Standard: $n = 72$) completed online surveys on self-efficacy and enjoyment at three points ([Table 11](#)). Within this chapter, because of the volume of data generated by this study, each of the two constructs has a quantified and qualitative results section, followed by discussion comparing findings to the body of existing literature on young learners' self-efficacy ([6.2](#), [6.3](#), [6.4](#)) and enjoyment ([6.5](#), [6.6](#), [6.7](#)). This discussion addresses the following research questions: 2a. What is the effect of each teaching approach on pupils' self-efficacy and enjoyment?; 2b. What are pupils' perceptions of the effect of each teaching approach on their progress in French?

6.1 General information on results

To avoid duplication, information pertaining to both self-efficacy and enjoyment measures is presented here.

6.1.1 Test quality

Test quality is presented at the start of each relevant section (self-efficacy, enjoyment), with text in bold indicating satisfactory results.

6.1.1.1 Reliability

Measures of reliability are reported using a combination of CFA and EFA as discussed in [4.3.4.2](#).

6.1.1.2 Handling missing data

[4.4.4.4](#) explains how missing data were handled.

6.1.1.3 Normality of distribution

Normality of distribution and assumption checks were conducted ([4.4.2](#), [4.4.4.5](#)); links to relevant appendices are provided in the reliability table for each outcome measure.

6.1.2 Effect sizes

Effect sizes were interpreted using Kraft's (2020) scale ([4.4.1.1](#)). In all tables of effect sizes, bold text indicates the time-point or approach with the higher scores. Confidence intervals were set at 95%, unless indicated otherwise.

6.1.3 Visualising descriptive data

Notched box visuals are presented for most findings (explained in [5.1.3](#)).

6.1.4 Predicted data: analyses of mixed-effects models

Details of mixed-effects models are provided in [4.4.4](#). For both measures, the type of model is outlined within the relevant section, with model codes in appendices.

6.2 Pupil self-efficacy: results from quantified data analysis

This section of the questionnaires PQ1, PQ2 and PQ3 ([4.3.12.4](#)) comprised seven questions each with the same matrix question, asking pupils to reflect on whether they thought they would be able to do certain things in French ([Appendix G.9](#)).

6.2.1 Test quality

Good levels of reliability ([Table 39](#)) were indicated at all time-points. Normality checks indicated non-normal distribution.

Table 39.

Quality of Measure Information for Self-efficacy Surveys

		Self-efficacy		
		PQ1	PQ2	PQ3
<i>N</i>		146	146	146
% cases complete prior to imputation		95	95	95
Reliability (CFA)	alpha	.83	.77	.82
	omega	.84	.78	.83
Normality of distribution link		Appendix P.1-P.2	Appendix P.1-P.3	Appendix P.1-P.4

6.2.2 Descriptive statistics for self-efficacy ratings

For each item, pupils rated their self-efficacy on a scale of 1 (lowest) – 4 (highest). For wording of response options and points system, see [Appendix G.10](#).

6.2.2.1 Between-groups

Ratings were lower in Standard than Novel at PQ1 but increased more than in Novel and were higher at PQ2 and PQ3 ([Table 40](#)). Comparing effect sizes, Standard was larger than Novel at both PQ2 and PQ3; effect sizes were medium at both time-points, but CIs crossed zero. As Standard had lower baseline ratings, the gains at each time-point were also calculated. Effect sizes for the gains PQ2-PQ1 and PQ3-PQ1 were large, and CIs only just passed through zero.

Table 40.

Descriptive Statistics and Effect Sizes for Self-efficacy Questions

Approach	Time-point (<i>k</i> = 7)	Descriptive statistics					Effect sizes		
		<i>n</i>	<i>Mean</i>	<i>SD</i>	Range	95% CI	Effect size	<i>d</i>	CI
Novel	PQ1	74	2.7	.75	1.0-4.0	2.5-2.9	Within-group:		
Standard	PQ1	72	2.6	.75	1.1-4.0	2.4-2.7	Novel: PQ1-PQ2	-.10	-.33-.12
Novel	PQ2	74	2.8	.59	1.6-4.0	2.7-2.9	Novel: PQ2-PQ3	-.03	-.27-.19
Standard	PQ2	72	2.9	.61	1.1-4.0	2.7-3.0	Novel: PQ1-PQ3	-.13	-.37-.09
Novel	PQ3	74	2.8	.60	1.3-4.0	2.7-3.0	Standard: PQ1-PQ2	-.45	-.70--.22
Standard	PQ3	72	2.9	.72	1.0-4.0	2.8-3.1	Standard: PQ2-PQ3	-.07	-.33-.15
							Standard: PQ1-PQ3	-.45	-.78--.20
							Between-group:		
							Novel-Standard: PQ1	.19	-.15-.51
							Novel-Standard: PQ2	-.14	-.48-.19
							Novel-Standard: PQ3	-.17	-.49-.14
							Novel-Standard: gains PQ2-PQ1	-.27	-.60-.06
							Novel-Standard: gains PQ3-PQ2	-.06	-.39-.27
							Novel-Standard: gains PQ3-PQ1	-.31	-.64-.02

6.2.2.2 Within-groups

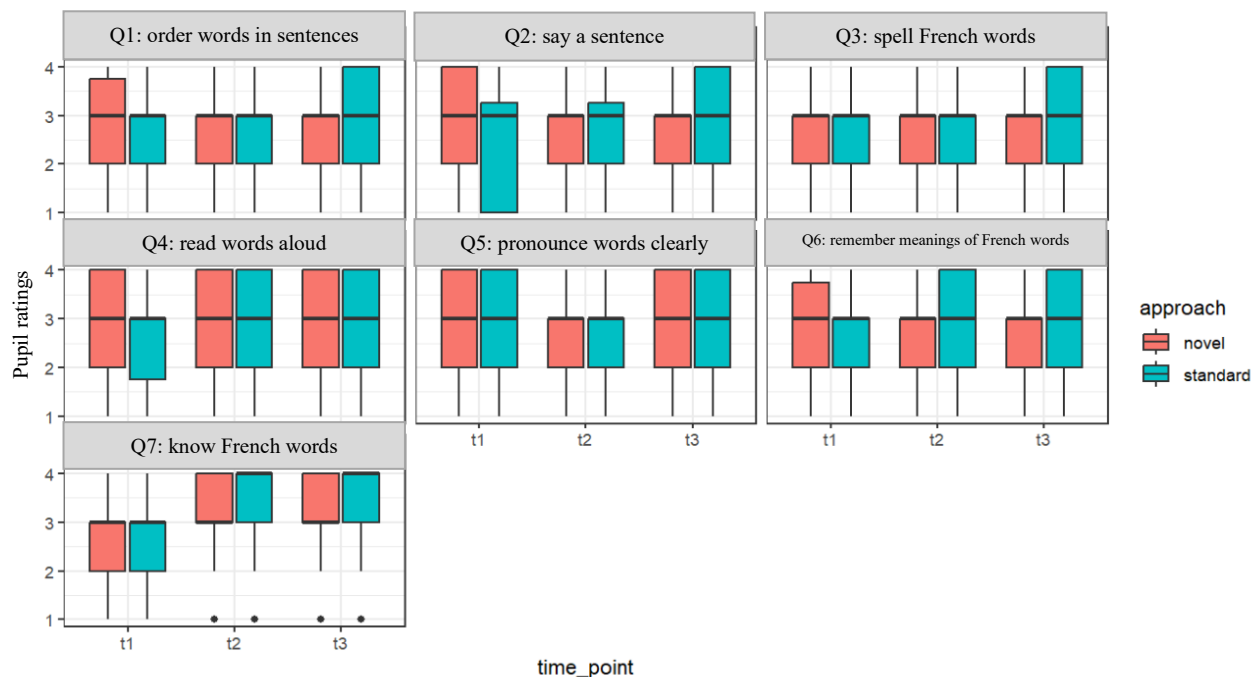
Means increased across the three time-points in both approaches ([Table 40](#)). Comparing Novel ratings at PQ1 and PQ3, the effect size was medium with CIs crossing zero. A larger increase was seen in Standard by PQ3, where the effect size was large, although Standard ratings, lower at PQ1, had more room for increase.

6.2.2.3 Individual item analysis

The box plot ([Figure 33](#)) shows participant responses (ratings 1-4, y axis) for each item in both approaches at the three time-points (t1, t2 and t3: x axis), illustrating the similar medians for each approach in all questions, except Q7. In Q1 (*Could you put the words in the right order in a sentence?*) and Q3 (*Could you spell the French words that you will learn?*), Novel and Standard medians were the same at PQ2 and PQ3, although the upper quartile extended to 4 points in Standard. In Q2, *Could you say a sentence that you want to say?*, medians were identical across approaches at both time-points, although pupils in Standard rated their ability to do this more highly by PQ3 than pupils in Novel, despite having exhibited lower levels of self-efficacy in this regard at the pre-study measure.

Figure 33.

Box Plot Showing Responses to Individual Items in Self-efficacy Questions by Approach



There were no between-group differences in responses at either time-point to Q4 (*Could you read words out loud in French?*) or Q5 (*Could you pronounce words clearly in French?*), with levels of self-efficacy increasing by PQ3 in the latter question in both approaches. Pupils in Standard approach appeared slightly more assured of their ability to remember the meaning of French words (Q6 – note the extended upper quartile) and to know the French for some English words (Q7 – note the higher median).

6.2.2.4 Comparison of individual item analysis with knowledge test outcomes

Although self-efficacy ratings were higher in Standard than Novel, this was not matched by pupils' performance in tests. For example, pupils in Standard rated their ability to spell words (Q3) slightly higher than pupils in Novel, but the outcomes of vocabulary recall in the written modality tests at post-test and delayed post-test were slightly higher for Novel (although the difference was not statistically significant) – see [5.5](#). Similarly, pupils in Standard also rated their ability to say a sentence (Q2) slightly higher than pupils in Novel, but their scores were descriptively slightly lower at both time-points in the open-ended PR task and in EIS, although differences were not statistically significant (see [5.15](#) and [5.16](#)).

In Q4 (reading aloud) and Q5 (pronouncing clearly), the median for pupils in both approaches was 3, but scores in Novel were higher for RA test at post-test (although not statistically significant – see [5.9](#)). Standard ratings for Q6 (remembering the meaning of French words) and Q7 (knowing the French for some English words) were slightly higher than Novel, but the descriptive trend was for higher scores in Novel in both vocabulary recognition tests (with a statistically significant difference in written recognition – see [5.3](#), [5.4](#)) and also in spoken recall ([5.6](#)).

6.2.3 Analyses using mixed-effects models

A cumulative link mixed-effects model was constructed (see [Appendix Q.1](#) for code).

6.2.3.1 Regression model assumptions

The Brant test Omnibus measure was $< .05$ ([Appendix Q.2](#)), with p -values for time-point, enjoyment ratings and approach also $< .05$. Thus, results should be interpreted with caution.

6.2.3.2 Self-efficacy

There appeared to be significant interactions between time-points and self-efficacy ratings, between positive anticipation baseline (whether pupils reported looking forward to French) and self-efficacy ratings, and also between approach and self-efficacy ratings ([Appendix Q.3](#)); these

are explored below. Fixed effects (marginal R^2) accounted for 5% of model variance, while fixed and random effects (conditional R^2) combined explained 29%.

(i) Between-groups: comparison of self-efficacy across approaches, categorised by positive anticipation baseline

Pupils' ratings were analysed in three categories based on the results from the question targeting positive anticipation (see [4.4.4.2](#)) in PQ1 (*Are you looking forward to learning French this year?*), which had three options: 'Yes'/'Unsure'/'No'. Pairwise comparisons are included in [Appendix Q.4](#). There was no statistically significant difference between approaches at any time-point for any of the three categories of pupils (who were, were unsure about, or were not looking forward to learning French).

(ii) Within-groups: change in self-efficacy over time, categorised by baseline positive anticipation

Baseline: pupils not looking forward to learning French ('No')

The sample size of pupils not looking forward to learning French (Novel, $n = 4$; Standard, $n = 3$) was too small to conduct statistical analysis.

Baseline: pupils unsure whether they were looking forward to learning French

There was no difference between PQ1 and PQ3 in Novel for these pupils ($n = 29$) (estimate = -.15, SE = .19, $z = .83$, $p = 1.00$). However, in Standard ($n = 23$), pairwise comparisons indicated a statistically significant difference between PQ1 and PQ3 (estimate = -.78, SE = .21, $z = -3.71$, $p < .001$), with this apparent by PQ2 (PQ1 – PQ2: estimate = -.79; SE = .21, $z = -3.75$, $p < .001$). The likelihood of these pupils in Standard scoring the highest ratings of 3 or 4 points at PQ1 was 48%, rising to 67% at both PQ2 and PQ3. (The probabilities in Novel of scoring 3 or 4 were 53% at PQ1 and PQ2, and 57% at PQ3).

Baseline: pupils looking forward to learning French ('Yes')

In this category, there was also a statistically significant difference between PQ1 and PQ3 in Standard ($n = 44$) (estimate = -.89, SE = .15, $z = -5.85$, $p < .001$), again evident by PQ2 (estimate = -.66; SE = .15, $z = -4.42$, $p < .001$). The likelihood of pupils in Standard scoring 3 or 4 points at PQ1 was 60%, rising to 74% at PQ2 and 78% at PQ3. There was a similar upward trend in Novel ($n = 40$), although this did not quite reach statistical significance (PQ1 – PQ3: estimate = -.38, SE = .16, $z = -2.43$, $p = .05$), but the starting point was higher. The probability of these pupils scoring 3 or 4 points was 70% at PQ1, rising to 75% at PQ2 and 76% at PQ3 ([Appendix Q.4](#)).

6.2.4 Pupil self-efficacy: summary of quantified data

Pupils in Standard had lower self-efficacy at the start of the project than learners in Novel. When comparing Novel and Standard approaches across time-points, pupils in Standard made descriptively larger gains with large effect sizes, but the model did not show the differences to be statistically significant. Considering pupil response to individual self-efficacy items, ratings in both approaches were similar, with Standard descriptively slightly higher. Ratings of pupils in individual self-efficacy items did not always relate to their actual performance in knowledge tests.

6.3 Pupil self-efficacy: results from qualitative data analysis

Twenty group interviews (10 Novel, 10 Standard, two groups from each of the five schools) were conducted ([4.3.13.3](#)), five at the mid-point (group interview 1) and five at the end of the project (group interview 2) in each approach. Pupils' responses in group interviews (see [G.7](#) for schedule) as well as to open-ended questions in PQ1 (*How well do you think you will be able to learn French in Year 3? Why do you think this is?*) were analysed broadly within the context of Dörnyei's framework (1994) – see sections [3](#) and [4.5.3](#). Qualitative findings address the *learner level* and are presented in four overarching themes: confidence ([6.3.1](#)), perceptions of future progress ([6.3.2](#)), perceptions of progress during the project ([6.3.3](#)) and spontaneous use of French ([6.3.4](#)). These are accompanied by evidence and a selection of supporting quotations - additional pupil quotations for all themes are provided in [Appendix R.1](#). Themes were constructed based on some of the a priori codes identified in the literature review: confidence and self-belief (Dörnyei, 1994); types of motivation; the importance of contact/connections with the foreign country (Gardner & Lambert, 1972; Tierney & Gallastegi, 2011); attitudes towards learning French (Courtney, 2017; Courtney et al., 2017; Hunt et al., 2008; Tierney & Gallastegi, 2005); and classroom factors (Dörnyei, 1994; Fenyvesi et al., 2020; Graham et al., 2016; Mihaljević Djigunović, 2009; Tierney & Gallastegi, 2011).

6.3.1 Theme of confidence

6.3.1.1 Start of the study: Pupil Questionnaire 1

PQ1 ($n = 146$ [Novel: 74; Standard: 72]) asked how well pupils thought they would learn French this year and why. The data suggested little difference between approaches. In terms of confidence and self-efficacy, some pupils from both approaches anticipated difficulty (Novel: *"It might be tricky"*), with one pupil from Standard referring to French as *"very hard to pronounce"* and another saying they were *"scared"*. Some pupils in Novel linked potential

difficulty with lack of prior knowledge (“*Because new is hard to me*”) and anticipated the demands likely to be placed on memory, concerned about forgetting. Equally, pupils from both groups expressed lack of confidence, with comments that they were “*not good at learning other languages*” (Standard) or French specifically. The number of *comments* indicating lack of confidence was greater in Standard (eight) than in Novel (four), although the number of *pupils* referring to having some prior knowledge of French was higher in Novel responses (12/74) than in Standard (6/72), which may have contributed to greater self-belief. This aligns with the more positive self-efficacy starting points identified among pupils in Novel in PQ1 quantitative data (6.2.2.1).

6.3.1.2 Mid-study: Pupil group interview 1

Pupils were asked in group interview 1 whether they believed they were good at learning French. Concept tables (Figures 34 and 35) present salient responses regarding pupils’ perceptions of their ability in each approach. The confidence category in the first column is the researcher’s own classification and was used to facilitate discussion/reference.

In Figure 34, the eight responses in Novel illustrate that confidence varied at this point. Three of the first five comments show how confidence related to word knowledge: inability to remember meanings of words was equated with failure or lack of ability. These responses all came from pupils with low-medium baseline levels of LAA (LLAA-MLAA) – see 4.3.6.4, ii and E.17 for explanation of these categories. In the ‘Medium’ confidence category, the pupil’s confidence was increasing with greater exposure to the language. The last two quotations illustrate the highest confidence levels in French in Novel.

Figure 34.

Impact of Approach on Confidence at Mid-point from Pupil Group Interview Data: Novel

Confidence Category	Supporting Participant Quotes	Analytic-ability	School
Very low	<i>I’m bad at it ... Because I can only remember one word.</i>	Low	B
	[What do you think you’ll be able to do in Year 4?] <i>Nothing.</i>	Low	E
Low	<i>I only know a few words ... not much at all.</i>	Medium	B
	[Do you think you’re good at learning French?] <i>I’m not sure about that.</i>	Low	E
Medium	<i>I think I understand ... French a little bit more, because sometimes I just get the words mixed up, so ... now that I’ve done a little bit of more French, I think I can guess at them.</i>	Medium	E
High	<i>I can pronounce words much better than I used to, but also, I can say some sentences.</i>	Low	B
Very high	<i>I think I’m amazing at it.</i>	High	A

	<i>I can think I'm a hundred per cent.</i>	Medium	B

Standard responses (Figure 35) suggest that, at group interview 1, there appeared little association between confidence levels and pupils' LAA. Quotations from pupils in Standard were mainly concentrated around the mid-levels, with few demonstrating extremes of confidence. Although two of the first five quotations refer to perceived difficulty related to vocabulary recall (as in Novel), three others cited pronunciation as a challenge. When asked to compare current progress with their starting point, one pupil (with high confidence) felt that they now knew considerably more, as noted in Novel.

Figure 35.

Impact of Approach on Confidence at Mid-point from Pupil Group Interview Data: Standard

Confidence Category	Supporting Participant Quotes	Analytic-ability	School
Very low	<i>I find it fun, but just sometimes words are very difficult, and sometimes I need a lot of time to think before I do the word.</i>	Low	C
	[Do you think you're good at learning French?] <i>No ... cause I always forget some stuff.</i>	High	F
Low	<i>It's like ... you're just really nervous to say it and ... you might say it wrong.</i>	High	C
Medium	<i>I think I'm kind of in the middle of good and ... not too bad or not too good, cause I find it hard to remember words and sometimes pronou-pronounce them.</i>	Medium	B
	[Do you think you're good at learning French?] <i>I'd say in the middle. Because of the sentences we use in French ... it's quite difficult ... how to pronounce the words. I find it difficult so I just ... don't volunteer cause it's a little tricky for me.</i>	Low	B
High	[do you think you're good at learning French?] <i>Oui. Because I usually learn most things quite quick.</i>	High	B
	[What things can you do now in French that you couldn't do before?] <i>Really, everything. I knew 'bonjour', that's probably the only word I actually knew at the start.</i>	Medium	C
	[And what can you do now?] <i>A lot.</i>		
Very high	<i>I'm good at learning French, cause it's very easy.</i>	High	B

6.3.1.3 End-study: Pupil group interview 2

In group interview 2, as in group interview 1, pupils in Novel (with both lower and higher confidence) referred to their knowledge of words as a measure of their perceived success and seemed adversely affected by struggling to remember (Figure 36). Both responses in the

‘Medium’ confidence category demonstrated pupils’ sense of progress compared with their starting points, although the comment made by HLAA (SchB) perhaps alluded to awareness of the long haul of language learning, which may ultimately be demotivating. In the ‘High’ confidence category, HLAA (SchB) referred to the potential impact of being able to work without support, suggesting increasing independence. Some pupils spontaneously demonstrated their knowledge, as illustrated here by MLAA (SchB) who very confidently and accurately summarised their learning about the different [e] sounds.

Figure 36.

Impact of Approach on Confidence at End-point from Pupil Group Interview Data: Novel

Confidence Category	Supporting Participant Quotes	Analytic-ability	School
Very low	<i>I don't think I'm good at it because ... I'm bad at it.</i>	Medium	B
	<i>I think I'm really bad because I ... never know what happens to my ... to my head cause ... at the start of the French lessons, I'm literally, like, 'I've got that word, I think it's going to be on one of those quiz sheets' and then, next, I don't know ... That's why I'm, like, terrible at it ... I think it's in my back box ... where everything usually is.</i>	Low	E
Low	<i>Er, no, not really. It's just because ... when we're doing the quiz it's, like, all the words just rub out of my mind. Go out the back of my head.</i>	Medium	E
Medium	<i>I don't think I'm good but I also don't think I'm bad. I think I'm in the middle. Cause ... I don't know as much as I could know, but I also know ... way more than I did in September.</i>	High	B
	<i>In the middle. Cause it was really harder before, but now it's really easy.</i>	Low	F
High	<i>There's [sic] loads of words I understand ... I can memorise them, and I don't have to look at the board.</i>	High	B
	<i>I know nearly all the words ... and it makes me feel like I'm ... good.</i>	Medium	E
Very high	<i>[Do you think you're good at learning French?] Yes. I say that because every Monday, I learnt something new ... about sounds. Cause the [e] sounds, there's three [e]s and they sound different. One of them sounds like [ez] ... one sounds like [è] and one of them sounds like [e].</i>	Medium	B
	<i>[Do you think you're good at learning French?] Fabulous. I'm a hundred per cent or infinite. I think I show the teachers how to speak French.</i>	High	F

In Standard group interviews, as in Novel, pupils affected by low confidence referred to poor recall of vocabulary (Figure 37). However, struggling with accent and pronunciation also apparently impacted negatively on confidence, as in group interview 1.

Figure 37.

Impact of Approach on Confidence at End-point from Pupil Group Interview Data: Standard

Confidence Category	Supporting Participant Quotes	Analytic-ability	School
Very low	[Do you think you're good at learning French?] <i>No, because when I hear the word, I remember them for at least, like, 10 seconds ... and then I forget them.</i>	High	A
	<i>Rubbish ... cause I try and then I forget.</i>	Low	A
Low	[Do you think you're good at learning French?] <i>Sort of ... just above the bottom, because I usually always forget the words, and then it's hard for me to say them out loud, because I forget them a lot, and when I s-see and hear them again, I usually remember them. And in fact, it's just really hard for me to remember because I have a bad memory.</i>	High	A
	<i>Let's say I'm not the best. Because I'm not. ... I can barely say ... hmm, not really.</i>	Low	C
Medium	<i>Well, yes and no. Because ... when you have to speak in front of the whole entire group ... you have to, like ... it's really hard to pronounce the words correctly, cause obviously we don't have, like, the right voice to talk in that accent, but I'm good at it but just not when it comes to pronouncing it correct.</i>	High	C
	<i>Probably, like, somewhere in the middle, cause I can't really do the accent.</i>	High	F
High	<i>I think I'm good at French because I'm always asked questions a lot and I help people.</i>	High	C
	[Do you think you're good at learning French?] <i>Yes. Because I know words and some people – it's a bit hard for people, so I help them and stuff ... So, that's why I think I'm good at French.</i>	Low	F
Very high	<i>I am really good at French ... because I know a lot of things ... Every time she asks us ... normally I get it right.</i>	High	B

Responses from both approaches show that young English learners of French express concern about their performance; as in group interview 1, there were fewer examples of very high confidence in Standard than Novel. Nevertheless, pupils in Standard were beginning to feel confident enough to support others (Figure 37 – ‘High’ confidence: HLAA, SchC; LLAA, SchF).

6.3.2 Pupils’ perceptions of future progress

During both rounds of group interviews, pupils were asked what things they thought they would be able to do in French the following year. Their responses are discussed here by time-point.

6.3.2.1 Mid-study

Pupils from both approaches conceptualised their potential progress similarly (Figure 38), referring generally to coping with increased linguistic challenge and being able to learn more – in Standard, contextualised within topics (e.g., “learn more animals”). Pupils from both

approaches mentioned being able to improve their written and oral production (Novel: “[I will be able to] write more French words”, MLAA, SchF), and, for both sets of pupils, being able to read longer texts was an appropriate, potentially achievable goal (Standard: “reading a whole book”, LLAA, SchA). Only pupils in Novel referred specifically to improving listening skills (“[being able to] understand what other people are saying ... sometimes”, HLAA, SchB) and recall – perhaps reflecting the emphasis placed on this in Novel. Pupils in Novel were also alone in mentioning being able to work independently (“do it without the Knowledge Organiser”, HLAA, SchB). For both sets of pupils, being able to speak at both word and sentence level were Year 4 goals, with some pupils in Standard targeting practical communication (“[I will be able to] go to France and ... talk to ... France people [sic]”, MLAA, SchA).

Figure 38.

Pupils’ Perceptions of What They Will Be Able to Do in French in Year 4 (Mid-point Pupil Group Interviews)

Novel (<i>n</i> = 29)	Standard (<i>n</i> = 29)
Remember French words (1)	Read in front of people (1)
Learn new French words (3)	Use a dictionary (1)
Work without support (3)	Learn more (1)
Understand French better (4): words (2)/what people say (2)	Write sentences (2)
Read in French (4): whole books (3)/pages in a book (1)	Learn new sentences (2)
Write French (5)	Learn trickier things (3)
Say [more] things in French (11): words (2)/sentences (5)/paragraphs (4)	Read French books (6)
	Have a conversation (7)
	Learn new words (8)

Note for Figures 38 and 39. The starting size for the font was 11 (reflecting one comment), increasing by one point with each mention. The number of comments also appears in brackets.

6.3.2.2 End-study

As in group interview 1, pupils in both approaches conceptualised their future abilities in terms of vocabulary acquisition by the end of the teaching input (Figure 39). References to potential progress in *topics* were now apparent in both approaches (Novel: “[I will be able to] pronounce every single animal that exists”, HLAA, SchE; Standard: “[I will be able to] know more fruits and vegetables and food, and ... count farther, to, like, twenty or thirty”, LLAA, SchE). The

language taught in Novel was not attached to topics, but seven pupils in Novel referred to being able to learn animals, foods, colours, and numbers (five in Standard). Although one pupil (in Standard) mentioned being able to read aloud, no participants mentioned reading for comprehension, whereas this had featured in both approaches (Novel: four; Standard: six) at the mid-point data collection.

Being able to pronounce French words better was mentioned in group interview 2 for both approaches, showing perhaps pupils' heightened awareness of the differences in sounds between French and English (Novel: "[I will be able to] do, like, long words and ... pronounce the sounds correctly", HLAA, SchF). For pupils in both approaches, being able to make practical use of the language to communicate with French-speakers was a clear goal (Standard: "talking to French people if I go to France", LLAA, SchE), reflecting the National Curriculum aim (DfE, 2013a:1): "communicate for practical purposes".

Figure 39.

Pupils' Perceptions of What They Will Be Able to Do in French in Year 4 (End-point Pupil Group Interviews)

Novel (n =29)	Standard (n= 29)
Know French names (1) Write paragraphs (1) Have conversations (1) Know the French alphabet (2) Communicate with French-speaking people (2) Speak French (3) Improve generally (3) Be able to pronounce words and sounds correctly (4) Learn new words (17)	Work without support (1) Know about different French-speaking countries (1) Read aloud (2) Improve generally (2) Have conversations (2) Be able to pronounce words and sounds correctly (3) Communicate with French-speaking people (4) Write in French (6): a book (1)/sentences (1)/words (2)/paragraphs (2)) Learn new words (8) Say sentences (10)

One pupil in Standard struggled to articulate what form progress in Year 4 might take, simply referring to being able to do "more of the same" (HLAA, SchB), while three pupils (Standard: one; Novel: two) were unsure what they would be able to do. A different pupil, also from Standard, was negative about their potential for progress: "I don't think I'll be able to underst-, remember anything ... cause I forgot everything from the start" (MLAA, SchC). A

participant from Novel was similarly negative about likely progress: “... *probably about two words, that’s it*” (LLAA, SchA).

6.3.3 Perceptions of progress during the project

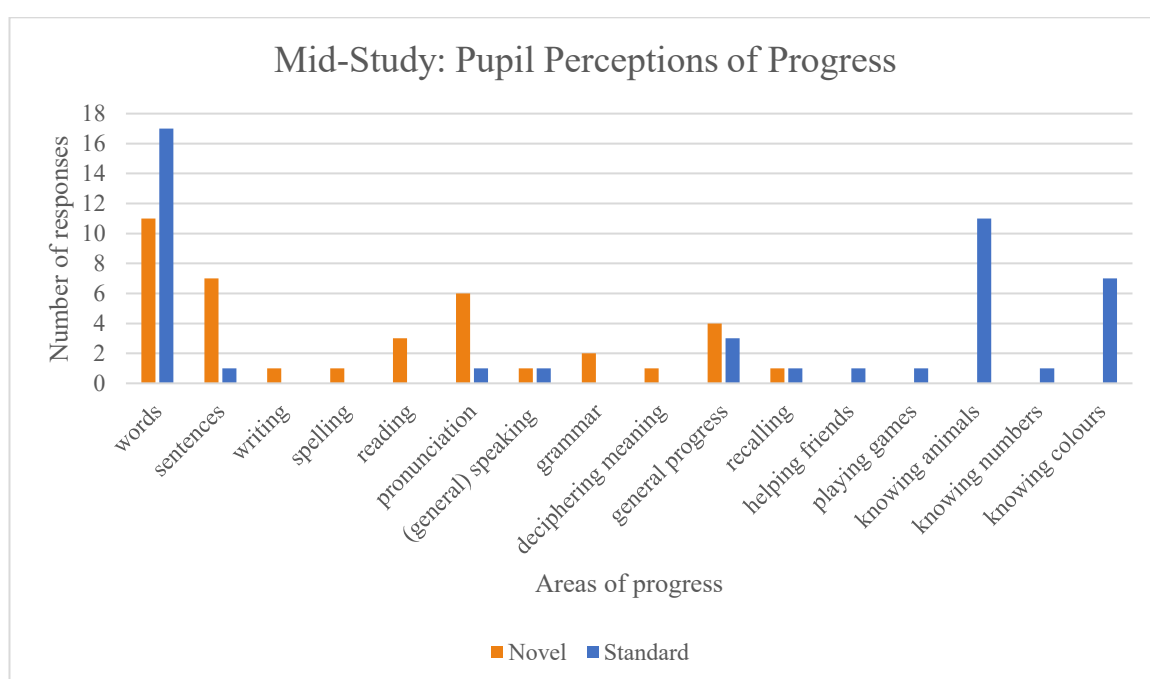
6.3.3.1 Mid-study

Additional contributions to an overview of pupils’ self-efficacy came from group interview discussion centred on their perceptions of progress since starting French. At group interview 1 (Figure 40), pupils in Standard recognised their achievements in topic-related words (e.g., animals, colours and numbers): “*I couldn’t say anything except for ‘bonjour’ before. [And what can you say now?] ‘Un chat’. ‘Un, deux, trois, quatre, cinq’. ‘Six’. And I know every single colour, except for some mixing colours*” (HLAA, SchB). Although pupils in Novel felt they had made progress in “words”, they also referred to sentences: being able to “say” (five comments); “understand” (one comment); or “read” them (one comment): “*I think I’ve got a lot of progress ..., because at the beginning I could only say ‘bonjour’, and now I can say, er, sentences and trick-, very, very hard words in French*” (MLAA, SchA).

Six pupils in Novel referred to their knowledge of pronunciation or the sounds of French, compared with one in Standard, suggesting that more pupils in Novel felt confident with the FL sounds than in Standard. “*I can pronounce words much better than I used to, but also, I can say some sentences*” (LLAA, SchB, Novel).

Figure 40.

Pupil Perception of Progress: Mid-point Pupil Group Interviews



6.3.3.2 End-study

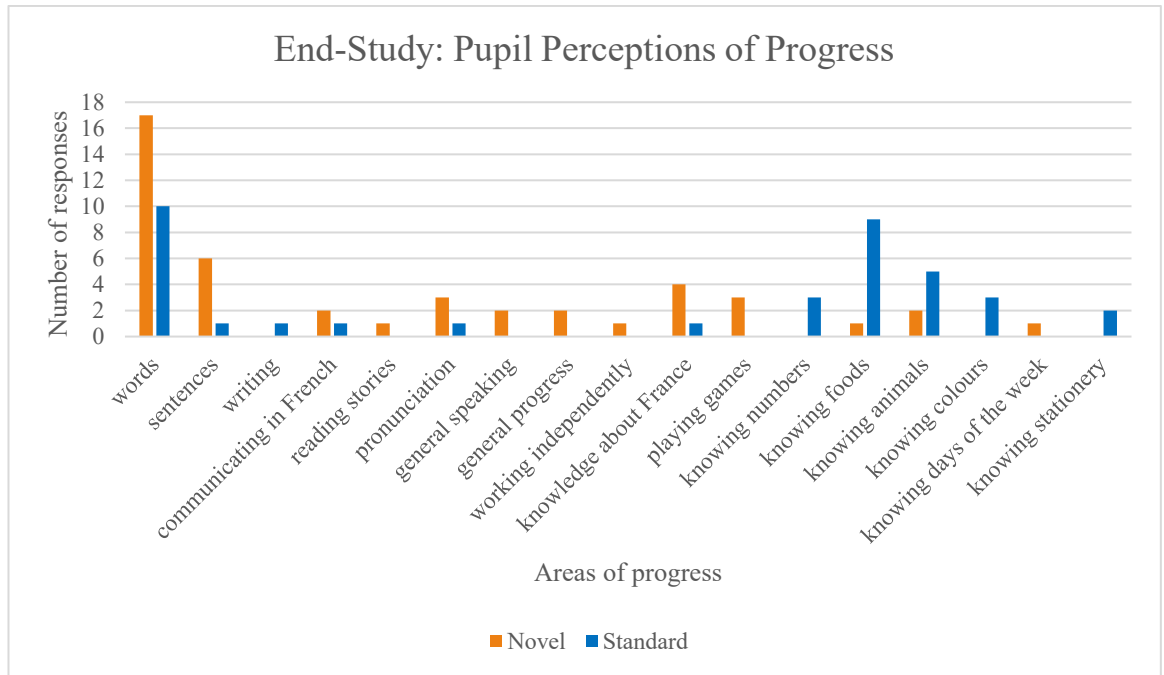
By group interview 2 ([Figure 41](#)), pupils in Standard still referred to progress made in topic-related vocabulary (“*knowing animals*”, “*knowing foods*”). Although most often referring to knowing vocabulary *generally*, pupils in Novel were also beginning to equate some individual items of high-frequency vocabulary that they had learnt with topics (e.g., “*chat*” and “*chien*” with animals).

More pupils in Novel than Standard referred to their knowledge of words and ability to form sentences at this point of the study. One pupil explained how they had made progress even within the field of vocabulary: “[*I can*] *say[ing], like, longer words ... like, ‘problème’, ‘un uniforme’. Words that I found tricky in the beginning*” (HLAA, SchF, Novel). The reference to “words” sometimes indicated a grasp of lexical items which would support sentence creation: “*I also learnt how to ... say ‘you are’ and ‘she is’ in French*” (LLAA, SchE, Novel). In group interview 2, more pupils in Novel cited sentence-level progress than in Standard (six mentions across 29 students versus one across 29 students), perhaps reflecting the increased emphasis in Novel on teaching learners component parts and encouraging sentence-building. Perceived progress in pronunciation was less apparent in both approaches in the final group interviews, although was again more frequently mentioned in Novel (five references) than Standard (one reference). There were indications that pupils in Novel had increased their understanding of the French way of life and early signs in both approaches of pupils’ awareness of the potential usefulness of the language for communicating with French-speaking people: “*... now, if I go to France, I know how to speak to people*” (LLAA, SchE, Novel).

Stories featured in both schemes of work (as required by the National Curriculum, DfE, 2013a), although more prominently in Standard, with series of lessons structured around Eric Carle’s *Brown bear, brown bear, what do you see?* (Carle, 1967) and *The very hungry caterpillar* (Carle, 1969). In Novel, stories served to broaden the range of vocabulary encountered and provided opportunities for revisiting and recycling. Interestingly, only one pupil in Novel (and none in Standard) referred to stories when discussing their progress: “*I learnt definitely how to, like, read French stories ... like the one with Lulu the tortoise*” (LLAA, SchE).

Figure 41.

Pupil Perception of Progress: End-point Pupil Group Interviews



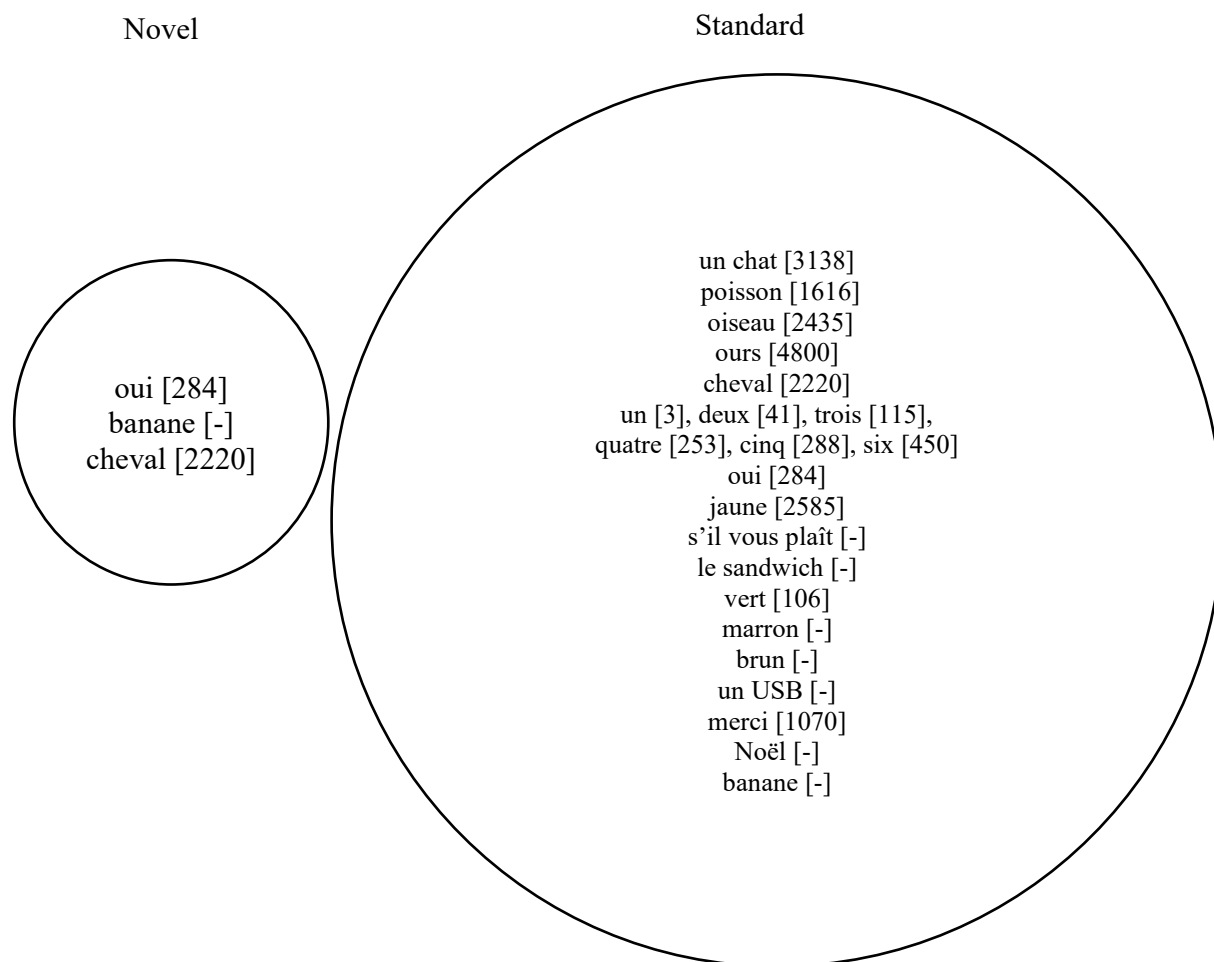
6.3.4 Spontaneous use of French

6.3.4.1 Mid-point of the study

During group interview 1, pupils from Standard were more likely than those in Novel to volunteer words in French (unprompted) and spontaneously demonstrate their knowledge of key topics such as colours, animals and numbers ([Figure 42](#)).

Figure 42.

Range of Different Words and Phrases Volunteered (Unprompted) in Each Approach: Mid-point Pupil Group Interviews



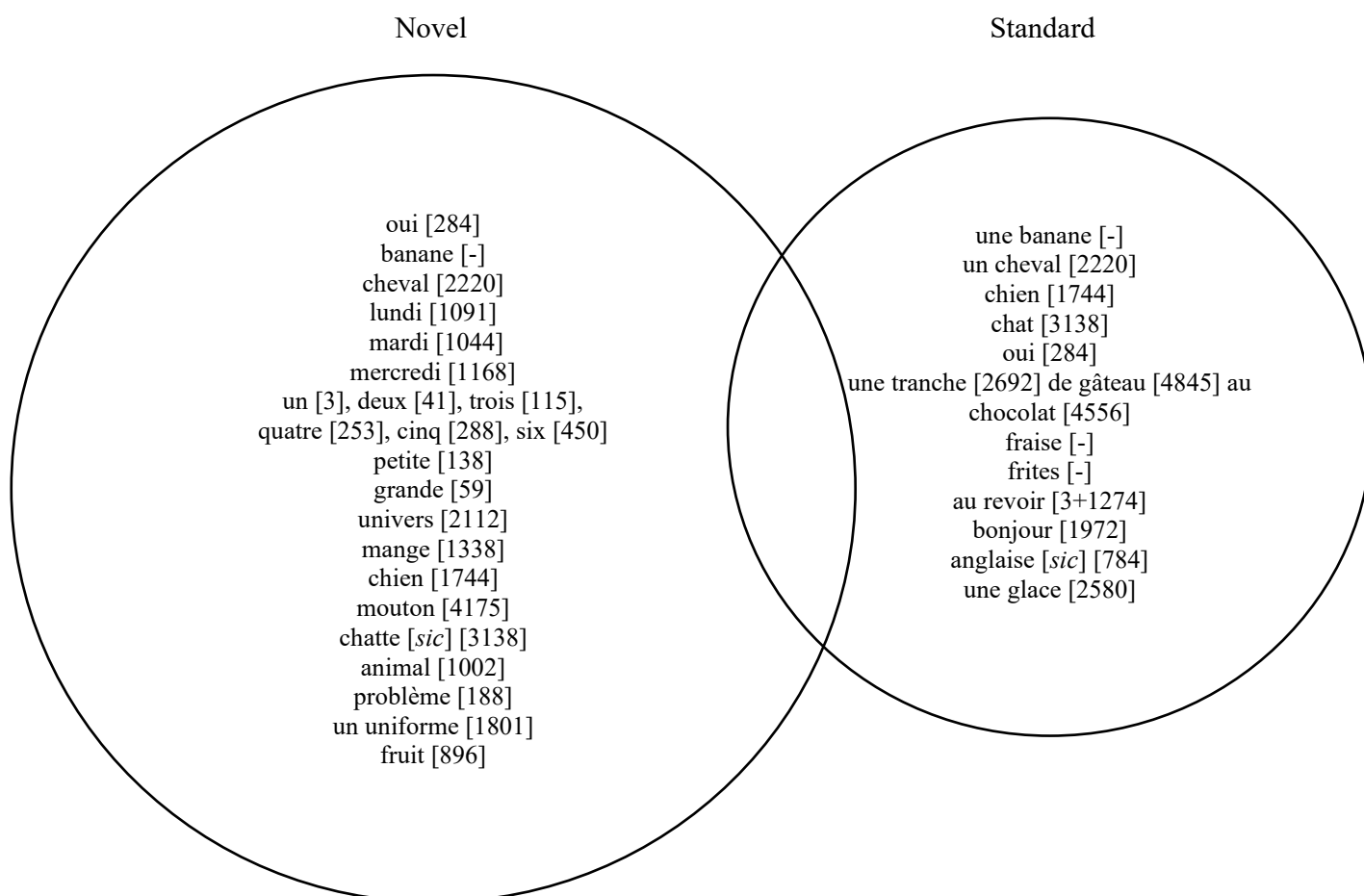
Note for Figures 42 and 43. Corpus-based frequency ranking according to Lonsdale and Le Bras (2009) is shown in brackets after each word. [-] indicates frequency ratings above 5000.

6.3.4.2 End-point of the study

Pupils in Novel provided more French beyond basic “oui” utterances by the end of the project (Figure 43). The number of different words and phrases volunteered by pupils in Standard was similar at both data collection points, with over half of these topic words (foods, animals), whereas examples in Novel reflected the range of word classes featured in that scheme, including adjectives and one verb, as well as category headers (“animal”, “fruit”). The average corpus-based frequency ranking for Novel words in group interview 2 was lower [1021] than Standard [2174].

Figure 43.

Range of Different Words and Phrases Volunteered (Unprompted) in Each Approach: End-point Pupil Group Interviews



6.3.5 Pupil self-efficacy: summary of qualitative data

Pupil comments at PQ1 evidenced similar starting points across approaches. Responses in group interviews indicated the importance of vocabulary knowledge to these young learners. For example, confidence and self-belief were associated in both approaches with self-reported success or failure in the learning of French. In both approaches future progress was also conceptualised as knowing more vocabulary, and reflections on progress made included self-assessment of lexical items learnt. Unprompted demonstrations of knowledge volunteered in group interviews were based on vocabulary learnt. In the Novel approach, greater experience of French throughout the project seemed to increase confidence and self-efficacy of some learners, and they referred more often to ability at sentence level. Pupils in both approaches envisaged practical use of the language as a realistic target. In Standard, pupils appeared slightly more aware of pronunciation difficulties than learners in Novel.

6.4 Pupil self-efficacy: discussion of quantitative and qualitative data analysis

This section discusses outcomes of the self-efficacy quantitative measures and qualitative findings from pupil group interviews and the open-ended questions in PQ1.

6.4.1 Pupil self-efficacy: quantitative trends

This section discusses trends emerging from between-group, then within-group, comparisons.

6.4.1.1 Between-groups

Pupils in Standard had a descriptively lower starting point than those in Novel, yet the medians in Standard were the same or higher on all seven questions ([6.2.2.3](#)) and their ratings were descriptively higher at both PQ2 and PQ3. However, none of these descriptive differences reached statistical significance. This indicates that pupils' self-efficacy levels were similarly impacted by being in Novel and Standard.

6.4.1.2 Within-groups

Regardless of approach, self-efficacy means increased from the baseline at both PQ2 and PQ3 ([Table 40](#)), suggesting that pupils generally felt more capable of completing certain tasks in French as they made progress during the year. Reporting on the early language learning project in Croatia, comparing attitudes of two groups of primary-aged learners, Mihaljević Djigunović (2009a) concluded that greater time spent learning the language contributed to their sense of self-concept and confidence; similarly, Graham et al. (2014) found that perceived progress made through *continued* study of the FL contributed to pupils' increased self-efficacy in transition from primary to secondary school. In the current study, pupils had only 45 minutes' French weekly for 30 weeks, yet this learning experience still impacted positively on their sense of self-efficacy. However, this could also reflect the general tendency for younger learners to demonstrate higher levels of self-efficacy than older learners (Eccles et al., 1993; Fenyvesi et al., 2020; Muenks et al., 2018; Wigfield & Eccles, 2000).

The upward trend in self-efficacy ratings between PQ1 and PQ3 shown by the model did not reach statistical significance for any group of pupils in Novel, regardless of their level of initial positivity about the prospect of learning French in Year 3. This possibly indicates that all pupils' self-efficacy was affected similarly by being in the Novel approach, regardless of baseline attitudes. In Standard, where pupils had lower baseline self-efficacy than in Novel, there was a statistically significant difference between PQ1 and PQ2, and between PQ1 and PQ3, for pupils unsure about or looking forward to learning French. This suggests that experiencing the Standard approach had a greater positive effect on pupils who expressed enthusiasm about learning French at the outset.

6.4.1.3 Comparison of individual item analysis with knowledge test outcomes

Pupils in Standard rated themselves higher on five items than learners in Novel (Q1: order words in a sentence; Q2: say sentences; Q3: spell French words; Q6: remember meanings; Q7: know

some French words for the English ones) – [Figure 33](#). However, the between-group difference for pupils’ self-efficacy ratings on individual items was not statistically significant. It is possible that the nature of Standard approach, which included less explicit teaching of SSCs, made pupils feel that they could do more in the way of making connections between the printed and spoken word and spell French words more accurately, simply because they were less aware of mistakes. Similarly, the regular revisiting and recycling of vocabulary in Novel may have increased pupils’ awareness of their own deficiencies in this respect relative to pupils in Standard, where learners worked most often on a closely related set of (topic-focused) lexical items (usually nouns, rarely having to recall previous knowledge). In these aspects, it is possible that pupils exemplified the first two stages of the Four Stages of Competence (generally attributed first to Broadwell, 1969, then developed by Noel Burch in the 1970s): knowing a little can give you confidence (the ‘unconscious competence’ stage), but you are unaware of what you do not know; knowing more can decrease your confidence because your awareness of what is still to be learned is greater (‘conscious incompetence’). This Dunning-Kruger effect was reported as a “mis-match between self- and other-assessments” (Saito et al., 2020:1).

6.4.2 Pupil self-efficacy: qualitative trends

This section discusses findings presented in [6.3](#) relating to confidence, self-efficacy, perceptions of progress and spontaneous use of French.

6.4.2.1 Theme of confidence

Pre-project comments from pupils in PQ1 indicated concern about anticipated difficulty in learning French, often associated with lack of existing knowledge, whereas prior knowledge was related to increased confidence at the outset. By group interview 1 and 2, more confident pupils (in both approaches) were able to draw positive comparisons between current progress and their starting points, realising that they knew more French than at the outset. These examples support the suggestion made in [6.3.1.2](#) and [6.3.1.3](#) that progress impacts positively on self-efficacy, appearing to corroborate the importance of perceived progress for pupils’ increased self-efficacy found by Graham et al. (2014). However, taking stock of current progress may have developed in some pupils a greater awareness of what lies ahead, which could perhaps have been daunting and demotivating, noted by Muijs et al. (2005) and Myles (2017).

In both group interviews, pupils in both approaches (regardless of confidence levels) referred to being negatively affected by not being able to recall words, supporting initial concerns expressed at PQ1 about not remembering. Knowledge of words was equated with success. This was also found by Muñoz (2014a:36), in which pupils’ qualitative data indicated that “learning more words equals learning ‘more’ English”.

The only reference to pronunciation in Novel was made by a pupil exhibiting high confidence and was accompanied by accurate demonstrations of knowledge (Figure 36, ‘Very high’ confidence category, MLAA, SchB). The lack of confidence in Standard may exemplify the negative responses to the FL found by Erler (2004) and Woore (2007) among learners struggling to decode words. Woore (2007:177) suggested that this reinforced learners’ lack of independence – a reminder that they were beginners.

In sum, concern about performance was expressed by pupils with varying levels of confidence and did not appear related to baseline LAA, at least in group interview data. This potentially contrasts the positive associations found by Courtney et al. (2017) between L1 literacy and attitudes towards FL learning and confidence in French learning ability at primary-secondary transition.

6.4.2.2 Pupils’ perceptions of future progress

Pupils’ responses to the question “*What things do you think you’ll be able to do in French next year?*” were analysed to see if there were differences between approaches, as conviction that they would succeed in future tasks may indicate commitment to language learning (Heinzmann, 2013). Pupils’ responses were similar across approaches. In both Novel and Standard, pupils commented on being able to learn more, write more, read longer texts and speak at both word and sentence level. In both group interviews, pupils in both approaches conceptualised their future abilities in terms of vocabulary acquisition by the end of the teaching input, as found by Muñoz (2014a). For pupils in Standard, this vocabulary consisted of topic-specific items, which is unsurprising given the focus in this approach. However, although not apparent in group interview 1, pupils in Novel also began to associate individual lexical items with topics, even though this had not been a feature in their approach. Being able to communicate practically (e.g., with French people) was a target for pupils in both approaches but featured slightly more often in responses in Standard. This was perceived as important by slightly older learners (aged 9-11) in Tierney and Gallastegi’s (2011) questionnaire-study and was also found in older primary-aged learners by Cable et al. (2010), Chambers (2019), Courtney (2017), Courtney et al. (2017) and Graham et al. (2014, 2017), but this study suggests that it was equally important for younger learners.

6.4.2.3 Perceptions of actual progress made

When pupils were asked about their likely future progress (6.4.2.2), there was little difference between approaches. However, when asked about *actual* progress (i.e., what they had actually learnt rather than what they felt they would be able to do), pupils’ perceptions of progress seemed more reflective of differences between approaches. Pupils in Standard referred most frequently

to progress made in topic-related vocabulary, specifying different topics. Pupils in Novel appeared to draw on progress in a broader range of aspects of learning in French (Figures 40 and 41). They commented on progress in words generally, but also in sentences (at both time-points).

Feeling able to work at sentence-level may indicate that pupils in Novel were aware of the ways in which their French knowledge was increasing. Mitchell and Myles (2019) highlighted the issues of perceptions of minimal progress among Year 3 learners having a negative impact on their self-efficacy and motivation, and both Muijs et al. (2005) and Martin (2012) called for greater linguistic challenge to avoid loss of enthusiasm and ensure sustained progress. Perceived progress is likely to impact positively on engagement, as demonstrated by Graham et al. (2016) and Mitchell and Myles (2019), in which the learners showing the highest levels of cognitive engagement were the most successful in French.

Progress in pronunciation and SSCs was also mentioned in Novel, as well as a developing understanding of French culture. Only one pupil (in Novel) mentioned stories, although these were suggested by DCSF (2005) as likely to promote pupil engagement. Traditionally it has been assumed (e.g., by Cable et al., 2010) that young learners are motivated by an oral approach. Novel included more activities in the written modality than Standard but qualitative findings did not indicate negative effects on pupils' self-efficacy by being taught through such an approach.

6.4.2.4 Spontaneous use of French

Pupils' spontaneous use of French is included as indicative of their desire to show off their knowledge, potentially reflecting their linguistic confidence; it also acts as testimony to their perceived progress. In group interviews, some pupils chose (unprompted) to spontaneously include examples of their French knowledge to illustrate their comments. In group interview 1, such examples were more numerous from pupils in Standard than Novel, but the opposite was the case by group interview 2. The lexical items volunteered by pupils in Novel in group interview 2 were generally higher-frequency and included a wider range of word classes than Standard, reflecting the inclusion of higher-frequency vocabulary in the language-driven approach.

6.4.3 Summary of self-efficacy findings

Based on tests with good levels of reliability, statistically, there appeared little difference between measures of pupil self-efficacy across the two approaches at either time-point. The same was true for individual self-efficacy items. Although pupils in Standard expressed descriptively slightly higher levels of self-efficacy for translating words to and from the FL, this was not reflected in higher ratings in these vocabulary tests. The models predicted that self-efficacy

levels were likely to increase over time for pupils in Standard, who had higher baseline levels of enthusiasm for learning French relative to Novel. This increase in self-efficacy over time was not apparent in Novel. However, when the approaches were compared, there were no statistically significant differences between them, regardless of pupils' baseline levels of positivity.

Qualitative findings indicated that pupils' awareness of progress was apparent. Across both conditions, progress was conceptualised as knowing more words, with practical communication a common goal. More pupils in Novel than Standard referred to their ability to cope with longer utterances, evidenced by progress since the outset. There were more references to pronunciation difficulties in Standard group interviews than in Novel, and, although this was in line with outcomes of the RA post-test (where Novel ratings were higher), it was not reflected in the self-efficacy ratings for Q4 (ability to read aloud French words) or Q5 (pronounce words clearly) where there were no differences. Overall, there appeared to be no association between pupils' self-efficacy ratings for the specific sub-questions and their corresponding knowledge ratings.

6.5 Pupil enjoyment: results from quantified data analysis

As part of the measurement of pupils' affective variables, their enjoyment of learning French was considered alongside their self-efficacy. Pupils (Novel: $n = 74$; Standard: $n = 72$) responded to three questions in PQ2 and PQ3: Q1: *Are you enjoying learning French this year?*; Q2: *Do you look forward to French lessons?*; Q3: *Would you like to have more French lessons?* (PQ1 provided a baseline measure by simply asking pupils if they were looking forward to learning French). Pupils selected one response from four, accompanied by smiley faces: *Yes, a lot*; *Yes, quite a lot*; *A bit*; *No* (see [Appendix G.11](#) for layout and scoring system). Additionally, PQ3 included two open-ended questions asking pupils what they had enjoyed most in Year 3 and if there were aspects of French lessons which they did not like; qualitative responses to these questions are analysed in [6.6](#).

6.5.1 Test quality

Good levels of reliability ([Table 41](#)) were indicated at both time-points. Data were not normally distributed at either PQ2 or PQ3 (links in [Table 41](#)).

Table 41.

Quality of Measure Information for Enjoyment Surveys

		Enjoyment		
		PQ1	PQ2	PQ3
<i>N</i>		146	146	146
% cases complete prior to imputation		95	95	95
Reliability (CFA)	alpha		.86	.85
	omega		.87	.86
Normality of distribution link			Appendix P.5-P.6	Appendix P.5 – P.7

6.5.2 Descriptive statistics for enjoyment ratings

For each question, responses were scored 1 (lowest) – 4 (highest).

6.5.2.1 Between-groups

(i) Pupil Questionnaire 1: establishing a baseline

Pupils were asked if they were looking forward to learning French this year, and selected a response from ‘Yes’, ‘Unsure’, ‘No’; this formed their positive anticipation baseline ([Table 42](#)). Responses from both groups were similar, with pupils in Standard slightly more positive than those in Novel.

Table 42.

Pupils’ Positive Anticipation Baseline by Approach

Approach	‘No’: pupils not looking forward to French		‘Unsure’: pupils unsure about French		‘Yes’: pupils looking forward to French		
	<i>n</i>	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
Novel	73	4	5.48	29	39.73	40	54.79
Standard	70	3	4.29	23	32.86	44	62.86

(ii) Pupil Questionnaires 2 and 3

Means in Standard were higher than in Novel at both time-points, with large effect sizes at both time-points ([Table 43](#)), although the CIs just crossed zero (though only just at PQ3). The notched box plot ([Figure 44](#), with ratings 1-4 shown in the y axis) shows that medians (the thick horizontal bar in the middle of the box, or interquartile range) were the same (= 3) across both

time-points (shown along the x axis), but ratings extended higher in Standard at PQ3. Whiskers show the maximum/minimum data points outside the interquartile range but still part of the normal range, in this case 1-4.

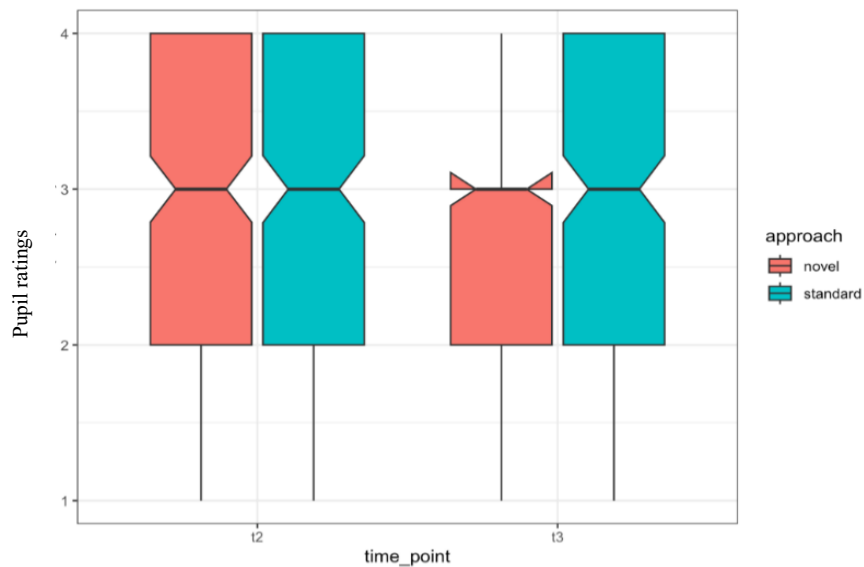
Table 43.

Descriptive Statistics and Effect Sizes for Enjoyment Ratings (Pupil Questionnaires)

Approach	Time-point	Descriptive statistics					Effect sizes		
	(<i>k</i> = 3)	<i>n</i>	<i>Mean</i> Response	<i>SD</i>	Range	95% CI	Effect size	<i>d</i>	CI
Novel	PQ2	74	2.7	.96	1.0-4.0	2.5-2.9	Within-group:		
Standard	PQ2	72	3.1	.92	1.0-4.0	2.8-3.3	Novel: PQ2-PQ3	.12	-.12-.33
Novel	PQ3	74	2.6	.90	1.0-4.0	2.4-2.8	Standard: PQ2-PQ3	.19	-.03-.44
Standard	PQ3	72	2.9	.92	1.0-4.0	2.6-3.1	Between-group:		
							Novel- Standard : PQ2	-.39	-.75-.07
							Novel- Standard : PQ3	-.32	-.67-.01

Figure 44.

Notched Box Plot Showing Pupil Enjoyment Ratings from Pupil Questionnaire Data by Approach and Time-point



Note. t2 = PQ2 ; t3 = PQ3.

6.5.2.2 Within-groups

Means for both approaches ([Table 43](#)) decreased slightly from PQ2 to PQ3. Effect sizes for within-group measures for PQ2-PQ3 were medium in both approaches, almost reaching large in Standard, although both crossed zero.

6.5.2.3 Individual item analysis

(i) *Q1: Are you enjoying learning French this year?*

At PQ2, the percentage of pupils scoring 3 or 4 points was higher in Standard (ca. 80%) than in Novel (ca. 65%), the equivalent of about 14 more pupils spread across the five classes ([Figure 45](#)), with little change by PQ3 (ca. 73% versus ca. 63%, the equivalent of about seven pupils spread across the five classes).

(ii) *Q2: Do you look forward to French lessons?*

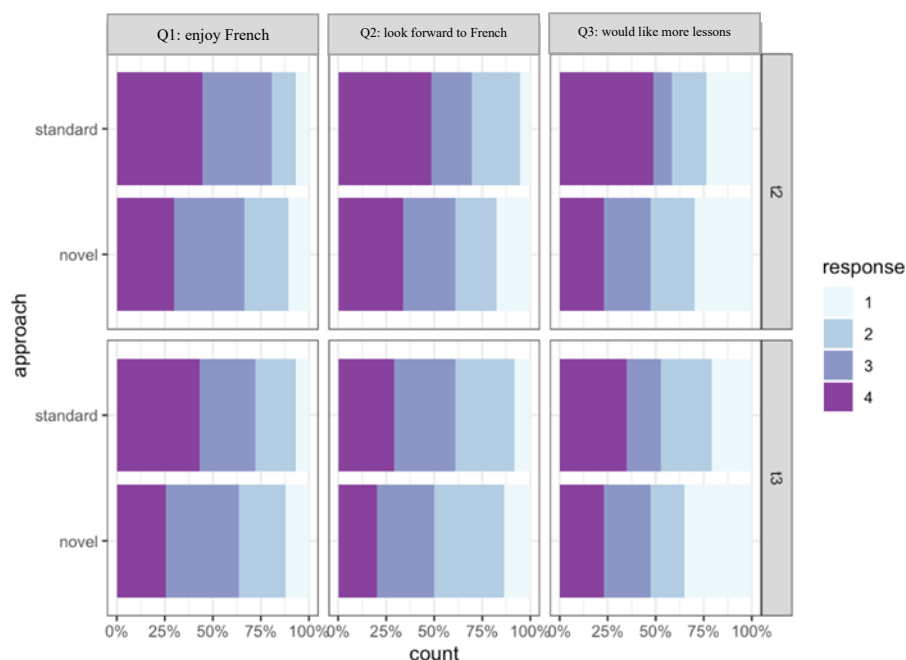
At PQ2, ca. 68% of pupils in Standard and ca. 60% in Novel responded ‘Yes, a lot’ or ‘Yes, quite a lot’. Both percentages decreased at PQ3: for Standard to ca. 60% and for Novel to 50% for ratings of 3 and 4 combined.

(iii) *Q3: Would you like to have more French lessons?*

At PQ2, nearly 50% of pupils in Standard and just under 25% in Novel scored 4 points (‘Yes, a lot’). When combining ratings of 3 (‘Yes, quite a lot’) and 4 (‘Yes, a lot’), there was less difference between approaches (Standard: ca. 60%; Novel: ca. 48%). By PQ3, the percentage in Standard responding ‘Yes, a lot’ dropped from nearly 50% to ca. 35%, whereas Novel ratings did not change. Again, at PQ3 there was little difference between approaches when combining ratings of 3 and 4 (Novel: ca. 48%; Standard: ca. 52%).

Figure 45.

Individual Enjoyment Question Analysis by Approach and Time-point



Note. t2 = PQ2 ; t3 = PQ3.

6.5.3 Analyses using mixed-effects models

A cumulative link mixed-effects model was constructed ([Appendix Q.6](#) provides code).

6.5.3.1 Regression model assumptions

In the Brant test, the Omnibus measure was $< .05$ ([Appendix Q.7](#)), with p -values for time-point and enjoyment ratings also $< .05$. Results should therefore be interpreted with caution.

6.5.3.2 Enjoyment

The model indicated a possible significant interaction between positive anticipation baseline responses ‘Yes’ and ‘No’ and between ratings of 3 and 4 ([Appendix Q.8](#) provides the model summary); these are explored below. Fixed effects (marginal R^2) explained 9% of variance in the enjoyment model; fixed and random effects (conditional R^2) accounted for 56% of variance.

Between-groups

Pupils’ enjoyment data are first considered across approaches across the two data collection points. The point range plot in [Figure 46](#) shows the likelihood of pupils scoring 1-4 (shown on the y axis) on any of the three enjoyment questions at PQ2 and PQ3, broken down by their positive anticipation baseline responses (‘No’, ‘Unsure’, ‘Yes’).

(i) *Baseline: pupils not looking forward to learning French ('No')*

The numbers in each approach (Novel: $n = 4$; Standard: $n = 3$) were too small to be able to draw any statistical conclusions (pairwise comparisons are provided in [Appendix Q.9](#)).

(ii) *Baseline: pupils unsure whether they were looking forward to learning French ('Unsure')*

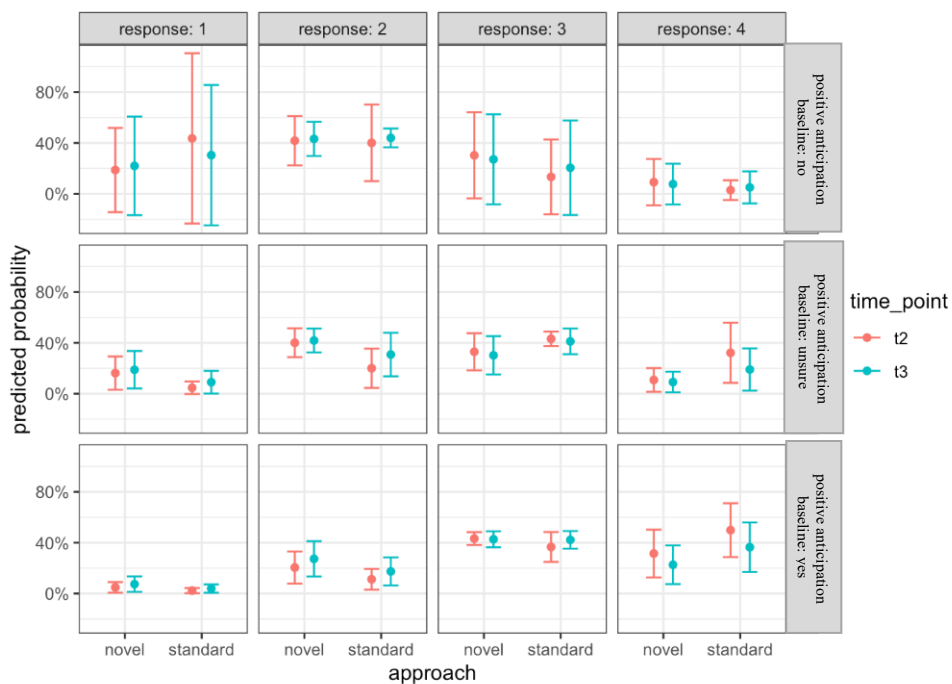
There was a statistically significant difference between approaches at PQ2 (estimate = -1.37, SE = .61, $z = -2.25$, $p = .02$), where 44% of pupils in Novel were likely to score 3 or 4 points, compared with 75% in Standard (pairwise comparison data shown in [Appendix Q.10](#)). The difference remained but was not statistically significant by PQ3 (estimate = .44, SE = 1.71, $z = .26$, $p = .80$), with 39% of these unsure pupils scoring 3 or 4 points in Novel, compared with 60% in Standard.

(iii) *Baseline: pupils looking forward to French ('Yes')*

There was no statistically significant difference between enjoyment ratings in the two approaches at either PQ2 or PQ3 (PQ2: estimate = -.77, SE = .47, $z = -1.64$, $p = .10$; PQ3: estimate = -.67, SE = .46, $z = -1.46$, $p = .15$) (for pairwise comparison data, see [Appendix Q.11](#)).

Figure 46.

Point Range Plot Showing Predicted Probability of Enjoyment Ratings on Any Item from Pupil Questionnaires by Baseline Positive Anticipation Rating, Approach and Time-point



Within-groups

Pupils' enjoyment data are now considered within approaches across the two data collection points, analysed according to baseline positive anticipation score.

(i) Baseline: pupils not looking forward to learning French at PQ1 ('No')

The numbers in each approach (Novel: $n = 4$; Standard: $n = 3$) were too small to be able to draw any statistical conclusions (see [Appendix Q.9](#) for pairwise comparisons).

(ii) Baseline: pupils unsure whether they were looking forward to learning French at PQ1 ('Unsure')

In Novel ($n = 29$), ratings of 3 or 4 decreased slightly from 44% at PQ2 to 39% at PQ3 ([Appendix Q.10](#)). This was not statistically significant (estimate = .18, SE = .29, $z = .63$, $p = .53$). In Standard ($n = 23$) 75% of pupils scored 3 or 4 points at PQ2, and 60% at PQ3. This decrease just reached statistical significance (estimate = .70, SE = .35, $z = 2.02$, $p = .04$) - ratings were higher at PQ2 and therefore more likely to regress to the mean.

(iii) Baseline: pupils looking forward to French at PQ1 ('Yes')

Enjoyment ratings were highest in this category in both approaches, with the likelihood in Standard ($n = 44$) of scoring 3 or 4 points 87% at PQ2 and 78% at PQ3, and in Novel ($n = 40$) 74% and 66% respectively ([Appendix Q.11](#)). This decrease was only statistically significant in Standard (Novel: estimate = .45, SE = .25, $z = 1.77$, $p = .08$; Standard: estimate = .55, SE = .25, $z = 2.19$, $p = .03$), where ratings were higher at PQ2 and therefore more likely to decrease to the overall mean (as per the RTM phenomenon).

6.5.4 Pupil enjoyment: summary of quantified data

Pupils' starting points for enjoyment were similar across approaches, although – unlike self-efficacy measures – Standard ratings were slightly higher than Novel. Pupils' PQ2 and PQ3 enjoyment ratings were descriptively higher in Standard than Novel, with the highest rating (4) more common in Standard than Novel. This only reached statistical significance for 'unsure' pupils at PQ2; by PQ3 there was no difference between approaches. Enjoyment decreased by PQ3 in both approaches.

6.6 Pupil enjoyment: results from qualitative data analysis

Qualitative data from PQ1, PQ3 (open-ended questions: *What things have you enjoyed the most about French lessons in Year 3? Are there things you didn't like about French lessons in Year 3?*) and group interviews (see [G.7](#) for schedule) were analysed using Dörnyei's (1994) concepts

of motivation at the *language level*, encompassing attitudes towards the target language, culture, and communities, and at the *learning situation level*, reflecting classroom experiences (3, 4.5.3). Part of the group interview approach may involve interactive discussion: it should be noted that occasionally pupils would amplify or expand on points made by their peers – something not present in (anonymous) open-ended questionnaire data. However, there were no instances of coercion or strong contradiction by pupils, and these data were not being analysed for reliability, but for the experiences they conveyed of each teaching approach in the five schools.

6.6.1 Motivation at the language level

This considers learners' comments from group interviews and PQ1 relating to the language, the country and its people.

6.6.1.1 Family matters

In both group interviews and pupils' responses in PQ1, references to family and personal contacts were positively associated with FL learning. At the start (PQ1), learners (Novel: seven pupils; Standard: three pupils) referred to relatives who had contacts with the country or spoke the language themselves and often shared this knowledge (Novel: *"I learnt French with my mum"*). In group interview 1, pupils in both approaches continued to refer positively to these contacts (Novel: five pupils; Standard: nine pupils), although roles seemed to be changing, with pupils assuming the role of experts: *"I really love French cause ... I get to learn a different language, and I can teach my mum, cause my mum always is, like, 'I don't know any French'"* (HLAA, SchB, Standard); *"I ... actually teach her [French friend] some things"* (LLAA, SchB, Novel). This seemed similarly apparent post-teaching in both approaches, as illustrated by these comments from group interview 2: *"[How do you feel about learning French?] ... good, because then we can ... tell our families new French words ... we can also teach our families French"* (LLAA, SchC, Novel); *"I taught my little sister a bit of French ... I taught her to count to 10 in French"* (LLAA, SchE, Standard).

6.6.1.2 Practical uses

In group interview 1 (29 pupils from each approach), the number of references to using French abroad was similar in both groups (Novel: eight references by seven pupils; Standard: 10 references by eight pupils): *"I really, really like it [French] because ... someday I want to go to France to try a new food, and I want to be able to talk in that language"* (MLAA, SchA, Novel); *"I just like learning French cause maybe one day if I want to go to France, erm, I'll ... be able to maybe talk to a few French people"* (HLAA, SchC, Standard). Illustrations of integrative motivation increased slightly in Novel by the end of the project. Ten references were made by

10 pupils: *“I probably think I would, like, speak something and then ... they can communicate with me ... and I can make more friends of, like, French people from France”* (MLAA, SchE, Novel). In Standard, eight pupils made such comments: *“... if I go to France ... I could speak to French people if they say something to me, I could understand them”* (HLAA, SchE).

6.6.2 Motivation at the learning situation level

This section presents positive and negative impacts of the classroom experience on motivation from pupil group interviews and PQ3 data.

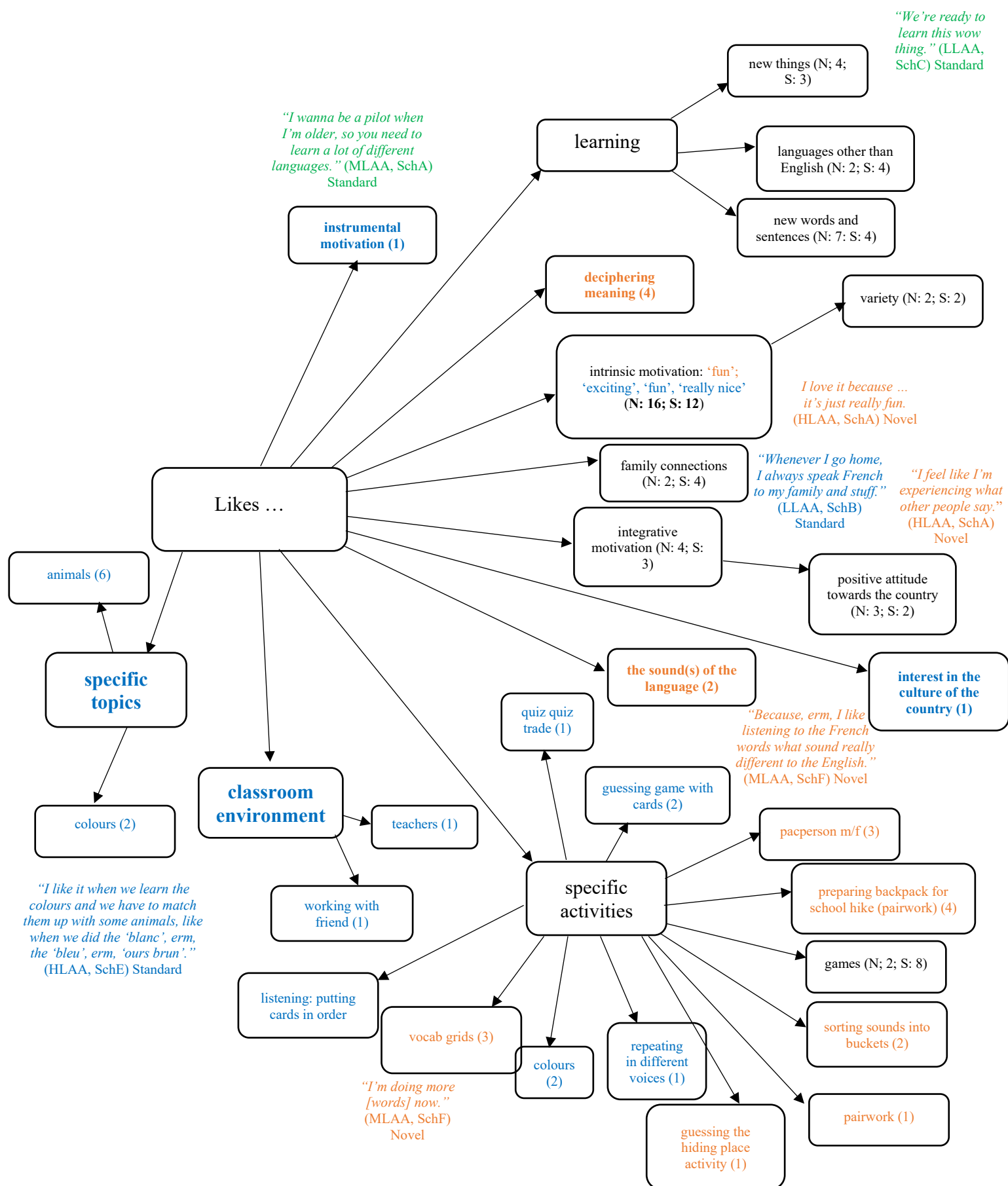
6.6.2.1 Positive impacts on motivation

(i) Pupil group interview 1

Reasons for liking French lessons were similar in both approaches ([Figure 47](#)).

Figure 47.

Pupils’ Reasons for Liking French: (Mid-point Pupil Group Interview)



Note. Text in black refers to common likes, orange uniquely to Novel, blue uniquely to Standard.

The general appeal of learning new things was apparent in comments made by pupils in Novel and Standard group interviews (PQ2 did not include qualitative data), and pupils from both approaches again demonstrated integrative motivation. Hearing the different sounds of French was cited as enjoyable by two pupils in Novel. Eight pupils in Standard referred specifically to enjoyment of topics such as animals and colours, whereas comments from the Novel group focused more on enjoyment of specific activities (e.g., vocabulary grids, SSC-related tasks, [Figure 47](#)).

(ii) Pupil group interview 2

In group interview 2, pupils in both Novel and Standard still cited intrinsic enjoyment of learning the language generally, and words specifically ([Figure 48](#)). Three pupils in Novel approach referred to the regularity of the accumulation of knowledge, with learning happening “*every time*” (MLAA, SchC) and “*every week*” (HLAA, SchF; LLAA, SchF). At PQ3, intrinsic enjoyment of language learning was more commonly referenced by pupils in Novel: out of 27 references, 18 were made by pupils in Novel and nine by Standard.

The same number of pupils (six from each approach) reported that they enjoyed “*everything*” in French lessons by group interview 2. The number of references specifically to ‘fun’ (or similar) had decreased in both approaches from group interview 1 (Standard: from 12 to three; Novel: from 16 to seven), perhaps aligning with the general trend for decreasing enjoyment ratings in both approaches seen in the descriptive statistics ([Table 43](#)) and in model predictions. This contrasted the findings from PQ3, where 33 pupils from Standard indicated that there was nothing they disliked, compared with 20 in Novel.

Considering both group interview 2 and PQ3 data, there were more references to integrative motivation in Novel than Standard: 12 participants from Novel referred to enjoying the work on Menton’s Fête du Citron, which involved pupils building sentences based on images from the festival ([Appendix O.2](#)); there was also one reference to enjoying “*finding out about French people*” (PQ3, SchA). One pupil in Standard enjoyed learning Christmas-related vocabulary. Enjoyment of sounds was mentioned by one learner in Standard approach (MLAA, SchA) in group interview 2, but - as in group interview 1 - this was slightly more prevalent in Novel groups.

A further reflection of the different foci in each approach was evident in activities enjoyed by pupils, as mentioned in group interview 2 and PQ3. For example, pupils in Standard appeared to refer more often to activities involving ‘doing’, for example, making a paper butterfly following French instructions (eight pupils in PQ3, 10 in group interview 2), whereas pupils’ responses in Novel related to skills, for example, translating. Pupils in Standard were more likely than learners in Novel to cite games as their favourite parts of French lessons.

References to the vocabulary grids that featured (only) in Novel increased from three in group interview 1 to 18 by the end, with 16 pupils (from all five schools) describing them as particularly enjoyable in group interviews, and 10 in PQ3.

Well, I really do like the vocabulary sheets, they're really fun; once I got the highest score you can get. Erm, they helped, like, bring back words to my memory, like: 'I need to know this word! Oh, yes, now I remember'. (HLAA, SchE)

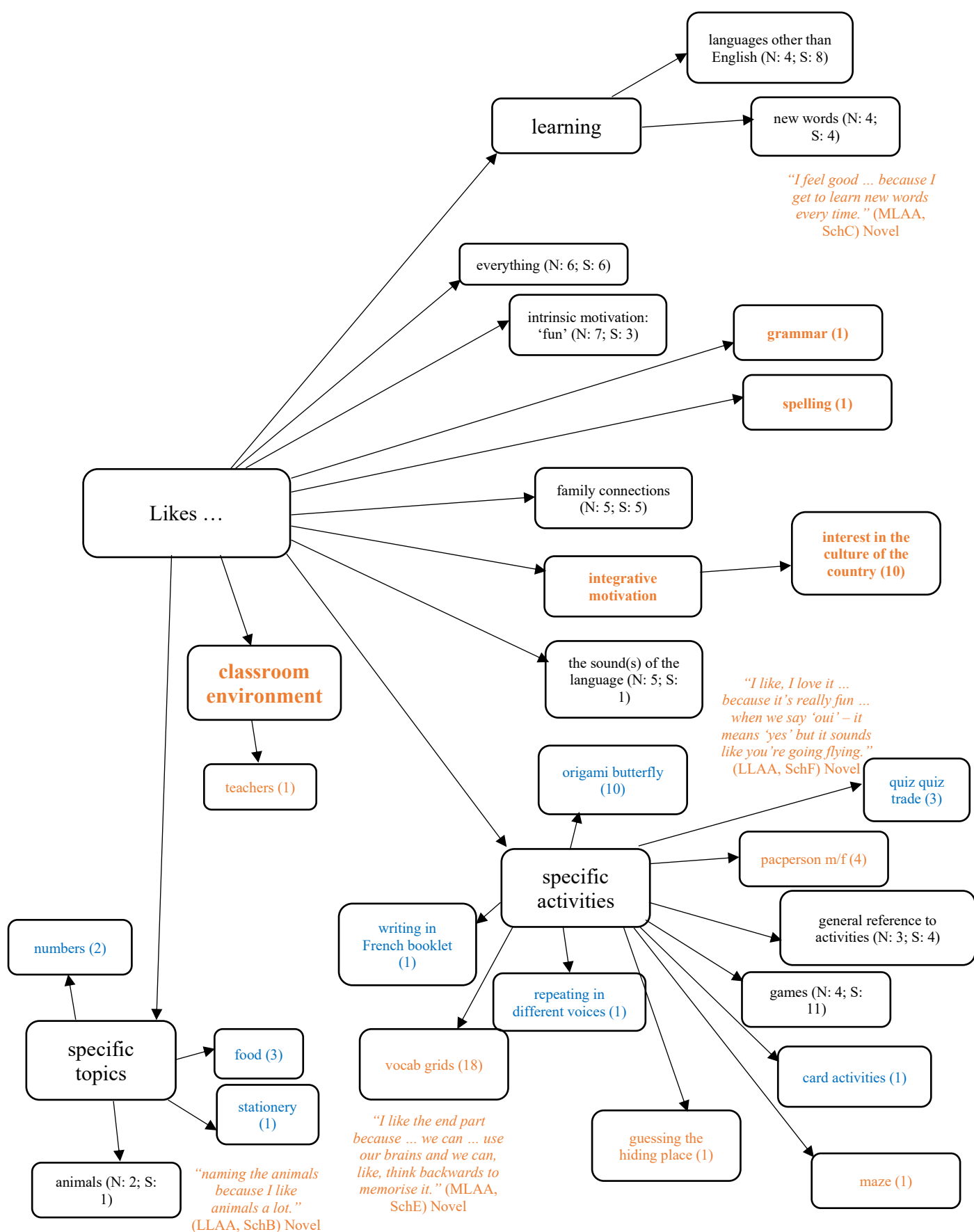
These pupils were not necessarily those with highest LAA: of the 16 citing the vocabulary grids in group interviews, three had high LAA ratings, seven had medium, and six had low. Participants' growing enjoyment of the vocabulary grid recall activities (see [Appendix O.1](#) for examples from and into English) was equally apparent in observed and recorded lessons, where pupils' responses to the grids became increasingly positive as the year progressed ([Appendix R.2](#)).

Participants in Standard (PQ3: 35 pupils; group interview 2: five) still referred to their enjoyment of concrete, tangible vocabulary taught through topics (animals, food, stationery, colours, days of the week) at the end of the study, with this occurring more often than in Novel (PQ3: no pupils; group interview 2: two).

References to teachers were relatively uncommon in both group interviews, with only one pupil in Standard (group interview 1) and one in Novel (group interview 2) citing their teacher as a particular reason for liking French. However, responses to the anonymous PQ3 *did* refer to teachers as an aspect of French lessons that pupils particularly liked (five pupils, all Novel). No pupils in either approach mentioned particularly enjoying songs, and only one in each approach referred to enjoying stories (in PQ3), although these featured in both schemes of work.

Figure 48.

Pupils' Reasons for Liking French: (End-point Pupil Group Interview)



Note. Text in black refers to common likes, orange uniquely to Novel, blue uniquely to Standard.

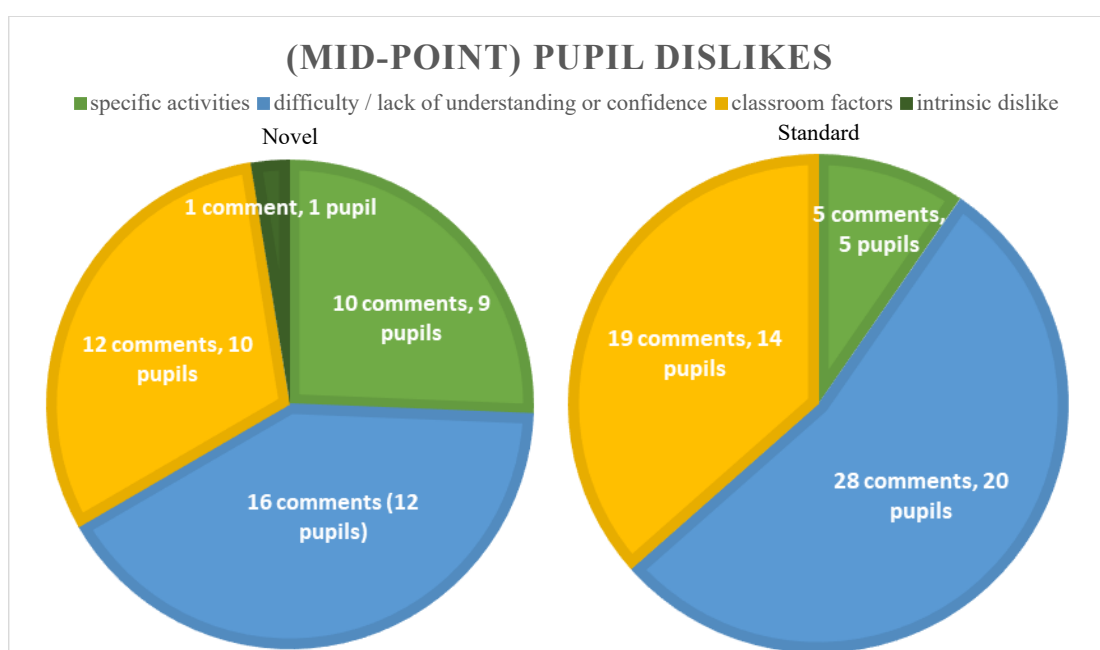
6.6.2.2 Negative impacts on motivation

(i) Mid-point: pupil group interview 1

Reasons given in group interview 1 for dislikes in each approach centred around four key areas (Figure 49): (a) dislike of content/specific activities; (b) difficulty or lack of understanding (also an a priori theme); (c) intrinsic dislike; (d) classroom factors (linking to the a priori theme of classroom experience). PQ2 did not include qualitative data.

Figure 49.

Negative Impacts on Pupil Motivation: (Mid-point – Pupil Group Interview 1)



(a) Dislike of content or activities

Pupils from both approaches expressed a dislike of learning French, with slightly more referring to specific activities in Novel (10 comments by nine pupils) than Standard (five comments by five pupils). In group interview 1 in Standard, these related to three specific activities (writing, saying words and pairwork), whereas in Novel, these referred to: vocabulary grid quizzes (four) (*“the tests I think are a little bit annoying”*, LLAA, SchE); oral repetition in class (one); writing (four) (*“it hurts my hand”*, LLAA, SchC); and songs (one).

(b) Perceived difficulty/lack of understanding

Forty-one per cent of the comments made in Novel (16 comments across 12 pupils) related to perceived difficulty or lack of understanding: *“Sometimes ... the words are really tricky ... sometimes when they explain, you don’t really get it”* (LLAA, SchB). This proportion was higher

in Standard, with 28 comments made by 20 pupils (51% of all comments relating to dislikes): *“When I find the questions hard ... I kind of get upset”* (LLAA, SchB).

(c) Intrinsic dislike

Only one pupil in Novel referred to an intrinsic dislike (*“It’s not very fun”*, MLAA, SchB) and none in Standard.

(d) Classroom factors

According to group interview 1 data, similar numbers of pupils in each approach (Novel: 10; Standard: 14) disliked aspects of classroom experience: 19 comments in Standard and 12 in Novel referred to factors impacting on their enjoyment of French. Ten comments in Novel related to classroom noise, and 14 in Standard (from three schools): *“I don’t like it when people start ... talking really loudly, cause I have, like, a sensitive, erm, ear”* (LLAA, SchE, Standard). In one school (Standard), this appeared to stem from overexcitement: *“Sometimes we play this game Bingo and then we would have to shout out and sometimes ... it would be nice to be a little bit quiet”* (LLAA, SchF).

In both approaches, isolated comments mentioned lesson disruption caused by poor behaviour, with a pupil (Novel) explaining that language then needed to be recapped ([Appendix R.1](#)). The other comments in Standard approach centred on being chosen to answer questions when unable to answer (*“when you don’t even want to get chosen”*, MLAA, SchC) or conversely, the teacher not asking them when they did know (*“she just picks the ... exact same person”*, HLAA, SchC).

(ii) End-point: pupil group interview 2 and Pupil Questionnaire 3

End-point data were collected from group interview 2 and through open-ended questions in PQ3. Responses relating to *content or activities* were similar across approaches, whereas differences emerged in the following area (in group interviews, PQ3 responses, or both): *perceived difficulty*; *intrinsic dislike*; and *classroom factors*.

(a) Dislike of content or activities

By the end of the study, the number of pupils in each approach disliking specific activities or content had increased since group interview 1 – particularly in Standard – and was now similar in both approaches, in both group interviews and PQ3 ([Figures 50](#) and [51](#)). In Novel, five of the 34 comments across both data sources referred to specific activities related to timed-condition activities (*“on the board they’re [the words] English and you’ve got to say the French words before the timer goes off”*, MLAA, SchA), where pupils had 60 seconds in which to translate

words into/from the FL, and six mentioned the vocabulary grids (similar to group interview 1). In Standard, nine of the 29 comments referred to writing (*“it was quite hard to write the sentences in French ... I liked the speaking better”*, MLAA, SchE), which was mentioned by only three pupils in Novel.

Whereas three pupils in Novel referred to a dislike of learning words generally and only one to a specific category of vocabulary (food), pupils in Standard were more specific: three disliked animals, one stationery items, one days of week and three food (with just one referring to disliking *“learning lots of words”*).

(b) Perceived difficulty/lack of understanding

In group interview 2 (29 learners in each approach), more learners in Standard (26) than in Novel (17) reported that they found aspects of the work challenging, or *“tricky”*. This was an increase from group interview 1 (Standard: 20; Novel: 12) and was again conceptualised by comments relating to difficulties encountered, or forgetting easily, and general lack of confidence. PQ3 responses, completed anonymously online, showed a similar number of pupils in each approach (Novel: 10; Standard: 11) referring to difficulty.

(c) Intrinsic dislike

At the end-point, the number of pupils in Novel expressing intrinsic dislike of aspects of French had increased from one (SchA) to five in group interviews (SchA: three; SchB: one; SchC: one) and five in PQ3, compared with none in Standard in group interview 1 increasing to two in group interview 2 (one each in SchA and SchC) and one in PQ3. In Novel, four pupils described aspects of learning French as *“boring”* in group interviews, compared with two in Standard, with two explaining that this was due to a dislike of revisiting content (*“I didn’t really like just going back and learning the same words that we’ve already learnt”*, MLAA, SchE, Novel).

Figure 50.

Negative Impacts on Pupil Motivation: (End-Point – Pupil Group Interview 2)

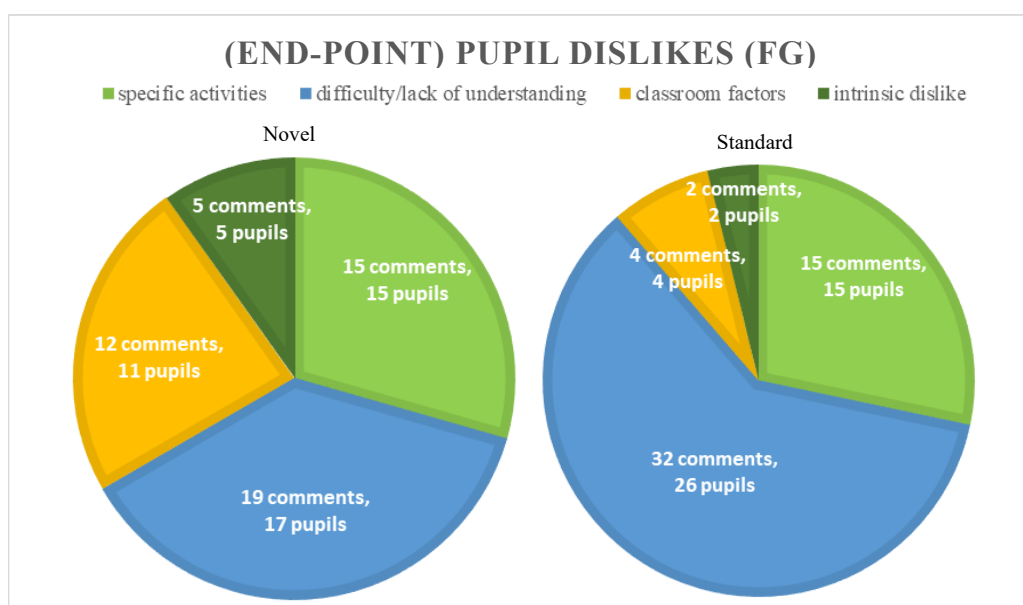
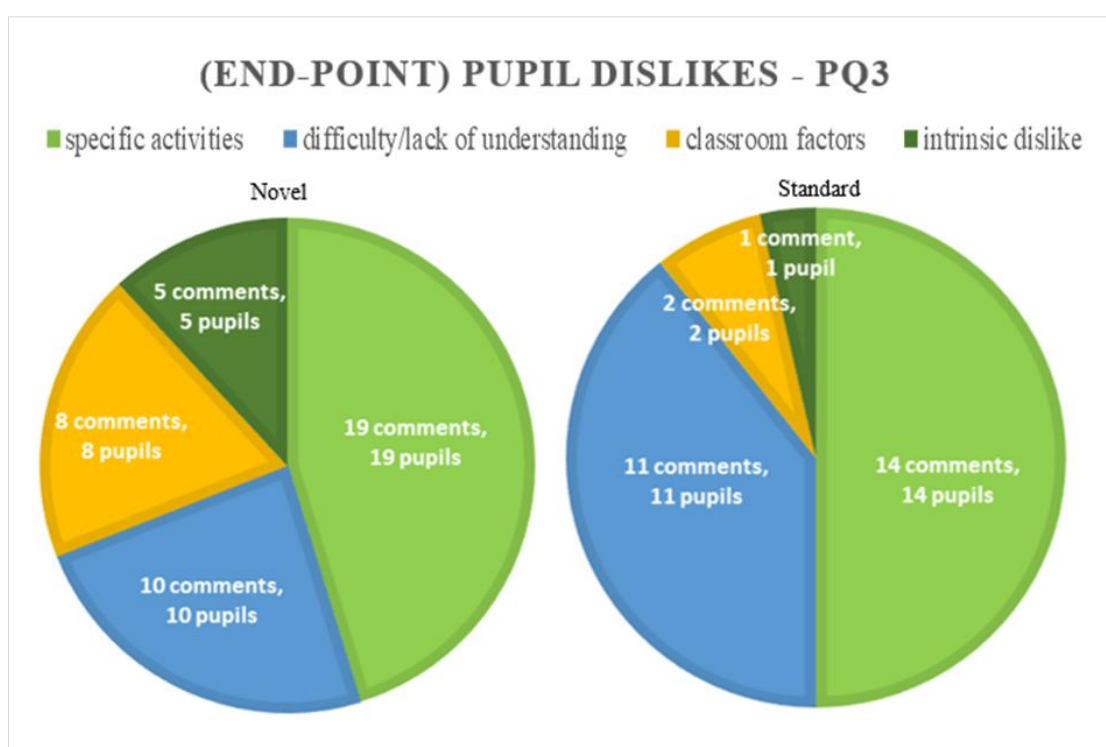


Figure 51.

Negative Impacts on Pupil Motivation: (End-Point – Pupil Questionnaire 3)



(d) Classroom factors

Classroom experiences also impacted on pupils' enjoyment of French in group interview 2, cited in both approaches as reasons for disliking French lessons. Eight pupils in Novel reported experiencing fatigue or complained about lesson length (SchA: six; SchE: two), compared with

two in Standard (both SchA): *“When you are very tired and it’s stuff like French, you might feel a bit angry”* (LLAA, SchA, Novel). One pupil in Standard did complain about the length, but because they felt lessons were too short. Two responses to PQ3 made by pupils in Standard indicated that they disliked the length of French lessons (*“it took a long time, they were long lessons”*), and this was mentioned by five pupils in Novel (*“it felt so long!!!!”*). In both approaches, lessons were scheduled to last 45 minutes, and this happened regularly ([4.4.5.1](#)).

These findings are also supported by PQ3 responses which indicated a greater importance accorded to classroom experience in Novel than Standard in the factors contributing to pupils’ enjoyment. These included references to the following: being with peers (one pupil: *“Because I have my friends”*), *“working with a partner”* (five pupils) or positive comments (from six pupils) about teachers (*“a fabulous teacher”*).

6.6.3 Pupil enjoyment: summary of qualitative data

In both approaches pupils expressed enjoyment of contact with the country, French people and the language. Examples of integrative motivation at this language level (Dörnyei, 1994) were slightly more frequent in Novel than Standard. Pupils in Standard referred more often than those in Novel to learning topic-related vocabulary, whereas an enjoyment of sounds was expressed slightly more frequently in Novel. Expressions of enjoyment of lessons based on content such as stories and songs or related to the teacher were absent or uncommon.

In terms of dislikes, reference to specific activities was more common in Novel at the mid-point of the study but featured in responses from pupils in both approaches by the end, with pupils in Standard more likely to express dislike of writing activities than in Novel. Pupils in Standard appeared more likely than in Novel to refer to dislike of French caused by difficulties or lack of understanding in group interview 1 and group interview 2, although in PQ3 this was similar across approaches. There was little expression of intrinsic dislike at group interview 1, but this was slightly more common in Novel than Standard by the end of the year; perceptions of longer lesson length were also more apparent in Novel than Standard.

6.7 Pupil enjoyment: discussion of quantified and qualitative data analysis

The results presented in [6.5](#) and [6.6](#) pertaining to pupil enjoyment are now discussed.

6.7.1 Pupil enjoyment: quantitative trends

Overall, there appeared to be a positive outlook initially and throughout the year, with considerable individual variation. Pupils in Standard looked forward to learning French slightly more than pupils in Novel at PQ1. Assuming that ratings of 3 (‘quite a lot’) and 4 (‘a lot’)

indicated positivity, questionnaire responses suggested that pupils in both approaches largely enjoyed French during the year (PQ2: overall: $M = 2.86$, $SD = 1.10$; Novel: $M = 2.68$, $SD = .96$; Standard: $M = 3.05$, $SD = .92$; PQ3: overall: $M = 2.69$, $SD = 1.09$; Novel: $M = 2.56$, $SD = .90$; Standard: $M = 2.86$, $SD = .92$), supporting earlier findings in this context that primary learners respond positively to early language learning (Cable et al., 2010; Muijs et al., 2005; OFSTED, 2008a). However, high standard deviations in both approaches indicated variation in individual responses. This broadly aligns with findings from the ELLiE project which found differences in attitudes and motivational levels from the outset (Enever, 2011), and also Courtney et al. (2017), in which 19% of learners at the end of Year 6 (aged 11) indicated that they did not enjoy learning French. In this current study, the descriptive statistics (6.5.2) suggested that the more positive the pupils' baseline anticipation of learning, the more likely they were to score the highest levels of enjoyment at PQ2 and PQ3.

There were minimal or no differences in enjoyment between the different approaches, with both decreasing over time. Ratings were higher for pupils in Standard than Novel for enjoyment at PQ2, where the model indicated statistically significant differences between approaches for pupils unsure about starting French. This suggests that these initially unsure pupils tended to respond more positively to the approach adopted in the Standard scheme than Novel in the first half of the project, but that this decreased in the second. Indeed, in both approaches, ratings for enjoyment decreased at the end of the study among pupils who had initially indicated that they were unsure about learning French or were looking forward to it, perhaps indicating that the novelty of learning a new language had begun to wane (see Mihaljević Djigunović & Lopriore, 2011). Other studies have also reported similar declines: a decrease was found in attitudinal levels ($M = 2.97$) in Graham et al. (2016), although not until the end of Year 7 (pupils aged 11-12), after means of 2.61 in Year 6 and 3.10 at the start of Year 7. In Courtney's (2017) study on transition in FL, 77% of learners reported enjoying French in Year 6 and this decreased to 68% at the start of Year 7, but data from their FL learning experiences *prior* to Year 6 were not part of the study. Pupils of a similar age to participants in this study (7-8-year-olds) completed surveys after one and two years of learning English as a FL in Denmark (Fenyvesi, 2020), and the percentage of pupils rating English as their first or second favourite subject statistically significantly decreased from the end of the first year of learning to the end of the second. Similarly, Cable et al.'s (2010) longitudinal study reported decreases in the percentage of learners reporting FL enjoyment in 2006-7 from 81.6% in Year 3 ($n = 136$) to 58.5% in Year 6 ($n = 142$), and in 2007-8 from 86.7% ($n = 144$) to 69.2% ($n = 133$). In the current study, this decrease in ratings between PQ2 and PQ3 was statistically significant only in Standard, possibly because the PQ2 ratings were higher in this approach than in Novel. By the end of the study, there was no statistically significant difference between approaches.

In sum, the only statistically significant difference between approaches was at PQ2, when those pupils in Standard unsure about learning French at the outset of the study (i.e., $n = 23$, 32.86% of Standard) scored higher on enjoyment measures than those in Novel ($n = 29$, 32.86% of Novel) at the same time-point. This difference was not apparent by the end of the study.

6.7.2 Pupil enjoyment: qualitative trends

Qualitative data presented in [6.6](#) are now discussed with reference to existing research.

6.7.2.1 Motivation at the language level

(i) Family matters

In both group interviews and responses in PQ1, pupils' references (in both approaches) to family members and personal contacts appeared positively associated with FL learning, seeming to reinforce the importance of the language learning 'milieu' represented by family support and friends' attitudes identified by Csizér and Dörnyei (2005).

(ii) Practical uses

During the coding process, the researcher reflected on whether pupils from either approach would envision opportunities to apply the language they had learnt, and whether this might reflect its perceived usefulness ([Appendix L.10](#)). Young learners in this project, regardless of approach, clearly saw a practical aim for learning French, identifying its potential usefulness in travel abroad. Martin (2012) also reported that pupils demonstrated instrumental motivation in their references to the usefulness of FL for travel and careers, as did Cable et al. (2010), Courtney (2017), and Graham et al. (2014, 2017). Similarly, pupils in both Novel and Standard referred to their desire to be able to communicate with native French speakers. Such integrative motivation was also shown by slightly older primary-aged learners through a desire to communicate with others in the FL (Tierney & Gallastegi, 2011), and was also found by Cable et al. (2010), Chambers (2019), Courtney (2017), Courtney et al. (2017) and Graham et al. (2014, 2017).

6.7.2.2 Motivation at the learning situation level

(i) Positive impacts on motivation

Positive attitudes towards or enjoyment of a task may support the development of intrinsic motivation, potentially creating enjoyment (Deci & Ryan, 1985). Reasons given for liking French were very similar in both approaches in group interview 1 ([Figure 47](#)). The general appeal of learning new things was apparent in comments made by pupils in both Novel and Standard

(echoing suggestions by Martin, 2012), and pupils from both approaches demonstrated integrative motivation.

In group interview 2, responses from pupils in Novel and Standard citing intrinsic enjoyment of learning the language generally, and words specifically (Figure 48) were similar, which is important given the positive correlation between intrinsic motivation and quality of learning (Deci & Ryan, 1985); this reflects findings of Graham et al. (2014), Muijs et al. (2005) and Myles (2017). In PQ3, intrinsic enjoyment of language learning was more commonly referenced by pupils in Novel: out of 27 references to this theme in PQ3, 18 were made by pupils in Novel and nine by Standard. Three pupils in Novel referred to the regularity of the accumulation of knowledge.

Overall, integrative motivation appeared slightly more evident in Novel, with 12 participants referring to enjoying the work on Menton's Fête du Citron either in group interview 2 or at PQ3. It is possible that the recency of the focus on this festival impacted learners – the Standard approach included cultural input on Christmas traditions in France, but it came earlier in the year (one pupil referred at both time-points to enjoying learning Christmas vocabulary and *“about what they eat and stuff for Christmas”*, LLAA, SchA). Culture also explained enjoyment of FL learning by primary-aged pupils in previous studies (Graham et al., 2014; Martin, 2012; Muijs et al., 2005; Myles, 2017).

Hearing the different sounds of French was cited as enjoyable by two pupils in Novel in group interview 1 (none in Standard), perhaps reflecting the growing awareness of differences and similarities between languages also apparent in Martin (2012). One pupil in Standard (MLAA, SchA) mentioned their enjoyment of sounds in group interview 2, but - as in group interview 1 - this was slightly more prevalent in Novel groups (five pupils). Pupils' enjoyment of getting to grips with different sound patterns perhaps reflected the greater emphasis in Novel on SSCs: *“I really like the word ‘chien’ ... it's a really fun word – I don't know why”* (MLAA, SchE).

More pupils in Standard than Novel mentioned games as an aspect of French lessons that they enjoyed, echoing suggestions that these promote engagement (DCSF, 2005); this reflects observations made by Cable et al. (2010), Courtney (2017) and Martin (2012). In group interview 1, only three pupils made positive references to the vocabulary grids in Novel, but this increased to 18 pupils of varying abilities by the end, across both group interview 2 and PQ3. This perhaps illustrates pupil recognition of the important contribution made by retrieval practice to learning, exemplifying desirable difficulty (Suzuki et al., 2019b) and possibly indicating an early awareness of progress. In contrast with comments made by participants in earlier research (Cable et al., 2010; Courtney, 2017; Martin, 2012) and with suggestions made by DCSF (2005), no pupils in either approach referred to enjoying songs. Indeed, one pupil cited this as a dislike

(6.6.2.2, a). Contrasting findings from the aforementioned research, only one pupil in each approach mentioned enjoying stories (at PQ3).

As found by Martin (2012), participants in Standard (PQ3: 35 pupils; group interviews: five) still referred to their enjoyment of concrete, tangible vocabulary taught through topics (animals, food, stationery, colours, days of the week) at the end of the study, with this occurring more often than in Novel (PQ3: no pupils; group interviews: two). For young learners, some age-appropriate topic-related vocabulary may be particularly important when it stimulates a positive response, and S. Webb and Nation (2017) suggested that animal- or environment-related words may be appropriate, even if lower-frequency. In Novel, one pupil commented on how much they liked learning the (higher-frequency) animal names, “*cause I really like animals a lot*” (MLAA, SchB). However, the topic was not universally popular: three pupils in Standard referred to disliking this vocabulary, illustrating the danger of trying to predict content appealing to all young learners.

Pupils’ references to teachers in both group interviews and PQ3 were scarce, with one pupil from each approach in group interview 2 and six at PQ3 (in Novel) citing their teacher as a particular reason for liking French. This appears to contradict the illustrations of affiliative motivation common in previous studies (Cable et al., 2010; Martin, 2012; Mihaljević Djigunović & Vilke, 2000; Muñoz, 2014a; Nikolov, 1999) and the importance accorded to the teacher by 8-year-old learners in a Danish study (Fenyvesi, 2020; Fenyvesi et al., 2020).

(ii) Negative impacts on motivation

(a) Dislike of content or activities

By the end of the study, more pupils in Standard referred to disliking writing than in Novel. Previous findings have indicated that writing is less positively perceived by young learners (Mihaljević Djigunović & Lopriore, 2011; Nikolov, 1999), considered more demanding (Courtney, 2017) or less appealing (Chambers, 2019). The greater emphasis placed on phonics in Novel possibly increased these pupils’ confidence in writing, compared with Standard: “*I don’t like trying to write the French words down. I don’t really know how to write them*” (HLAA, SchB, Standard). Indeed, facilitating written expression should be an aim of phonics instruction rather than it being viewed as “an end in itself” (Woore, 2022:182).

(b) Perceived difficulty/lack of understanding

The number of pupils disliking French because of perceived difficulties or lack of understanding increased from group interview 1 to 2, in both approaches. However, in both group interviews, pupils in Standard referred to experiencing negative challenges in learning French more often than learners in Novel, with the difference in the number of comments between approaches

increasing from group interview 1 (Novel: 16 comments made by 12 pupils; Standard: 28 comments made by 20 pupils) to group interview 2 (Novel: 19 comments made by 17 pupils; Standard: 32 comments made by 26 pupils). It is possible that pupils' confidence benefited from the staged repetition featuring in Novel which perhaps made them more aware of their progress. However, this difference was less apparent in PQ3 responses, where a similar number of pupils from each approach (Novel: 10; Standard: 11) made a similar number of comments (Novel: 10; Standard: 11). PQ3 was completed anonymously online, whereas group interviews encouraged interactive discussion, which led to pupils picking up on, elaborating and extending peers' comments (see discussion in [6.6](#)). Additionally, these young learners were more at ease with speaking than writing/typing, and their responses in PQs tended to be very short, often one-word answers, whereas group interviews encouraged more detailed reflection.

(c) Intrinsic dislike

Intrinsic dislike was expressed slightly more often by Novel pupils than Standard in both group interview 2 and PQ3. This round of data collection took place immediately after the final lesson of the teaching input, shortly after pupils in Standard had experienced a practical lesson following French instructions to make a paper butterfly. Teachers commented that this was positively received by pupils, and this possibly led to a degree of envy on the part of learners in Novel. Equally, there may have been some awareness of the higher number of games included in Standard than Novel. Indeed, two pupils' comments from Novel indicated that they perceived a general lack of fun or fewer fun activities in Novel than Standard (*"groop s [sic] got more fun things"*, PQ3).

(d) Classroom factors

The greater awareness of lesson length in Novel was possibly due to pupils' perception (expressed in group interviews) that some elements of lesson content were uninteresting, as six pupils from three schools in Novel group interviews commented that they *"sometimes get bored"* (example from LLAA, SchA), whereas only one pupil in Standard referred to lack of excitement. One pupil in Standard commented that it was boring at the start when they *"learnt four words"* (MLAA, SchC). Although these interviews were conducted at the end of the summer term during a period of very warm weather, when pupils were tired and anticipating the holidays, this would have affected all pupils equally.

PQ3 responses also indicated that classroom factors appeared to impact pupils in Novel to a greater extent than in Standard. This is salient given the importance of a positive classroom experience deemed to play a pivotal role in sustaining motivation in primary-secondary

transition (Graham et al., 2016) and the absence of which was found to contribute to a dislike of FL learning (Tragant, 2006).

Previous research with primary-aged learners identified that pupils can express negative attitudes towards repetition of content (Cable et al., 2010; Fenyvesi, 2020; Graham et al., 2016), but in this study, revisiting previously learnt material was only seen as demotivating by three pupils, two from Novel (*“really loads of the times we’re learning things what we’ve already learned in French”*, MLAA, SchB) and one from Standard in group interviews, and one from Novel in PQ3, commenting on what they had disliked in the project: *“learning the same french [sic] words”*. Similarly, choral repetition was perceived to contribute to pupils’ dislike of FL learning by both Courtney’s (2017) and Martin’s (2012) group interview data but was only identified by two pupils (Novel) in this study, although it is possible that, as these learners were in their first year of learning, choral repetition was still appealing.

6.7.3 Summary of enjoyment findings

Based on tests showing good reliability, descriptive statistics showed that pupils’ enjoyment ratings were higher in Standard than Novel at both PQ2 and PQ3 (and across all three items), with large (but not secure) effect sizes at both time-points. There were too few learners not looking forward to learning French (‘No’) to conduct statistical analyses. According to descriptive statistics, those whose baseline was unsure or more positive enjoyed French more in Standard, and this reached statistical significance in the model for unsure pupils at PQ2, but not PQ3. In all categories, pupils were more positive in the first half of the project than the second, with the enjoyment ratings of pupils who were unsure or positive about learning French decreasing by PQ3. The decreases in Standard at PQ2 and PQ3 (for unsure and positive pupils) reached statistical significance, but these were higher at PQ2 than Novel and therefore more likely to decrease to the mean. Overall, the most positive pupils at PQ1 were most likely to enjoy learning French, regardless of approach.

In both approaches, pupils’ qualitative responses acknowledged the importance of family connections with the language and recognised the value of French for practical communication. Intrinsic enjoyment of language learning was apparent in both approaches, although slightly more evident in Novel at PQ3. By PQ3 there appeared to be fewer references to total enjoyment of French and to “fun” in the open-ended questions, across both groups, which supports the decrease in enjoyment ratings seen in PQs. Examples of integrative motivation based on French culture were slightly more apparent in Novel, as was enjoyment of the sounds of the language, based on comments in group interviews and PQ3.

Mid-study, pupils in Novel cited specific activities as dislikes more often than learners in Standard, although the difference between the approaches had decreased by the end. At both

time-points, pupils in Standard were more likely to refer to difficulties encountered (although only in group interviews, not in PQs), whereas pupils in Novel appeared to be more affected by fatigue and lesson length, referring more often to lack of interest than those in Standard.

6.8 Pupil perceptions of teacher enjoyment: results from quantified and qualitative data analysis

Pupils were asked in PQ2, PQ3 and group interviews whether they felt that teachers enjoyed teaching them French and whether they tried to make them excited about learning the language (4.4.8). The quantified and qualitative data are presented below and discussed in 6.8.3.

6.8.1 Pupil perceptions of teacher enjoyment: results from quantified data analysis

Pupils (Novel: $n = 74$, Standard: $n = 72$) rated their perceptions of teacher enjoyment on a scale of 1 (lowest) – 4 (highest). Responses to the two questions (Q1: “Do you think your teachers like teaching you French?”, Q2: “Do they try to make you excited to learn French?”) were seen as two slightly different constructs rather than a scale. Data were not normally distributed (Appendix S.1-S.6).

6.8.1.1 Descriptive statistics

(i) Between-groups

Pupil perceptions of teacher enjoyment as measured by both items (Q1, Q2) were similar across both approaches, with median ratings the same in both approaches at PQ2, and slightly higher in Standard than in Novel at PQ3 (Table 44).

(ii) Within-groups

Median ratings remained the same in Standard across both time-points for both items but decreased slightly from PQ2 to PQ3 in Novel in both Q1 and Q2 (Table 44).

Table 44.

Descriptive Statistics for Pupil Perceptions of Teacher Enjoyment Question1 and Question2

Approach	Time-point ($k = 7$)	Descriptive statistics				
		n	Q1: Median Response	Q1: Range	Q2: Median Response	Q2: Range
Novel	PQ2	74	4.0	1.0-4.0	4.0	1.0-4.0
Standard	PQ2	72	4.0	1.0-4.0	4.0	1.0-4.0
Novel	PQ3	74	3.3	1.0-4.0	3.3	1.0-4.0
Standard	PQ3	72	4.0	1.0-4.0	4.0	1.0-4.0

6.8.1.2 Inferential statistics

Non-parametric unmatched two-samples Wilcoxon rank sum tests with continuity correction were performed to compare pupil perceptions of teachers' enjoyment of teaching each approach (Table 45). This suggested no statistically significant difference between pupils' perceptions of teacher enjoyment on Q1 or Q2 at either time-point.

Table 45.

Comparison of Pupil Perceptions of Teacher Enjoyment at Pupil Questionnaire 2 and Pupil Questionnaire 3

Approach	Inferential statistics			
	Item	Time-point	Wilcoxon W	p-value
Novel-Standard	Q1	PQ2	2592.50	.87
Novel-Standard	Q2	PQ2	2335.00	.19
Novel-Standard	Q1	PQ3	2417.00	.36
Novel-Standard	Q2	PQ3	2249.00	.10

6.8.2 Pupil perceptions of teacher enjoyment: results from qualitative data analysis

There was little difference between approaches in pupils' comments in group interview 1 and 2 regarding teacher enjoyment. In group interview 2, comments in both approaches indicated that they generally thought their teachers enjoyed teaching them French: 19 pupils in Novel expressed positive opinions, with five negative and six unsure; 19 were positive in Standard, three negative and eight unsure. Pupil comments appeared to relate more to general teaching than French and, as they were not approach-specific, they are summarised briefly below.

- Pupils from both approaches expected teachers to enjoy teaching French: *"I don't really think they won't enjoy it though"* (MLAA, SchE, Standard, group interview 2).
- They referred to teachers' patient, helpful demeanour (*"They like teaching and helping you"*, MLAA, SchF, Novel, group interview 2).
- Teachers were perceived as knowledgeable - *"I think she knows a lot of French and I think she wants to share it with us"* (MLAA, SchC, Novel, group interview 2) - although pupils were clearly aware that they were learning alongside the teachers in some cases

(equally mentioned in both approaches): “... *they’ll learn it themselves as well as us*” (HLAA, SchE, Standard, group interview 2).

- Pupils imagined that teachers were positively impacted by pupils’ success (“*French is ... her favourite thing and seeing other people actually understand it, that makes her happy*”, HLAA, SchB, Standard, group interview 2) but equally affected by their failures, with one commenting that the teacher did not enjoy teaching French “*because I forget everything*” (HLAA, SchF, Standard, group interview 1).
- Pupils in both approaches identified negative impacts of poor behaviour and a difficult learning environment on teachers’ enjoyment.
- Generally, pupils from both approaches appreciated the effort invested by teachers to ensure their enjoyment of French and felt that this then impacted positively on teachers’ experiences, leading to a perception of infectious excitement: “... *they’re [the teachers] excited and also they’re trying to make me excited all at the same time*” (MLAA, SchC, Novel, group interview 2).

6.8.3 Pupil perceptions of teacher enjoyment: discussion of quantified and qualitative data

Quantitative and qualitative findings are now discussed.

6.8.3.1 Perceptions of teacher enjoyment: quantitative trends

Although descriptively learners perceived very slightly higher levels of enjoyment in Standard teaching, generally teachers appeared to pupils to be similarly enjoying teaching French through the two approaches, and they appeared to make similar efforts to make learners excited about learning French. As discussed in [7.8.2.3](#), some teachers preferred one approach, but this was not overtly conveyed to pupils.

6.8.3.2 Perceptions of teacher enjoyment: qualitative trends

In this study, the same teachers taught both approaches, which should have reduced potential ‘teacher effect’. However, it was impossible to eliminate the teacher variable completely, especially when teachers were expected to innovate with one of the approaches, but the other was more familiar; teachers’ preferences for one approach over the other may have inadvertently impacted on learners and influenced some of the group interview/PQ3 comments.

Teachers were seen as striving to provide a positive classroom environment which encouraged pupils to enjoy French. This reflects expectations of teachers in the Croatian primary FL experiment, when teachers were to ensure a “happy and relaxed” learning atmosphere

(Mihaljević Djigunović & Vilke, 2000:68). Pupils were aware of the reciprocal impact of their own and teachers' excitement and positivity, supporting actual findings from Frenzel et al. (2018), Frenzel, Goetz, Lüdtke, et al., (2009) and Mierzwa (2019).

6.8.4 Summary of pupil perceptions of teacher enjoyment findings

Although pupils in Novel seemed to be slightly more affected by their classroom experience than those in Standard ([6.7.2.2, Classroom factors](#)), pupils' perceptions of teacher enjoyment were similar across approaches, suggesting that it was unlikely that pupils were differently affected by how their teachers felt about teaching that particular approach. Ratings were slightly higher in Standard, but the difference was not statistically significant. Means decreased slightly from PQ2 to PQ3 but the overall impression was that pupils felt that teachers enjoyed teaching them French, even if negatively impacted by instances of misbehaviour in lessons. Reasons offered by pupils for teachers' apparent enjoyment included professional obligation, reciprocal enthusiasm and a sense of pride in pupils' achievement.

6.9 Summary of results and discussion of pupil affective variables

Considering both self-efficacy and enjoyment, [Figure 52](#) illustrates key trends arising from this study. In both approaches, pupils' self-efficacy ratings showed upward trends across the academic year, suggesting that pupils' confidence and belief in their ability to achieve certain tasks in French increased with experience of the language. Self-efficacy ratings were slightly higher in Standard, although between-group differences were not statistically significant at either time-point.

In terms of enjoyment, increasing experience of the language impacted positively on learners' perceptions of progress but did not appear to be associated with changes in ratings, which peaked mid-way through the year and then dropped by PQ3 in both approaches. Again, ratings in Standard were higher but this was only statistically significant for the (approximately) third of pupils whose initial self-assessment indicated that they were unsure about learning French, and only at the mid-point – not post-teaching. This slight decline in enjoyment, apparent in both approaches as the year progressed, may reflect an initial effect of novelty which gradually lessened as pupils grew accustomed to what was involved in language learning. These patterns of findings, for self-efficacy and enjoyment, were similar regardless of whether the pupils had reported looking forward to learning French or had been unsure ([Table 46](#)).

Figure 52.

Self-efficacy and Enjoyment: Overview of Mean Ratings over Time

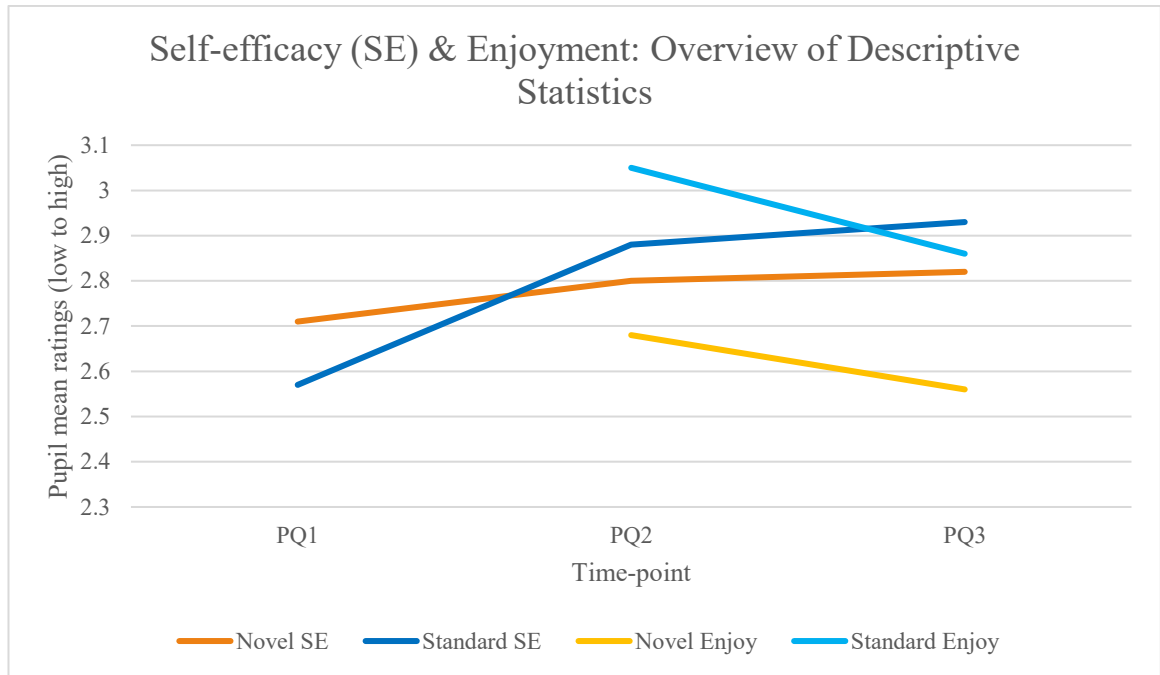


Table 46.

Between-Group Overview of Trends in Self-efficacy and Enjoyment

Measure	Report of anticipated enjoyment of learning French prior to study	Between- group	
		PQ2	PQ3
Self-efficacy	All pupils	N < S (trend)	N < S (trend)
	Pupils not looking forward to learning French*	-	-
	Pupils unsure about learning French	N < S (trend)	N < S (trend)
	Pupils looking forward to learning French	N < S (trend)	N < S (trend)
Enjoyment	All pupils	N < S (trend)	N < S (trend)
	Pupils not looking forward to learning French*	-	-
	Pupils unsure about learning French	N < S (statistically significant)	N < S (trend)
	Pupils looking forward to learning French	N < S (trend)	N < S (trend)

Note. *These figures are not reported due to small sample size.

Pupils' qualitative contributions did not indicate particularly strong differences between approaches. Pupils in Novel group interviews referred more often to slightly higher levels of confidence in French than those in Standard. Pupils' conceptualisations of future progress were similar in both approaches, although there were slight differences in how pupils explained the progress they had made in French, with references in Novel appearing more skills-related and in Standard more vocabulary-focused, reflecting the priorities of each approach. In terms of enjoyment, pupils in both approaches referred to negative as well as positive factors.

Given the slight descriptive trends in favour of Standard apparent in measures of self-efficacy, enjoyment, and pupil perceptions of teacher enjoyment, it would be beneficial to develop further longitudinal research to explore the continued impact of learning in such approaches on affective variables beyond the 22.5 hours afforded by this study.

Chapter 7: Results and discussion III. Teacher findings

This final results and discussion chapter ascertains fidelity to treatment and sets out the seven participant teachers' starting points in terms of their experience, training, support, emotions, and attitudes towards FL teaching and learning. The chapter then presents findings from post-lesson evaluations ([7.4](#)), teacher questionnaires ([7.6](#)) and interviews ([7.8](#)), before comparing teacher perceptions with pupil outcomes and opinions ([7.10](#)). Main themes and trends are discussed after each results section (i.e., in [7.5](#), [7.7](#) and [7.9](#)), addressing, for each approach, teachers' perceptions of the effect on pupils' progress in French (RQ2b), and the effect on teachers' affective variables (RQ3a) and on their developing subject knowledge (RQ3b). Discussion is occasionally informed by brief reference to lesson observations and audio-recordings.

7.1 Comparing fidelity to treatment across the two approaches

This section presents analysis of quantified data from audio-recorded and in-person sample lesson observation schedules ([7.1.1](#)) and teacher perceptions from interview data ([7.1.2](#)).

7.1.1 Analysis of quantified data relating to teachers' fidelity to treatment

For a description of the process of checking fidelity to treatment, see [4.4.5.1-4.4.5.2](#). Teachers' fidelity ratings were normally distributed in both approaches ([Appendix K.6](#)), as assessed by the Shapiro-Wilk test (Novel: $W = .97$, $p = .92$; Standard, $W = .85$, $p = .08$). The Welch two-sample t -test was run using the statistical software package R version 4.3.2 (R Core Team, 2023) and suggested no significant difference ($t [12.56] = 1.23$, $p = .24$) between measures of fidelity to treatment for Novel and Standard approaches.

Occasionally, the five-minute interval measure of fidelity to treatment ([K.4](#)) contained more than one deviation. Where this happened, the measure recorded once that delivery was not as intended. However, the descriptive deviations chart ([Appendix K.7](#)) indicates whether deviations were minor or more serious and categorises them. There were 32 occasions in all monitored Novel lessons when delivery was not as intended (/483 checks): 20 minor, 12 more serious. More serious deviations related to omission of activities either due to time (see [7.1.2](#)) or circumstances beyond teachers' control (e.g., lesson interruptions). Time-related omissions nearly all occurred within the first term, when teachers were adjusting to a novel treatment. Although the number of minor deviations was higher in Standard (44/460 checks), only two were more serious. Minor deviations in Standard appeared to be caused less by time constraints and more by slight changes made by teachers to activities.

7.1.2 Teacher perceptions of fidelity to treatment

Fidelity to treatment was additionally monitored in interviews through teachers' perceptions of whether their delivery was as intended.

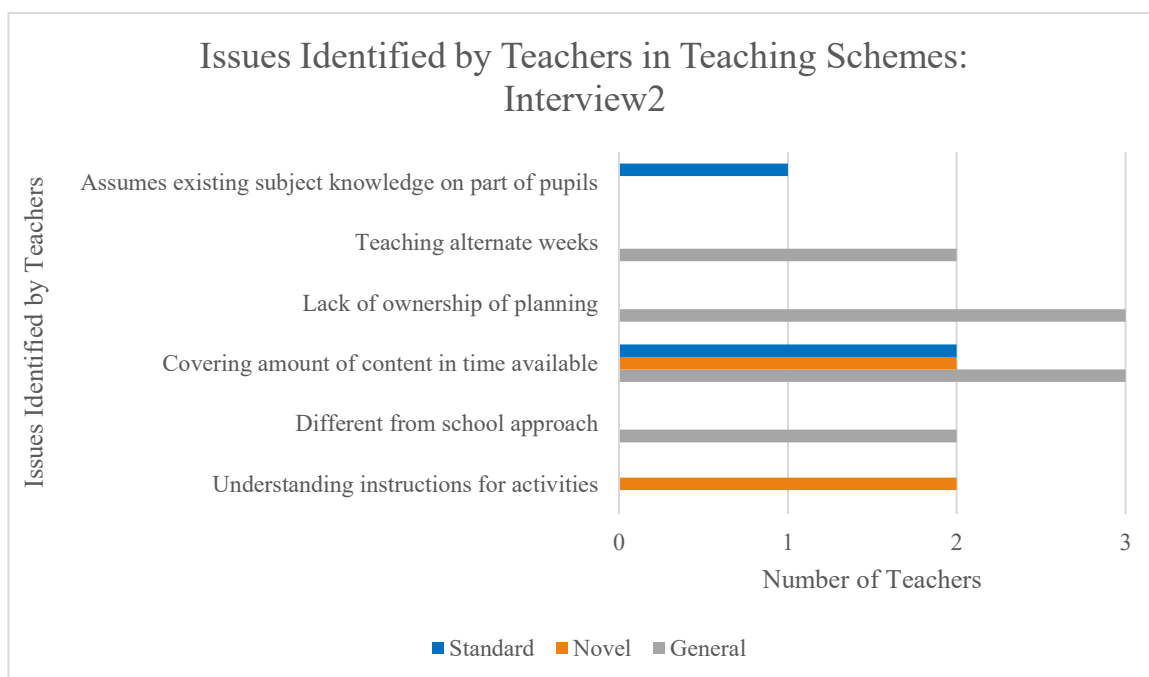
7.1.2.1 Teacher Interview 2 (Interview2)

At Interview2, teachers generally felt that they were following schemes very closely: *"we try to follow exactly what's on the plan"* (T1). For T7, who reported following plans *"to the 't'"*, this was partly due to needing support for insecure subject knowledge, *"because of my confidence"*. T3 acknowledged getting *"slightly muddled"* between approaches and saying *"something which I shouldn't have done"* but acted on the researcher's feedback.

Teachers also identified issues encountered in teaching the two schemes (Figure 53). From these, two themes were constructed: (i) New ways of working; (ii) It's not my own!

Figure 53.

Teacher-identified Issues Arising from Teaching Schemes (Interview2)



Note. Entries refer to the *number of teachers* citing an issue: grey indicates issues common to both approaches; orange - Novel-specific; blue - Standard-specific.

(i) Theme 1: New ways of working

Teachers experienced challenges related to working differently. The most common issue was *"too much content ... more so with the Novel approach"* (T3). In both approaches, teachers perceived insufficient time for writing activities (T7) and lack of flexibility to *"go back over*

something again” for learners’ benefit (T9). Understanding task instructions in Novel appeared challenging for T4 and T7 (*“because of all the grammar and everything”*, T4) – unsurprising, given its novelty: *“when I’m reading the planning ... it takes me longer sometimes to fully understand something or work things out in my head”* (T7). Both schemes were less *“free-flow and ... child-led”* (T1) than the usual approach for School B, requiring adaptation to different ways of working. Teachers also sometimes struggled to find time to acquaint themselves with material covered in the alternate weeks when they were not teaching a particular approach: *“I’ve always missed what they’ve done the lesson before”* (T4).

(ii) *Theme 2: It’s not my own!*

Plans were necessarily detailed to ensure clarity and provide support ([K.1](#), [K.2](#)), but *“following that amount of detail in someone else’s plans is quite difficult”* (T4) in both approaches. Pre-planned lessons meant that teachers lacked ownership of lessons, making them feel that they were *“delivering it rather than planning it”* (T9), unable to put their *“own personality on something”* (T7).

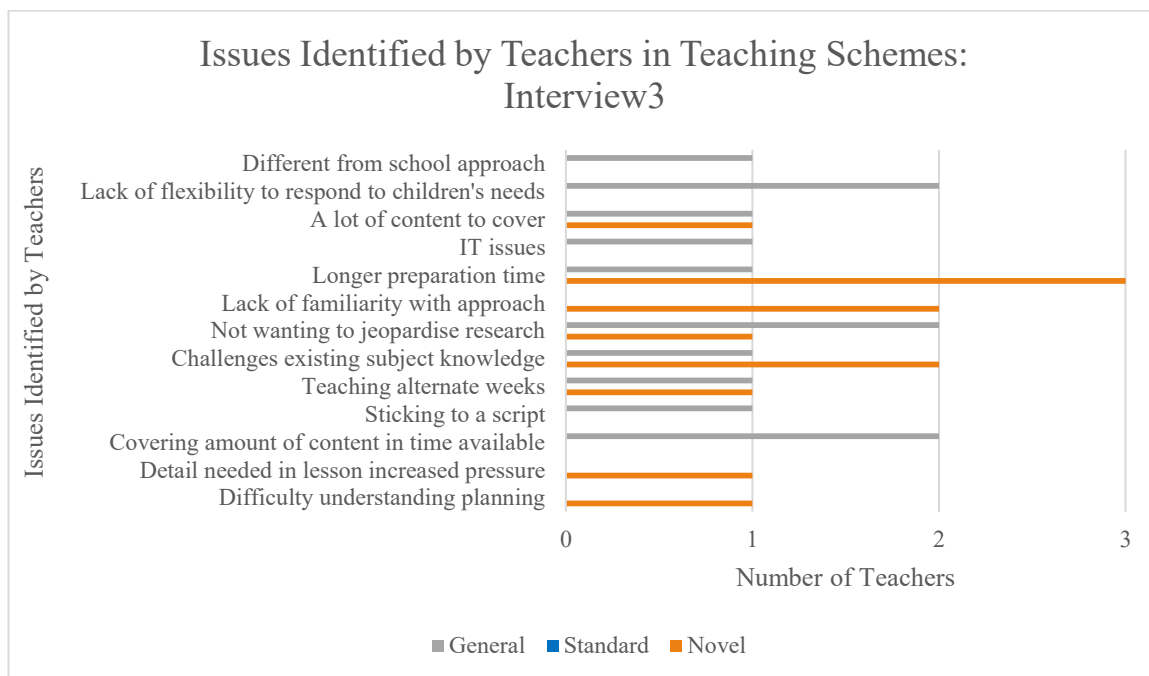
7.1.2.2 Teacher Interview 3 (Interview3)

At the end of the project, all teachers maintained that they were delivering lessons as per instructions: *“it [lesson plan] became almost my Bible really”* (T8). There was no reported confusion between schemes from any teacher. Although teachers had taught both schemes for the academic year, issues still arose ([Figure 54](#)) related to working differently (theme 1: *‘New ways of working’*); these issues either related solely to Novel or to both schemes. T1, T2 and T3 still found it *“tricky just to try and get the timings”* (T2) or work in ways which contradicted the school’s approach (T1). Technical issues were reported due to teaching in unfamiliar contexts (T3), and teachers struggled with the restricted flexibility to respond to emerging needs (T1, T9): *“Occasionally, you see they [pupils] haven’t got something, so something that should take, say, three minutes might end up taking six ... because they didn’t quite get it the first time”* (T9).

Some teachers struggled to understand task requirements in Novel (T3, T2, T9): *“you had to be careful that it was a reading ..., writing, ... or ... listening activity ... You know, I’d highlight it ... so that ... it was true to the plan”* (T3). This was because teachers had to *“think a bit harder”* (T3) about this approach: *“the lesson plans [in Standard] were a bit more forgiving ... I could follow it a little bit easier”* (T2). Alternating approaches still caused issues (T7, T9), particularly teaching Novel: *“because you haven’t had that week of seeing how something’s gone”* (T9). At this later point, lack of ownership of planning was only identified by T9 (*“keeping to a script is quite restrictive”*), with teachers keen to ensure that the research process was not jeopardised (T1, T2, T3): *“I wanted to teach it as you wanted it taught”* (T1).

Figure 54.

Teacher-identified Issues Arising from Teaching Schemes (Interview3)



Note. Entries refer to the *number of teachers* citing an issue: grey indicates issues common to both approaches; orange - Novel-specific; blue - Standard-specific.

7.1.3 Summary of findings on fidelity to treatment

From the statistical analysis and teachers' own perceptions of the strength of their adherence to the planning intentions, it seemed that both schemes were delivered as intended. Where there were occasional deviations from intention, these appeared to apply in both schemes.

7.2 Teacher starting points: results

The initial semi-structured interviews (Interview1, [7.2.2](#)) and first questionnaire (TQ1, [7.2.1](#)), provided background information. Results are presented here.

7.2.1 Teacher Questionnaire 1: results

Teachers' experiences of learning French ranged from holding a GCSE qualification to none ([Table 47](#)); two had taught primary French for 15 years while others had little or no experience. Teachers responded to five enjoyment and four tension questions, with each item asking separately about (a) their general teaching and (b) specifically their French teaching to date ([4.3.11.3](#)). The items were (with the French-specific version in parentheses):

- (i) Enjoyment
 - I generally enjoy teaching (French).
 - I look forward to teaching (French).
 - Teaching (French) is so enjoyable that I like preparing the lessons for the class.
 - I often have reasons to be happy when teaching (French).
 - I generally teach (French) with enthusiasm.
- (ii) Tension
 - I generally feel tense when teaching (French).
 - I often worry that my (French) teaching isn't going well.
 - Preparing to teach (French) often causes me to worry.
 - I feel uneasy when I think about teaching (French).

Teachers responded using a 6-point Likert scale, with 1 indicating very strong agreement, and 6 very strong disagreement; *tension* ratings were then reverse coded so that lower ratings indicated lower tension (4.3.11.1). Averaging each teacher's scores across all enjoyment and all tension items converted the ordinal data from each item into interval data, enabling the mean to be used (University of St Andrews, n.d.).

Table 47 shows that, for all teachers except T3, enjoyment ratings were higher for general teaching than French teaching, and tension ratings were higher for French teaching than general teaching.

Table 47.

Teacher Starting Points (Teacher Questionnaire 1)

Teacher (school)	Number of years' teaching experience	Level of French	Number of years' experience teaching French	General teaching average enjoyment rating /6 (1 = high enjoyment)	French teaching average enjoyment rating /6 (1 = high enjoyment)	General teaching average tension rating /6 (1 = low tension)	French teaching average tension rating /6 (1 = low tension)
T1 (B)	>5<10	Pre-GCSE	6	1.0	2.8	2.3	3.8
T2 (B)	>10 < 20	GCSE	8	1.2	3.4	3.0	4.8
T3 (A,C,E)	20+	GCSE	15	2.3	2.0	3.0	2.3
T4 (A)	>10 < 20	GCSE	15	2.0	3.2	3.0	3.3
T7 (E)	>10 < 20	Zero	5	2.8	4.4	2.5	5.3
T8 (F)	>1 <5	Zero	1	1.5	4.6	2.0	6.0

T9 (F)	<1	Pre-GCSE	N/A	1.3	N/A	3.0	N/A

7.2.2 Teacher Interview 1: results

Building on teacher responses to TQ1, this section sets out teachers' starting points in five key areas: (a) previous training to teach French, including ITT and CPD; (b) existing sources of support; (c) positive and (d) negative emotions at the prospect of participating in the project; and (e) perceived ability to support learners. It discusses teachers' views on support required and the prospect of teaching each scheme.

7.2.2.1 Previous training to teach French

(i) ITT

Teachers' ITT experiences ranged from no FL input (T3, T4) to regular input each year of a three-year course (T8); it generally featured whole (T1) or half days (T9, T2).

(ii) CPD

Most teachers had received CPD (except T1, T4) focusing mainly on pedagogy (T2, T8) or linguistic and pedagogical content combined (T3, T7). T4 had not received ITT or CPD training.

7.2.2.2 Existing sources of support

Where used, schemes of work were considered helpful, enabling T7 to "*relax a little bit*", but difficult to source (T4). Teachers used online support (google translate, pronunciation support, social media, websites), books, local experts and colleagues.

7.2.2.3 Pre-project positive emotions

To simplify presentation, emotions were grouped together into five categories ([Figure 55](#)): (i) security (feeling reassured, relaxed, comfortable, expressing a sense of adequacy); (ii) enthusiasm, encompassing excitement, happiness and positivity; (iii) curiosity/interest; (iv) empathy; and (v) hope. For both positive and negative emotions ([7.2.2.4](#)), the most appropriate label was identified by teachers themselves or interpreted by the researcher from notes made during the interview or transcription process; each emotion was logged and a cause identified (see sample, [Appendix L.2](#)), with totals tallied to enable comparisons across causes. In [Figure 55](#), the inner circle presents the emotion type; the outer circle shows the reported cause of that emotion, grouped according to the 'topic' categories shown in the middle circle. For example, for the emotion type 'Enthusiasm', one topic header was 'Past personal/professional

experiences', under which the causes 'Receiving training'/'Positive experiences of teaching FL'/'Seeing learners make progress' are categorised.

The most common positive emotion was a sense of security, stemming most often from participants' conviction that their subject knowledge (abbreviated to SK in [Figure 55](#)) was sufficient to teach French effectively. This was partly due to learners' *"very, very basic level"* (T2), and was also provided by holding a basic qualification in the FL. Teachers felt most secure in vocabulary, then grammar, above writing and pronunciation. Feeling well supported (primarily by schemes of work and resources, including video support) increased their sense of adequacy: *"you would just kind of follow that scheme of work. It was really straightforward"* (T2).

Thorough preparation was also supportive for teachers: T4 explained that *"as long as I know what I'm doing ... as long as I've... planned the sequence of lessons, and it's not just kind of thrown at me ... then I'm fine"*. This extended to their confidence to respond to learners' enquiries in lessons, provided they were *"based around something that we were learning ... and I had that in the planning"* (T7).

Some less secure participants explained that experience of teaching French had helped develop their confidence:

I was seeing the French being spoken on videos so I could build my confidence over the years a bit more about how to pronounce words ... using that resource for a while became actually quite fun ... I kind of relaxed a bit more with it. (T7)

Other support strategies included using knowledge of other languages: *"the grammar, I know through other languages. It's quite similar, I can ... cope with the grammar"* (T8) and drawing on general pedagogical strategies which *"kind of translate across subjects"* (T9).

7.2.2.4 Pre-project negative emotions

Negative emotions were grouped into three categories: insecurity (uncertainty, inadequacy, a sense of failure); worry (an umbrella term covering nervousness, apprehension, fear, discomfort, panic, desperation, self-consciousness); and frustration (including annoyance). The sunburst chart ([Figure 56](#)) is structured in the same way as [Figure 55](#).

The largest influence on teachers' negative emotions was their sense of insufficient subject knowledge (SK in [Figure 56](#)): 57 mentions, compared to 38 referring to adequate subject knowledge – [4.5.4.4](#) explains the description of the quantification of non-numeric data. This contributed to all three types of negative emotions. It manifested itself in feelings of general inadequacy and specific references to insecurity in pronunciation/speaking and writing/grammar, although not vocabulary ([7.2.2.5](#)). Teachers generally worried about getting things wrong and making mistakes that children would learn: *"My worry would be that, if I'm*

pronouncing something wrong, they'll ... go to secondary school ... and I've taught them the wrong thing – I don't want to give them the wrong information, as it were" (T2).

Figure 55.

Teachers' Pre-project Positive Emotions and Their Causes (Interview1)

■ Sense of security (38) ■ Enthusiasm (14) ■ Curiosity & interest (12) ■ Empathy (2) ■ Hope (4)



Figure 56.

Teachers' Pre-project Negative Emotions and Their Causes (Interview1)



Teachers' frustration with their lack of progress in French subject knowledge was evident. There was a desire to try something different, a recognition that historical approaches to teaching FL have not been sufficiently effective (full quotation in [Appendix R.3](#)):

I am a complete novice ... I just can't understand how in this country we've not nailed it ... something isn't right ... we're doing something wrong ... I find it really frustrating ... The Standard approach ... is very topic-based ... what we do here at school ... just from memory – literally, 'You need to learn, you need to learn ... why are you not learning?' And it's probably why I'm not learning, because it's the wrong way of learning. So, I'm hoping that the Novel approach will just kind of cushion us with a foundation. (T1)

Teachers' emotions were negatively impacted by lack of involvement in FL teaching: *"I'd probably say 'fear', yeah. It is probably my lack of knowledge having not been teaching it regularly"* (T2). In contrast, regular French teaching appeared to impact positively on teachers' confidence and sense of security ([Figure 55](#)): *"I would say actually I've improved ... over the years"* (T7). Teachers were also affected by personal learning experiences. Although a basic qualification such as GCSE often boosted their security, feelings of failure as FL learners increased their nervousness and insecurity: *"I did French at school ... I wasn't very good at it ... There were children there who were clearly much better at it than I was. And ... you know how you rank yourself against other people"* (T9).

Participating in the study increased teachers' sense of disquiet. For the visiting teacher (T3), this was caused by the prospect of working in unfamiliar environments. For some teachers with French teaching experience, nervousness arose from the recognition that, while Standard was *"very similar to what we're doing now"* (T1), Novel would involve teaching sometimes unfamiliar content differently: *"I really need to get my head around this before I think about teaching it"* (T1). There was apprehension about delivering schemes as intended (*"sticking to a particular ... script, as it were. And being careful not to deviate from that"*, T3), especially after having autonomy in French teaching: *"I've had free rein ... up until now"* (T3). The process of being monitored, including audio-recordings and observations, increased worry *"because you want it done in such a certain way"* (T4).

7.2.2.5 Perceived ability to support learners

Teachers' perceptions of their ability to assist learners varied according to the support required and are now presented in three sections: writing/grammar; pronunciation; explaining vocabulary.

(i) Writing/grammar

Four teachers (T1, T4, T8, T9) reported confidence pre-project to support learners' writing, including simple sentence creation (T7, T9). However, some elements were acknowledged as more demanding: *"the spelling of the word ... the accents ... the extra parts, I'll just say!"* (T8). T2, T3 and T4 expressed concern about grammar teaching, making 16 references specifically to grammatical gender (four), sentence construction (three)/word order (two), and general difficulties (seven), although *"current teaching at the moment doesn't include a great deal of grammar"* (T3).

(ii) Pronunciation

Correcting pronunciation or supporting with phonics was only seen as manageable by T3 and T4, both with > 10 years' experience of teaching French, although for T4 this depended *"on what the word was and how confident I was about saying it"* (T4). For T9, pronunciation was the most daunting aspect of teaching French: *"All of it's a worry, but not as much of a worry as the pronunciation ... to a non-specialist teacher, the rules of how you write at this level ... seem clearer than the pronunciation"* (T9).

(iii) Explaining vocabulary

Explaining vocabulary was not perceived as overly worrying. T3, T8, and T9 expressed confidence, depending on individual words and the children's level (with Years 5 and 6 considered more challenging).

7.2.2.6 Teachers' perceptions of an effective scheme of work (Interview1)

Teachers described in Interview1 their preferred FL curriculum and desired support. Common priorities included:

- ensuring sequential progression.
- focusing on oral communication.
- support for planning and subject knowledge development, including pronunciation.

T9 wanted a scheme *"accessible for somebody who hasn't got a clue about what they're doing"*, revealing the low starting point in linguistic competence and confidence of some of the teachers, and was supported by T7's comment that a scheme required: *"the things I need to be able to deliver it confidently ... the actual words that I need, the pronunciations ..."*.

7.2.2.7 Teachers' reflections on the prospect of teaching both schemes

Teachers were asked their impressions of the schemes in Interview1, when the first half-term's lesson plans and PowerPoints had been shared with them, but no teaching had yet taken place.

Their reflections on the prospect of teaching each approach supported the construction of the theme ‘Ready for change’. Standard was “*very similar to what we’re doing now*” (T1), “*something that I’ve come across before*” (T7), supporting its embodiment of current practice and reinforcing its familiarity. T3 had previously taught a topic-based approach:

There have been times when I have taught that way where I’m thinking, ‘I want them to be able to say ... a sentence’, which would require a verb, and ... that hasn’t been part of the scheme ... I wonder whether with the new approach the sentence level will progress more quickly. (T3)

Teachers recognised the value of providing a secure foundation in language learning: “*the very, very basics ... the Novel approach seems a lot better set up to ensure that ... they’re [pupils] successful long-term ... they won’t just pronounce a certain phrase*” (T2).

There was also a general appeal of pedagogical experimentation: “*it’s quite difficult because you get used to being introduced to language in a fairly standard way all of the time, so the opportunity to break out from that ...*” (T9): for T9, Novel was “*a shiny new thing!*”. However, this novelty increased the challenge, with T3 feeling “*a little bit more daunted about the Novel approach ... probably because I’ve only been used to the topic-based approach*”.

7.3 Teacher starting points: summary and brief discussion

Except for T3, teachers generally enjoyed French teaching less than general teaching, experiencing higher levels of tension. Where available, they found schemes supportive. Feeling prepared and secure in subject knowledge increased positive emotions (as found by Dewaele & MacIntyre’s statistical analysis, 2014), as did successful experience of French teaching (previous studies reported increases in self-perceived teacher confidence through experience of teaching - Cable et al., 2010; Cooke & Faez, 2018; Driscoll, 1999; Driscoll, Jones, & Macrory, 2004; Muijs et al., 2005; Rowe et al., 2012). Negative emotions arose from fear of getting things wrong and insecure subject knowledge (also reported by Barnes, 2006; Finch et al., 2018; Muijs et al., 2005; Woodgate-Jones, 2008).

A lack of confidence was more apparent when supporting learners with writing, grammar and pronunciation than when explaining vocabulary. Teaching writing and developing pupils’ grammatical understanding were identified as challenging in national surveys (Board & Tinsley, 2014, 2015; Tinsley & Board, 2012). Teachers in the current study raised the concern that pupils’ phonological outcomes may be limited by their inaccurate models of pronunciation, also noted by Tinsley and Board (2017) and reported by Muñoz (2009) and Larson-Hall (2008).

Teachers wanted support for their own pronunciation, planning and subject knowledge development as well as in ensuring pupil progression. This reflected the difficulty often encountered by non-specialist FL teachers in moving beyond basic language noted by Driscoll

(1999) and McLachlan (2009). It was also found by Graham et al. (2017), which speculated that primary-aged learners taught by teachers with lower levels of proficiency in French made less progress because of the challenge of planning for progression beyond beginner levels. Some of the teachers were ready for a change and hoped that Novel could bring this but were aware of the anxiety that unfamiliarity might engender.

Having described the starting points for the participant teachers, this chapter now reports their reactions to experiences of teaching each scheme.

7.4 Teacher responses to each approach: results from post-lesson evaluations

Post-lesson evaluations captured teachers' emotional responses to taught lessons, including perceptions of pupil progress and pupil affective variables. Evaluations consisted of eight items to which teachers responded using a 6-point Likert scale, with 1 indicating very strong agreement, and 6 very strong disagreement ([Appendix F.1](#); also [4.3.10](#), [4.4.6](#)).

7.4.1 Test quality

Normality of distribution for individual items is presented in [Appendices T.1-T.8](#). No items were normally distributed.

7.4.1.1 Number of responses

With six of 36 lessons allocated to testing, questionnaires, and group interviews, the total possible evaluations per teacher was 30 (15 from each approach). However, due to teacher absence and the removal of evaluations completed more than two hours after the lesson ([4.3.10.3](#)), not all teachers completed the same number of evaluations ([Appendix F.5](#)). T3 (the visiting teacher) completed evaluations for three schools.

7.4.1.2 Reliability of post-lesson evaluation survey

Cronbach's alpha, run both manually using Excel and the 'psych' package (Revelle, 2024) in R, showed reliability of $\alpha = .89$. Removing item 2 would have improved reliability to .92, but, with non-specialist teachers' levels of tension when teaching each approach considered important to their overall satisfaction, it was retained.

7.4.2 Item-by-item analysis: descriptive and inferential statistics

[Table 48](#) presents descriptive statistics for individual items in post-lesson evaluations. In all cases, lower means indicate stronger levels of teacher agreement with the item, except for Item 2, where lower ratings indicate lower levels of tension.

Table 48.*Descriptive Statistics for Individual Items in Teachers' Post-lesson Evaluations*

Item	Approach ($k = 1$)	Descriptive statistics					Inferential statistics			
		Number of data points	Mean (total/6)	CI	SD		Effect size d	Effect size CI	Wilcoxon W	p -value
1. I enjoyed this lesson.	Novel	139	2.9	2.8-3.1	.8	Novel - Standard	.31	.07-.55	11004.00	.009
	Standard	137	2.7	2.6-2.8	.7					
2. I was tense during this lesson.	Novel	139	3.5	3.4-3.6	.9	Novel - Standard	.25	.009-.49	10876.00	.02
	Standard	137	3.3	3.1-3.4	1.0					
3. Students understood the material.	Novel	139	2.9	2.8-3.0	.6	Novel - Standard	.45	.21-.69	11676.00	< .001
	Standard	137	2.6	2.5-2.7	.7					
4. Students were motivated.	Novel	139	2.9	2.8-3.0	.8	Novel - Standard	.34	.10-.58	11371.00	.001
	Standard	137	2.7	2.6-2.8	.7					
5. Students were focused.	Novel	139	3.1	2.9-3.2	.8	Novel - Standard	.37	.13-.61	11290.00	.001
	Standard	137	2.8	2.7-2.9	.7					
6. I was satisfied with what students knew.	Novel	139	2.9	2.8-3.0	.7	Novel - Standard	.26	.02-.50	10684.00	.04
	Standard	137	2.7	2.6-2.8	.7					
7. The French was useful.	Novel	139	2.7	2.6-2.9	.7	Novel - Standard	.15	-.09-.39	10312.00	.15
	Standard	137	2.6	2.5-2.8	.7					
8. Activities gave me sufficient evidence about learning.	Novel	139	2.8	2.7-2.9	.7	Novel - Standard	.18	-.05-.42	10292.00	.16
	Standard	137	2.7	2.6-2.8	.7					

Note. Bold text indicates the approach with the lower mean rating (and therefore stronger agreement with the statement), except for item 2, where it indicates lower levels of tension.

7.4.3 Factor analysis of teacher satisfaction items from post-lesson evaluations

Three factors were extracted from the factor analysis (4.4.6.2) based on eigenvalues > 1.00 (Kaiser, 1960, cited in Field et al., 2012) and, in combination, accounted for 75% of the data variance (Table 49).

Table 49.

Exploratory Factor Analysis of the Teacher Post-lesson Evaluation Items

Item (I)	Factor		
	1: perceptions of pupil outcomes and linguistic usefulness	2: perceptions of classroom enjoyment	3: teacher tension
I1 (Teacher enjoyment)	.37	.39	.17
I2 (Teacher tension)	-.01	-.01	1.00
I3 (Perceptions of pupil understanding)	.68	.13	.03
I4 (Perceptions of pupil motivation)	.05	.87	-.02
I5 (Perceptions of pupil focus)	-.04	.90	.00
I6 (Satisfaction with pupil learning)	.87	-.02	.05
I7 (Satisfaction with usefulness of French)	.90	-.06	-.04
I8 (Satisfaction with evidence of learning)	.80	.05	-.03
% of variance explained	38.00	24.00	13.00

Note. Loadings larger than .40 are shown in bold.

Factors were as follows (with loadings indicated in brackets):

(i) Factor 1, represented by data from items 3 (.68), 6 (.87), 7 (.90), and 8 (.80), clustered to give an overall indication of teachers' perceptions of pupil knowledge and understanding, as well as of usefulness of the language taught.

(ii) Factor 2, represented by data from items 4 (.87) and 5 (.90), provided a measure of reported pupil focus and motivation; item 1 (teacher enjoyment) also appeared most closely

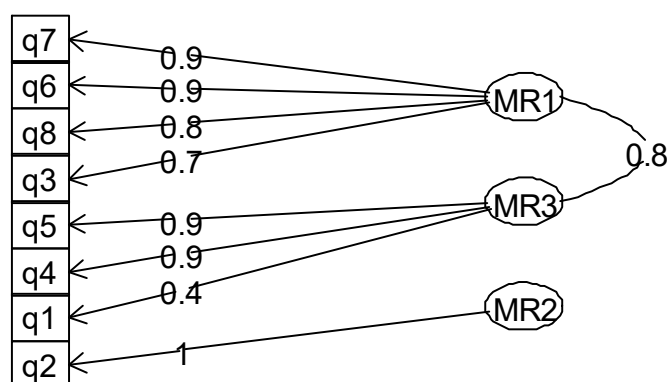
related to this factor and, with a loading of .39, was above the .30 typically considered important (Field et al., 2012).

(iii) Factor 3: teachers' tension levels (item 2) teaching lessons in each approach constituted the third factor with a loading of 1.0.

In summary, three distinct factors arose from the exploratory factor analysis: perceptions of pupil outcomes and linguistic usefulness; perceptions of classroom enjoyment; and teacher tension (Figure 57).

Figure 57.

Factor Analysis of Teachers' Post-lesson Evaluations



Descriptive and inferential statistics for the three factors are now presented, with overall calculation of all teacher responses. Here, the items representing similar questions were combined into a single score, thus enabling Likert scale data to be analysed as interval data (University of St Andrews, n.d.).

7.4.3.1 Factor 1: Teacher perceptions of pupil outcomes and linguistic usefulness

Individual teacher responses to items 3, 6, 7 and 8 (Appendix F.1) were summed and analysed to determine whether individuals exhibited higher levels of satisfaction in either approach. This factor showed very good reliability: $\alpha = .90$ (checked using the 'psych' package [Revelle, 2024] in R). Data were not normally distributed (Appendix T.9). For each lesson, a maximum rating of 24 was possible (four items, 6-point Likert scale), with any rating below 12 indicating general satisfaction or positivity (i.e., an average rating of ≤ 3 on each of the four items). Lower ratings reflect more positive opinions.

(i) *All teachers' ratings*

Teachers appeared generally satisfied with pupil outcomes and linguistic usefulness, although the lower mean in Standard indicated perceptions of greater satisfaction than in Novel ([Table 50](#)). The effect size was large; CIs did not cross zero. There was a significant difference between approaches, with teachers more satisfied with pupil understanding and knowledge in Standard lessons than in Novel, with greater conviction that the French was useful.

Table 50.

Teacher Perceptions of Pupil Outcomes and Linguistic Usefulness (Post-lesson Evaluations)

Approach	$k = 4$	Descriptive statistics				Inferential statistics	
		Number of data points	<i>Mean</i> of totals /24	<i>SD</i>	Effect size d	95% CI	Wilcoxon W
Novel	556	11.4	2.3	Novel – Standard: .30	.06-.54	11194.00	< .008
Standard	548	10.6	2.6				

(ii) *Individual teacher ratings*

Individual teacher ratings were higher for Standard than Novel, except for T1 and T3 (in School C), whose ratings were slightly higher in Novel than Standard with medium and large effect sizes respectively, although these differences were not statistically significant. Data for the construct of teachers' perceptions of pupil knowledge and understanding, as well as of usefulness of the language taught, appear in [Appendix U.1](#), with normality of distribution in [V.1-V.8](#).

7.4.3.2 Factor 2: Classroom enjoyment

From the factor analysis, the construct of perceived classroom enjoyment was represented by items 4, 5 and 1 ([Appendix F.1](#)), with individual teachers' responses summed and analysed to determine whether enjoyment ratings were higher in either approach. Reliability was very good: $\alpha = .86$ (checked using the 'psych' package [Revelle, 2024] in R). Normality of distribution checks are provided in [Appendix T.10](#). A maximum rating of 18 was possible per lesson (three items, 6-point Likert scale), with any rating below 9 indicating general satisfaction or positivity (i.e., an average rating of ≤ 3 on each of the three items). Lower ratings reflect higher levels of perceived (pupil) motivation or real (teacher) enjoyment.

(i) *All teachers' ratings*

The mean was slightly lower in Standard than Novel, with a large effect size. There was a statistically significant difference between approaches, suggesting that teachers' own enjoyment and their perception of pupil motivation were higher in Standard than Novel ([Table 51](#)). This was unsurprising, as this was the case in all three individual items making up this construct ([Table 48](#)).

Table 51.

Classroom Enjoyment: Teacher Enjoyment and Perceptions of Pupil Motivation (Post-lesson Evaluations)

Approach	$k = 3$	Descriptive statistics			Inferential statistics		
		Number of data points	Mean of totals /24	SD	Effect size d	95% CI	Wilcoxon W p -value
Novel		417	8.9	2.0	Novel – Standard: .38	.14-.62	11662.00 < .001
Standard		411	8.2	1.8			

(ii) *Individual teacher ratings*

For the construct of enjoyment, individual teacher ratings appear in [Appendix U.2](#), with normality of distribution in [V.9-V.17](#). These all show the same pattern (higher in Standard than Novel), though there may have been a trend in the opposite direction for T1 and T3 (School C), both with large effect sizes and the latter reaching statistical significance; however, these CIs crossed zero.

7.4.3.3 Factor 3: tension

From the factor analysis, the construct of teacher tension was represented by item 2 ([Appendix E.1](#)). As the construct consisted of a single item, reliability was not checked. Normality of distribution checks are provided in [Appendix T.11](#). Teachers' responses to the statement "*I was tense during this lesson*" were reverse coded, so higher ratings indicate higher levels of tension. A maximum rating of six was possible (one item, 6-point Likert scale), with any rating below three indicating lower tension.

(i) *All teachers' ratings*

Teachers clearly experienced some tension when teaching lessons in both approaches (item 2, [Table 48](#)), with the full range of ratings selected. Standard deviations were quite high, especially in Standard, indicating individual variation. Means were lower in Standard than Novel,

indicating lower tension (with some overlap in CIs around the means) and a large effect size, although at the lower end these CIs were only just above zero. There was a significant difference in teachers' levels of tension during lessons in each approach ([Table 48](#)), with teachers experiencing higher levels of tension when teaching Novel than Standard.

(ii) Individual teachers' ratings

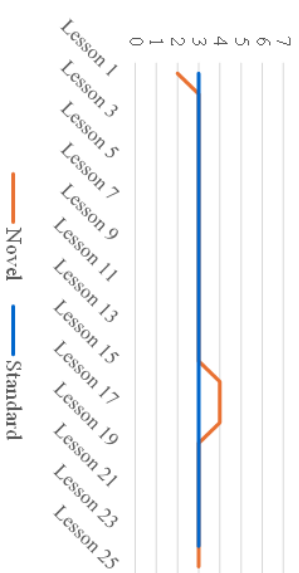
For the construct of tension, [Figure 58](#) shows individual teacher post-lesson evaluation ratings and [Table 52](#) presents descriptive and inferential statistics. These are included here rather than in the appendices because of individual teacher differences. For normality of distribution checks, see [Appendices V.18-V.26](#).

Higher means in Novel than Standard for all teachers except T1, T3 (School A) and T9 indicated higher tension, and reached statistical significance for T8, with a large effect size. T1 appeared to experience slightly less tension when teaching Novel than Standard (with a medium effect size); for T3 and T9 there was little or no difference.

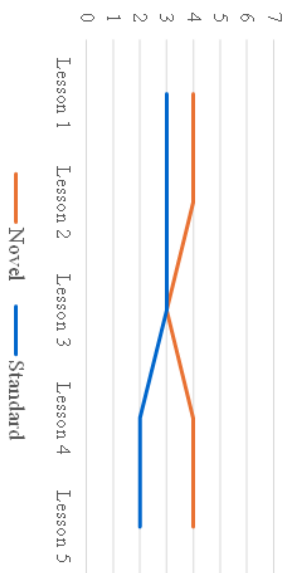
Figure 58.

Individual Teacher Ratings for Tension Construct by Lesson (Post-lesson Evaluations)

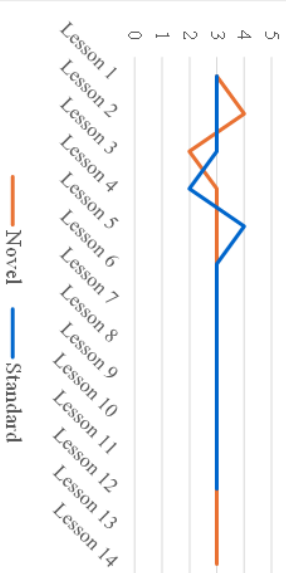
T1: Teacher Tension



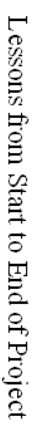
T3: (School A) Teacher Tension



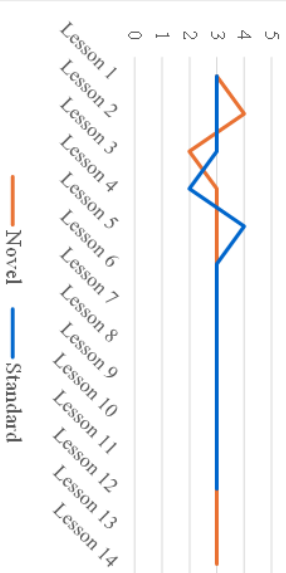
T3: (School C) Teacher Tension



T7: Teacher Tension



T8: Teacher Tension



T9: Teacher Tension



Table 52.*Individual Teachers' Tension Ratings in Post-lesson Evaluations*

Teacher	Approach (<i>k</i> = 1)	Descriptive statistics				Inferential statistics				
		# data points	Mean /6	SD		# data points	Cohen's <i>d</i>	CI	Wilcoxon W	<i>p</i> -value
T1	Novel	15	4.4	.8	Novel - Standard	30	-.19	-.94-.56	102.00	.65
	Standard	15	4.5	.5						
T2	Novel	11	4.2	1.2	Novel - Standard	22	.62	-.29-1.54	80.00	.20
	Standard	11	3.2	1.9						
T3 (Sch A)	Novel	25	3.1	.4	Novel - Standard	49	.28	-.30-.86	324.00	.32
	Standard	24	3.0	.0						
T3 (Sch C)	Novel	29	3.1	.3	Novel - Standard	57	-.009	-.54-.52	405.00	.99
	Standard	28	3.1	.3						
T3 (Sch E)	Novel	16	3.1	.3	Novel - Standard	32	.21	-.52-.93	136.00	.58
	Standard	16	3.1	.3						
T4*	Novel	5	3.8	.5	Novel - Standard	10				
	Standard	5	2.6	.6						
T7	Novel	14	4.6	.5	Novel - Standard	28	.20	-.58-.98	114.00	.43
	Standard	14	4.4	.9						
T8	Novel	10	3.2	1.1	Novel - Standard	22	1.00	.05-1.95	91.50	.03
	Standard	12	2.2	.9						
T9	Novel	14	3.00	.4	Novel - Standard	26	.00	-.81-.81	84.00	1.00
	Standard	12	3.00	.4						

Note. Bold text indicates the approach with the lower mean rating (and therefore lower tension).

*T4 left at the end of the first term.

7.4.4 Summary of results from post-lesson evaluations

Based on post-lesson evaluations with good or better levels of reliability, when considered item-by-item and according to the three factors arising from the factor analysis, the following points should be noted:

- teachers were largely satisfied with pupils' outcomes and linguistic usefulness in both approaches, but more so in Standard than Novel.
- teachers experienced and perceived enjoyment teaching both approaches, but at higher levels in Standard than Novel.

- five teachers experienced some tension teaching both approaches, but at higher levels in Novel than Standard. The reverse was true for one teacher. For the visiting teacher, the level of tension in each level varied according to the school context.

These results are now discussed.

7.5 Teacher responses to each approach: discussion of results from post-lesson evaluations

This section discusses findings according to the three factors identified in the factor analysis: teacher perceptions of pupils' outcomes and linguistic usefulness (7.5.1); classroom enjoyment (7.5.2); and teacher tension (7.5.3).

7.5.1 Teacher perceptions of pupils' outcomes and linguistic usefulness: discussion

This section addresses teachers' perceptions of the effect of each teaching approach on pupils' progress in French (RQ2b).

The individual item ratings (Table 48) illustrate teachers' general satisfaction with pupils' progress in *both* approaches, although this appeared greater in Standard than Novel in all items, reaching statistical significance in terms of teachers' perceptions of pupils' knowledge and understanding (Table 53). Means < 12 (out of a total rating possible of 24) indicated general satisfaction. Only T7's ratings were > 12 ($M = 13.5$), in Novel. T1 exhibited fluctuations, whereas T3 and T9 were more consistent throughout the year (Appendix U.1, Figure 86). Post-teaching, ratings for T8 appeared consistently higher for Standard lessons.

The provision of resources in both approaches (the importance of which was documented in national surveys - Cable et al., 2010; Tinsley & Board, 2012; Tinsley & Doležal, 2018) possibly supported teachers to feel that all lessons enabled, and provided feedback on, progress (requested by teachers in Interview1), although teachers generally still preferred Standard.

Table 53.

Overarching Summary of Post-lesson Evaluations: Teacher Perceptions of Pupil Outcomes and Linguistic Usefulness

Measure	Between-group
Teachers' perceptions of pupil understanding	N < S (statistically significant $p < .001$, $d = .45$ [CI .21-.69]: large effect size)
Teachers' satisfaction with pupils' knowledge	N < S (statistically significant $p = .04$, $d = .26$ [CI .02-.50]: large effect size)

Teachers' satisfaction with the usefulness of French taught	N < S (trend)
Teachers' perceptions of the evidence provided about pupils' learning	N < S (trend)
Factor 1: Overall teacher perceptions of pupils' outcomes and linguistic usefulness	N < S (statistically significant $p < .008$, $d = .30$ [CI .06-.54]: large effect size)

7.5.2 Classroom enjoyment: discussion

For each approach, this section addresses the (perceived) effect on pupils' (RQ2a) and primary teachers' affective variables (RQ3a).

Teachers generally enjoyed teaching both approaches and perceived pupils as motivated and focused in both ([Table 48](#)). However, in all three measures of enjoyment and (therefore inevitably) the overall factor, lower mean ratings in Standard than Novel indicated higher experienced and perceived enjoyment (see summary, [Table 54](#)), reaching statistical significance in all measures with large effect sizes.

Table 54.

Overarching Summary of Post-lesson Evaluations: Classroom Enjoyment

Measure	Between-group
Teachers' enjoyment	N < S (statistically significant, $p < .009$, $d = .31$ [CI .07-.55]: large effect size)
Teachers' perceptions of pupil motivation	N < S (statistically significant, $p < .001$, $d = .34$ [CI .10-.58]: large effect size)
Teachers' perceptions of pupil focus	N < S (statistically significant, $p < .001$, $d = .38$ [CI .14-.62]: large effect size)
Factor 2: Overall classroom enjoyment	N < S (statistically significant $p < .001$, $d = .30$ [CI .06-.54]: large effect size)

The effect sizes of the difference between Standard and Novel ([Table 62](#), [Appendix U.1](#)) were highest for T7 and T8, with, at TQ1, the lowest levels of French, lowest levels of enjoyment and highest starting points for tension whilst teaching French ([Table 47](#)). Low levels of linguistic proficiency can impact on confidence (reported in Muijs et al.'s Pathfinder evaluation, 2005) and engender reluctance in teachers (see Tinsley & Board's questionnaire-study, 2016), with greater FL proficiency correlating positively with higher levels of FL enjoyment (see Dewaele & MacIntyre's statistical analysis, 2014). It is therefore possible that, especially for T7 who had

greater experience teaching French than T8, the Novel approach was less similar to (and therefore more daunting than) previous schemes, as anticipated pre-project, and this decreased the sense of enjoyment. Indeed, Du Plessis et al.'s (2015) interview-study and observations found that feeling overwhelmed affected out-of-field (non-subject specialist) teachers' ability to enjoy their teaching.

The impact of insufficient linguistic competence applied less to T3, with a GCSE in French, 15 years' experience of French teaching, the highest level of enjoyment and lowest level of tension while teaching French pre-project: their responses varied according to the school context. Whereas their overall measure of enjoyment was no different in either approach according to the TQs (7.6), post-lesson evaluations revealed higher perceptions of enjoyment in Novel lessons in School C (statistically significant), but higher in Standard in Schools A (also statistically significant) and E (with a large effect size but not significant; CIs crossed zero). This raises the possibility that, in this study, perceived pupil enjoyment and real teacher enjoyment were to some extent related to teaching context, as found by Frenzel et al.'s (2015, 2016) statistical analyses of teacher emotions and Keller et al.'s (2018) empirical investigation of teacher diaries. This appears supported by the findings that pairs of teachers teaching in the same school also gave similar ratings (T3/T7, School E; T3/T4, School A; T8/T9, School F), suggesting that the school itself can have an effect.

However, T1 and T2, both teaching in School B, did not rate similarly for enjoyment. T1's results varied according to *data source*: they perceived higher enjoyment in Standard in TQs (7.6), but their enjoyment ratings in post-lesson evaluations were higher in Novel. T2's enjoyment ratings were higher in Standard. For both teachers, *p*-values in post-lesson evaluations were $> .05$ and the CIs around the effect sizes crossed zero. [Appendix U.2](#) shows fluctuations in T1's ratings: for the first two lessons, T1 rated teacher and pupil enjoyment higher in Novel than Standard; by Lesson 3 there was less difference; and high/low enjoyment ratings were given in both approaches throughout the year ([Figure 87](#)). Possibly, T1 completed TQ3 after a less enjoyable Novel lesson or a particularly enjoyable Standard lesson and this impacted on ratings. This supports the idea that emotional responses may differ according to the individual teacher or the group being taught (Frenzel et al., 2015) and may be context-dependent (Keller et al., 2018), even to the level of individual lessons.

There was a possible positive association between teachers' real enjoyment and their perceptions of pupils as focused and motivated, as pupils are more likely to enjoy lessons if they perceive the teacher as positive (see Frenzel et al.'s investigation, 2018, of emotional transmission in the classroom). Pupils' enjoyment levels can be impacted by teachers' instructional behaviour (found in Dewaele et al.'s questionnaire-study, 2019, and Keller et al.,

2018). There may additionally be an element of reciprocity, as teachers may respond positively to signs of pupil engagement (Frenzel et al., 2018), explored further in [7.8.2.2](#).

7.5.3 Teacher tension: discussion

This section discusses teachers' levels of tension as reported in the post-lesson evaluations and thus addresses for each approach the effect on primary teachers' affective variables (RQ3a) and their perceptions of the effect on their developing subject knowledge (RQ3b).

Teachers experienced some tension while teaching both approaches. Means were lower in Standard than Novel, indicating lower levels of tension and reaching statistical significance with a large effect size (see [Table 55](#)).

Table 55.

Overarching Summary of Post-lesson Evaluations: Teacher Tension

Measure	Between-group
Teachers' tension/Factor 3: Overall teacher tension	N > S (statistically significant, $p = .02$, $d = .25$ [CI .00-.49]: large effect size)

Note. Only one item constituted the teacher tension factor, therefore there is only one measure.

7.5.3.1 Relationship between tension experienced while teaching French and linguistic competence/experience teaching French

Lack of confidence has been reported as arising from low levels of linguistic proficiency (Barnes, 2006; Finch et al., 2018; Muijs et al., 2005; Woodgate-Jones, 2008). Frenzel, Goetz, Stephens, et al. (2009) reported that anxiety may occur when teachers believe they lack competence to achieve classroom goals, and FL teaching anxiety was negatively predicted by language proficiency in Amengual Pizarro's (2019) questionnaire-study of 75 prospective primary EFL teachers in Spain. It is therefore possible that (self-perceived) insufficient linguistic competence may increase tension when teaching FLs. T1, T7, T8 and T9 had the lowest levels of competence in French. Of these, T1 and T7 experienced the highest levels of tension of all teachers when teaching either Standard (T1) or Novel (T7), but their ratings were also high when teaching the alternative condition, indicating tension regardless of approach. This was despite their previous experience teaching French (six and five years respectively). T8 and T9's tension ratings were lower, even though they had similar levels of competence in French to T1 and T7 (zero or pre-GCSE) and little or no experience of teaching French (T8: one year; T9: no experience). Means for T9 were 3.0 in both approaches and showed little fluctuation throughout the year ([Figure 58](#)).

Some researchers have suggested that *perceived* competence matters more than *actual* competence (Dörnyei & Csizér, 1998; Horwitz, 1996; Pattison, 2014; Sharpe, 1995); alternatively, this tension may be partly driven by personality. T8's ratings were consistently lower in Standard than Novel from the start of the project, as evidenced by both post-lesson evaluations and TQs ([Table 52](#) and [Figure 60](#)); post-teaching, these ratings reflected little to no tension when teaching Standard, indicating that teachers' emotional responses may differ according to the individual (Frenzel et al., 2015) and their personalities.

A combination of a GCSE qualification in French and considerable experience perhaps supported T3, whose ratings for both approaches indicated some of the lowest tension levels in all three schools, shown in both post-lesson evaluations and questionnaires ([Figure 60](#)). This teacher, visiting three schools, also had most experience of teaching Novel, perhaps explaining why they experienced lower levels of tension in Novel than Standard in one school. The familiarity of Standard possibly explains T2's and T4's low tension ratings in this approach. T4's tension ratings were consistently lower in Standard, whereas T2 showed lesson-to-lesson fluctuation ([Figure 58](#)).

7.5.4 Discussion of results from post-lesson evaluations: summary

Teachers were positive about pupil progress even when their own subject knowledge was insecure. Holding a GCSE qualification appeared to afford some benefits in terms of decreasing teachers' tension when teaching French, with this more evident in Standard than Novel; this is perhaps unsurprising, as teachers' previous experience of teaching French was of an approach more similar to Standard than Novel.

(Self-perceived) lack of linguistic competence appeared to impact negatively on levels of tension and decrease enjoyment, with this more apparent in Novel. Overall, teachers seemed less positive about Novel, possibly because of its unfamiliarity – as, pre-project, they were generally positive about its potential. Rogers (2003) identified characteristics of innovations likely to impact on how quickly they would be adopted, one of which was compatibility, “the degree to which an innovation is perceived as consistent with the existing values, past experiences, and needs of potential adopters” (2003:225). Rogers (2003) suggested that greater compatibility may increase the familiarity of an innovation to potential adopters. However, teaching context and the individual teachers themselves also seemed instrumental in determining emotional responses.

7.6 Teacher responses to each approach: results from teacher questionnaires

In addition to TQ1, six teachers responded to a short survey mid-project (TQ2) and post-teaching (TQ3), reporting their enjoyment and tension using a 6-point Likert scale, with 1 indicating very

strong agreement, and 6 very strong disagreement ([Appendix F.4](#); [4.3.11](#), [4.4.7](#)). No reliability measure is provided for TQs due to the small sample size ($n = 6$), although the internal consistency for the questionnaire when it was used in the original study was $\alpha = .90$ (Frenzel, Goetz, Stephens, et al., 2009). Results are presented here with cross-reference to post-lesson evaluations where relevant.

7.6.1 Teacher questionnaires: enjoyment results

Teachers' responses to the five enjoyment items ('I generally enjoy teaching French'; 'I look forward to teaching French through this approach'; 'Teaching French is so enjoyable that I like preparing the lessons for the class'; 'I often have reasons to be happy when teaching French'; 'I generally teach French with enthusiasm') were summed to give a possible total out of 30, with low ratings indicating strong agreement and therefore high enjoyment.

The ratings of T1, T3 and T8 decreased between TQ2 and TQ3 ([Figure 59](#)), indicating greater enjoyment, with preferences emerging for Standard from T1 and T8 (contradicting post-lesson evaluation findings in the case of T1 but confirming those of T8). T2 enjoyed Standard more than Novel at both points, although the extent of this preference decreased by TQ3. T7's responses were consistent at both time-points, with the preference for Standard also supporting results of post-lesson evaluations; for this teacher, the difference between approaches appeared less pronounced in TQs than in evaluations, where they reached statistical significance. T9's results also showed higher levels of enjoyment when teaching Standard, supporting findings in post-lesson evaluations – particularly the decrease in ratings in the latter part of the project ([Appendix U.2](#), [Figure 87](#)). When considering all teachers' ratings together ([Table 56](#)), compared with the mean value of their starting points for French teaching (TQ1, [Table 47](#)), enjoyment appeared to increase in both approaches by TQ3.

In [Figure 59/Table 56](#), low ratings indicate strong agreement, therefore high enjoyment.

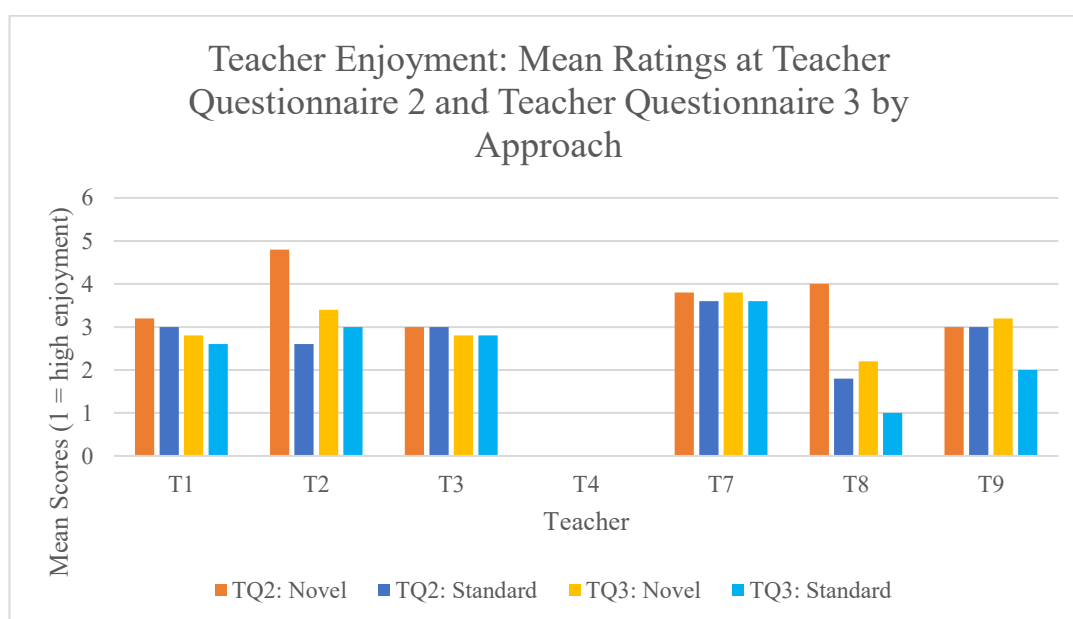
Table 56.

All Teachers' Enjoyment: Mean and Median Ratings at Teacher Questionnaire 1, Teacher Questionnaire 2 and Teacher Questionnaire 3

Teacher	Approach	$k = 5$	
		Descriptive statistics	
		Mean of totals /6	Median /6
All Ts	TQ1 French teaching	3.4	3.00
	TQ2 Novel	3.5	3.00
	TQ2 Standard	2.8	3.00
	TQ3 Novel	3.0	3.00

Figure 59.

Individual Teacher Enjoyment: Mean Ratings at Teacher Questionnaire 2 and Teacher Questionnaire 3



Note. T4 left the school after the first term of the project and did not complete TQ2 or TQ3.

7.6.2 Teacher questionnaires: tension results

Teachers' responses to the four tension items ('I generally feel tense when teaching French'; 'I often worry that my French teaching isn't going well'; 'Preparing to teach French often causes me to worry'; 'I feel uneasy when I think about teaching French') were reverse coded and summed to give a possible total /24, with low ratings indicating lower levels of tension.

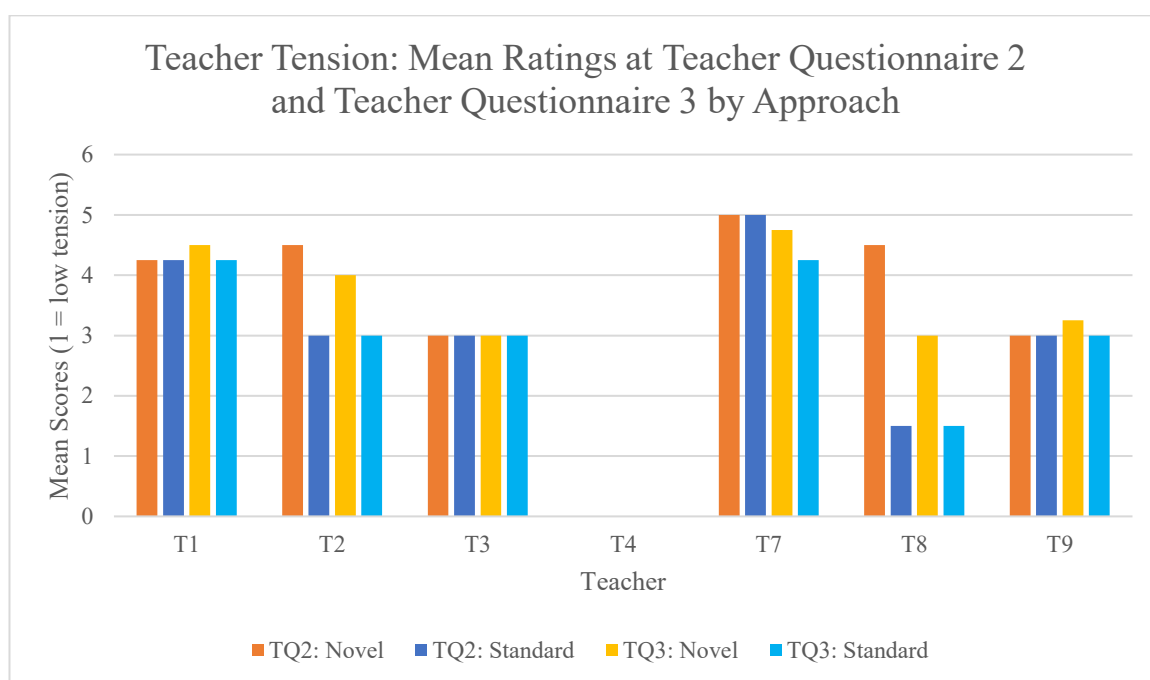
TQs broadly supported the results of the post-lesson evaluations ([Figure 60](#)). Compared with their tension starting points in general French teaching ([Table 47](#)), T1 and T3 showed higher levels of tension (and lower enjoyment) at both time-points in this project, whereas T2, T7 and T8 all reported lower levels of tension at TQ2 and TQ3 than for their previous French teaching. The ratings for T1, T3 and T9 illustrated the similarity in tension levels in both approaches throughout the year, with T1's ratings increasing slightly in Novel at TQ3, contradicting the findings of the overall post-lesson evaluations (slightly lower in Novel than Standard). T7 reported lower tension in both approaches by TQ3 than TQ2, and their lowest ratings in Standard. T8 reported the lowest level of tension of all teachers when teaching Standard at both time-points. T2's tension levels decreased slightly in Novel by TQ3 but remained higher than both

measures in Standard. When considering all teachers' ratings ([Table 57](#)), tension appeared to decrease from the starting point by TQ3 in all three measures, particularly in Standard.

In [Figure 60](#) and [Table 57](#), low ratings indicate lower tension.

Figure 60.

Teacher Tension: Mean Ratings at Teacher Questionnaire 2 and Teacher Questionnaire 3



Note. T4 left the school after the first term of the project and did not complete TQ2 or TQ3.

Table 57.

All Teachers' Tension: Mean and Median Ratings at Teacher Questionnaire 1, Teacher Questionnaire 2 and Teacher Questionnaire 3

Teacher	Approach	$k = 5$	Descriptive statistics	
			Mean /6	Median /6
All Ts	TQ1 French teaching		4.2	4.00
	TQ2 Novel		4.0	4.00
	TQ2 Standard		3.3	3.00
	TQ3 Novel		3.8	4.00
	TQ3 Standard		3.2	3.00

7.6.3 Summary of results from teacher questionnaires

Teachers broadly enjoyed teaching in this project more than their previous French teaching, reporting higher levels of enjoyment in Standard than Novel, except for T3. Similarly,

involvement in teaching appeared to decrease levels of reported tension for most teachers. Overall, teachers were less tense post-teaching than in previous French teaching and experienced less tension when teaching Standard than Novel. There were some discrepancies between post-lesson evaluations and TQs, notably for T1.

7.7 Teacher responses to each approach: discussion of teacher questionnaire results

Discussion of TQ results for enjoyment and tension now follows.

7.7.1 Teacher enjoyment

The general increases in enjoyment ratings across the year for T2, T7, T8 and T9 indicated that experience of teaching French increased teachers' positive emotional responses. This was possibly due to the development of their French proficiency, resulting from regular teaching of the language, as found in Dewaele and McIntyre's statistical analysis (2014). Having fully resourced, progressive plans for the year with additional support available may have impacted positively on their levels of confidence and security, thus increasing enjoyment (as identified pre-project, [7.2.2.3](#)). However, T1 and T3 reported lower levels of enjoyment during the study compared with previous French teaching: these teachers were possibly adversely affected by their concerns about fidelity to treatment ([7.8.2.5](#)). This was not due to any constraint imposed by either scheme, rather an impact resulting from their desire to deliver content as intended.

TQs were completed only twice during the project, which may explain the discrepancy in T1's results (a preference emerged for Standard in TQ3, contradicting the post-lesson evaluations), as teachers' emotions can vary according to context (Frenzel et al., 2015, 2016). TQs possibly also provided a measure of *trait-like* rather than *state-like* emotions (Frenzel, Goetz, Stephens, et al., 2009) – see [4.3.10.2](#).

7.7.2 Teacher tension

T2, T7 and T8 all experienced less tension during the project than in previous French teaching. Given their low levels of linguistic competence in French ([Table 47](#)), they possibly initially considered themselves insecure in their competence to teach it ([Figure 56](#)), which may have caused anxiety (see Frenzel, Goetz, Stephens, et al., 2009; Woolhouse et al., 2013). In Woolhouse et al.'s case-study project, teachers' experience of CPD designed to improve both subject knowledge and pedagogical competence was linked to increases in confidence in French teaching; in Rowe et al.'s (2012) questionnaire-study, non-specialist trainee teachers with classroom opportunities to implement their French training were more confident than those

without. In the current study, although teachers were not following CPD, the detailed lesson plans and reflection on pupil outcomes may have produced similar effects (and were considered to impact positively on their subject knowledge, [7.8.2.4](#)). The teaching experience may have boosted confidence levels (as found by Cable et al.'s national survey [2010], Cooke & Faez's mixed-methods study [2018], and Muijs et al.'s questionnaire and case-study project [2005]), thus decreasing tension.

This was not the case for T1 and T3, who reported higher tension in the project than previous French teaching, possibly due to their concerns over fidelity to treatment. This level of control was an artefact of the experimental design rather than a characteristic of either scheme. T3's original starting points for tension were low ([Figure 58](#)), with their tension levels teaching French lower than for general primary teaching ([Table 47](#)).

7.7.3 Discussion of results from teacher questionnaires: summary

Overall, teachers seemed to find participation in the project a positive experience, particularly when teaching Standard. Generally, enjoyment levels increased, and tension decreased, for both approaches throughout the project. The snapshot nature of the TQs possibly explains occasional discrepancies between TQs and post-lesson evaluations, highlighting the role of the teaching context in influencing enjoyment. Tension appeared to increase through involvement in the project for some teachers, possibly because of the demands of ensuring fidelity to treatment, but generally practitioners reported less tension this year than in previous teaching, and less gradually over the year of the project itself when considering the teachers as a group, particularly in Standard. This was perhaps due to the impact of regular French teaching on their subject knowledge, providing a CPD effect, especially as teachers had previously reported low confidence related to (perceived) insufficient subject knowledge at Interview1.

7.8 Teacher responses to each approach: results from teacher interviews (Interview2/Interview3)

Having considered the quantitative data for teachers, this section addresses qualitative findings. Six teachers participated in semi-structured individual interviews mid- (Interview2) and post-teaching (Interview3). Additionally, T4 completed Interview2 before leaving the school (December, 2022). For each approach, responses contributed to findings on the *perceived* effect on pupils'/teachers' affective variables (RQ2a/RQ3a) and pupil progress in French (RQ2b), and on teachers' developing subject knowledge (RQ3b). This section presents findings relating to teachers' views on pupil progress and the usefulness of the French taught, before exploring the

impact of each approach on teacher emotions. As explained in [4.5.4.4](#), qualitative data have sometimes been quantified by counting non-numeric data after extraction of themes.

7.8.1 Results: teacher perceptions of pupils' outcomes

This section presents qualitative findings related to teachers' opinions on pupil progress and the usefulness of the language taught, addressing RQ2b.

7.8.1.1 Interview2

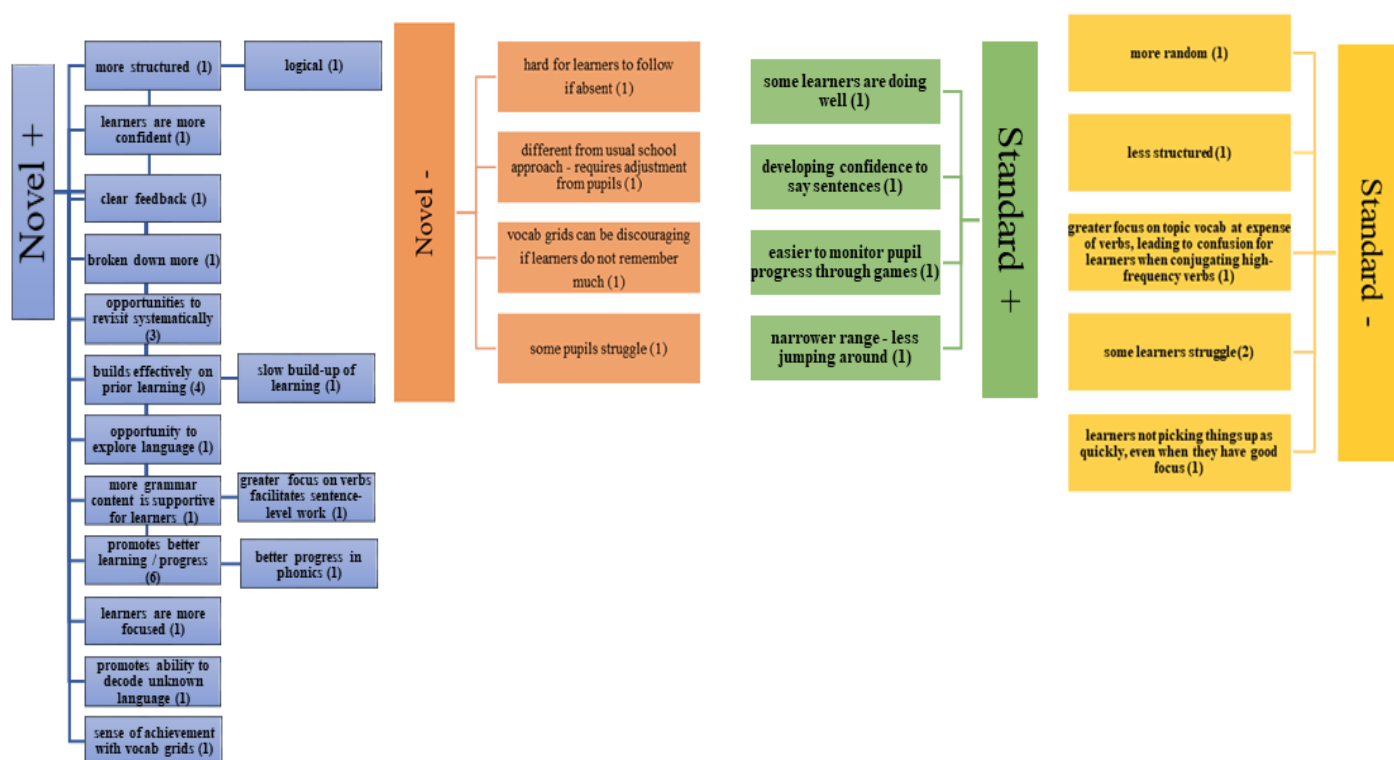
Mid-project, all teachers felt that progress was apparent in both schemes: “[pupils] seem to have generally made a lot of progress in each” (T1). T9 thought that games in Standard made it “easier to monitor the progress”. When asked about learners' progress, there were 18 positive comments made by five teachers on progress in Novel, compared to one teacher's single reference to progress by some learners in Standard ([Appendix W.1](#); summary, [Figure 61](#)). There were repeated references to opportunities in Novel for revisiting language previously taught:

You revisit ... To begin with, I thought: ‘This is too much for the children ... They’ve got these six words, or eight words, now they’ve got another four’, but it’s amazing how they do ... it’s dripped in sort of thing. They do pick it up quite quickly (T3).

Comments were also made about Novel's logical structure, building upon prior learning: “it's building upon their language skills a little bit better, whereas Standard ... can be ... a bit more random” (T2). Explanations for this progress appeared grounded in Novel's explicit focus on verbs and sentence structure, which was “helping the children. They seem to be able to do more than I thought they would, like, say whole sentences, for example” (T3). In contrast, learners in Standard, consisting of “more topic-y vocabulary”, were “muddled up between ‘avoir’ and ‘être’” (T3).

Figure 61.

Infographic Summarising Teachers' Positive (+) and Negative (-) Comments about Pupil Progress in Each Scheme (Interview2)



Note. The numbers in brackets indicate the number of comments.

7.8.1.2 Interview3

As at Interview2, teachers generally reported good pupil progress across the year: *“I was really impressed with the progress all of them made actually with both ... they've really been ... pushed but were able to do it”* (T3). Perceptions of progress remained more numerous in Novel than Standard: 13 positive comments made by five teachers in Novel, compared with three comments made by three teachers in Standard ([Appendix W.2](#); summary, [Figure 62](#)). Three themes illustrating key differences were created from the data: (i) ‘A good grounding’; (ii) ‘Sentence-level achievements’; and (iii) ‘Surpassing expectations’.

(i) Theme: A good grounding

As T1 anticipated in Interview1 ([7.2.2.7](#)), Novel was perceived to provide better foundational knowledge in French than Standard, particularly in grammar and phonics, which would serve learners well in continued language learning (T1, T3, T9). This was recognised by T9, despite acknowledging their personal preference for teaching Standard: *“the Novel, obviously, I think gives you a better grounding”*.

(ii) Theme: Sentence-level achievements

Strengths in vocabulary knowledge were perceived in pupils in Standard (T3, T8), but also noted in Novel in terms of rapid acquisition (T8) and effective recall (T2). The difference in pupil progress between approaches was identified from five comments (T1, T3, T8) relating to learners' ability to work at sentence level in Novel (none in Standard). The explicit focus on grammar and relevant high-frequency vocabulary apparently equipped pupils to create spoken and written sentences:

... the frequency of words, you can apply it to the language, whereas just learning words is OK, but you're not learning the sentence structure ... as we got towards the end ... they were really good at learning the sentence structure. (T1)

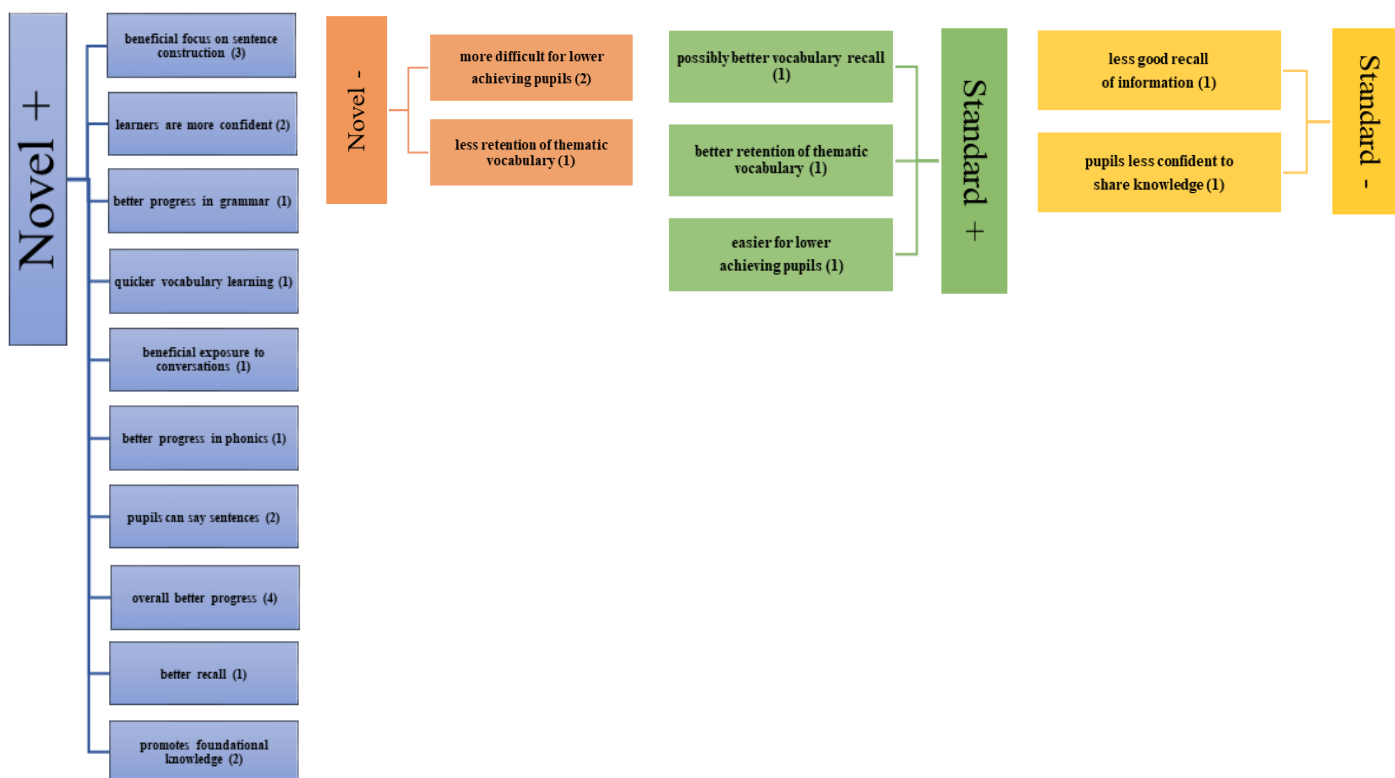
However, by the end of the teaching, some learners began to struggle in Novel (as some had in Standard - see [Figure 61](#)), noted by T3 and T9: *"I think the less able ones find the Standard approach easier than the less able ones in the Novel"* (T9).

(iii) Theme: Surpassing expectations

Teachers saw the benefits of *"regular teaching"* (T7) in *both* approaches: frequency of contact with French impacted positively on pupil progress (T3), enabling learners to exceed expectations based on teachers' experiences ([Appendix W.2](#) provides additional supporting comments). The only approach-specific comment in this theme was to learner progress in Novel, made by T8 (who preferred teaching Standard): *"their confidence with French, their understanding surpassed anything I would have expected for a Year 3 French lesson ... By the end, they were really, really confident in lessons"*.

Figure 62.

Infographic Summarising Teachers' Positive (+) and Negative (-) Comments about Pupil Progress in Each Scheme (Interview3)



Note. The numbers in brackets indicate the number of comments.

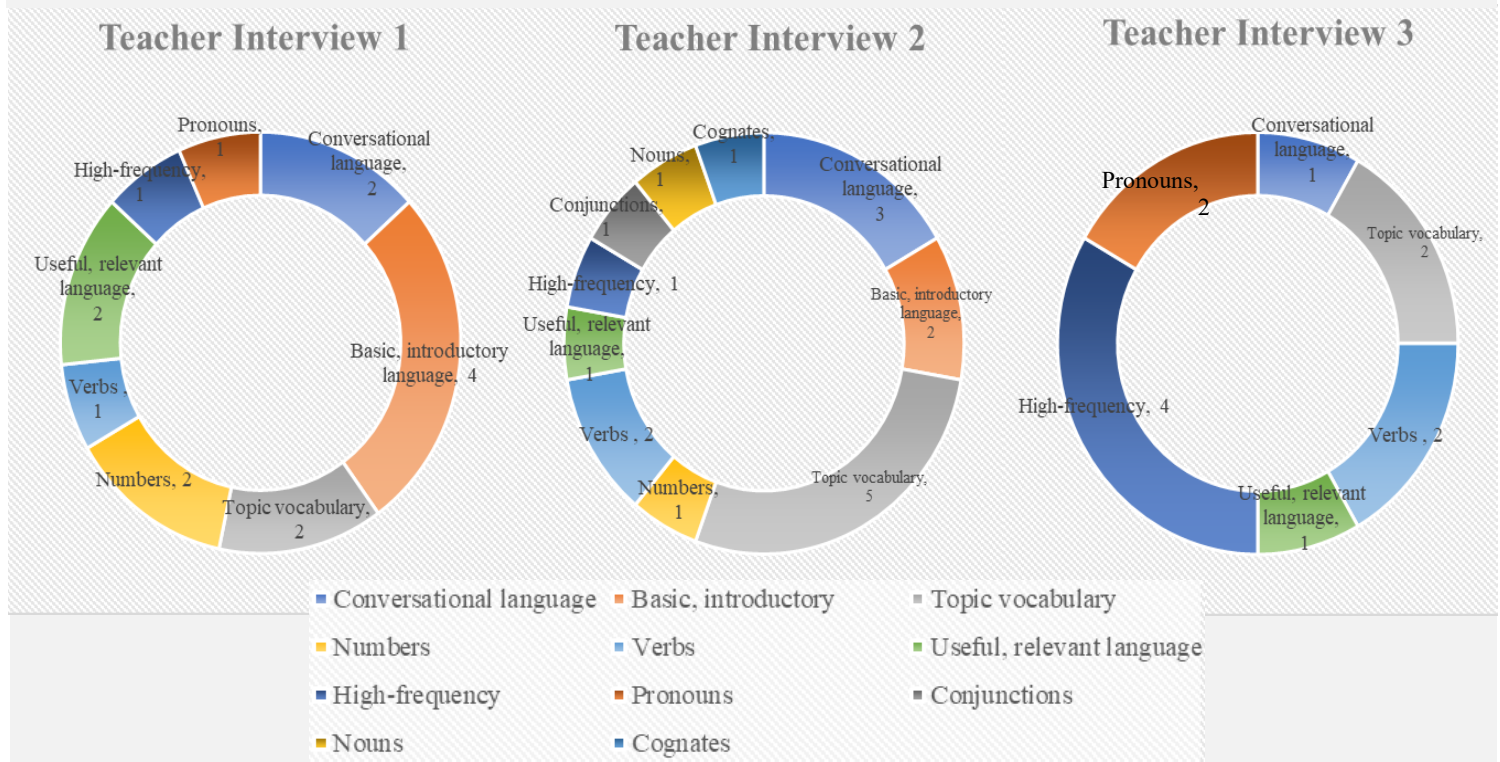
7.8.1.3 Results: teacher perceptions of linguistic usefulness

In all interviews, teachers were asked which words they considered important for children to learn. Non-approach-specific responses ([Figure 63](#)) were categorised using teachers' own terminology. The time-ordered meta-matrix ([Figure 64](#)) summarises teachers' perceptions across the academic year of the usefulness of French taught.

Figure 63.

Teacher Perceptions of Usefulness of French Taught (Interview1/2/3)

Teacher Perceptions of Usefulness of French Taught



Note. [Figure 63](#) indicates the number of teachers referring to each category on one or more occasions.

Figure 64.

Time-ordered Meta-matrix Summarising Individual Teacher Perceptions of the Usefulness of French Taught (Interview1, Interview2, Interview3)

	Interview1	Interview2	Interview3
Teacher 1:	Conversational language Colours Fruit Numbers Days of the week	Conversational language Greetings Colours Numbers Animals <i>"a really good basis to ... build upon, ... to move up the school and then on to secondary school"</i>	High-frequency vocabulary <i>"I think ... the Novel with the frequency of words because you can apply that"</i>
Teacher 2:	Basic language High-frequency – e.g., pronouns	Colours Animals Exceptions to rules	High-frequency vocabulary
Teacher 3:	Verbs, e.g., 'to be' Greetings, introductions Numbers <i>"things that interest them"</i>	Verbs, e.g. 'avoir', 'être' <i>"It's surprised me how quickly they're picking that up but ... if you're doing a particular topic, it's useful to have vocabulary for that particular topic ... they've enjoyed learning about animals for example"</i> Classroom language	Verbs: <i>"'avoir' and 'être' are the most important things, probably"</i> Pronouns <i>"... because ... if they continue, that'll [verbs and pronouns] be really useful to them where they might forget the vocabulary for"</i>

			<i>a caterpillar or whatever, mightn't they?"</i>
Teacher 4:	Transactional language useful for pupils: <i>"if they were to go to France, they could actually use the French that they have learned"</i>	Conversational language Greetings Classroom objects Days of the week Language <i>"relevant to them [pupils] on a day-to-day basis"</i>	
Teacher 7:	Useful transactional language: <i>"words that can help you get by if you were visiting that place"</i> Greetings, introductions	Verbs, e.g., 'it is' High-frequency language Conjunctions Nouns	Topics – e.g., animals, colours – using these in sentences High-frequency language – <i>"it gives them an understanding of the language ... rather than the sort of 'getting by in French', which is like the other approach [Standard]. It's kind of showing the reasoning for it"</i>
Teacher 8:	Everyday objects Greetings, introductions	Animals Classroom objects Cognates	<i>"The ones that they use every day, so, foods ... lots of them like that they know the foods, and I think that's quite important ..."</i> High-frequency language, verbs <i>"But especially, I think actually the way that Novel did approach the language and vocabulary. It was nice because ... it was the frequent words that were coming up ... learning how to say 'I', 'me' and 'my'"</i>
Teacher 9:	Conversational language	Conversational language	Conversational, transactional language

Note. T4 left the school before Interview3 was conducted. Colour-coded boxes indicate the approach with which teachers' comments broadly align (decided by the researcher): blue - Standard; orange - Novel; green – both.

(i) *Teacher Interview 1/Teacher Interview 2*

Considering linguistic usefulness, some teachers were influenced by personal experience as learners or teachers (particularly T1), with T9 explaining the difficulty in thinking differently about vocabulary (7.2.2.7), rather than *"in a fairly standard way all of the time"* (Interview1). This perhaps explains the five references to the usefulness of topic-based vocabulary such as colours, fruits, animals and days of the week at Interview2 (Figure 63). Using language for communication, personal introductions, and transactional purposes was considered important, perhaps while visiting France (*"you might want to go and order an ice-cream"*, T7, Interview1).

Aside from potential practical application, teachers prioritised relevant language. Perhaps because travelling to France was unlikely for some pupils (T2, Interview2), animal vocabulary featured in three teachers' (T1, T2, T8) suggestions of useful language, considered

“relatable to everyone, so it’s a bit easier” (T2). Teachers were possibly influenced by pupils’ clear enjoyment of such language, which *“obviously holds, attracts their attention”* (T1, Interview2).

There were indications at Interview1 and Interview2 (and also Interview3, discussed below) that teachers were beginning to consider the value of high-frequency language such as conjunctions (T7), pronouns (T2), and verbs (T3, T7).

(ii) Teacher Interview 3 compared with Teacher Interviews 1/2

[Figure 64](#) supports the assertion that, at Interview1 and Interview2, teachers generally considered topic-related and conversational language (characterising Standard) most important for these young learners. However, by Interview3, all teachers except T9 acknowledged the importance of high-frequency language, noting its general applicability (T1, T3, T8) and its support for understanding how language works (T7). For T8 and T9, what mattered was practical application of the language, whether *“holiday-making French”* (T9) or in the classroom:

... they’ll [pupils] ... say ‘Ooh, I’ve got this for break-time’, and they’ll say it in the French word. The verbs ... they were using them, and they were then able to say in ... an English lesson: ‘Oh, we need to ... ‘écouter’”. (T8)

7.8.2 Results from interviews: teachers’ emotions

Evaluating the impact of each approach on teachers’ affective variables when teaching French included considering their emotional responses. This section encompasses: teachers’ perceptions of impact on teacher-pupil relationships; teacher enjoyment, enthusiasm and preferences for either scheme (RQ3a), including perceptions of pupil enjoyment (RQ2a); teachers’ perceptions of the impact on their subject knowledge (RQ3b); and feelings of worry and underconfidence, particularly related to pronunciation (RQ3a, RQ3b).

7.8.2.1 Results: perceived impact of teaching approach on teacher relationships with learners

This section contributes to an understanding of the perceived impact on pupils’ and teachers’ affective variables based on findings from Interview1 and Interview2.

(i) Teacher Interview 2

There was little difference in teachers’ perceptions of impact from each approach on teacher-pupil relationships: it *“depends on the lesson content”* (T1) or *“the make-up of the group”* (T3). Teaching in a different way (through Novel) meant that French lessons became a shared learning experience: *“we’re kind of learning together”* (T7).

(ii) Teacher Interview 3

Post-teaching, T2 and T7 reported no difference in impact of each approach on teacher-pupil relationships. T1 felt that in Novel pupils were “*more focused*”, with T9 noticing “*more participation*” in Novel initially. For T7 and T8, learning alongside pupils was positive, particularly in Novel: “*I felt I was just delivering it [Standard] to them, whereas with the Novel children I felt a bit more, that we were learning together*” (T8). According to T3, learners sensed “*a bit more fun*” in Standard than Novel by the end of the year.

7.8.2.2 Results: enjoyment and enthusiasm

This section addresses teachers’ personal enjoyment, then summarises perceptions of impact from each approach on learners’ positive engagement.

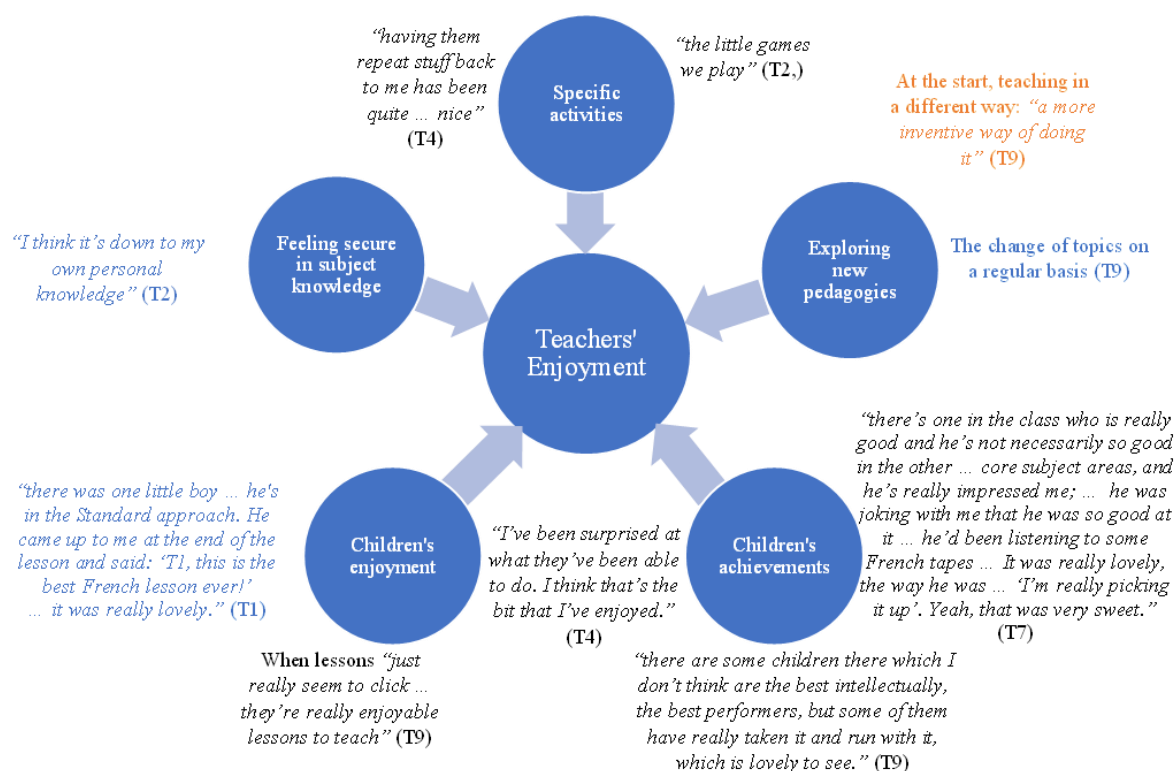
(i) Teachers’ enjoyment and enthusiasm

Interview2

All teachers were positive about teaching French: “*I’ve certainly enjoyed it*” (T3). This resulted from specific activities (games, T2; repetition, T4), as well as opportunities to familiarise themselves with different approaches ([Figure 65](#)). T1, T4 and T9 derived pleasure from learners’ positive reactions as well as their achievements. For T2, enjoyment came from confidence teaching Standard.

Figure 65.

Main Sources of Teacher Enjoyment (Interview2)



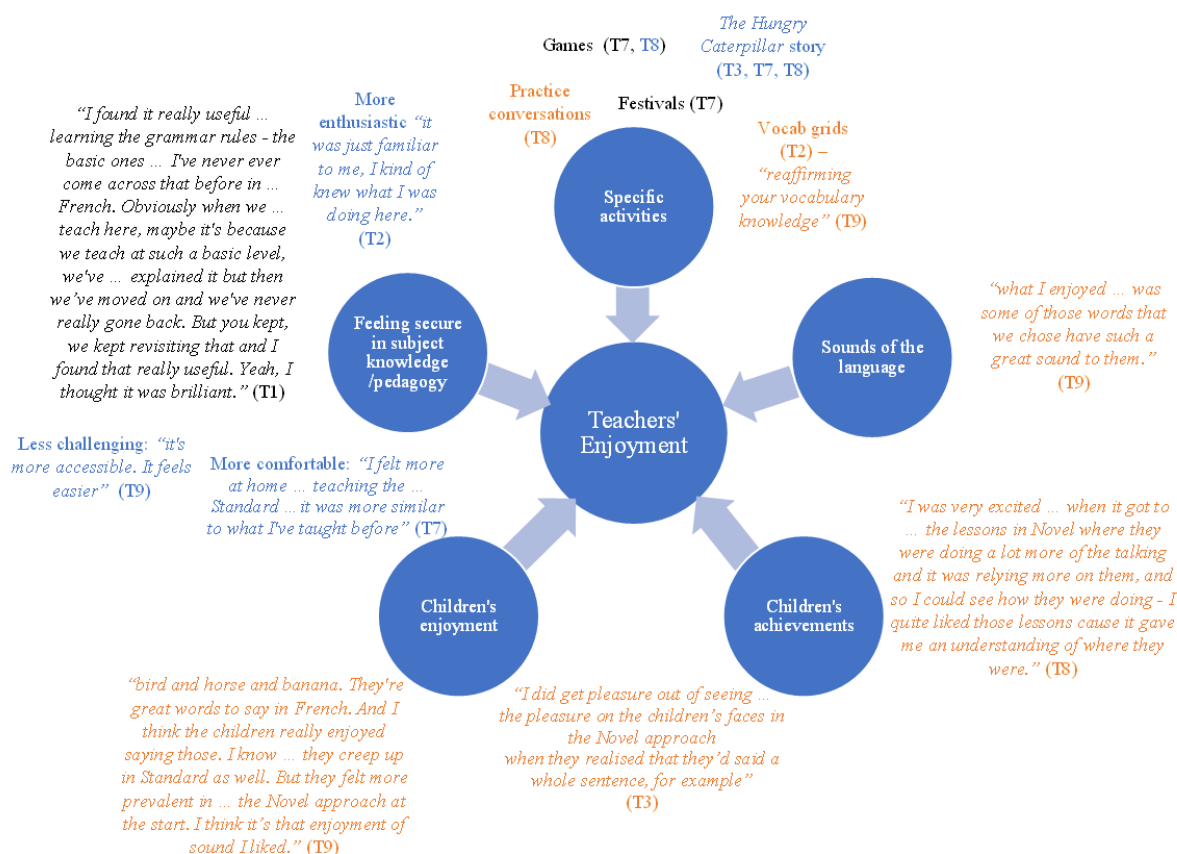
Note. Comments made by teachers with specific reference to Novel are orange and to Standard, blue; black indicates non-approach-specific comments.

Interview3

By Interview3, all teachers were still positive about both approaches: T3 and T1 enjoyed both similarly (*"definitely equally enthusiastic"*, T1); T2, T7, T8 and T9 enjoyed Standard more than Novel. The appeal of discovering new schemes apparent at Interview2 was replaced by enjoyment of the familiarity of the approach and content (Figure 66). T1 appreciated opportunities to improve their subject knowledge, but generally enjoyment came from feeling comfortable with content. Teachers still enjoyed particular activities, with the children's story *The Very Hungry Caterpillar* (Carle, 1969) in Standard being popular. T9 enjoyed the Novel focus on the words chosen to teach particular SSCs, enjoying the sounds and also the children's response to them. As at Interview2, teachers responded positively to aspects of lessons which appealed to learners and enabled them to demonstrate learning.

Figure 66.

Main Sources of Teacher Enjoyment (Interview3)



Note. Comments made by teachers with specific reference to Novel are orange and to Standard, blue; black indicates non-approach-specific comments.

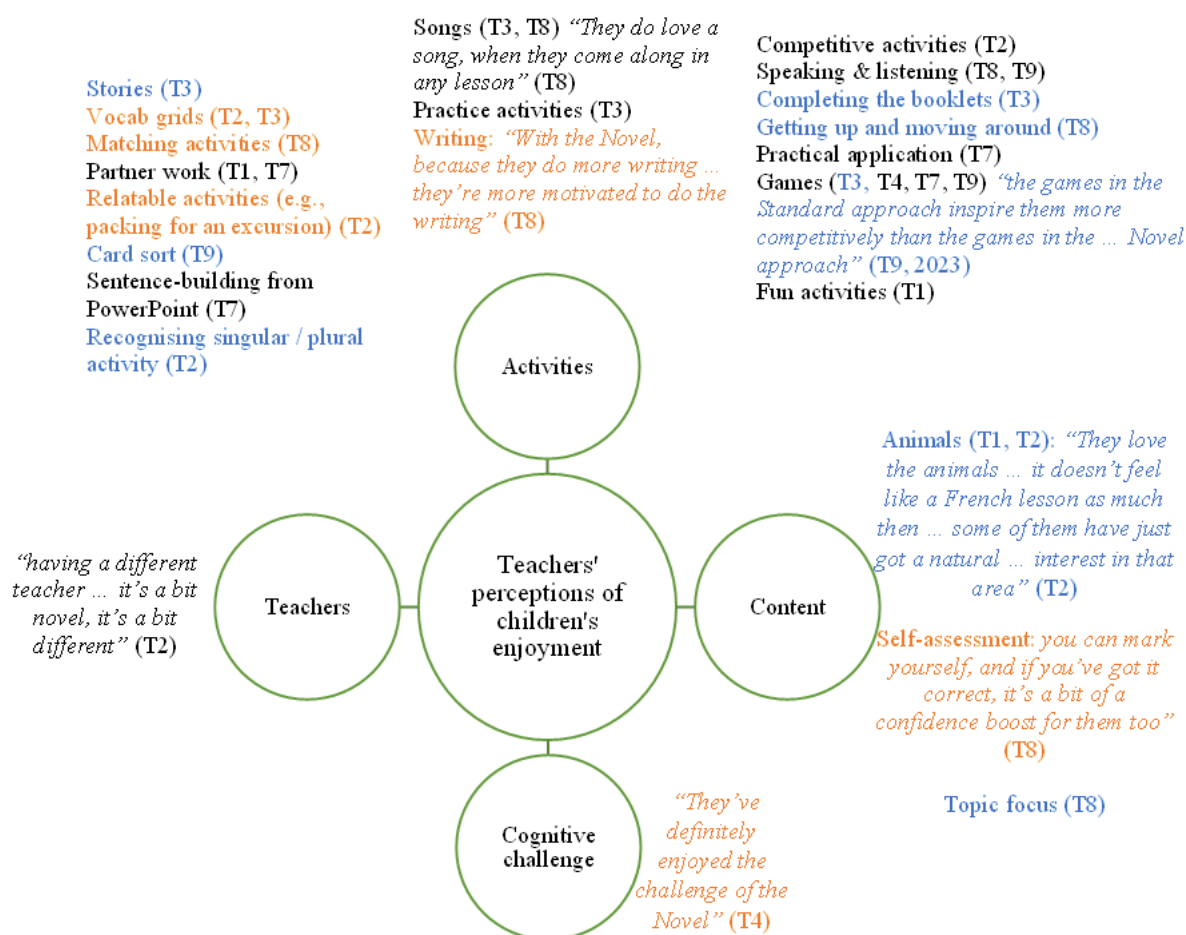
(ii) Teachers' perceptions of pupil enjoyment

Interview2

The most common perceived source of *pupil* enjoyment was specific activities, which – whilst having some commonalities - differed according to approach (Figure 67). Non-approach-specific activities such as pairwork, speaking and listening were considered enjoyable for all learners; other approach-specific tasks such as Novel vocabulary grids and accompanying booklets in Standard appeared to engage pupils. T3 and T8 felt that pupils enjoyed songs, but this was refuted by T4 (“they’ve not been that ... interested”); indeed, no pupils in either approach reported enjoying songs. T8 admitted that pupils’ responses may have been influenced by their own musicality: “it might be reliant on me”. Teachers reported the appeal of topics (e.g., animals) in Standard, and the positive challenge of high expectations in Novel.

Figure 67.

Teachers' Perceptions of Sources of Pupil Enjoyment (Interview2)



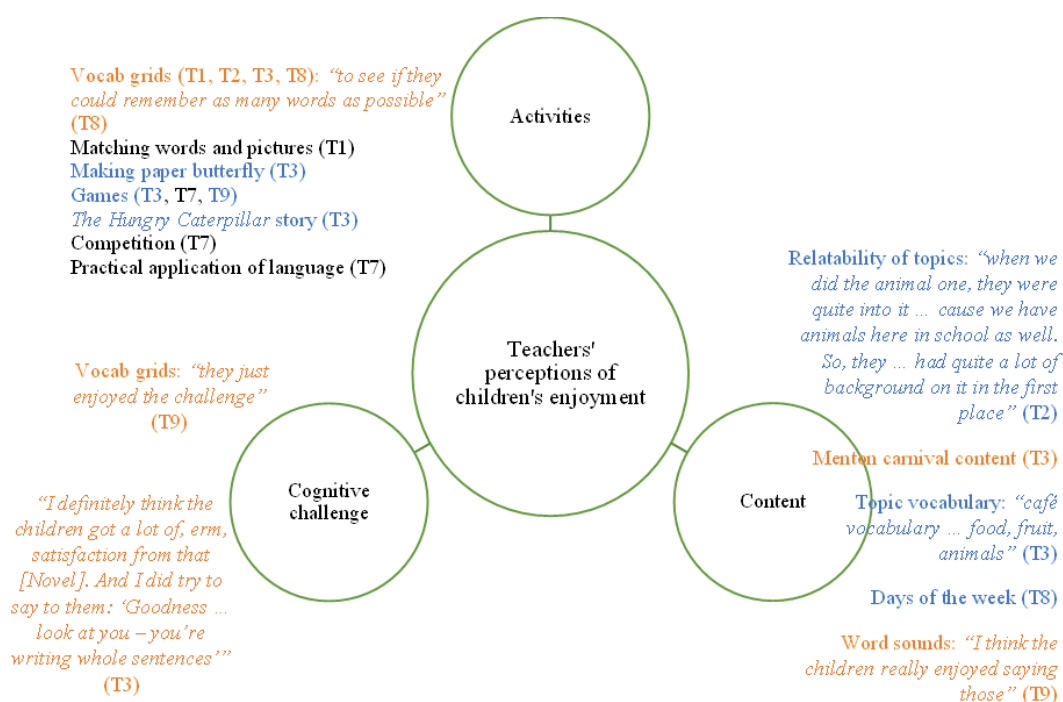
Note. Novel - orange; Standard - blue; black - non-approach-specific references.

Interview3

Specific activities remained a reported source of enjoyment, although less so than at Interview2 (Figure 68). Teachers (T1, T2, T3, T8) mentioned greater popularity of Novel vocabulary grids among learners than at Interview2. Pupils were considered positive about almost all topic vocabulary (days of the week, animals, food). The challenge posed by Novel was still apparently enjoyable, although not for all learners, as observed by one teacher: *"children that were struggling a bit ... perhaps the less able, SEN, etc, lost interest a bit in the Novel, towards the end"* (T3).

Figure 68.

Teachers' Perceptions of Sources of Pupil Enjoyment (Interview3)



Note. Novel approach - orange; Standard - blue; black - non-approach-specific references.

7.8.2.3 Results: teachers' preferences for schemes

(i) Teacher Interview 2

Preferences across approaches were evenly distributed ([Figure 69](#)): T1, T7 and T9 expressed no preference; T3 (who taught three classes) and T4 preferred Novel, due to its positive impact on pupil progress; T2 and T8 preferred Standard, considering it familiar and easier to teach.

(ii) Teacher Interview 3

All teachers except T3 expressed overall preferences for Standard, whilst acknowledging that they enjoyed Novel ([Figure 70](#)). Teachers found it familiar (T7), easier to teach (T2), with appealing content and a gamified approach (T9); Novel's grammatical focus appealed less (T2, T7, T9).

Figure 69.

Teachers' Preferences for Schemes (Interview2)

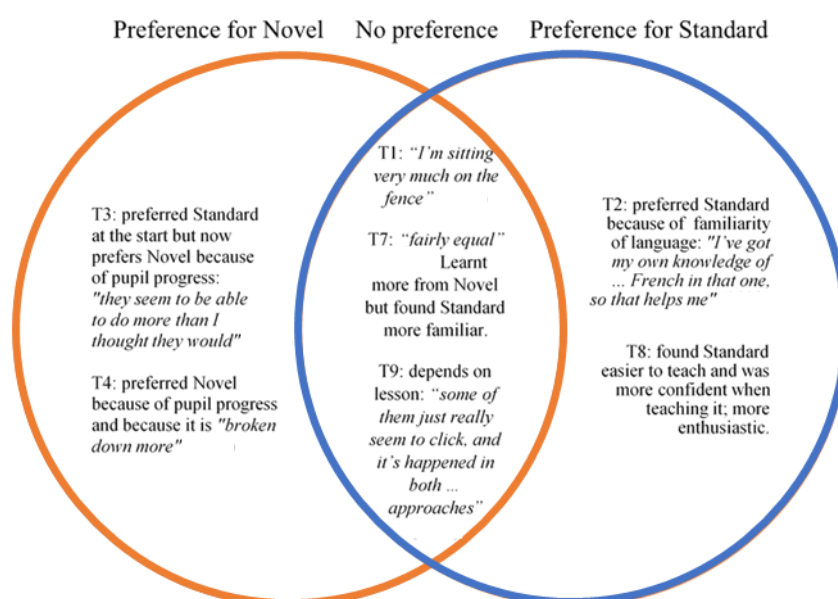
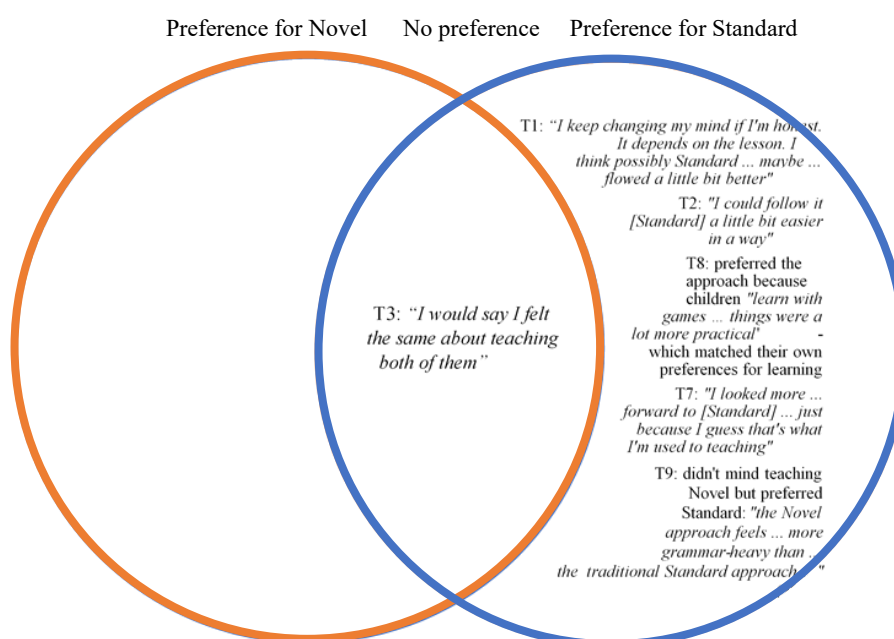


Figure 70.

Teachers' Preferences for Schemes (Interview3)



7.8.2.4 Results: teachers' perceptions of the impact of each approach on subject knowledge

(i) Teacher Interview 2

Overall preferences for the two schemes (described just above) appeared unrelated to teachers' perceptions of the impact from each scheme on their subject knowledge ([Figure 71](#)). While four teachers acknowledged positive impacts from both schemes, and T9 developed greater understanding of grammatical gender through teaching Standard, all seven teachers recognised that their subject knowledge had improved through Novel, whether generally or in specific areas (grammar, SSCs, retention).

Figure 71.

Teachers' Perceptions of the Positive Impact of Each Approach on Subject Knowledge (Interview2)

Impact on subject knowledge from Novel	Impact from both	Impact on subject knowledge from Standard
Reinforced (T3)/improved grammatical understanding: <i>"just being able to see how the language works"</i> (T8); better understanding of sentence structure (T7)	T1 T2 T4 T7	Increased understanding of gender <i>"It just feels as if it's reinforced more in the Standard approach"</i> (T9)
Improved general understanding: <i>"I think I understand what I'm teaching more with Novel than Standard"</i> (T1)		Learnt new vocabulary (T4, T8)
Improved understanding of SSCs (T2, T7)		
Supports language retention: <i>"more, like, precision ... [in lesson plans]. I guess because of the grammar"</i> (T4)		
Greater impact initially: <i>"I noticed it most starkly with the Novel approach at the very start but that might be ... just because ... I hadn't taught French before"</i> (T9)		
Clarity of explanation in lesson plans which <i>"seem to explain it ... a bit more simply ... for me as a non-specialist teacher"</i> (T1)		

(ii) Teacher Interview 3

Despite the shift from balanced distribution to increased preferences for Standard ([Figure 70](#)), teachers again recognised greater impact on their subject knowledge from Novel than Standard ([Figure 72](#)). All teachers except T3 acknowledged impact from both, but references to increased understanding of all aspects of French related specifically to Novel, with impact on progression, vocabulary, SSCs, grammar and knowing how language works.

Figure 72.

Teachers' Perceptions of the Positive Impact of Each Approach on Subject Knowledge (Interview3)

Impact on subject knowledge from Novel	Impact from both	Impact on subject knowledge from Standard
Supported progression: <i>"from a progression point of view, the Novel has made me ... more secure"</i> (T1)	T1: <i>"from a subject knowledge point of view ... teaching both the schemes has helped me"</i>	
Vocabulary knowledge: <i>"because of ... the phonics"</i> (T1)	Improved phonics knowledge (T2)	
Improved general understanding of French: <i>"I think in terms of my own kind of CPD ... I probably benefited more off the Novel approach"</i> (T2); <i>"there was quite a lot of things that ... I was being refreshed on myself"</i> (T3); <i>"probably I learned more in N ... the S was kind of more relatable to what I normally do"</i> (T7); <i>"N has probably given me more of the subject knowledge"</i> (T7); <i>"Novel definitely ... taught me a lot more French than I had before"</i> (T8)	Increased vocabulary knowledge (T7); <i>"my vocabulary from both has bloomed massively"</i> (T8) Sentence-building (T7) Grammatical understanding (e.g., gender) (T7) T9	
Increased knowledge about language: <i>"it just helped me appreciate language more ... I think teaching-wise ... Novel has definitely ... helped my own understanding of French a lot better than the ... Standard"</i> (T2)		
Preparation for teaching grammar: <i>"when I was posed questions by the children, I felt more confident I could answer"</i> (T2); T3. Better understanding of sentence structure: <i>"it's definitely helped me understand how to structure sentences"</i> (T8)		
Logical structure (T2)		
Improved understanding of SSCs (T3) and <i>"how to form words"</i> (T8)		

7.8.2.5 Results: teachers' general feelings of worry

The focus on teachers' emotional responses to each approach included general feelings of worry and tension, now explored for Interview2 and Interview3.

(i) Teacher Interview 2

Teachers' main areas of concern were represented by five themes, constructed from Interview2 comments and presented below, with supporting quotations in [Figure 73](#). Although these are representative quotations, more worries were associated with Novel than Standard.

Figure 73.

Causes of Teachers' Feelings of Worry (Interview2)



Note. Blue text denotes worry related to Standard, orange to Novel, and black are not approach-specific (as categorised by teachers).

Theme 1: 'Doing right by them'

T1 wanted to ensure a positive experience of French for learners and not let them down (non-specific to approach).

Theme 2: 'Doing it in the right way'

Delivering the schemes as intended (“*making sure, wanting to get it right*”, T4) created a sense of worry (T3, T4) - more pronounced in Novel (T7) because it differed from previous French schemes.

Theme 3: 'Doing it right myself'

General worries about making mistakes and teaching the wrong thing permeated responses from T1, T2, T7 and T8 (though all with the highest tension ratings *pre-project*, [Table 47](#)). This theme encompasses references to a general desire to get French right as well as subject-knowledge-specific concerns about making mistakes.

Theme 4: 'Managing the class'

Teaching unknown pupils in unfamiliar contexts was a particular concern for T3, the visiting teacher.

Theme 5: 'Out of my depth now'

Feeling insecure due to insufficient subject knowledge was a common experience for all teachers except T3 (with the lowest tension rating *pre-project*, [Table 47](#)). These feelings related to three areas of linguistic competence: remembering words (T7), understanding and explaining grammar (T7, T8, T9), and SSCs/pronunciation (T1, T2, T4, T7, T8, T9). The greater grammatical focus in Novel than Standard meant that more concerns were expressed about subject knowledge in Novel. Assessing learners was considered challenging in both approaches due to insecure subject knowledge (T8).

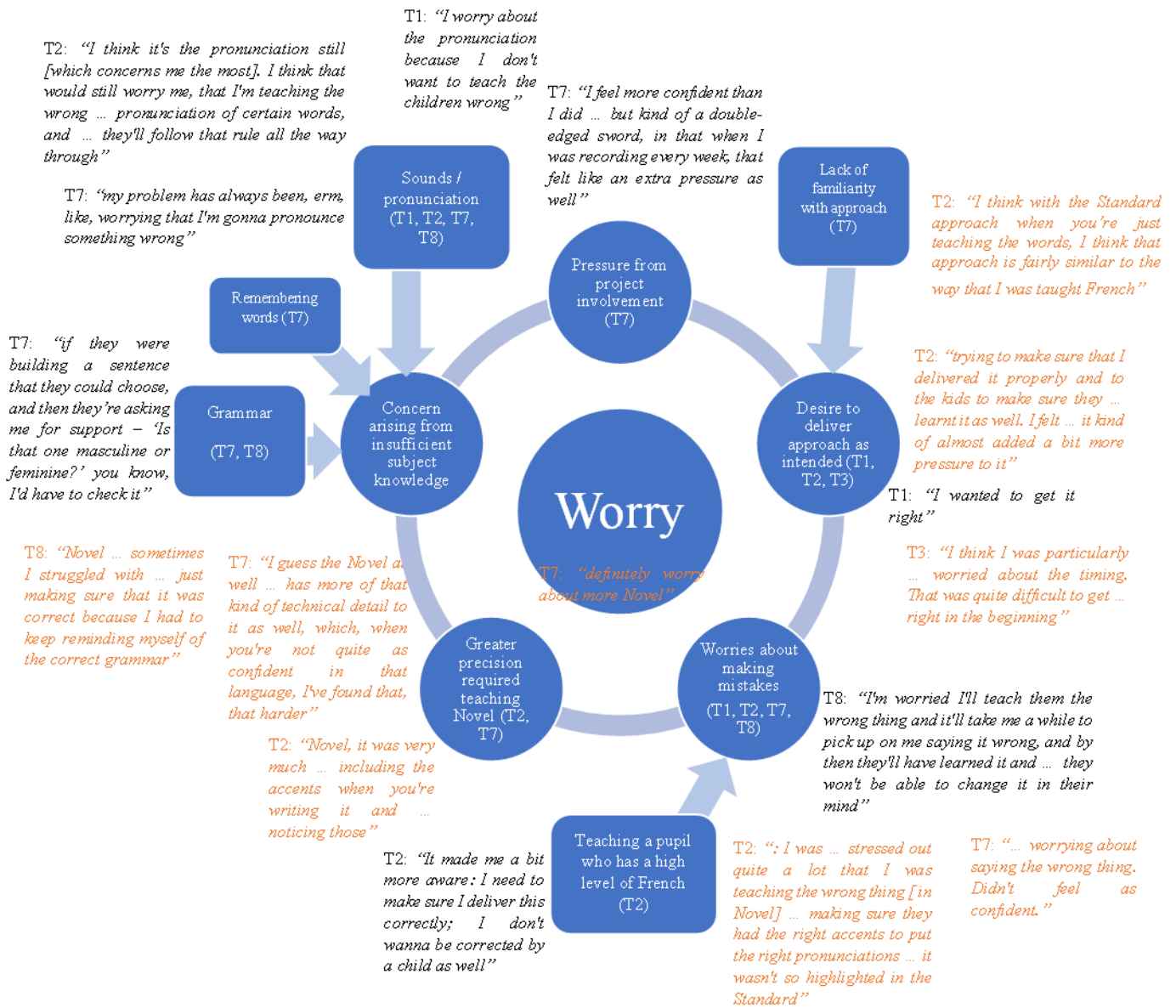
(ii) Teacher Interview 3

Similar concerns were expressed by teachers at Interview3, with illustrative quotations included in [Figure 74](#). Overall, Novel appeared to provide more causes for concern than Standard.

T1, T2 and T3 continued to express concerns about fidelity to treatment (Theme 2: ‘Doing it in the right way’), particularly given Novel’s unfamiliarity (T7). T1, T2, T7 and T8 still worried about teaching pupils the wrong thing (Theme 3: ‘Doing it right myself’). Novel appeared to demand greater levels of precision and attention to detail, causing teachers (T2, T7) to worry about delivery. All teachers except T3 and T9 still expressed insecurity about subject knowledge (Theme 5: ‘Out of my depth now’), related to the same three areas as at Interview2: remembering words (T7), understanding and explaining grammar (T7, T8), and SSCs/pronunciation (T1, T2, T7, T8).

Figure 74.

Causes of Teachers' Feelings of Worry (Interview3)



Note. Orange text relates to worry associated with Novel; black - non-approach-specific worry as categorised by teachers.

7.8.2.6 Results: teachers' overall and aspect-specific references to (under)confidence

As non-specialists, much of teachers' emotional response related to feelings of (under)confidence when teaching French. This section provides brief context for overall confidence, before examining *aspect-specific* comments.

(i) *Teacher Interview 2*

Confidence increased with experience of teaching this project for the four teachers with highest tension levels pre-project: *“I’m the most confident that I’ve felt with French ... I’ve learnt a lot”* (T7) ([Appendix W.3](#) provides additional supporting quotations). Teachers’ views on which approach increased their confidence *generally* are collated in [Figure 75](#).

Familiarity with content increased T7’s confidence: Standard, topic-based and similar to previous approaches they had taught, was therefore less daunting. Teaching Novel involved doing things differently (T3); ensuring correct delivery therefore engendered some anxiety. T9 benefited from the detailed support that both schemes offered. For T1 and T8, confidence vacillated according to the lesson, rather than approach. Where teachers expressed greater confidence in one approach, this was always qualified by *“slightly”*.

Figure 75.

Teachers’ Confidence According to Approach (Interview2)

More confident teaching Novel	No difference	More confident teaching Standard
T8: <i>“[in Standard] we’re getting into kinds of things that I’m not as familiar with, so presently I’m slightly more confident with the content of Novel”</i>	T1: <i>“sometimes I think, yes, I find it easier with the ... Novel approach, but then the same will happen with the Standard”</i> T9: <i>“I know ... I’m going to get the support from the slides”</i>	T3: <i>“[Novel] that’s slightly more challenging ... possibly because of the grammar focus... And, you know, making sure, wanting to get it right”</i> T7: <i>“the topic one ... feels ... maybe more familiar ... so that’s possibly why I feel slightly maybe more confident with that one”</i> T8: <i>“at the moment, I’m finding Standard slightly easier ... when I do those lessons, I feel myself more confident ... I’m still happy with Novel ... but sometimes I don’t know, I get a bit frazzled when I teach that one, I’m not sure why yet”</i>

Teachers’ concerns often arose from self-perceived insufficient subject knowledge. When asked which approach increased their confidence to accomplish certain tasks with learners (explain vocabulary, support with grammar/writing/simple sentence-building, correct pronunciation errors and create a task based on SSCs), teachers reported feeling slightly better

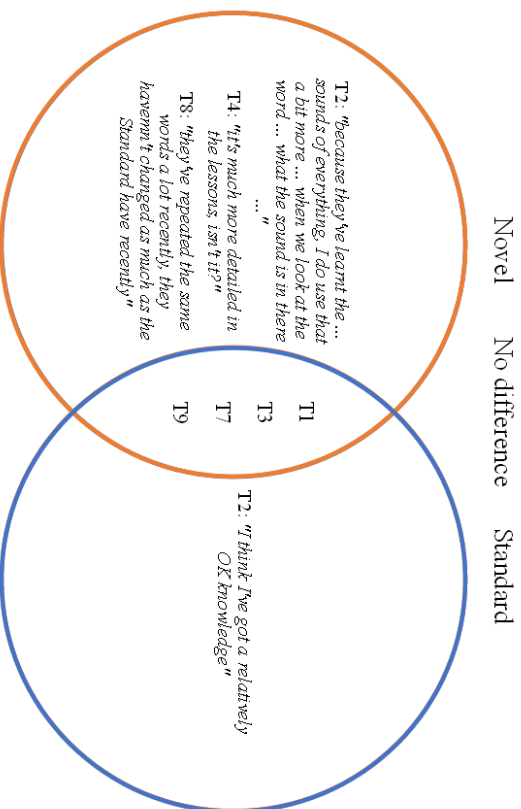
supported in Novel ([Figure 76](#)). This increased confidence appeared to stem from Novel's: progressive, consistent nature; detailed planning; frequent opportunities to revisit and build on prior learning. T2's reference to the use of sounds when explaining vocabulary in Novel indicated its holistic approach, interweaving phonics, vocabulary and grammar.

Figure 76.

Teachers' Feelings of Confidence Completing Certain Tasks in Each Approach (Interview2)

Which Approach Makes You Feel Most Confident to ...? (Interview2)

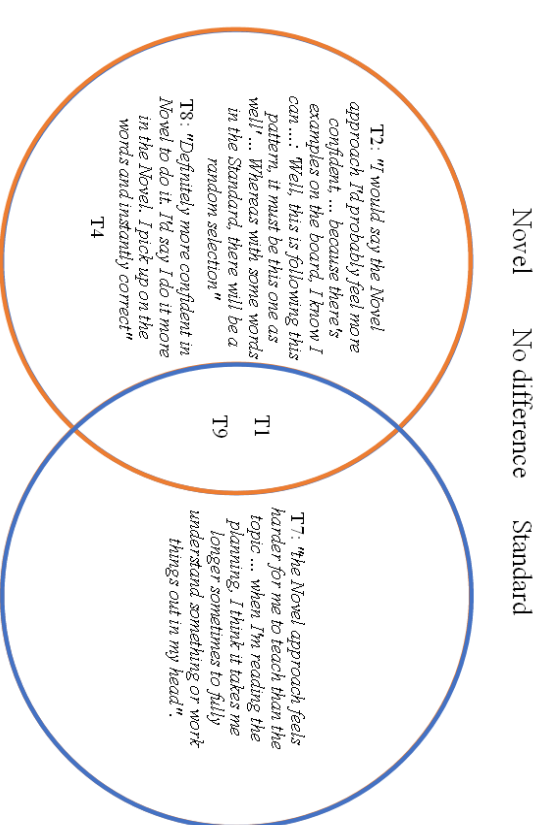
Explain Vocabulary



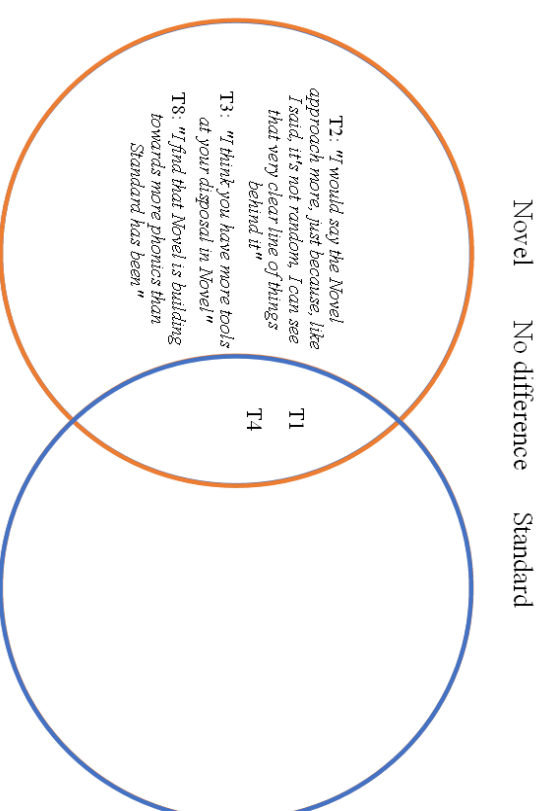
Support with Grammar, Writing and Simple Sentence-building



Correct Pronunciation Errors



Support with Symbol-Sound Correspondences



Note. If a teacher did not refer to a specific approach or explicitly to there being no difference between approaches, they are not included. If they responded indicating explicitly that there was no difference in how each approach affected their confidence, they appear in the diagram overlap.

(ii) *Teacher Interview 3*

Teaching French regularly continued to impact positively on confidence (T7, T8 and T9):

I'm probably the most confident I've ever been in French ... this year. I'm not saying that I'm good at it. But ... I feel more confident than I did ... I feel better about French than I've ever done. (T7)

([Appendix W.3](#) provides additional supporting quotations). Teachers' views on which approach increased their *overall* confidence are collated in [Figure 77](#).

For T1, confidence varied according to “*what the content was*”. Familiarity of content and pedagogy continued to make Standard easier to teach (T2, T7, T9): “*I was a bit more at ease with it, I would say ... primarily because it was very similar to my own experience of learning French*” (T7); “*I looked more forward to [Standard], just because I guess that's what I'm used to teaching*” (T7). T9, more at ease teaching Standard, speculated that this could change if they were to re-teach Novel: “*It'd be quite interesting going back and doing the Novel again, all the way through to see if that changes ... having done ... half of it already ... I imagine it'd feel easier. That familiarity*”.

The number of teachers reporting no difference in confidence to explain vocabulary or correct pronunciation errors ([Figure 78](#)) in each approach increased slightly from six at Interview2 to nine at Interview3. The explicit phonics focus in Novel appeared to increase teachers' confidence (T1, T2, T3, T8) compared with Standard, whereas the heavier, explicit focus on grammar made T8 and T9 feel less confident in Novel.

Figure 77.

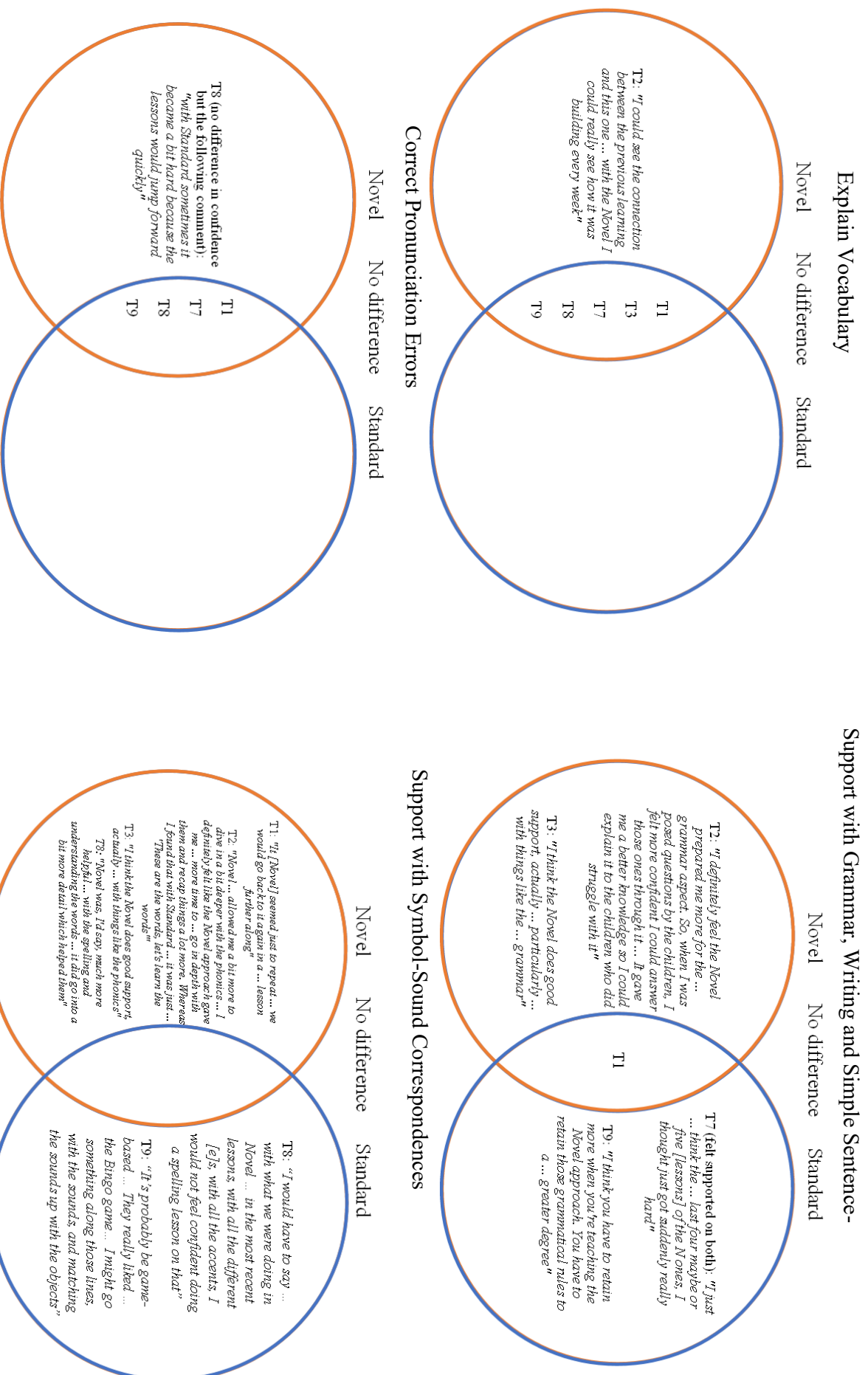
Teachers' Confidence According to Approach (Interview3)

More confident teaching Novel	No difference	More confident teaching Standard
	T3: <i>"I don't think [I felt more confident in one approach than the other], in all honesty. I think my subject knowledge was ... enough for ... that not to be an issue. But I could imagine it could be ... for others"</i>	T2: <i>"primarily because it was very similar to my own kind of experience of learning French ... it was a bit more familiar ... so I felt a bit more comfortable teaching it ... learning just those ... kind of standard ... words that I knew as well, that kind of put me a bit more at ease cause I ... knew it a bit more"</i> T7: <i>"I guess the Novel ... has more of that kind of technical detail to it as well, which, when you're not quite as confident in that language, I've found that ... harder ... I didn't feel as confident teaching the N. Definitely not"</i> T8: <i>"[Novel] felt more advanced than the Standard. I don't know if it was in any way, but ... it felt like I needed to be more secure in French to help them ... so, I had to brush up a bit whenever I was doing the Novel lesson. Make sure I knew"</i> T9: <i>"[Standard] makes more sense to me quicker than looking at the Novel approach ... as someone who doesn't speak French at all, ... it's easier for me to access and because it's easier for me to access, I feel more comfortable delivering it ... it makes more sense to me. It probably hearkens back more to the way I was taught French"</i>

Figure 78.

Teachers' Feelings of Confidence Completing Certain Tasks in Each Approach (Interview3)

Which Approach Makes You Feel Most Confident to ...? (Interview3)



Note. If a teacher did not refer to a specific approach or explicitly to there being no difference between approaches, they are not included. If they responded indicating explicitly that there was no difference in how each approach affected their confidence, they appear in the diagram overlap.

7.8.2.7 Results: teachers' concerns with pronunciation

As teachers expressed concern about pronunciation, this section establishes a *general* context before addressing how teachers felt supported by each approach.

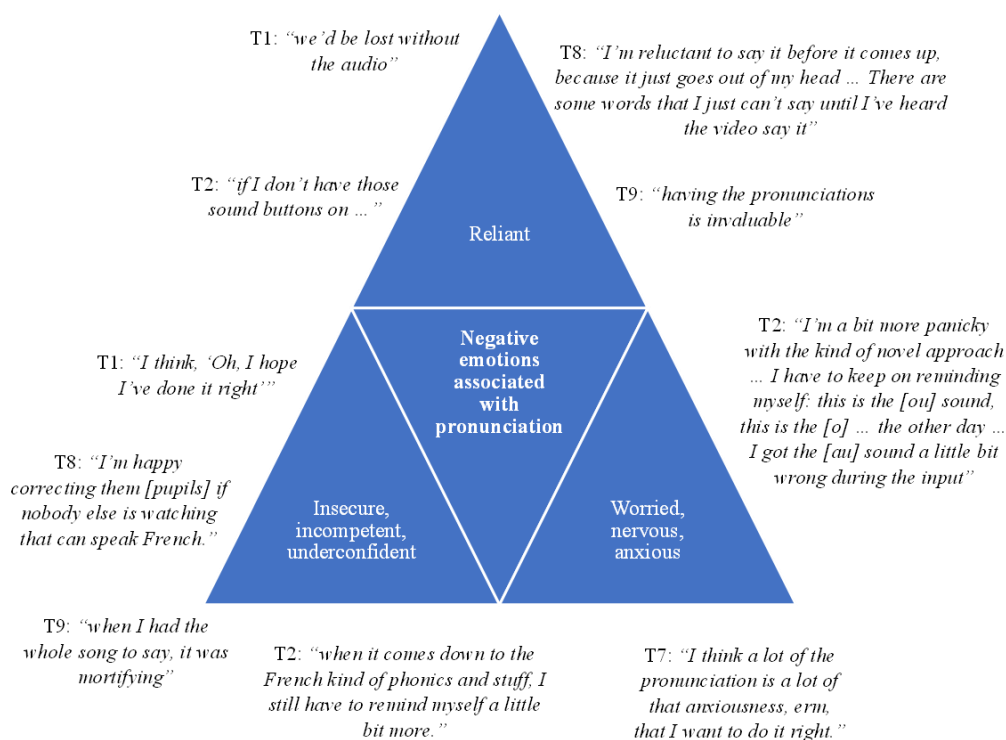
(i) Teachers' negative emotions associated with pronunciation

Interview2

Apart from T3 and T4, teachers felt insecure in their ability to pronounce French accurately and in phonics, highlighting their dependence on audio support (see also [Figure 73](#)). [Figure 79](#) summarises these negative emotions.

Figure 79.

Teachers' (Non-approach-specific) Negative Emotions Associated with Pronunciation (Interview2)

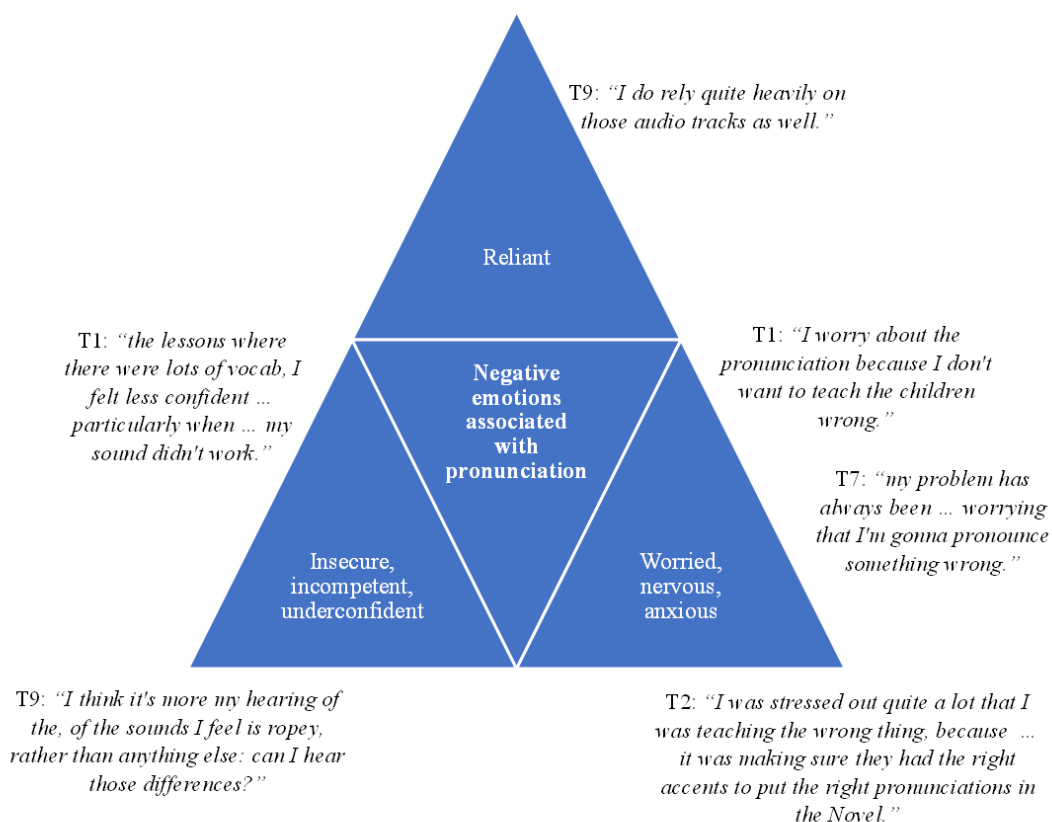


Interview3

Similar negative emotions (insecurity, dependence, anxiety) were apparent at Interview3 as at Interview2, although less numerous, with pronunciation still the most common cause of worry for these non-specialist teachers ([Figure 80](#)).

Figure 80.

Teachers' (Non-approach-specific) Negative Emotions Associated with Pronunciation (Interview3)



(ii) Approach-specific support for pronunciation

Interview2

Support was integral to both approaches, but was perhaps deemed more comprehensive in Novel, particularly because of its explicit teaching, regular practice and revision, which increased teachers' confidence to correct pupils' errors ([Figure 81](#)).

Figure 81.

Teachers Perceptions of Support for Pronunciation and Symbol-Sound Correspondences in Each Approach (Interview2)

Area of support	Support from Novel	Support from Standard
Audio	- audio (T1/T4/T8/T9) - visual showing sounds made by the mouth (T3)	- audio (T1/T4/T8/T9)
Explanation	- information on SSCs/vowel sounds (T3) - can refer back to information on "actual sounds" (T2)	- information on SSCs (T3) - enables learners to make connections between spoken/written word (T9)

	- explicit comparisons between French and English (T8)	
Pattern-spotting	- supports identification of patterns in SSCs (T2)	
Progression	- builds effectively on prior learning (T2)	
Practice	- regular reinforcement through games and activities (T3) - practice of SSCs (T4)	- practice of SSCs (T4)
Assessment & Feedback	- supports correction of learners' pronunciation errors: <i>"I pick up on the words and instantly correct"</i> (T8)	

Interview3

As at Interview2, Novel appeared to provide more pronunciation support than Standard ([Figure 82](#)), with regular revision and a progressive approach helping teachers identify and correct pupils' errors.

Figure 82.

Teachers' Perceptions of Support for Pronunciation and Symbol-Sound Correspondences in Each Approach (Interview3)

Area of support	Support offered by Novel	Support offered by Standard
Audio	- video support showing sounds being mouthed (T3) - audio; <i>"little videos"</i> (T9)	- audio (T9)
Explanation	- clear examples (T2) - explicit focus: <i>"more time to go in depth"</i> (T2) - detail increased understanding of sounds (T8)	
Revision	- opportunities to revisit: <i>"we would go back to it again in ... a lesson further along"</i> (T1); T2	
Progression	- builds upon prior learning (T2)	
Assessment & Feedback	- supports correction of learners' pronunciation errors: <i>"[in] Novel approach, I could correct those a little bit more than the Standard"</i> (T2)	
Practice		- practice through games (T9)

7.8.3 Summary of results: teacher responses to each approach

Teachers acknowledged pupil progress in both approaches, partly due to the regularity of French lessons. Qualitative comments supported the value of Novel, considered to provide a thorough

foundational knowledge, increase language awareness and support sentence-level work more effectively than Standard.

Teachers were affected by their own experience of language learning when considering the usefulness of the French taught. Language for communication was prioritised, as were interest and relevance to learners. Being exposed to Novel seemed to increase teachers' awareness of the value of high-frequency vocabulary.

Teachers did not report any particular differences in impact of either approach on teacher-pupil relationships. They derived personal enjoyment from specific activities and from children's enjoyment and achievements. Teachers reported enjoying both approaches, although greater enjoyment was evident in Standard than Novel post-teaching, partly due to their confidence and reported familiarity with content. This aligned with their expressed preferences for each scheme: they did not mind teaching either, but preferences were more numerous for Standard by Interview3. Standard was perceived as familiar and easier to teach than Novel, where the grammatical focus was less appealing, despite the positive comments about its usefulness and about pupil progression in making sentences.

Teachers identified activities enjoyed by pupils which were common to both approaches and also approach-specific. Pupils were perceived to be generally positive about the challenge posed by Novel, although this became, in some teachers' views, somewhat daunting for lower-achieving pupils by the end of the project.

Teachers consistently acknowledged the greater impact on their subject knowledge of Novel than of Standard in all areas of French. Their confidence improved overall through experience of regular teaching, although they reported higher levels of general confidence in Standard than Novel because of the latter's novelty and grammatical focus. However, they recognised that Novel increased their confidence in teaching specific areas of French. Teachers still worried about getting things wrong (generally; in lesson delivery; and through making mistakes due to insecure subject knowledge). They seemed particularly underconfident in pronunciation and SSCs, where Novel was perceived to offer more effective support.

7.9 Teacher responses to each approach: discussion of results from interviews

The results presented in [7.8](#) relating to teachers' perceptions of pupil outcomes and linguistic usefulness and their emotions are now discussed.

7.9.1 Discussion of teacher perceptions of pupils' outcomes

Teachers were positive about pupil progress in both approaches, due in part to the regularity and stability of lessons: time availability for FL lessons has repeatedly been noted as problematic in national surveys and reports (Holmes & Myles, 2019; OFSTED, 2005, 2008a; Tinsley & Board,

2017). Graham et al.'s (2017) study showed that primary-aged learners receiving 60 minutes of instruction each week significantly outperformed groups experiencing shorter French lessons on measures of vocabulary and grammar (see also preliminary findings from Kasprovicz, 2025). Similarly, Unsworth et al. (2015) found that amount of teaching time predicted scores in receptive grammar and vocabulary in pupils learning EFL. Teachers' commitment to this current study ensured that teaching time was protected (perhaps to a greater extent than usual provision).

Although teachers noted pupil progress in both approaches, the higher number of comments relating to Novel at both Interview2 and Interview3 suggested that aspects of this approach were considered to facilitate pupil learning more than Standard. This perception of progress is significant as Mitchell and Myles's (2019) longitudinal study of Year 3 learners of French warned of the danger of minimal progress impacting negatively on pupils' motivation and self-efficacy. Similarly, Martin (2012) reported decreased pupil engagement and frustration at their lack of progress. The aspects of Novel considered to facilitate this progress are now discussed.

7.9.1.1 Revisiting and improved recall

Teachers noticed improved recall in learners in Novel compared with Standard. Retrieval practice, characteristic of Novel, included systematic recapping of vocabulary, SSCs and grammatical concepts. Recapping was found by Roediger and Karpicke (2006) to impact positively on future retention when in test-form, rather than re-studying. In Novel, this took the form of pupils working without reference materials (e.g., when completing vocabulary grid challenges). Teachers' perceptions appear to support E. Peters's (2014) experimental study that found that frequency of occurrence of lexical items resulted in higher vocabulary recall scores. Regular use of vocabulary grids in Novel, a form of cumulative testing (Maie et al., 2025) which engaged learners in recognising and recalling words individually under test conditions, possibly also promoted recall. This cumulative approach included lexical items learnt both recently and previously. Such an approach was two or three times more effective (in receptive and recall respectively) than non-cumulative tests in Nakata et al.'s (2021) experimental study (see also Maie et al., 2025, which found particular benefits for lower-achieving learners).

Teachers also seemed to recognise the value of spaced learning conditions inherent in Novel, as recommended by R.A. Schmidt and Bjork (1992). The superiority of spaced over massed conditions was found by both S. Kim and Webb's (2022) meta-analysis ($g = .58$, 95% [CI .16, 1.00]) and Nakata and Suzuki (2019a), in which a spaced approach was found to be 1.60 times as effective as a massed approach in promoting retention of English-Japanese word pairs in a study of 133 university students.

7.9.1.2 Focus on verbs and sentence structure

Teachers frequently referred to the benefit in Novel of pupils working at sentence level, being taught and having opportunities to manipulate verbs. This aligns with Schmitt's (2008) recommendation that manipulating lexical items supports and promotes learning effectively. It appears to confirm that the Novel approach offers a different way of teaching French, contrasting with practice observed by OFSTED (2008a, 2011) (which noted little emphasis on sentence manipulation) and the observed prioritisation of oracy over literacy reported by Graham et al. (2014) in their longitudinal study measuring pupil outcomes and motivation levels across Years 5, 6 and 7. Teachers reported that learners in Novel demonstrated better verb knowledge and less reliance on memorised language and set phrases, contrasting observed practice in Cable et al.'s (2010) national survey. Novel may address issues with verb production as identified in upper-primary lessons in Courtney's (2017) longitudinal mixed methods case-study. Teachers' observations of sentence-level work in Novel indicate that even young learners can produce sentences using verbs (more commonly seen in 17-18-year-olds' responses in Marsden and David's analysis, 2008, of L2 French and Spanish learner language oral corpora).

T3's surprise at pupils' ability to cope with grammar content (*"there's more grammar content in the Novel and that's helping the children. They seem to be able to do more than I thought they would, like, say whole sentences, for example"*, Interview2) aligns with Lichtman's (2016) 'instructional hypothesis', which found that young learners can develop and apply grammatical understanding when instructed explicitly to focus on key forms.

7.9.1.3 Foundational knowledge

According to national reviews, schools have struggled to teach writing including grammatical understanding (Board & Tinsley, 2015; Tinsley & Board, 2012), particularly in provision by non-specialist teachers (Board & Tinsley, 2014). Pre-project, teachers recognised that teaching children how language works was necessary to *"set the foundation"* (T1) and support longer-term success (T2). By Interview3, this was evident according to T1's comment: *"I think possibly ... with the Novel approach they have learned a good foundation"*. Enabling teachers to provide such foundational linguistic competence in learners is important, because studies (e.g., Cable et al., 2010) have identified issues in grammatical accuracy among non-specialist teachers with limited linguistic proficiency and acknowledged the challenge for them in ensuring learners' progress beyond basic language (McLachlan, 2009).

7.9.1.4 Challenge

Teachers' identification of pupils' generally positive response to the challenge in Novel appears to support the concept of desirable difficulty (R. Bjork, 2017). Whilst experiencing success on

such tasks appeared beneficial, there were suggestions that for some learners the level of challenge was not optimal. Faced with failure, the challenge ceased to be desirable, and these learners began to disengage from the learning, noted by one teacher (T3).

7.9.2 Discussion of teacher perceptions of linguistic usefulness

Personal experiences of language learning influenced teachers' assessment of the usefulness of language for pupils, resulting in their references to topic-related vocabulary as the kind of language children should learn. Potential practical application of language was valued, with teachers keen for pupils to be able to use it in situations such as travel and personal introductions. Relevant language was also considered important by teachers, as were topics which learners enjoyed, such as animals. In this respect, teachers appeared to advocate that age and context should play a part in vocabulary selection (see Cobb & Laufer, 2021; Vilkaitė-Lozdienė & Schmitt, 2020; S. Webb & Nation, 2017), recognising that children "speak and read about different things than adults do" (Nation, 2016:9).

However, exposure to high-frequency vocabulary in Novel appeared to make teachers reassess the type of French language they deemed important for young learners. By Interview3, the importance of verbs was acknowledged (as stipulated in the National Curriculum, DfE, 2013a), but also function words (conjunctions, pronouns), in supporting learners' understanding of how language works and its application.

7.9.3 Discussion of teacher emotions

Findings relating to teachers' perceptions of impact on their relationships with pupils from, and their enjoyment and preferences for, either scheme (RQ3a) are now discussed. This section also includes discussion of: teacher perceptions of the impact on their subject knowledge (RQ3b) (7.9.3.4); their feelings of enjoyment (7.9.3.2, i), worry and underconfidence (7.9.3.5, 7.9.3.6), particularly related to pronunciation (RQ3a, RQ3b) (7.9.3.7); and teachers' perceptions of pupil enjoyment (RQ2a) (7.9.3.2, ii).

7.9.3.1 Discussion of the perceived impact of teaching approach on teacher-pupil relationships

Durable differences were not apparent in teachers' perceptions of the impact of each teaching approach on classroom relationships, with the latter fluctuating according to lesson content and class dynamic on the day. Du Plessis's (2015) qualitative investigation found that teachers teaching outside their field of expertise experienced less positive relationships with students than when secure in their subject knowledge, but in this current study teachers reported positive collaborative working as they learnt alongside learners (see Myles, 2017). This is important, as

Du Plessis et al. (2015:14) reported the view from a school principal that teachers feeling outside their “comfort zone” (which teachers acknowledged more frequently in Novel than Standard in this current study) were unlikely to form connections with pupils. However, it should be acknowledged that, with the exception of T3 and T9 (and to a lesser extent, T1 and T2), teachers were teaching pupils whom they taught on a regular basis in their general teaching.

7.9.3.2 Discussion of teacher and perceived pupil enjoyment

(i) Teacher enjoyment

Teachers’ overall responses to teaching both approaches were positive, with teachers taking pleasure in pupils’ enjoyment and achievements: “*I’ve been surprised at what they’ve been able to do. I think that’s the bit that I’ve enjoyed*” (T4, Interview2). The importance of teachers’ enjoyment was highlighted by Martin (2012) as, in their status as role-models, their enthusiasm may impact on learners’ positivity.

Security arising from familiarity of pedagogical approach and scheme content appeared to impact on teacher enjoyment: T2 enjoyed teaching Standard more as it was less grammar-focused and presented less linguistic challenge than Novel. Tierney and Gallastegi (2005) reported that the Scottish MLPS initiative had avoided including too much grammar as it was perceived to induce anxiety in *learners*; although this was not apparent among pupil participants in this study, as FL learners themselves, these *teachers* may have experienced similar emotions. The challenge in the innovative Novel approach is representative of complexity, an additional characteristic identified in innovations by Rogers (2003) (see [7.5.4](#)). The “degree to which an innovation is perceived as relatively difficult to understand and use” (Rogers, 2003:240), complexity can impact on individuals’ willingness to adopt innovation.

(ii) Teacher perceptions of pupil enjoyment

Specific activities

Teachers believed that pupils in both approaches enjoyed specific activities, such as pairwork, speaking and listening. This aligns with the desire expressed by older primary-aged learners to communicate with others through the FL reported by Cable et al. (2010), Chambers (2019), Courtney (2017), Courtney et al. (2017), and Graham et al. (2014, 2017). T9 also reported positive pupil responses to competition and games (in both approaches), supporting the suggestion in the *Key Stage 2 Framework for Languages* (DCSF, 2005) that young learners are motivated by games, and aligning with Finch et al.’s interview-study findings (2018:7), in which teachers reported positive pupil response to this “ludic dimension”. Other studies have suggested that by Year 5 and Year 6, older pupils responded less well to such activities, desiring greater linguistic challenge (Butler, 2017; Chambers, 2019; Fenyvesi, 2020). T3 and T8 described a

positive reaction from pupils towards songs (in both approaches), supporting pedagogical recommendations made by DCSF (2005) and Nguyen (2018), and matching widespread practice in the primary FL classroom (Courtney, 2017; Graham et al., 2014; Mitchell & Myles, 2019; Muijs et al., 2005), although this was contradicted by T4, who reported lack of pupil interest in songs. T3 described pupil enjoyment of the craft activity in Standard, involving making a paper butterfly following French instructions, which prompted some envious responses from pupils in Novel. This was part of a general feeling identified by T3 that pupils perhaps perceived Standard as more “*fun*” than Novel by the end of the project. This reflects observations made by Muijs et al. (2005:113) and Martin (2012:360) in references to the Pathfinders project that primary FL lessons have often tended to be a “light-hearted and fun experience”. However, these researchers went on to note that, in order to sustain enthusiasm, secure progression, and raise the status of the subject in learners’ minds, greater challenge was needed.

Experiencing success

As discussed in [7.9.1.1](#), pupils were seen to respond more positively to the vocabulary grids in Novel as the year progressed. Initially, the challenge was high: “*they didn’t like at first the grids where they had to write down ... sometimes it’s the French, sometimes it’s the English, and they have to get 15 points*” (T3, Interview2), but this became more manageable mid-project: “*they’re beginning to like those now, because they realise that they can get 15 points, quite a lot of them*” (T3, Interview2). This supports the importance of challenge advocated by Mitchell and Myles (2019), showing the perceived positive impact of desirable difficulty on learners and illustrating that motivation can result from satisfaction derived from intellectual challenge (Bandura, 1986; Deci & Ryan, 1985; Edelenbos & Kubanek, 2009; Ryan & Deci, 2000; Wu, 2003). However, as noted by one teacher (T3 – see [7.9.1.4](#)), too great a challenge resulted in learner disengagement, perhaps arising from the degraded short-term performance possible during the acquisition stage of challenging learning (R. Bjork, 2017; R.A. Schmidt & Bjork, 1992; Suzuki et al., 2019b).

Specific content

In terms of lexical items, teachers perceived positive responses from pupils to topic-based vocabulary, particularly animals. This is not to say that learners reacted negatively to Novel content, just that teachers reported noticing greater engagement with topic vocabulary (see discussion in section [7.9.2](#)).

7.9.3.3 Discussion of teachers’ preferences for schemes

At Interview2, teachers expressed preferences for Novel because of the perceived impact on pupil progress, and for Standard because of its familiarity and linguistic accessibility. By

Interview3, all teachers except T3 (who taught six different classes across three schools) preferred Standard for three main reasons: familiarity; a less explicit grammar focus; the gamified approach (each discussed below). T3 seemed the most positive about the Novel approach; it is possible that, teaching it more often than the other teachers (five lessons a fortnight compared to once a fortnight), they developed greater familiarity more quickly with this way of teaching, and it therefore appeared less daunting, less ‘novel’.

(i) Familiarity

Teachers (except T3) preferred Standard partly because it was a known entity, perhaps supporting Rogers’s (2003) reference to the importance of compatibility (see [7.5.4](#)). Torgerson and Torgerson (2008) explained that teachers can experience a sense of loyalty to existing treatments, although these teachers appeared willing to try something new ([7.2.2.7](#)). Introducing an innovative approach which has to “operationalise the learning theory ... as classroom practice” (Marsden, 2007:574) and yet differ sufficiently to be able to observe differential effects (Gueron, 2002) can be difficult, as too much difference can alienate teachers, but too much similarity makes it difficult to investigate the effects of innovation. FL pedagogy has historically been viewed as different from other subject areas (see Finch et al.’s interview-study, 2018), sometimes daunting (Board & Tinsley, 2014), and teachers had developed familiarity with Standard-type approaches prior to this project: for this reason, Standard’s familiarity may have appealed.

(ii) A less explicit grammar focus

Teachers’ enjoyment was negatively impacted by Novel’s grammar focus. Whilst acknowledging the positive effect of this on learner progress, it was challenging and contributed to teachers’ sense of worry and lack of confidence. If we consider that one contributory factor in teachers’ feelings of (in)security when teaching FL as outlined in Interview1 ([Figures 55, 56](#)) was a sense of subject knowledge adequacy, and specifically in grammar, pronunciation and writing, it is unsurprising that this security should be challenged by an approach which heightened their awareness of aspects of the language they did not know. Additionally, teachers were initially reassured by teaching at a very basic level, yet T8 referred to the fact that Year 3 learners in Novel approach ([7.8.1.2, iii](#)) “*surpassed anything ... expected*” for this age group - again challenging teachers’ linguistic security. Du Plessis (2015) reported that one consequence of teachers’ sense of reduced competence was the avoidance of certain concepts in lessons. The preference for Standard appears to reflect this, with teachers ‘voting with their feet’ to avoid the grammar content in Novel and opt for Standard, where they felt more at ease.

(iii) The gamified approach

For some teachers, the use of games appealed as it aligned with their preferred teaching approach. Initially, when FLs were first introduced in English primary schools, such gamification of learning set the subject apart from others in the curriculum and contributed to the sense of pedagogical unfamiliarity (see Finch et al., 2018), whereas this now appears accepted, even welcomed. However, what is perceived as ‘fun’ by teachers (e.g., songs) may not be viewed as such by learners.

7.9.3.4 Discussion of teachers’ perceptions of the impact of each approach on their developing subject knowledge

Teachers repeatedly acknowledged positive impacts from Novel on their understanding of progression, general language awareness, ability to support learners, and understanding of grammar, sentence construction and SSCs. Novel appeared to provide a form of CPD for teachers, addressing the lack of training specific to non-specialist teachers’ needs identified by Holmes and Myles (2019), notably in supporting them to plan for progression as required by the National Curriculum (DfE, 2013b) and in developing their understanding of grammar and the FL-L1 literacy relationship. Improving subject knowledge can impact positively on teacher confidence (found by Barnes, 2006; Finch et al., 2018; Woodgate-Jones, 2008). CPD can alter teachers’ views of their linguistic proficiency (Machida, 2016; Woolhouse et al., 2013), and perceived increased competence can be positively associated with decreased anxiety (Dewaele & MacIntyre’s questionnaire-study, 2014). Teachers in this current study possibly needed more experience of teaching Novel for their confidence to benefit from sustained exposure to more challenging subject knowledge.

7.9.3.5 Discussion of teachers’ general feelings of worry

Teachers demonstrated high levels of professionalism in wanting to deliver schemes as intended and not let learners down, but this contributed to their sense of worry, greater in Novel than Standard because of its novelty. Such commitment to doing their best for learners was found in Du Plessis’s (2015) qualitative investigation into the lived experiences of out-of-field teachers.

Teachers’ worries about insufficient subject knowledge causing them to teach the wrong thing aligned with findings from Legg’s (2013:60) interview-study, in which participants expressed concern that they were teaching learners “bad habits”. Teachers’ concerns in the areas of recalling vocabulary, pronunciation, grammar and assessment are now discussed considering previous literature.

(i) Vocabulary

In Tsui's case-studies (2003), non-specialist teachers were found to provide less meaningful explanations of vocabulary. However, although T7 expressed concern about their ability to remember lexical items, vocabulary generally caused fewer worries in this current study than other aspects of French teaching; concern about vocabulary was not more prevalent in either approach.

(ii) Assessment

T8 worried about their ability to correct errors and provide feedback, although this was common to both approaches, and not unusual: H. Richards et al. (2013) observed non-specialist teachers' inaccurate correction of pupil error and restricted feedback on grammatical mistakes outside the immediate scope of the grammar focus being taught. Less accurate and less spontaneous feedback by non-specialist teachers was also noted by Andrews (2007) and T. Farrell and Richards (2007).

(iii) Grammar

Teachers (except T1 and T3) repeatedly expressed concern about their insufficient grammatical knowledge. Similar concerns were also found in interviews with trainee teachers in Macrory and McLachlan's (2009) study. Although grammar was identified as a concern pre-project by three teachers (T2, T3, T4), this had become more common post-teaching. The perceived greater linguistic demands of Novel and the requirement to teach grammar explicitly to learners possibly increased teachers' awareness of deficiencies in grammatical knowledge.

(iv) Pronunciation

Teachers' concerns relating to pronunciation are discussed separately in [7.9.3.7](#) to avoid repetition.

7.9.3.6 Discussion of teachers' approach-specific references to (under)confidence

Teachers' regular references to lack of confidence in French teaching (regardless of approach) are unsurprising: it was repeatedly identified as significantly concerning in national studies and reviews discussing class teacher provision in England (Board & Tinsley, 2014; Tinsley, 2019; Tinsley & Board, 2016, 2017); Du Plessis et al. (2015) also reported stress in teachers teaching outside their field of specialism.

In this current study, there was some discrepancy in teachers' perceptions of the impact of each teaching approach on their confidence. At both Interview2 and Interview3, when discussing their *overall* confidence teaching French, teachers (T1, T8) acknowledged the

(negative) impact of individual lesson content (regardless of approach). However, they also described feeling more at ease teaching Standard than Novel, because of the familiarity of content and pedagogy in the former and the challenge of the subject knowledge demands made by the latter. Several studies and reviews found that self-perceived teacher confidence, reported by T7, T8 and T9, increased with experience of teaching (Cable et al., 2010; Cooke & Faez, 2018; Driscoll, 1999; Driscoll, Jones, & Macrory, 2004; Muijs et al., 2005; Rowe et al., 2012). This appeared to apply more to Standard than Novel in the current study, with higher confidence reported teaching Standard, although it was acknowledged by T9 in reference to likely increased confidence if teaching Novel again the following year.

However, as shown in [Figure 76](#), teachers referred more often to increased confidence when teaching Novel than Standard in *individual* areas (explaining vocabulary, supporting with grammar/writing/sentence-building, correcting pronunciation and generally with teaching related to SSCs). Such support is important, as non-specialists were observed correcting errors less (than would be associated with specialists) by Cable et al. (2010) and Muijs et al. (2005), and making generally less use of assessment (Driscoll, 1999; Muijs et al., 2005). Teachers' *overall* perceptions of confidence may have been coloured by their generally greater enjoyment of and preference for Standard, whereas when asked about specific aspects of teaching (related to how they felt able to support learners), they reported greater feelings of confidence from the support offered by Novel.

7.9.3.7 Discussion of teachers' concerns with pronunciation

The importance of audio support in both approaches identified by the teachers has previously been acknowledged (see DCSF, 2005; Driscoll, Jones, Martin, et al., 2004; Kasprowicz, 2025; OFSTED, 2005), particularly because of the accurate, authentic model it provides for non-specialists (Barton et al., 2009; Collen & Duff, 2025). Such support is intended to address the impact of teachers' insufficient linguistic skills limiting pupil outcomes in terms of accuracy of pronunciation, intonation and oral fluency (noted by Cable et al., 2010; DCSF, 2005; Moon, 2009; Muijs et al., 2005; OFSTED, 2011; Tinsley & Board, 2017), as well as the impact of accented input negatively affecting their phonological ability (found by Muñoz, 2009, in the Barcelona Age Project). In this current study, teachers considered Novel to offer more support than Standard, partly because of the explicit SSC-focus. Through its content and pedagogical approach, Novel appears to respond to Muñoz's (2014a) recommendations to provide training for FL teachers to develop pupils' grapho-morphological awareness. Such an approach may ultimately increase teachers' positive response to the FL, as Uchida and Sugimoto's (2019) questionnaire-survey of Japanese EFL teachers revealed a causal relationship between greater confidence in teachers' pronunciation and positive attitudes towards teaching pronunciation to

learners.

7.9.4 General discussion of teachers' attitude to change

Enacting change can be challenging in any field (as demonstrated by the findings of T. Webb & Sheeran's 2007 meta-analysis of 47 experimental studies: a medium-to-large change in intention brought about a small-to-medium behavioural change). It is perhaps therefore unsurprising that curriculum innovation is difficult. For example, Decorte et al. (2024) found that, even when teachers acknowledged the benefits of adopting metacognitive oracy instruction in the FL classroom, they experienced difficulties in implementing long-lasting changes in practice. Teachers' attitudes to change in FL teaching can be usefully compared to other areas of the curriculum where innovations have been introduced or aspects of classroom practice targeted for development. One example here is the implementation of technology (perhaps a comparable area to primary FL provision in terms of existing teacher confidence and competence at the point of introduction).

7.9.4.1 Barriers to change

With reference to technology implementation in education, Ertmer (1999) cited Brickner's (1995) concepts of first- and second-order changes. First-order changes are external to teachers; they involve changing practice but do not challenge teachers' beliefs. Second-order changes relate to teaching and learning approaches, possibly "confront[ing] fundamental beliefs about current practice" (p.48). According to Ertmer (1999), *any* barrier (first- or second-order) arising from factors such as personal fears, logistical issues or organisational and pedagogical concerns could impact on adoption of changed practice. Indeed, Howard (2013) – researching in the same field - found in detailed qualitative analysis of teacher interviews that feelings of stress arose from unfamiliarity with technology equipment and a low sense of self-efficacy. Such findings may be relevant to teachers implementing Novel, with the more explicit and analytical approach to teaching vocabulary, phonics and grammar requiring change in FL teaching habits. It was also evident, particularly from comments made by T1, that aspects of both approaches differed considerably from the more child-led pedagogy adopted in School B ([7.1.2.1](#)).

Personal fears (Ertmer, 1999) related to linguistic insufficiency have been clearly established in this current study. Most teachers appeared to struggle to surmount their linguistic anxiety, and this impacted on their attitude to Novel in particular. Konakli and Akdeniz (2022) reported from interview analysis that teachers resisted change when it caused anxiety. Similarly, Howard (2013) found that levels of anxiety in technology use led to restricted classroom implementation of technology and avoidance behaviour. One participant in Howard's interview analysis described *risk-taking* as "trying something you are either not comfortable or familiar

with, and doing something different to what you would normally do” (2013:367). Connecting this to the instructional change in this study, the Novel approach led to a sense of discomfort for some teachers and highlighted the areas of language competence in which they felt lacking. Viewing this as *‘risk’* perhaps contextualises the preference shown by some teachers towards Standard.

Some teachers’ comments in interviews indicated that Novel was more challenging to teach than Standard, both because of the linguistic demands it posed and because its intentions were unfamiliar ([7.1.2.2](#)). Explaining factors which impacted on users’ acceptance of computers, Davis (1989) found that two determinants of technology acceptance were (i) perceived usefulness (i.e., the extent to which users believed technology would make their job easier) and (ii) perceived ease of use - “the degree to which a person believes that using a particular system would be free of effort” (p.320) (i.e., even when recognising the benefits of use, users felt it would be too difficult to use). Extending this theoretical Technology Acceptance Model (see Davis et al., 1989), Teo et al.’s (2008) investigation into pre-service teachers’ attitudes towards computers found that their perceived ease of use had significant effects on their perceived usefulness and teachers’ computer attitudes. It is possible that the difficulty in using Novel (its complexity - see Rogers, 2003:240, [7.9.3.2](#)) perceived by some of these non-specialist teachers impacted on their attitudes towards this approach.

Suggestions for supporting teachers through change in instructional habits are discussed in [8.4.4.1](#).

7.9.5 Discussion of results from teacher responses to each approach – summary

Teachers’ surprise at and pleasure from learners’ achievements after 22.5 hours of learning French offer important insight into realistic potential expectations for Year 3 learners when taught regularly for 45 minutes each week. In teachers’ views, learners’ progress did not appear disadvantaged in any way by being taught language more explicitly (i.e., through Novel). Indeed, the teachers thought that the inclusion of regular opportunities for retrieval practice (with a spaced approach and cumulative testing) appeared to support recall effectively. Learners’ ability to manipulate the component parts of sentences possibly resulted from Novel’s explicit focus on verbs, suggesting that young learners *can* learn the way they are taught.

Teachers acknowledged the role of age and context in vocabulary selection. Although initially influenced by personal experiences as learners and teachers, they were impacted by exposure to the high-frequency vocabulary characteristic of Novel, with this being regarded as important by the end of the study.

Teachers often reported feeling out of their depth when teaching French, particularly in Novel, but there did not appear to be any negative impact on classroom relationships. Indeed, a

positive benefit arose from teachers and pupils learning alongside each other. Teachers' enjoyment appeared affected by their status as FL learners themselves, particularly when considering the anxiety induced by learning and teaching grammar. Teaching unfamiliar language in an unfamiliar way was challenging, and for this reason teachers preferred teaching Standard: familiarity increased their confidence.

Teachers reported pupil enjoyment in both approaches, although this was considered higher in Standard than Novel by the end of the study. Several of the teachers thought that games (more common in Standard than Novel) appealed to these young learners but also reported that pupils responded generally positively to the challenge presented in Novel, which (when optimal) promoted progress from which pupils seemed to derive satisfaction. In this respect, Novel approach appeared to embody the challenge of providing a light-hearted experience designed to promote enjoyment, whilst at the same time ensuring linguistic progression within a time-limited context.

Teachers consistently acknowledged the positive impact of Novel on all areas of their subject knowledge, to the extent that it appeared to provide a form of CPD. Even though teachers reported greater overall confidence when teaching Standard, when asked specifically about individual aspects of teaching French, the Novel approach appeared to boost their confidence and offered greater support, particularly in the areas of pronunciation, SSCs, and supporting learners to develop grammatical understanding. Adapting to Novel may have needed a longer period of adjustment: Rogers (2003) highlighted that, even when innovations have obvious advantages, acceptance of these new ideas can take significant time.

7.10 Comparing teacher reports/perceptions with pupil data: discussion drawing together key findings from Results I, II and III

This section compares teacher perceptions and reports of pupil progress ([7.10.1](#)) and enjoyment ([7.10.2](#)) with pupil outcomes and pupil views of their own learning and of French. It concludes by linking these discussions to Guskey's (2002) model of teacher change ([7.10.3](#)).

7.10.1 Comparing teacher reports and teacher/pupil perceptions of pupil progress with actual outcomes

The factor analysis of post-lesson evaluations identified one factor relating to teachers' satisfaction with pupils' knowledge and understanding, as well as with linguistic usefulness ([7.4.3.1](#)). Teachers reported that pupils made progress in both approaches, but evaluations evidenced higher levels of teacher satisfaction with this factor in Standard than Novel. This was, however, contrary to data emerging from teacher interviews ([7.8.1](#), [7.9.1](#), [7.9.2](#)), where

references to progress were more numerous in Novel than Standard at both Interview2 and Interview3, with teachers repeatedly referencing the foundational knowledge in Novel that appeared to enable pupils to work at sentence level. It would appear that teachers' qualitative comments were more accurate than their post-lesson evaluation scores, better reflecting the *actual* progress made by learners, given that several of the tests seemed to suggest a benefit for pupils in Novel ([5.7.6](#), [5.10.5](#), [5.14.7](#), [5.17.5](#)).

PQs also showed that, descriptively at least, pupils in Standard appeared to rate their self-efficacy higher than those in Novel, both in overall ratings and with reference to individual elements of FL learning (i.e., ordering words correctly in a sentence, spelling words, saying sentences). However, teacher perceptions and pupil ratings were not borne out by test outcomes, with, by and large, a descriptive trend for pupils in Novel to outperform those in Standard.

As discussed in [6.4.2.2](#), pupils in both approaches conceptualised their future progress similarly, but, in discussing their achievements since the start of the project ([6.4.2.3](#)), they referred to different strengths according to the approach adopted in each scheme. In this respect, teacher perceptions of pupils in Novel being able to work at sentence level appeared to be supported by pupils' own references to this, as these were more common in Novel group interviews than Standard.

7.10.2 Comparing teacher/pupil reports and perceptions of pupil motivation and focus with actual measures

Teachers generally reported pupils to be more focused and motivated in Standard than Novel, according to their post-lesson evaluations ([7.4.3.2](#)), with a couple of exceptions (T1 reported in Interview3 that they found pupils more focused in Novel than Standard and T9 referred to their higher levels of participation at least initially). Similarly, self-ratings for enjoyment were descriptively higher in Standard than Novel for all categories of pupils (though only reached statistical significance for pupils who were initially unsure about learning French). Pupils' responses in group interview 2 and PQ3 appeared to support the suggestion that pupils in Novel were more affected by classroom factors, with pupils commenting negatively on fatigue or lesson length ([6.7.2.2, d](#)).

In terms of pupil enjoyment, teachers largely identified aspects of lessons which pupils also mentioned enjoying in group interviews and PQ3, with occasional anomalies. These similar and differing aspects are summarised below.

- Teachers reported increasing engagement with vocabulary grids in Novel as the year progressed, and this was supported by pupils' own positive references to these.
- Teachers referred to pupil enjoyment of topic vocabulary (animals in particular), and this was evident in pupil responses (particularly) in Standard.

- Individual activities such as card sorting and the butterfly-making task (both Standard) were perceived as enjoyable for learners by both teachers and pupils.
- Teachers identified games as an enjoyable aspect of lessons, especially in Standard, and this aligned with pupil responses.
- T3 suggested that cultural elements were positively received by pupils, and this was supported by pupil responses (particularly) in Novel.
- T9 referred to enjoying witnessing pupils' positive responses to the sounds of the language (in Novel) and this was apparent in pupil references.
- T3 and T8 felt that pupils enjoyed songs, but this was not mentioned by any pupils in either approach (and indeed was contradicted by T4).
- T3 felt that young learners generally enjoy stories, but this was not regularly cited as a reason for enjoying French by pupils in either approach.

7.10.3 Teacher attitudes to change and the accuracy of teacher judgements

Guskey (2002) presented a model of teacher change consisting of four steps: i) initial professional development input; ii) enactment of changed classroom practices; iii) impact observed in student outcomes; iv) impact seen in changed teacher beliefs and attitudes. In this model, teacher beliefs and attitudes change *as a result of* observing the positive impact on student outcomes (rather than through experiencing professional learning opportunities): “They believe it works because they have seen it work, and that experience shapes their attitudes and beliefs” (Guskey, 2002:383). Changed practice not perceived as beneficial for student outcomes will likely be discarded. Although impact on teachers' beliefs in the current study was varied, they did not really benefit from all stages of Guskey's (2002) model: for example, there was no initial formal training in the new approach, and they did not have the outcome evidence from the test results at the point of Interview3. Both these steps may have facilitated great change in teacher beliefs and attitudes.

Guskey (2002) also acknowledged that the model was a simplification of a complex process. In the current study, it could be that some teachers' responses to Novel would have been different, given greater training in this approach and increased awareness of its impact on learning outcomes. Indeed, teacher assessment of ‘what works’ appeared mixed, depending on when and how they were asked, and therefore they were making up their minds based on incomplete evidence. Within the limited time frame of this study, teachers may not have judged pupil affective variables or academic achievement reliably: Novel appeared to be supporting increased learning but this was not absolutely clear.

Observable measures such as academic achievement *can* be easier to judge than pupils' emotional responses and motivational levels (see Urhahne & Wijnia's 2021 literature review on the accuracy of teacher judgements). However, although Südkamp et al.'s (2012) meta-analysis of 75 studies found a positive correlation between teacher assessment of learner outcomes and students' actual achievements, they asserted that "there is plenty of room for improvement" (p.755) in terms of how accurately these judgements reflect actual student progress. In this current study, teachers' judgements on pupil progress were indeed inconsistent. Qualitative data indicated that teachers almost unanimously perceived greater progress among pupils in Novel, yet their quantitative post-lesson evaluations reported greater perceived pupil learning in Standard than Novel. Teachers were asked for an impressionistic, global overview of their perceptions of pupil progress in these post-lesson evaluations, whereas some studies have suggested that, to assess pupil achievement accurately, teachers need to be provided with clear assessment criteria and be involved in developing these (Harlen, 2005), with a greater level of specification potentially increasing the accuracy of teacher judgement (Südkamp et al., 2012). Given that such specific criteria were not provided to teachers when they were asked to make judgements on pupil progress, and given the absence of CPD and clear data from tests to evidence progress over time into subsequent school years, it is perhaps unsurprising that teachers did not fully embrace innovation, in the short time frame here, given its complexity and unfamiliarity.

7.11 Teacher findings: summary

Teacher findings exemplify the importance of mixed methods research. While post-lesson evaluations and questionnaires indicated a preference for Standard over Novel (with individual exceptions), interviews presented a more nuanced view. Teachers' responses in Interview1 appeared to indicate readiness for change, acknowledging the need for a different pedagogical approach, but change takes time; the linguistic and operational demands of Novel, whilst providing valuable CPD in many areas, challenged (most of) these non-specialist practitioners, making them feel out of their comfort zone. Similarly, whereas in post-lesson evaluations teachers' perceptions of the feedback on pupil learning were more positive in Standard than Novel, this was not supported by test data, which generally indicated a trend for greater learning in Novel. Additionally, teachers' qualitative comments reinforced the value of Novel in supporting pupils' ability to understand the language and work at sentence level and the positive challenge it provided. While generally accurately identifying sources of pupils' enjoyment, teachers occasionally perceived some activities to be more enjoyable than they actually were, according to pupils' own reports.

Chapter 8: Conclusions and recommendations

This chapter reviews the study's design before presenting and discussing key findings relating to the research questions (RQs). It outlines the study's wider contribution to the field of primary foreign language (FL) education in this and similar contexts (and language learning more generally), acknowledging its limitations and discussing potential areas for further research.

8.1 Study context

FLs became a statutory part of the English primary National Curriculum in 2014 (DfE, 2013b), but beyond promoting a language competence model, teaching guidance is limited and content not prescribed. (Prevalent) non-specialist provision is not acknowledged: insufficient detail and advice on assessment and expectations (noted by Ayres-Bennett & Carruthers, 2019; Holmes & Myles, 2019; Tinsley & Board, 2017) means that generalist class teachers, constituting two thirds of the workforce (Collen & Duff, 2024), are largely unsupported. The lack of a recommended minimum time allocation has created disparity in provision and pupil progress.

“Substantial progress” is expected of learners (DfE, 2013b:194) – an ambitious aim requiring language manipulation and practical communication, both in oral and written forms. However, primary FL teachers in England operate in time-limited contexts (Collen & Duff, 2024), with little guidance on using that time effectively. Holmes and Myles (2019) called for further research into age-appropriate pedagogies which provide challenging activities to promote greater long-term motivation and engagement, a need identified by Muijs et al. (2005) and Martin (2012).

8.2 Summary of study

This study, a small-scale impact and process evaluation of classroom practice within a Randomised Controlled Trial, evaluated two approaches to teaching French to 7-8-year-old learners in their first year of learning. From the south coast of England, 149 pupils from five schools received 22.5 hours of instruction over 30 weeks in the academic year 2022-23, taught by seven non-specialist teachers, none of whom held a qualification in French higher than a GCSE. The two approaches, termed for the purposes of this thesis Novel and Standard, were both created by Dr Rachel Hawkes (Hawkes, 2013; Hawkes & Guillou, 2022). Standard, a phrase-based approach, represented practice common in English primary schools; it prioritised topic-based vocabulary, taught phonics non-systematically and included little explicit grammar teaching. Novel, drawing on the work of the National Centre for Excellence in Language Pedagogy (2018-2023) (Hawkes et al., 2019; Marsden & Hawkes, 2023; Marsden & Hawkes, in press; Marsden et al., 2023), was based on language-driven principles, featuring careful lexical

selection (including high-frequency vocabulary), systematic phonics and sequenced explicit grammar.

Pre-project language-analytic ability (LAA) tests (Kasprowicz et al., 2019) provided a baseline measure of pupils' potential aptitude for language learning. From ranked scores, pupils were matched in pairs, then randomly assigned to Novel or Standard treatment, following either the topic-driven or language-driven approach for the academic year. There were no significant differences between the LAA scores of each group. Teachers alternated teaching Standard and Novel approaches, thus experiencing both approaches and hopefully reducing a 'teacher effect'. There was no significant difference between teachers' fidelity to treatment in Novel and Standard approaches, monitored through observations, audio-recordings and interview discussion and additionally supported by measurement of pupil perceptions of teachers' enjoyment of each approach.

Data were collected through pupil tests, questionnaires and group interviews, and teacher questionnaires, post-lesson evaluations and semi-structured interviews. These addressed, for each approach: pupils' performance in vocabulary learning, symbol-sound correspondences (SSCs), and simple sentence manipulation; pupils' and teachers' perceptions of pupils' self-efficacy, enjoyment and progress in French; pupils' and teachers' perceptions of teachers' motivation; and teachers' perceptions of their developing subject knowledge.

Pupils were tested at two points: post-instruction and eight weeks after instruction had ceased. Only content aligning with National Curriculum requirements (DfE, 2013b) and common to both approaches was tested. This included the following: vocabulary recognition and recall in oral and written modalities; phonics transcription and Read Aloud (RA); grammar recognition in the written modality and recall in the oral modality, as well as in Elicited Imitation (also measuring general oral proficiency); and open-ended Picture Response (PR). Pupil self-efficacy and enjoyment were measured through questionnaires, enhanced by qualitative data from group interviews. Teachers' affective variables were measured through their opinions of and emotional response to taught lessons and gauged through interview discussion.

As well as descriptive statistics and effect sizes, mixed-effects models were used to analyse results from tests and questionnaires and enabled assessment of whether the impact of the teaching approach could be generalised beyond the study participants to the wider population, while accounting for any random variation of both participants and test/questionnaire items.

8.3 Summary of findings in relation to research questions

Key findings are now summarised according to RQs. For each test area, headline findings are presented and discussed.

8.3.1 RQ1: What are the effects of each French teaching approach on Year 3 pupils' performance in vocabulary learning, symbol-sound correspondences, and simple sentence manipulation in one academic year?

8.3.1.1 Vocabulary learning

- According to the statistical models, gains were similar across both approaches in recognition and recall in the oral modality, and recall in the written modality, although the descriptive trend was for pupils in Novel to slightly outperform pupils in Standard in these three vocabulary measures.
- In the recognition test in the written modality at post-test, Novel scores were higher than Standard, reaching statistical significance.
- Effect sizes for between-group (Novel-Standard) comparisons of all four measures were all medium/large (in favour of Novel) using Kraft's (2020) scale, except in recognition in the oral modality at post-test (small).
- Learning in Novel was durable, with scores higher at delayed post-test than post-test in recognition and recall in the oral modality (medium effect sizes).

As this study evaluated a whole approach, it is impossible to attribute gains to a single component of the intervention: rather, we should acknowledge the contributory elements. Pupils appeared to benefit from the systematic revision inherent in Novel through a spaced practice schedule (see Marsden & Hawkes, 2023; Nakata & Suzuki, 2019a), as well as its systematic presentation and analysis of orthographic forms (Hawkes et al., 2019). Retrieval practice appeared to afford learners an advantage in reinforcing form-meaning relationships (see Conklin, 2020), with regular opportunities for forced memory retrieval (van den Broek et al., 2018) and cumulative testing (Maie et al., 2025) provided through the vocabulary grids.

8.3.1.2 Symbol-sound correspondences

- RA test scores were descriptively slightly higher in Novel than Standard at post-test (large effect size but not statistically significant); there was no difference between approaches at delayed post-test.
- The transcription test was unreliable.

Pupils' low scores in both approaches evidenced the challenging nature of both transcription and RA tests, with the influence of learners' knowledge of English SSCs apparent. The higher RA

scores in Novel at post-test indicated a possible benefit of the explicit and systematic focus on SSCs (which may have become more apparent with a larger sample size), but this was not durable at delayed post-test. Pupils apparently needed a longer period of instruction for L1 influence to be suppressed (see also Woore et al., 2018, for 11-12-year-old learners).

8.3.1.3 Grammar

- Pupils in Novel outperformed Standard at both time-points in spoken grammar production tests (medium-large effect sizes, reaching statistical significance).
- Novel scores were descriptively higher than Standard in the Elicited Imitation Grammar (EIG) test at both time-points (medium-large effect sizes, not reaching statistical significance).
- Reliability was low in grammar recognition tests.

The Novel scheme explicitly addressed the DfE (2013a) requirement to teach grammatical features of the FL (gender, linguistic patterns, sentence-building and high-frequency verbs) and enable the use of these to support both comprehension and written/oral expression. The statistically significant difference between groups in spoken grammar production and the higher scores in Novel in the EIG test at both time-points suggest that pupils were not disadvantaged in any way when taught explicitly. Indeed, their pronoun knowledge, tested in spoken grammar production tests, was superior in Novel. (Some) young learners may learn FL implicitly (DeKeyser, 2003) when time and exposure are unlimited, but the current study's ecologically valid, classroom-based findings support Lichtman's (2016) instructional hypothesis in suggesting that children learn in the way in which they are taught.

8.3.1.4 Open-ended speaking

- The Novel approach appeared descriptively more beneficial than Standard for learners in PR and Elicited Imitation Speaking (EIS) tests at both time-points (medium-large effect sizes but not statistically significant).
- A larger number of pupils in Novel than Standard made greater use of verbs in PR at both time-points.

These findings make an important contribution to our understanding of how children learn FL in instructed contexts. Pupils' more frequent spontaneous production of verbs in Novel than Standard – at least when describing unfamiliar pictures - appeared to demonstrate greater verb awareness. It is possible that, as Häcker (2008) suggested, combining vocabulary learning with

explicit grammar instruction enabled learners to better understand individual sentence components. It suggests that young learners do not just begin to use elements of unanalysed language (rote-learnt as chunks) more creatively and autonomously after a period of imitation and reproduction (as observed by Myles et al., 1998, with 11-13-year-old learners), but that some are able to construct meaningful sentence-level utterances at a younger age using individual syntactic components. Porter and Graham (2022) highlighted the importance of pedagogical approaches providing pupils with a sense of progress; such progress appears to have been evidenced by the autonomous production of sentence-level utterances in this open-ended speaking task.

8.3.2 RQ2a. What is the (perceived) effect of each teaching approach on pupils' self-efficacy and enjoyment?

This section draws on both quantified and qualitative data on self-efficacy ([8.3.2.1](#)) and enjoyment ([8.3.2.2](#)).

8.3.2.1 Self-efficacy

Throughout the project, pupils' self-efficacy increased in both conditions. The mixed-effects model showed that pupils' ratings in Novel did not increase as much as in Standard, but the former had higher starting points, and the between-group difference was not statistically significant. Pupils' self-efficacy data in Standard did not correspond to their test performances. For example, their perceptions of their ability to spell words, say sentences, remember the meaning of French words (and vice versa) – higher than self-perceptions in Novel – were not borne out by performance in the corresponding knowledge tests. This may reflect a lack of metacognition on the part of these learners in Standard, who were perhaps unaware – having less opportunity to analyse sentence components, write words from memory, and encounter vocabulary outside tightly grouped semantic topic fields – of how challenging these areas could be.

Pupils' group interview comments on their ability to perform certain tasks were similar across groups. Qualitative data suggested a possibly stronger sense of self-efficacy in phonics among pupils in Novel than Standard, in which learners seemed slightly more negative about their ability to read aloud and pronounce French words accurately. This was not replicated in phonics tests, but possibly the more explicit, ongoing SSC focus in Novel had at least awareness-raising benefits and potentially promoted self-belief among learners.

8.3.2.2 Enjoyment

Pupils' enjoyment appeared to decrease as the year progressed, regardless of approach, perhaps as they became more aware of the long haul of language learning and the challenges involved, although large standard deviations reflected high levels of individual variation in enjoyment. Ratings were higher in Standard than Novel but there was no statistically significant difference post-teaching (and mid-project only among those pupils in Standard who had been unsure about learning French pre-project). Regardless of approach, pre-project positivity about the prospect of learning French seemed to make these learners most likely to enjoy learning the language during the year.

Reflecting Dörnyei's 'language level' (1994), pupils' motivation in both approaches appeared partly intrinsic, enhanced by having family or other contacts with the language/country and being able to use the language to communicate with representatives from the French-speaking community or travel there themselves. In terms of the 'learning situation' level, pupils from both approaches were motivated by particular activities or vocabulary. By PQ3, specific examples of intrinsic and integrative motivation were slightly more common among Novel responses, whereas more pupils in Standard were universally positive about learning French than in Novel. As with self-efficacy, pupils in Novel seemed slightly more positive about phonics in French, with positive comments about enjoying the sounds of the language; this may reflect Novel's greater emphasis on teaching SSCs explicitly.

Teacher perceptions indicated that pupils appeared generally positive about learning French in both approaches with slightly higher levels of perceived enjoyment in Standard, supporting quantified pupil data. Teachers seemed broadly supportive of the positive nature of the desirable difficulty experienced by pupils in Novel. Pupils themselves were more likely to refer to difficulty in French learning in Standard than Novel, although this was only apparent in group interviews, not in PQ3 responses; such comments may have been influenced by peers in group discussions.

Overall, numbers of pupils expressing an intrinsic dislike of French were small in both approaches. Although grammar has been empirically observed (Cenoz, 2003) and noted (Tierney & Gallastegi, 2005) to be unappealing to young learners, this was not evident. Indeed, although it has traditionally been assumed that young learners are motivated by an oral approach (see, e.g., Cable et al., 2010) rather than a more literacy-based one, pupils in Novel were not generally negative about writing in French.

8.3.3 RQ2b. What are pupils' and teachers' perceptions of the effect of each teaching approach on pupils' progress in French?

In terms of how pupils conceptualised future progress ([8.3.2.1](#)), there was little difference between approaches, whereas, when reflecting on progress made in French since the start of the year, pupils in Novel referred more often to working at sentence level and spontaneously produced a greater range of word classes than pupils in Standard. The latter were more likely to refer to having learnt topic-related vocabulary.

Teachers were positive about pupils' progress in both approaches but considered it higher in Standard than Novel (according to post-lesson evaluations). This was inconsistent with test data and their qualitative comments, again highlighting the importance of mixed methods research. Positive comments about pupil progress in Novel outnumbered those about Standard, with teachers appreciating Novel's structure and opportunities for sentence construction, viewing it as providing better foundational knowledge for learners in phonics and grammar than Standard. Novel seemed to provide support to ensure pupil progress, challenging for non-specialist teachers. Its in-built retrieval practice was perceived as beneficial, particularly that provided by the vocabulary grid activities, which were considered to have a positive impact on pupil progress by teachers and pupils in teacher and pupil interviews and pupil questionnaires.

8.3.4 RQ3a. What is the (perceived) effect of each teaching approach on non-specialist primary teachers' motivation?

Teachers' self-reports evidenced enjoyment of both approaches, but generally greater enjoyment of Standard. The variation among teachers in terms of preference for approaches seemed related to their level of linguistic competence and teaching experience, with teachers with lower levels of French and less French teaching experience preferring Standard. Tension levels increased when teachers felt insecure in subject knowledge and unfamiliar with pedagogical approaches - most apparent when teaching grammar explicitly in Novel. Again, there were anomalies in the data: teachers expressed greater confidence teaching Standard because it felt more familiar *overall*, yet when discussing *specific* aspects of teaching and supporting learners, they reported feeling better supported and more confident in Novel. Generally, gaining experience of French teaching (in both approaches) appeared to increase teachers' enjoyment of teaching the language, possibly because of the positive impact on their subject knowledge (see Dewaele & MacIntyre, 2014).

Pupils' perceptions of teachers' enthusiasm did not appear to differ across approaches, indicating perhaps that teachers' strong sense of professionalism ensured that both groups of pupils viewed them as positive, even though the majority of the teachers preferred Standard. In

qualitative data, both teachers and pupils recognised the positive impact of perceived pupil progress on teachers' enjoyment, regardless of approach.

8.3.5 RQ3b. What are non-specialist teachers' perceptions of the effect of each teaching approach on their developing subject knowledge?

Teachers all acknowledged the positive impact of Novel on their developing subject knowledge. Its explicit focus on grammatical understanding and SSCs constituted a kind of CPD for teachers. Similarly, being exposed to high-frequency vocabulary in Novel made them reassess the importance of such language. Teachers felt that Novel supported them to enable pupil progression, by providing a clearly structured scheme and increasing their confidence to correct pupil error, perhaps enabling the "substantial progress" required by the DfE (2013b:194). However, with greater requirement to teach explicitly came increased awareness of their own linguistic deficiencies, particularly in grammatical understanding.

8.4 Wider contribution of findings and implications for the primary foreign languages classroom and policy

This section considers the ways in which this study contributes to existing knowledge relating to the teaching of FL to young learners and the measurement of their learning outcomes.

8.4.1 Realistic expectations of what young learners can achieve in a time-limited context

This study makes a valuable contribution to understanding what young learners can achieve in one academic year, given weekly provision of 45 minutes, addressing the paucity of evidence in this area (Mitchell & Myles, 2019) particularly for the youngest learners in Key Stage 2. Such awareness could then contribute to a smoother transition into secondary FL learning (DfE, 2025). Whilst acknowledging that success can be achieved in different ways (Marsden, 2007; Teaching Schools Council, 2016; Tragant Mestres & Lundberg, 2011), this study has explored possible responses to the 'what' and the 'how' of primary FL provision, measuring the impact of different pedagogical approaches and content on pupils' learning outcomes and affective variables, when taught exclusively by non-specialist teachers.

8.4.2 Demonstrating the value of applied empirical research

Marsden (2007) outlined the challenge in testing an educational hypothesis and concurrently informing practice. This study exemplifies the possibilities and difficulties of conducting valuable quasi-experimental research in field settings with young learners from a range of contexts. Roehr-Brackin (2020) noted that research involving younger learners is uncommon in

classroom studies of L2 acquisition, and this study attempts to redress this balance. It addresses a similarly neglected area, that of L2 French rather than English as a FL. It demonstrates a range of methods to help safeguard validity, such as matched pair randomisation of participants to condition, rigorous monitoring of provision to ensure fidelity to treatment, and allocation of teacher to both conditions to minimise teacher effect (see Marsden, 2007), whilst maintaining high ecological validity. Using mixed methods research to develop a well-rounded approach to evaluation, it considers all aspects of provision: pupil outcomes (in varied knowledge types); pupil experience; teacher voice. Data from various sources support the creation of rich, beneficial insight. Different test types contribute to our understanding of the challenges inherent in language testing (e.g., in designing appropriate measures of phonics knowledge in such young learners) as well as the potential of some tests (e.g., in the popularity and effectiveness of Elicited Imitation which served a dual purpose of informing on grammatical knowledge and overall proficiency).

8.4.3 Language-analytic ability (LAA)

In tests, LAA provided a baseline indication of pupils’ potential to learn a FL, included in the mixed-effects models as a fixed effect. This study appears to support the influence of LAA on learners’ language learning aptitude (see Roehr-Brackin & Tellier, 2019): pupils with lower LAA scores (ca. < 25) generally performed less well in tests than those with higher LAA (as also found by Kasprowicz et al., 2019), with some nuanced effects. However, there was a descriptive differential impact according to approach - indications of an interaction between approach and LAA ([Table 58](#)). For pupils with higher LAA, approach appeared to vary in impact according to test type, whereas Novel seemed particularly beneficial for learners with lower levels of LAA in almost all tests. The explicit instruction characterising Novel possibly stimulated explicit learning processes in these young learners (see Tellier & Roehr-Brackin, 2013) and appeared to minimise (some) individual differences expected to impact on FL learning, supporting learners who needed analytical, explicit explanations the most. This supports Skehan (1989), Erlam (2003) and Roehr-Brackin (2020).

Table 58.

Summary of the Descriptive Impact of Approach on the Achievement of Pupils with Higher and Lower Language-analytic Ability Scores in Different Test Types

Test type	Pupils’ LAA scores	
	Pupils with lower LAA	Pupils with higher LAA
PT: Vocabulary recognition, oral modality	Advantage in Novel	Advantage in Standard

DPT: Vocabulary recognition, oral modality	Advantage in Novel	No difference
PT: Vocabulary recognition, written modality	Advantage in Novel	No difference
DPT: Vocabulary recognition, written modality	Advantage in Novel	No difference
PT: Vocabulary recall, oral modality	Advantage in Standard	Advantage in Novel
DPT: Vocabulary recall, oral modality	Advantage in Novel	Advantage in Novel
PT: Vocabulary recall, written modality	Advantage in Novel	Advantage in Standard
DPT: Vocabulary recall, written modality	Advantage in Novel	Advantage in Standard
PT: Phonics Read Aloud	Advantage in Novel	No difference
DPT: Phonics Read Aloud	No difference	No difference
PT: Spoken grammar production	Advantage in Novel	Advantage in Novel
DPT: Spoken grammar production	Advantage in Novel	Advantage in Novel
PT: Elicited Imitation Grammar	Advantage in Novel	Advantage in Novel
DPT: Elicited Imitation Grammar	Advantage in Novel*	Advantage in Novel
PT: Elicited Imitation Speaking	Advantage in Novel	Advantage in Standard
DPT: Elicited Imitation Speaking	Advantage in Novel	Advantage in Standard

*For all but pupils with the very lowest scores, where there was no difference.

Note. PT stands for post-test, and DPT for delayed post-test. Phonics transcription and written grammar recognition tests had low reliability scores and are therefore not included. In the open-ended PR speaking task, assumptions were not met for the model; these are also excluded.

In this respect, this study contributes to the debate centring on age-appropriate pedagogies, demonstrating that 7-8-year-old learners were not disadvantaged by being taught explicitly.

8.4.4 Contributions to the debate surrounding non-specialist teachers of foreign languages

If we acknowledge that continuing professional development in FL does not seem to have been a priority for policy-makers (Tinsley & Board, 2016), then alternative ways of upskilling non-specialist practitioners should be found, otherwise their lack of pedagogical knowledge is predictable (Porter & Graham, 2022). In this study, teachers unanimously recognised the positive impact of being involved in the evaluation study: experiencing both pedagogical approaches first-hand supported their reflection on what worked best for pupils and themselves, and why. In a context where practitioners may prefer not to be involved in FL delivery, teachers' comments also illustrate the value of personal experience of teaching French, with positive impact on their confidence and subject knowledge over the course of the study.

This study exemplifies how difficult innovation and educational change can be. Even when teachers acknowledged the benefits for both pupil progress and their own subject knowledge, they still found the adjustment to Novel approach challenging. Whilst involvement

in this project served as useful development of teachers' linguistic competence and pedagogical understanding, it was not intended to replace CPD. A dual focus on the development of linguistic competence and pedagogical understanding should be prioritised. Marsden et al. (2023) highlighted the need for a multi-faceted approach to nurturing lasting change, involving CPD, provision of resources and sustained engagement with relevant research. This is explored below.

8.4.4.1 Supporting teachers through change

Various strategies have been suggested as supportive for teachers involved in innovation in any area of classroom practice. For example, with reference to technology implementation, Ertmer (1999) suggested the importance of developing a clear vision of its usefulness in realising objectives; the importance of shared goals was also acknowledged by Howard (2013), as was the need for a “clear rationale for change” (Marsden et al., 2023:11). Once such a vision is established, support can be offered to practitioners through modelling by experienced practitioners. Collaboration with colleagues is also deemed effective (Decorte et al., 2024; Ertmer, 1999; Howard, 2013; B. Taylor et al., 2005), along with clear leadership and communication, with practitioner involvement (Konakli & Akdeniz, 2022; Lomba-Portela et al., 2022; Marsden & Hawkes, in press; Marsden et al., 2023; B. Taylor et al., 2005) and reflection on practice (Ertmer, 1999; Lomba-Portela et al., 2022). Regular feedback over time, addressing particularly the impact on learners' outcomes, will likely support changes in practice (Guskey, 2002).

Overall, it should be recognised that change takes time (Decorte et al., 2024; Guskey, 2002; B. Taylor et al., 2005), and in this case it was acknowledged that the changes recommended in the *Modern Foreign Languages Pedagogy Review* (Teaching Schools Council, 2016) represented “fundamental shifts in thinking about principles that underpin curriculum design, pedagogy, and assessment” (Marsden et al., 2023:11). Any change brings uncertainty and a risk of failure (Guskey, 2002; Howard, 2013), heightening anxiety – and this may increase when teachers already feel insecure about a key aspect of FL teaching, namely, their linguistic competence. Howard (2013), referring to innovations with educational technology, advocated that teachers should develop familiarity with new approaches to counteract the sense of dread and fear of failure; this would appear congruent with the repeated references made by teachers in this current study to the importance of familiarity with an approach (which occurred more often in Standard than Novel).

8.5 Limitations and future research

8.5.1 Increasing study length

The short length of the study (at least, the low intensity of the language lessons, totalling just 22.5 hours over a whole year) may have had an adverse impact on the possibility of measuring progress reliably, because pupils did not score very highly in some of the tests. For example, low scores in vocabulary recognition in the oral modality and narrow standard deviations suggested possible floor effects. Also, low scores were probably at the root cause of poor reliability in phonics transcription and assumptions not being met for the mixed-effects model in the PR task (it is possible that a longer study, facilitating more extensive piloting, which was impossible pre-study due to the novel nature of an innovative pedagogical approach, could have been beneficial). These issues illustrate the difficulty of constructing tests for this young age group. A longer intervention may have helped avoid floor effects in testing and would perhaps have avoided the losses at delayed post-test relative to post-test which suggested the importance of more time to sustain early gains. A longer intervention may also have made it possible for the tests to draw on additional language content, facilitating more robust test design (which permitted insight into the differences between the two approaches, at the same time as ensuring equitable tests for both groups).

A longer study period overall would have provided data on the long-term progress and impact of the two approaches on pupils and teachers. For example, T9 felt that Novel promoted better long-term progress in pupils. It would also have enabled more longitudinal measures of pupils' self-efficacy and enjoyment to see whether the slight descriptive trends in favour of Standard developed into a stronger difference. Similarly, some teachers were beginning to feel slightly overwhelmed by the linguistic demands of Novel post-teaching, but the scheme actually revisited much content in the following year. This may have boosted teachers' confidence and reduced its novelty - especially important as familiarity had played such an important role in their preferences for Standard over Novel.

8.5.2 Increasing sample size and scope

Larger sample sizes may have made a difference (Cook & Campbell, 1979; Larson-Hall, 2016) in some tests (e.g. EIS, phonics RA), where differences between Novel and Standard approaches had medium-large effect sizes but did not reach statistical significance. Increasing the teacher sample to include specialist teachers (at least with qualifications above GCSE) could provide greater insight into what young learners could achieve in each approach. Replicating the study with Spanish or German groups would enable parallels to be drawn across the most commonly taught languages in England.

8.5.3 Increasing range of measures

Although LAA served as the pre-test baseline measure, other measures of participants' cognitive ability (such as memory ability or phonetic coding ability in L1 – see, e.g., Sparks & Ganschow, 1991, 1993 - or metalinguistic awareness) could be included from the outset (recommended by Roehr-Brackin, 2020), especially given that these young learners were still at a dynamic stage of cognitive development. Equally, although in this study pupils' qualitative feedback was useful in indicating potential (lack of strong) association between self-efficacy and actual performance, correlation tests could determine the extent to which self-efficacy ratings are associated with achievement scores at this early age.

8.6 Final concluding remarks

The aim of this study was to evaluate two approaches to teaching French to young learners. Rather than prescribe one specific pedagogical approach that non-specialist teachers should adopt, this study has aimed to provide a detailed and informative account of teachers' and learners' experiences and outcomes. In so doing, it is hoped that such findings can “influence teachers' underlying theories of learning and also influence and widen the repertoire of possible actions and choices open to the teacher” (Mitchell, 2000:298).

Reference List

- Alamer, A., Alrabai, F., & Sparks, R. (2025). Reducing language anxiety by increasing language achievement: A new experimental study. *Language Teaching Research*, 00(0), 1–22. <https://doi.org/10.1177/13621688251322840>
- Alamer, A., & Lee, J. (2024). Language achievement predicts anxiety and not the other way around: A cross-lagged panel analysis approach. *Language Teaching Research*, 28(4), 1572–1593. <https://doi.org/10.1177/13621688211033694>
- Alderson, J. C. (1992). Guidelines for the evaluation of language education. In J. C. Alderson, & A. Beretta (Eds.), *Evaluating second language education* (pp. 274–299). Cambridge University Press.
- Alderson, J. C., & Beretta, A. (1992). *Evaluating second language education*. Cambridge University Press.
- Alderson, P. (2004). Ethics. In S. Fraser, V. Lewis, S. Ding, M. Kellett, & C. Robinson (Eds.), *Doing research with children and young people* (pp. 97–112). Sage Publications Ltd.
- Alderson, P. (2005). Doing ethical research with children. In A. Farrell (Ed.), *Ethical research with children* (pp. 27–36). Open University Press.
- Alderson, P., & Morrow, V. (2004). *Ethics, social research and consulting with children and young people*. Barnardo's.
- Allred, P., & Burman, E. (2005). Analysing children's accounts using discourse analysis. In S. Greene & D. Hogan (Eds.), *Researching children's experience: Approaches and methods* (pp. 175–198). SAGE.
- Allen, G. (2005). Research ethics in a culture of risk. In A. Farrell (Ed.), *Ethical research with children* (pp. 15–26). Open University Press.
- Amengual Pizarro, M. (2019). Do prospective primary school teachers suffer from foreign language anxiety (FLA) in Spain? *Vigo International Journal of Applied Linguistics*, 16, 9–30. <https://doi.org/10.35869/vial.v0i16.91>
- Anderson, G., & Arsenault, N. (1998). *Fundamentals of educational research* (2nd ed.). The Falmer Press.
- Andrews, S. (2007). *Teacher language awareness*. Cambridge University Press.
- Andringa, S., & Godfroid, A. (2020). Sampling bias and the problem of generalizability in applied linguistics. *Annual Review of Applied Linguistics*, 40, 134–142. <https://doi.org/10.1017/S0267190520000033>
- Araújo, L., & Dinis da Costa, P. (2013). *The European survey on language competences: School-internal and external factors in language learning*. Publications Office of the European Union. <https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research->

- Assor, A., & Connell, J. P. (1992). The validity of students' self-reports as measures of performance affecting self-appraisals. In D. Schunk & J. L. Meece (Eds.), *Student perceptions in the classroom* (pp. 25–50). Lawrence Erlbaum Associates, Inc.
- Atkins, L., & Wallace, S. (2012). *Qualitative research in education*. SAGE.
- Ayres-Bennett, W., & Carruthers, J. (2019). *Policy briefing on modern languages educational policy in the UK*. Multilingualism: Empowering Individuals, Transforming Societies (MEITS) and the Centre for Science and Policy (CSaP).
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.
- Bahrack, H. P., Bahrack, L. E., Bahrack, A. S., & Bahrack, P. E. (2016). Maintenance of foreign language vocabulary and the spacing effect. *Psychological Science*, 4(5), 316–321. <https://doi.org/10.1111/j.1467-9280.1993.tb00571.x>
- Baidak, N., Balcon, M.-P., & Motiejunaite, A. (2017). *Eurydice brief – Key data on teaching languages at school in Europe*. https://eacea.ec.europa.eu/national-policies/eurydice/content/eurydice-briefkey-data-teaching-languages-school-europe_en
- Bandura, A. (1986). *Social foundations of thought and action*. Prentice-Hall.
- Bandura, A. (2006). Guide for constructing self-efficacy scales. In F. Pajares & T. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (pp. 307–337). Information Age Publishing.
- Barcroft, J. (2007). Effects of opportunities for word retrieval during second language vocabulary learning. *Language Learning*, 57(1), 35–56. <https://doi.10.1111/j.1467-9922.2007.00398.x>
- Barcroft, J. (2020). Key issues in teaching single words. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 479–492). Routledge.
- Barnes, A. (2006). Confidence levels and concerns of beginning teachers of modern foreign languages. *The Language Learning Journal*, 34(1), 37–46. <https://doi.org/10.1080/09571730685200221>
- Barton, A., Bragg, J., & Serratrice, L. (2009). 'Discovering Language' in primary school: An evaluation of a language awareness programme. *The Language Learning Journal*, 37(2), 145–164. <https://doi/abs/10.1080/09571730902928029>
- Bassetti, B., Mairano, P., Masterson, J., & Cerni, T. (2020). Effects of orthographic forms on second language speech production and phonological awareness, with consideration of speaker-level predictors. *Language Learning*, 70(4), 1218–1256. <https://doi.org/10.1111/lang.12423>

- Bates, D., Mächler, M., Bolker, B., Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. <https://doi:10.18637/jss.v067.i01>
- Baumert, J., Fleckenstein, J., Leucht, M., Köller, O., & Möller, J. (2020). The long-term proficiency of early, middle, and late starters learning English as a foreign language at school: A narrative review and empirical study. *Language Learning*, 70(4), 1091–1135. <https://doi.org/10.1111/lang.12414>
- Becker, E. S., Goetz, T., Morger, V., & Ranellucci, J. (2014). The importance of teachers' emotions and instructional behavior for their students' emotions - An experience sampling analysis. *Teaching and Teacher Education*, 43, 15–26. <https://doi.org/10.1016/j.tate.2014.05.002>
- Bell, A. (2007). Designing and testing questionnaires for children. *Journal of Research in Nursing*, 12(5), 461–469. <https://doi.org/10.1177/1744987107079616>
- Ben-Eliyahu, A. (2019). Academic emotional learning: A critical component of self-regulated learning in the emotional learning cycle. *Educational Psychologist*, 54(2), 84–105. <https://doi.org/10.1080/00461520.2019.1582345>
- Bennett, J. (2003). *Evaluation methods in research*. Continuum.
- Bennett, J., Hanley, P., Abrahams, I., Elliott, L., & Turkenburg-van Diepen, M. (2019). Mixed methods, mixed outcomes? Combining an RCT and case studies to research the impact of a training programme for primary school science teachers. *International Journal of Science Education*, 41(4), 490–509. <https://doi.org/10.1080/09500693.2018.1563729>
- Beretta, A. (1992). Evaluation of language education: An overview. In J. C. Alderson & A. Beretta (Eds.), *Evaluating second language education* (pp. 5–24). Cambridge University Press.
- Bielick, S. (2017). Surveys and questionnaires. In D. Wyse, N. Selwyn, E. Smith, & L. Suter (Eds.), *The BERA/SAGE handbook of educational research* (pp. 640–658). SAGE Reference.
- Bjork, E. & Bjork, R. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). Worth Publishers.
- Bjork, R. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). MIT Press.
- Bjork, R. A. (2017). Being suspicious of the sense of ease and undeterred by the sense of difficulty: Looking back at Schmidt and Bjork (1992). *Association for Psychological Science*, 13(2), 146–148. <https://doi.org/10.1177/1745691617690642>

- Board, K., & Tinsley, T. (2014). *Language trends 2013/14. The state of language learning in primary and secondary schools in England*. British Council. <https://edtlive.b-cdn.net/livenew/media/f5zl4al2/language-trends-2013-14.pdf>
- Board, K., & Tinsley, T. (2015). *Languages trends 2014-15. The state of language learning in primary and secondary schools in England*. British Council. https://www.britishcouncil.org/sites/default/files/language_trends_survey_2015.pdf
- Boers, F. (2020). Factors affecting the learning of multiword items. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 143–157). Routledge.
- Bolster, A., Balandier-Brown, C., & Rea-Dickins, P. (2004). Young learners of modern foreign languages and their transition to the secondary phase: A lost opportunity? *The Language Learning Journal*, 30(1), 35–41. <https://doi.org/10.1080/09571730485200211>
- Borgers, N., & Hox, J. (2001). Item nonresponse in questionnaire research with children. *Journal of Official Statistics*, 17(2), 321–335.
- Borgers, N., Leeuw, E. de, & Hox, J. (2000). Children as respondents in survey research: Cognitive development and response quality. *Bulletin de Méthodologie Sociologique*, 66(1), 60–75. <https://doi.org/10.1177/075910630006600106>
- Boruch, R. F., De Moya, D., & Snyder, B. (2002). The importance of randomized field trials in education and related areas. In F. F. Mosteller & R. F. Boruch (Eds.), *Evidence matters: Randomized trials in education research* (pp. 50–79). Brookings Institution Press.
- Bouffard, L. A., & Sarkar, M. (2008). Training 8-year-old French immersion students in metalinguistic analysis: An innovation in form-focused pedagogy. *Language Awareness*, 17(1), 3–24. <https://doi.org/10.2167/la424.0>
- Brant, R. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics*, 46, 1171–1178. <https://doi.org/10.2307/2532457>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp063oa>
- British Educational Research Association (BERA). (2018). *Ethical guidelines for educational research*, (4th ed.). <https://www.bera.ac.uk/publication/ethical-guidelines-for-educational-research-2018-online>
- Broadwell, M. M. (1969). Teaching for learning. *The Gospel Guardian*, 20(41), 1–3a. https://www.wordsfityspoken.org/gospel_guardian/v20/v20n41p1-3a.html
- Brown, D. (2011). What aspects of vocabulary knowledge do textbooks give attention to? *Language Teaching Research*, 15(1), 83–97. <https://doi.org/10.1177/1362168810383345>
- Burić, I., Slišković, A., & Macuka, I. (2018). A mixed-method approach to the assessment of teachers' emotions: Development and validation of the Teacher Emotion Questionnaire.

- Educational Psychology*, 38(3), 325–349.
<https://doi.org/10.1080/01443410.2017.1382682>
- Burstall, C. (1975). Primary French in the balance. *Educational Research*, 17(3), 193–198.
<https://doi.org/10.1080/0013188750170304>
- Burtless, G. (2002). Randomized field trials for policy evaluation: Why not in education? In F. Mosteller & R. F. Boruch (Eds.), *Evidence matters: Randomized trials in education research* (pp. 179–197). Brookings Institution Press.
- Butler, Y. G. (2017). The dynamics of motivation development among young learners of English in China. In Enever J. & Lindgren (Eds.), *Early language learning: Complexity and mixed methods* (pp. 189–207). Multilingual Matters.
- Butler, Y. G. (2025). *Children's additional language learning in instructional settings: Implications for teaching and future research*. Multilingual Matters.
- Cable, C., Driscoll, P., Mitchell, R., Sing, S., Cremin, T., Earl, J., Eyres, I., Holmes, B., Martin, C., & Heins, B. (2010). *Languages learning at Key Stage 2. A longitudinal study*. Department for Children, Schools and Families.
<https://eprints.soton.ac.uk/143157/1/DCSF-RR198.pdf>
- Camburn, E. M., Goldring, E., Sebastian, J., May, H., & Huff, J. (2016). An examination of the benefits, limitations, and challenges of conducting randomized experiments with principals. *Educational Administration Quarterly*, 52(2), 187–220.
<https://doi.org/10.1177/0013161X15617808>
- Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1(1), 1–47.
<https://doi.org/10.1093/applin/I.1.1>
- Carey, E. G., Ridler, I., Ford, T. J., & Stringaris, A. (2023). Editorial perspective: When is a ‘small effect’ actually large and impactful? *Journal of Child Psychology and Psychiatry*, 64(11), 1643–1647. <https://doi.org/10.1111/jcpp.13817>
- Carle, E. (1967). *Brown bear, brown bear, what do you see?* Puffin.
- Carle, E. (1969). *The Very Hungry Caterpillar*. Puffin.
- Carreira, J. M. (2006). Motivation for learning English as a foreign language in Japanese elementary schools. *JALT*, 28(2), 135–158. <https://doi.org/10.37546/JALTJJ28.2-2>
- Cenoz, J. (2003). The influence of age on the acquisition of English: General proficiency, attitudes and code-mixing. In M. del P. García Mayo & M. L. García Lecumberri (Eds.), *Age and the acquisition of English as a foreign language* (pp. 77–93). Multilingual Matters Ltd. <https://doi.org/10.21832/9781853596407-005>
- Cerezo, L., Caras, A., & Leow, R. P. (2016). The effectiveness of guided induction versus deductive instruction on the development of complex Spanish “gustar” structures. *Studies*

- in *Second Language Acquisition*, 38(2), 265–291.
<https://doi.org/10.1017/S0272263116000139>
- Chacón, C. T. (2005). Teachers' perceived efficacy among English as a foreign language teachers in middle schools in Venezuela. *Teaching and Teacher Education*, 21(3), 257–272. <https://doi.org/10.1016/j.tate.2005.01.001>
- Chambers, G. (2019). Pupils' perceptions of Key Stage 2 to Key Stage 3 transition in modern foreign languages. *The Language Learning Journal*, 47(1), 19–33.
<https://doi.org/10.1080/09571736.2016.1172331>
- Chapelle, C. A. (2007). Challenges in evaluation of innovation: Observations from technology research. *Innovation in Language Learning and Teaching*, 1(1), 30–45.
<https://doi.org/10.2167/illt041.0>
- Choi, E., & Lee, J. (2016). Investigating the relationship of target language proficiency and self-efficacy among nonnative EFL teachers. *System*, 58, 49–63.
<https://doi.org/10.1016/j.system.2016.02.010>
- Christensen, R. (2023). *Ordinal - Regression models for ordinal data*. R package version 2023.12-4.1. <https://CRAN.R-project.org/package=ordinal>
- Cloutier, C., & Ravasi, D. (2021). Using tables to enhance trustworthiness in qualitative research. *Strategic Organization*, 19(1), 113–133.
<https://doi.org/10.1177/1476127020979329>
- Cobb, T., & Laufer, B. (2021). The nuclear word family list: A list of the most frequent family members, including base and affixed words. *Language Learning*, 71(3), 834–871.
<https://doi.org/10.1111/lang.12452>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- Cohen, L., Manion, L., & Morrison, K. (2018). *Research methods in education* (8th ed.). Routledge.
- Collen, I. (2020). *Language trends 2020: Language teaching in primary and secondary schools in England*.
https://www.britishcouncil.org/sites/default/files/language_trends_2020_0.pdf
- Collen, I. (2022). *Language trends 2022: Language teaching in primary and secondary schools in England*.
https://www.britishcouncil.org/sites/default/files/language_trends_report_2022.pdf
- Collen, I. (2023) *Language trends England 2023: Language teaching in primary and secondary schools in England*.
https://www.britishcouncil.org/sites/default/files/language_trends_england_2023.pdf

- Collen, I., & Duff, J. (2024). *Language trends England 2024: Language teaching in primary, secondary and independent schools in England*.
https://www.britishcouncil.org/sites/default/files/language_trend_england_2024.pdf
- Collen, I., & Duff, J. (2025). *Language trends England 2025: Language teaching in primary, secondary and independent schools in England*.
https://www.britishcouncil.org/sites/default/files/language_trends_england_2025.pdf
- Collins, K. M. T. (2017). Sampling decisions in educational research. In D. Wyse, N. Selwyn, E. Smith, & L. Suter (Eds.), *The BERA/SAGE handbook of educational research* (pp. 280–292). SAGE Publications Ltd.
- Conklin, K. (2020). Processing single-word and multiword items. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 174–188). Routledge.
- Connolly, P., Biggart, A., Miller, S., & O'Hare, L. (2017). *Using randomised controlled trials in education*. Sage Publications Ltd. <https://doi.org/10.4135/9781473920385>
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Houghton Mifflin.
- Cooke, S., & Faez, F. (2018). Self-efficacy beliefs of novice French as a second language teachers: A case study of Ontario teachers. *Canadian Journal of Applied Linguistics*, 21(2), 1–18. <https://doi.org/10.7202/1057963ar>
- Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Council of Europe Publishing.
- Courtney, L. (2017). Transition in modern foreign languages: A longitudinal study of motivation for language learning and second language proficiency. *Oxford Review of Education*, 43(4), 462–481. <https://doi.org/10.1080/03054985.2017.1329721>
- Courtney, L., & Graham, S. (2019). 'It's like having a test but in a fun way': Young learners' perceptions of a digital game-based assessment of early language learning. *Language Teaching for Young Learners*, 1(2), 161–186. <https://doi.org/10.1075/ltyl.18009.cou>
- Courtney, L., Graham, S., Tonkyn, A., & Marinis, T. (2017). Individual differences in early language learning: A study of English learners of French. *Applied Linguistics*, 38(6), 824–847. <https://doi.org/10.1093/applin/amv071>
- Crichton, H., & Templeton, B. (2010). Curriculum for Excellence: the way forward for primary languages in Scotland? *The Language Learning Journal*, 38(2), 139–147. <https://doi.org/10.1080/09571731003790391>
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334. <https://doi.org/10.1007/BF02310555>

- Csizér, K., & Dörnyei, Z. (2005). The internal structure of language learning motivation and its relationship with language choice and learning effort. *The Modern Language Journal*, 89(1), 19–36. <https://doi.org/10.1111/j.0026-7902.2005.00263.x>
- Cunnings, I. (2012). An overview of mixed-effects statistical models for second language researchers. *Second Language Research*, 28(3), 369–382. <https://doi.org/10.1177/0267658312443651>
- Cunnings, I., & Finlayson, I. (2015). Mixed effects modeling and longitudinal data analysis. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 159–181). Routledge.
- Daftari, G. E., & Tavit, Z. M. (2017). The impact of non-native English teachers' linguistic insecurity on learners' productive skills. *The Journal of Language and Linguistic Studies*, 13(1), 379–399.
- Dang, T. N. Y., & Webb, S. (2016). Making an essential word list for beginners. In I. S. P. Nation (Ed.), *Making and using word lists for language learning* (pp. 153–167). John Benjamins Publishing Company.
- Dang, T. N. Y., Webb, S., & Coxhead, A. (2020). Evaluating lists of high-frequency words: Teachers' and learners' perspectives. *Language Teaching Research*, 26(4), 617–641. <https://doi.org/10.1177/1362168820911189>
- David, A. (2008). Vocabulary breadth in French L2 learners. *The Language Learning Journal*, 36(2), 167–180. <https://doi.org/10.1080/09571730802389991>
- David, T., Tonkin, J., Powell, S., & Anderson, C. (2005). Ethical aspects of power in research with children. In A. Farrell (Ed.), *Ethical research with children* (pp. 124–137). Open University Press.
- Davis, F. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3), 319–340. <https://www.jstor.org/stable/249008>
- Davis, F., Bagozzi, R., & Warshaw, P. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science*, 35(8), 982–1003. <https://www.jstor.org/stable/2632151>
- Dearing, R., & King, L. (2007). *Languages review 6970-DfES*.
- DeBruine, L. (2023). *faux: Simulation for factorial designs*. R package version 1.2.1. <http://doi.org/10.5281/zenodo.2669586>
- Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behaviour*. Plenum Press.
- Decorte, R., Colognesi, S., Dachet, D., & Meunier, F. (2024). Perceptions and practices of L2 teachers in metacognitive oracy instruction: A longitudinal case study. *Australian Journal of Applied Linguistics*, 7 (3), 1–36. <https://doi.org/10.29140/ajal.v7n3.1531>

- De Groot, A. M. B., & Keijzer, R. (2000). What is hard to learn is easy to forget: The roles of word concreteness, cognate status, and word frequency in foreign-language vocabulary learning and forgetting. *Symposium Abstracts*, 50(1), 1–56. <https://doi.org/10.1111/0023-8333.00110>
- DeKeyser, R. (1997). Beyond explicit rule learning. Automatizing second language morphosyntax. *Studies in Second Language Acquisition*, 19(2), 195–221. <https://doi.org/10.1017/S0272263197002040>
- DeKeyser, R. (2003). Implicit and explicit learning. In C. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 313–348). Blackwell Publishing Ltd.
- DeKeyser, R. (2017). Knowledge and skill in ISLA. In S. Loewen & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition* (pp. 15–33). Routledge.
- DeKeyser, R., & Prieto Botana, G. (2015). The effectiveness of processing instruction in L2 grammar acquisition: A narrative review. *Applied Linguistics*, 36(3), 290–305. <https://doi.org/10.1093/applin/amu071>
- Department for Children, Schools & Families. (2005). *Key Stage 2 Framework for languages – Parts 1, 2 and 3*. <https://webarchive.nationalarchives.gov.uk/20110511211850/http://nationalstrategies.standards.dcsf.gov.uk/node/85274>
- Department for Education. (n.d.). Explore education statistics - GOV.UK. Retrieved 30/08/2025 from <https://explore-education-statistics.service.gov.uk/>
- Department for Education. (2013a). *Languages programmes of study: Key stage 2 national curriculum in England*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/239042/PRIMARY_national_curriculum_-_Languages.pdf
- Department for Education. (2013b). *The national curriculum in England: Key stages 1 and 2 framework document September 2013*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/425601/PRIMARY_national_curriculum.pdf
- Department for Education. (2014). *The national curriculum in England. Key stages 3 and 4 framework document*. DfES. <https://www.gov.uk/national-curriculum/overview>
- Department for Education. (2025). *Curriculum and assessment review: Building a world-class curriculum for all. Final report*. https://assets.publishing.service.gov.uk/media/690b96bbc22e4ed8b051854d/Curriculum_and_Assessment_Review_final_report_-_Building_a_world-class_curriculum_for_all.pdf
- Department for Education and Skills. (2002). *Languages for all: Languages for life. A strategy for England*. <https://www.education-uk.org/documents/pdfs/2002-languages-for-all.pdf>

- Dewaele, J.-M. (2019). The effect of classroom emotions, attitudes toward English, and teacher behavior on willingness to communicate among English foreign language learners. *Journal of Language and Social Psychology*, 38(4), 523–535. <https://doi.org/10.1177/0261927X19864996>
- Dewaele, J.-M. (2022). Enjoyment. In S. Li, P. Hiver, & M. Papi (Eds.), *The Routledge handbook of second language acquisition and individual differences* (pp. 190–206). Routledge. <https://doi.org/10.4324/9781003270546-16>
- Dewaele, J.-M., & MacIntyre, P. D. (2014). The two faces of Janus? Anxiety and enjoyment in the foreign language classroom. *Studies in Second Language Learning and Teaching*, 4(2), 237–274. <https://doi.org/10.14746/ssllt.2014.4.2.5>
- Dewaele, J.-M., Magdalena, A. F., & Saito, K. (2019). The effect of perception of teacher characteristics on Spanish EFL learners' anxiety and enjoyment. *The Modern Language Journal*, 103(2), 412–427. <https://doi.org/10.1111/modl.12555>
- Dewaele, J.-M., Witney, J., Saito, K., & Dewaele, L. (2018). Foreign language enjoyment and anxiety: The effect of teacher and learner variables. *Language Teaching Research*, 22(6), 676–697. <https://doi.org/10.1177/1362168817692161>
- Diamond, I., & Jefferies, J. (2001). *Beginning statistics: An introduction for social scientists*. SAGE.
- Dörnyei, Z. (1994). Motivation and motivating in the foreign language classroom. *The Modern Language Journal*, 78(3), 273–284. <https://doi.org/10.2307/330107>
- Dörnyei, Z. (1998). Motivation in second and foreign language learning. *Language Teaching*, 31(3), 117–135. <https://doi.org/10.1017/S026144480001315X>
- Dörnyei, Z. (2013). Motivation in second language learning. In M. Celce-Murcia, D. Brinton, & M. A. Snow (Eds.), *Teaching English as a second or foreign language* (4th ed.) (pp. 518–531). Cengage Learning.
- Dörnyei, Z., & Csizér, K. (1998). Ten commandments for motivating language learners: Results of an empirical study. *Language Teaching Research*, 2(3), 203–229. <https://doi.org/10.1177/136216889800200303>
- Dörnyei, Z., & Taguchi, T. (2010). *Questionnaires in second language research: Construction, administration, and processing* (2nd ed.). Routledge. <https://doi.org/10.4324/9780203864739>
- Driscoll, P. (1999). Teacher expertise in the primary modern foreign languages classroom. In D. Frost & P. Driscoll (Eds.), *Teaching modern languages in the primary school* (pp. 27–47). Taylor & Francis Group.
- Driscoll, P., Jones, J., & Macrory, G. (2004). *The provision of MFL learning for pupils at Key Stage 2. Research Report 572*. DfES.

- Driscoll, P., Jones, J., Martin, C., Graham-Matheson, L., Dismore, H., & Sykes, R. (2004). *A systematic review of the characteristics of effective foreign language teaching to pupils between the ages 7 and 11*. <http://eppi.ioe.ac.uk/>
- Dudley, A., & Marsden, E. (2024). The lexical content of high-stakes national exams in French, German, and Spanish in England. *Foreign Language Annals*, 57(2), 311–338. <https://doi.org/10.1111/flan.12751>
- Duff, P. (1991). Innovations in foreign language education: An evaluation of three Hungarian-English dual-language schools. *Journal of Multilingual and Multicultural Development*, 12(6), 459–476. <https://doi.org/10.1080/01434632.1991.9994476>
- Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, 105(3), 399–412. <https://doi.org/10.1111/bjop.12046>
- Du Plessis, A. (2015). Effective education: Conceptualising the meaning of out-of-field teaching practices for teachers, teacher quality and school leaders. *International Journal of Educational Research*, 72, 89–102. <https://doi.org/10.1016/j.ijer.2015.05.005>
- Du Plessis, A., Carroll, A., & Gillies, R. M. (2015). Understanding the lived experiences of novice out-of-field teachers in relation to school leadership practices. *Asia-Pacific Journal of Teacher Education*, 43(1), 4–21. <https://doi.org/10.1080/1359866X.2014.937393>
- Eccles, J. (1999). The development of children ages 6 to 14. *The Future of Children*, 9(2), 30–44. <https://doi.org/10.2307/1602703>
- Eccles, J., Wigfield, A., Harold, R. D., & Blumenfeld, P. (1993). Age and gender differences in children's self- and task perceptions during elementary school. *Child Development*, 64(3), 830–847. <https://doi.org/10.1111/j.1467-8624.1993.tb02946.x>
- Edelenbos, P., Johnstone, R., & Kubanek, A. (2006). *The main pedagogical principles underlying the teaching of languages to very young learners: Languages for the children of Europe*. https://ec.europa.eu/assets/eac/languages/policy/language-policy/documents/young_en.pdf
- Edelenbos, P., & Kubanek, A. (2009). Early foreign language learning: Published research, good practice and main principles. In M. Nikolov (Ed.), *The age factor and early language learning* (pp. 39–59). De Gruyter, Inc.
- Edelenbos, P. & Suhre, C. (1994). A comparison of courses for English in primary education. *Studies in Educational Evaluation*, 20(4), 513–534. [https://doi.org/10.1016/0191-491X\(94\)00040-N](https://doi.org/10.1016/0191-491X(94)00040-N)
- Edwards, A., & Talbot, R. (1994). *The hard-pressed researcher*. Longman Group Ltd.

- Ehri, L. C. (2005). Learning to read words: Theory, findings, and issues. *Scientific Studies of Reading*, 9(2), 167–188. https://doi.org/10.1207/s1532799xssr0902_4
- Ellis, R. (1994). *The study of second language acquisition*. Oxford University Press.
- Enever, J. (2009). Can today's early language learners in England become tomorrow's plurilingual European citizens? In Nikolov, M. (Ed.), *Early learning of modern foreign languages: Processes and outcomes* (pp. 15–29). Multilingual Matters.
- Enever, J. (2011). *ELLiE: Early language learning in Europe*. British Council. https://www.researchgate.net/publication/265383696_ELLiE_Early_language_learning_in_Europe/link/56681d4d08ae8905db8df8a6/download
- Erlam, R. (2003). The effects of deductive and inductive instruction on the acquisition of direct object pronouns in French as a second language. *The Modern Language Journal*, 87(2), 242–260. <https://doi.org/10.1111/1540-4781.00188>
- Erlam, R. (2005). Language aptitude and its relationship to instructional effectiveness in second language acquisition. *Language Teaching Research*, 9(2), 147–171. <https://doi.org/10.1191/1362168805lr161oa>
- Erler, L. (2004). Near-beginner learners of French are reading at a disability level. *Francophonie*, 30, 9–15.
- Erler, L., & Macaro, E. (2011). Decoding ability in French as a foreign language and language learning motivation. *Modern Language Journal*, 95(4), 496–518. <https://doi.org/10.1111/j.1540-4781.2011.01238.x>
- Ertmer, P. (1999). Addressing first- and second-order barriers to change. *Educational Technology Research and Development*, 47(4), 47–61. <https://doi:10.1007/BF02299597>
- Ethical research involving children (ERIC). (n.d.). Retrieved 08/01/2022 from <https://childethics.com/>
- European Commission. (2018). Proposal for a Council recommendation on a comprehensive approach to the teaching and learning of languages COM(2018) 272 final. <https://eur-lex.europa.eu/>
- European Council Barcelona. (2002). *European council Barcelona conclusions of the presidency*. <http://www.europarl.ep.ec/bulletins>
- Farrell, A. (2005). New times in ethical research with children. In A. Farrell (Ed.), *Ethical research with children* (pp. 166–175). Open University Press.
- Farrell, T. S. C., & Richards, J. C. (2007). Teachers' language proficiency. In T. S. C. Farrell (Ed.), *Reflective language teaching: From research to practice* (pp. 55–66). Continuum.
- Farrimond, H. (2017). The ethics of research. In D. Wyse, N. Selwyn, E. Smith, & L. Suter (Eds.), *The BERA/SAGE handbook of educational research* (pp. 72–89). Sage Publications.

- Fauth, B., Decristan, J., Decker, A.-T., Büttner, G., Hardy, I., Klieme, E., & Kunter, M. (2019). The effects of teacher competence on student outcomes in elementary science education: The mediating role of teaching quality. *Teaching and Teacher Education*, 86, 102882. <https://doi.org/10.1016/j.tate.2019.102882>
- Fein, E. C. J., Gilmour, J., Machin, T., & Hendry, L. (2022). *Statistics for research students: An open access resource with self-tests and illustrative examples*. University of Southern Queensland.
- Fenyvesi, K. (2020). English learning motivation of young learners in Danish primary schools. *Language Teaching Research*, 24(5), 690–713. <https://doi.org/10.1177/1362168818804835>
- Fenyvesi, K., Hansen, M. B., & Cadierno, T. (2020). The role of individual differences in younger vs. older primary school learners of English in Denmark. *International Review of Applied Linguistics in Language Teaching, IRAL*, 58(3), 289–322. <https://doi.org/10.1515/iral-2017-0053>
- Field, A. P., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. SAGE. <http://prism.librarymanagementcloud.co.uk/chi-ac/items/521863>
- File, K. & Adams, R. (2010). Should vocabulary instruction be integrated or isolated? *TESOL quarterly*, 44(2), 222-249. <https://doi.org/10.5054/tq.2010.219943>
- Finch, K., Theakston, A., & Serratrice, L. (2018). Teaching modern foreign languages in multilingual classrooms: An examination of Key Stage 2 teachers' experiences. *The Language Learning Journal*, 1–15. <https://doi.org/10.1080/09571736.2018.1448432>
- Flora, D. B. (2020). Your coefficient alpha is probably wrong, but which coefficient omega is right? A tutorial on using R to obtain better reliability estimates. *Advances in Methods and Practices in Psychological Science*, 3(4), 484–501. <https://doi.org/10.1177/2515245920951747>
- Frenzel, A. C., Becker-Kurz, B., Pekrun, R., & Goetz, T. (2015). Teaching this class drives me nuts! Examining the person and context specificity of teacher emotions. *PloS One*, 10(6), e0129630. <https://doi.org/10.1371/journal.pone.0129630>
- Frenzel, A. C., Becker-Kurz, B., Pekrun, R., Goetz, T., & Lüdtke, O. (2018). Emotion transmission in the classroom revisited: A reciprocal effects model of teacher and student enjoyment. *Journal of Educational Psychology*, 110(5), 628–639. <https://doi.org/10.1037/edu0000228>
- Frenzel, A. C., Goetz, T., Lüdtke, O., Pekrun, R., & Sutton, R. E. (2009). Emotional transmission in the classroom: Exploring the relationship between teacher and student enjoyment. *Journal of Educational Psychology*, 101(3), 705–716. <https://doi.org/10.1037/a0014695>

- Frenzel, A. C., Goetz, T., Stephens, E. J., & Jacob, B. (2009). Antecedents and effects of teachers' emotional experiences: An integrated perspective and empirical test. In P. A. Schutz (Ed.), *Advances in teacher emotion research: The impact on teachers' lives* (pp. 129–151). Springer. https://doi.org/10.1007/978-1-4419-0564-2_7
- Frenzel, A. C., Pekrun, R., Goetz, T., Daniels, L. M., Durksen, T. L., Becker-Kurz, B., & Klassen, R. M. (2016). Measuring teachers' enjoyment, anger, and anxiety: The teacher emotions scales (TES). *Contemporary Educational Psychology*, 46, 148–163. <https://doi.org/10.1016/j.cedpsych.2016.05.003>
- Fulcher, G. (2010). *Practical language testing*. Hodder Education.
- Gamer, M., & Lemon, J. (2019). *irr: Various coefficients of interrater reliability and agreement*. R package version 0.84.1. <https://CRAN.R-project.org/package=irr>
- Gardner, R. C., & Lambert, W. E. (1972). *Attitudes and motivation in second-language learning*. Newbury House Publishers, Inc.
- Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.
- Gibson, W. J., & Brown, A. (2011). *Working with qualitative data*. SAGE.
- Gillham, B. (2000). *Developing a questionnaire*. Continuum.
- Gillham, B. (2008). *Developing a questionnaire* (2nd ed.). Continuum.
- Giordano, M. J. (2021). Lexical coverage in dialogue listening. *Language Teaching Research*, 27(6), 1504–1531. <https://doi.org/10.1177/1362168821989869>
- Godfroid, A., & Andringa, S. (2023). Uncovering sampling biases, advancing inclusivity, and rethinking theoretical accounts in Second Language Acquisition: Introduction to the special issue “SLA for All?” *Language Learning*, 73(4), 981–1002. <https://doi.org/10.1111/lang.12620>
- Goldacre, B. (2013). *Building evidence into education*. Department for Education. <https://doi.org/10.13140/RG.2.1.5101.8967>
- Gorard, S. (2013). *Research design: Creating robust approaches for the social sciences*. SAGE.
- Gorard, S. (2017). An introduction to the importance of research design. In D. Wyse, N. Selwyn, E. Smith, & L. Suter (Eds.), *The BERA/SAGE handbook of educational research* (pp. 203–212). SAGE Reference.
- Gorard, S., & Taylor, C. (2004). *Combining methods in educational and social research*. Open University Press.
- Goswami, U. (2005). Synthetic phonics and learning to read: A cross-language perspective. *Educational Psychology in Practice*, 21(4), 273–282. <https://doi.org/10.1080/02667360500344823>

- Graham, S., Courtney, L., Marinis, T., & Tonkyn, A. (2014). *Primary modern languages: The impact of teaching approaches on attainment and preparedness for secondary school language learning*. The Nuffield Foundation.
- Graham, S., Courtney, L., Marinis, T., & Tonkyn, A. (2017). Early language learning: The impact of teaching and teacher factors. *Language Learning*, 67(4), 922–958. <https://doi.org/10.1111/lang.12251>
- Graham, S., Courtney, L., Tonkyn, A., & Marinis, T. (2016). Motivational trajectories for early language learning across the primary–secondary school transition. *British Educational Research Journal*, 42(4), 682–702. <https://doi.org/10.1002/berj.3230>
- Greene, S., & Hill, M. (2005). Researching children’s experience: Methods and methodological issues. In S. Greene & D. Hogan (Eds.), *Researching children’s experience: Approaches and methods* (pp. 1–21). SAGE.
- Greig, A., Taylor, J., & MacKay, T. (2013). *Doing research with children: A practical guide* (3rd ed.). SAGE.
- Gries, S. (2021). (Generalized linear) Mixed-effects modeling: A learner corpus example. *Language Learning*, 71(3), 757–798. <https://doi.org/10.1111/lang.12448>
- Gu, P. Y. (2020). Strategies for learning vocabulary. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 271–287). Routledge.
- Gueron, J. M. (2002). The politics of random assignment: Implementing studies and affecting policy. In F. F. Mosteller & R. F. Boruch (Eds.), *Evidence matters: Randomized trials in education research* (pp. 15–49). Brookings Institution Press.
- Guskey, T. (2002). Professional Development and Teacher Change. *Teachers and Teaching*, 8(3), 381–391. <https://doi.org/10.1080/135406002100000512>
- Häcker, M. (2008). Eleven pets and 20 ways to express one’s opinion: The vocabulary learners of German acquire at English secondary schools. *The Language Learning Journal*, 36(2), 215–226. <https://doi.org/10.1080/09571730802393183>
- Haight, C. E., Herron, C., & Cole, S. P. (2007). The effects of deductive and guided inductive instructional approaches on the learning of grammar in the elementary foreign language college classroom. *Foreign Language Annals*, 40(2), 288–310. <https://doi.org/10.1111/j.1944-9720.2007.tb03202.x>
- Hajduk, G. K., & Gallois, E. (2023, December 3). *Introduction to linear mixed models*. Coding Club <https://Ourcodingclub.Github.io/Tutorials/Mixed-Models/>
- Hamada, M., & Koda, K. (2008). Influence of first language orthographic experience on second language decoding and word learning. *Language Learning*, 58(1), 1–31. <https://doi.org/10.1111/j.1467-9922.2007.00433.x>

- Hammersley, M. (2017). Research ethics. In R. Coe, M. J. Waring, L. V Hedges, & J. Arthur (Eds.), *Research methods and methodologies in education* (2nd ed.) (pp. 57–68). Sage Publications Ltd.
- Harlen, W. (2005). Trusting teachers' judgement: Research evidence of the reliability and validity of teachers' assessment used for summative purposes. *Research Papers in Education*, 20(3), 245-270. <https://doi.org/10.1080/02671520500193744>
- Harris, J. (2009). Late-stage refocusing of Irish-language programme evaluation: Maximizing the potential for productive debate and remediation. *Language Teaching Research*, 13, 55–76. <https://doi.org/10.1177/1362168808095523>
- Hart, R. (1992). *Children's participation: From tokenism to citizenship*. Innocenti Essays No.4 Series. UNICEF International Child Development Centre.
- Hashimoto, B. J., & Egbert, J. (2019). More than frequency? Exploring predictors of word difficulty for second language learners. *Language Learning*, 69(4), 839–872. <https://doi.org/10.1111/lang.12353>
- Hattie, J. (2008). *Visible Learning*. Routledge. <https://doi.org/10.4324/9780203887332>
- Hawkes, R. (2013). *rachelhawkes.com - Yr 3 French*.
https://www.rachelhawkes.com/Resources/Y3_French/Yr3French.php
- Hawkes, R., & Guillou, M.-O. (2022). *rachelhawkes.com - Key stage 2 languages*.
<https://www.rachelhawkes.com/Resources/KS2Lang.php>
- Hawkes, R., Marsden, E., Avery, N., Kasprowicz, R., & Woore, R. (2019). Making language learning make sense at the National Centre for Excellence for Language Pedagogy. *Languages, Society & Policy*, 1–6. <https://doi.org/10.17863/CAM.40160>
- Haynes, L., Service, O., Goldacre, B., & Torgerson, D. (2012). *Test, learn, adapt: Developing public policy with randomised controlled trials*. Cabinet Office Behavioural Insights Team.
- He, X. (Stella), & Godfroid, A. (2019). Choosing words to teach: A novel method for vocabulary selection and its practical application. *TESOL Quarterly*, 53(2), 348–371. <https://doi.org/10.1002/tesq.483>
- Hedges, L. (2017). Design of empirical research. In R. Coe, M. Waring, L. Hedges & J. Arthur (Eds.), *Research methods and methodologies in education* (pp. 25-33). Sage Publications Ltd.
- Heinzmann, S. (2013). *Young language learners' motivation and attitudes*. Bloomsbury.
- Hennessy, E., & Heary, C. (2005). Exploring children's views through focus groups. In S. Greene & D. Hogan (Eds.), *Researching children's experience: Approaches and methods* (pp. 236–252). SAGE.

- Henry, N., Culman, H., & VanPatten, B. (2009). More on the effects of explicit information in instructed SLA. *Studies in Second Language Acquisition*, 31(4), 559–575. <https://doi.org/10.1017/S0272263109990027>
- Higgins, S. (2017). Impact evaluation. In R. Coe, M. J. Waring, L. V. Hedges, & J. Arthur (Eds.), *Research methods and methodologies in education* (2nd ed.) (pp. 145–152). Sage Publications Ltd.
- Hill, M. (2005). Ethical considerations in researching children’s experiences. In S. Greene & D. Hogan (Eds.), *Researching children’s experience: Approaches and methods* (pp. 61–86). Sage.
- Hilton, H. (2017). Individual differences and English L2 learning in two primary classrooms in France. In Enever J. & Lindgren (Eds.) *Early language learning: Complexity and mixed methods* (pp. 148-168). Multilingual Matters.
- Hobbs, V., & Kubanyiova, M. (2008). The challenges of researching language teachers: What research manuals don’t tell us. *Language Teaching Research*, 12(4), 495–513. <https://doi.org/10.1177/1362168808097162>
- Holmes, B., & Myles, F. (2019). *White paper: Primary languages policy in England - The way forward*. www.rpl.uk/policy/
- Hopman, E. W. M., & MacDonald, M. C. (2018). Production practice during language learning improves comprehension. *Psychological Science*, 29(6), 961–971. <https://doi.org/10.1177/0956797618754486>
- Horwitz, E. K. (1996). Even teachers get the blues: Recognizing and alleviating language teachers’ feelings of foreign language anxiety. *Foreign Language Annals*, 29(3), 365–372. <https://doi.org/10.1111/j.1944-9720.1996.tb01248.x>
- Horwitz, E. K., Horwitz, M. B., & Cope, J. (1986). Foreign language classroom anxiety. *The Modern Language Journal*, 70(2), 125. <https://doi.org/10.2307/327317>
- Howard, S. (2013). Risk-aversion: Understanding teachers’ resistance to technology integration. *Technology, Pedagogy and Education*, 22(3), 357-372. <https://doi.org/10.1080/1475939X.2013.802995>
- Hu, C.-F. (2008). Use orthography in L2 auditory word learning: Who benefits? *Reading and Writing*, 21(8), 823–841. <https://doi.org/10.1007/s11145-007-9094-6>
- Huang, S., Willson, V., & Eslami, Z. (2012). The effects of task involvement load on L2 incidental vocabulary learning: A meta-analytic study. *The Modern Language Journal*, 96(4), 544–557. <https://doi.org/10.1111/j.1540-4781.2012.01394.x>
- Hunleth, J. (2011). Beyond ‘on’ or ‘with’: Questioning power dynamics and knowledge production in ‘child-oriented’ research methodology. *Childhood*, 18(1), 81–93. <https://doi.org/10.1177/0907568210371234>

- Hunt, M., Barnes, A., Powell, B., & Martin, C. (2008). Moving on: The challenges for foreign language learning on transition from primary to secondary school. *Teaching and Teacher Education*, 24, 915–926. <https://doi.org/10.1016/j.tate.2007.08.005>
- Hwang, H. (2025). Undesirable difficulty of interleaved practice: The importance of initial blocked practice for declarative knowledge development in low-achieving adolescents. *Language Learning*, 75(1), 5–41. <https://doi.org/10.1111/lang.12659>
- Issa, B. I., & Morgan-Short, K. (2019). Effects of external and internal attentional manipulations on second language grammar development. *Studies in Second Language Acquisition*, 41(2), 389–417. <https://doi.org/10.1017/S0272226311800013X>
- Jaekel, N., Schurig, M., Florian, M., & Ritter, M. (2017). From early starters to late finishers? A longitudinal study of early foreign language learning in school. *Language Learning*, 67(3), 631–664. <https://doi.org/10.1111/lang.12242>
- Johnstone, R. (2000). Context-sensitive assessment of modern languages in primary (elementary) and early secondary education: Scotland and the European experience. *Language Testing*, 17(2), 123–143. <https://doi.org/10.1177/026553220001700202>
- Johnstone, R. (2002). *Addressing the age factor: Some implications for languages policy. Guide for the development of language education policies in Europe from linguistic diversity to plurilingual education reference study*. <https://rm.coe.int/addressing-the-age-factor-some-implications-for-languages-policy-/1680886e92>
- Johnstone, R. (2008). Teaching a modern language to children at primary school: Making it work for all. *English Teaching & Learning*, 32(4), 1–40. <https://doi.org/10.6330/ETL.2008.32.4.01>
- Johnstone, R. (2009). An early start: What are the key conditions for generalized success? In J. Enever, J. Moon, & U. Raman (Eds.), *Young learner English language policy and implementation: International perspectives* (pp. 31–42). Garnet Publishing Ltd. https://www.researchgate.net/publication/265383704_Young_Learner_English_Language_Policy_and_Implementation_International_Perspectives
- Jones, N., Gille, E., Kortes, J., & Ashton, K. (2012). *First European survey on language competences: Final report*. http://www.surveylang.org/media/ExecutivesummaryoftheESLC_210612.pdf
- Jung, J. (2016). Effects of glosses on learning of L2 grammar and vocabulary. *Language Teaching Research*, 20(1), 92–112. <https://doi.org/10.1177/1362168815571151>
- Kang, S. H. K., Gollan, T. H., & Pashler, H. (2013). Don't just repeat after me: Retrieval practice is better than imitation for foreign vocabulary learning. *Psychonomic Bulletin & Review*, 20(6), 1259–1265. <https://doi.org/10.3758/s13423-013-0450-z>

- Karavas-Doukas, K. (1998). Evaluating the implementation of educational innovations: Lessons from the past. In P. Rea-Dickins & K. P. Germaine (Eds.), *Managing evaluation and innovation in language teaching: Building bridges* (pp. 25–50). Addison Wesley Longman Ltd.
- Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science (American Association for the Advancement of Science)*, 319(5865), 966–968. <https://doi.org/10.1126/science.1152408>
- Kasprowicz, R. (2025). Is progression in primary languages possible? Reflections from a large-scale longitudinal research study. *The Language Learning Journal*, 1-9. <https://doi.org/10.1080/09571736.2025.2558994>
- Kasprowicz, R., & Marsden, E. (2018). Towards ecological validity in research into input-based practice: Form spotting can be as beneficial as form-meaning practice. *Applied Linguistics*, 39(6), 886–911. <https://doi.org/10.1093/applin/amw051>
- Kasprowicz, R., Marsden, E., & Sephton, N. (2019). Investigating distribution of practice effects for the learning of foreign language verb morphology in the young learner classroom. *The Modern Language Journal*, 103(3), 580–606. <https://doi.org/10.1111/modl.12586>
- Keller, M. M., Becker, E. S., Frenzel, A. C., & Taxer, J. L. (2018). When teacher enthusiasm is authentic or inauthentic: Lesson profiles of teacher enthusiasm and relations to students' emotions. *AERA Open*, 4(2), 1–16. <https://doi.org/10.1177/2332858418782967>
- Keller, M., Woolfolk Hoy, A., Goetz, T., & Frenzel, A. C. (2016). Teacher enthusiasm: Reviewing and redefining a complex construct. *Educational Psychology Review*, 28(4), 743–769. <https://doi.org/10.1007/s10648-015-9354-y>
- Kellett, M., & Ding, S. (2004). Middle childhood. In S. Fraser, V. Lewis, S. Ding, M. Kellett, & C. Robinson (Eds.), *Doing research with children and young people* (pp. 161–174). Sage Publications Ltd.
- Kemp, L. S., & McDonald, J. L. (2021). Second language vocabulary acquisition: The effects of semantic relatedness, form similarity, and translation direction. *Language Learning*, 0(0), 1–41. <https://doi.org/10.1111/lang.12449>
- Kiely, R. (2006). Evaluation, innovation, and ownership in language programs. *The Modern Language Journal*, 90, 597–601. https://doi.org/10.1111/j.1540-4781.2006.00466_8.x
- Kiely, R. (2009). Small answers to the big question: Learning from language programme evaluation. *Language Teaching Research*, 13(1), 99–116. <https://doi.org/10.1177/1362168808095525>
- Kiely, R., & Rea-Dickins, P. (2005). *Program evaluation in language education*. Palgrave.

- Kim, K. M., & Godfroid, A. (2019). Should we listen or read? Modality effects in implicit and explicit knowledge. *The Modern Language Journal*, 103(3), 648–664. <https://doi.org/10.1111/modl.12583>
- Kim, S. K., & Webb, S. (2022). The effects of spaced practice on second language learning: A meta-analysis. *Language Learning*, 72(1), 269–319. <https://doi.org/10.1111/lang.12479>
- King, N. & Horrocks, C. (2010). *Interviews in qualitative research*. Sage.
- Kolb, A. (2007). How languages are learnt: Primary children's language learning beliefs. *International Journal of Innovation in Language Learning and Teaching*, 1(2), 227–241. <https://doi.org/10.2167/illt033.0>
- Konakli, T. & Akdeniz, R. (2022). The emergence, reasons and results of resistance to change in teachers. *International Journal on Lifelong Education and Leadership*, 8(1), 1-19. <https://doi.org/10.25233/ijlel.1107137>
- Kormos, J. (2020). How does vocabulary fit into theories of second language learning? In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 207–222). Routledge.
- Kostromitina, M., & Plonsky, L. (2022). Elicited imitation tasks as a measure of L2 proficiency - A meta-analysis. *Studies in Second Language Acquisition*, 44(3), 886–911. <https://doi.org/10.1017/S0272263121000395>
- Kraft, M. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, 49(4), 241–253. <https://doi.org/10.3102/0013189X20912798>
- Kremmel, B. (2020). Measuring vocabulary learning progress. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 406–418). Routledge.
- Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, 28(5), 1-26. [doi:10.18637/jss.v028.i05](https://doi.org/10.18637/jss.v028.i05), <https://www.jstatsoft.org/index.php/jss/article/view/v028i05>
- Larson-Hall, J. (2008). Weighing the benefits of studying a foreign language at a younger starting age in a minimal input situation. *Second Language Research*, 24(1), 35–63. <https://doi.org/10.1177/0267658307082981>
- Larson-Hall, J. (2016). *A guide to doing statistics in second language research using SPSS and R* (2nd ed.). Routledge. <https://doi.org/https://doi-org.libproxy.york.ac.uk/10.4324/9781315775661>
- Laufer, B. (2020). Evaluating exercises for learning vocabulary. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 351–368). Routledge.
- Laufer, B., & Aviad-Levitsky, T. (2017). What type of vocabulary knowledge predicts reading comprehension: Word meaning recall or word meaning recognition? *The Modern Language Journal*, 101(4), 729–741. <https://doi.org/10.1111/modl.12431>

- Laufer, B., & Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics*, 22(1), 1–26. <https://doi.org/10.1093/applin/22.1.1>
- Laufer, B., & Nation, I. S. P. (2012). Vocabulary. In S. M. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 163–176). Routledge.
- Lee, S. (2025). *7 advanced Shapiro-Wilk test techniques for data scientists*. Number Analytics. <https://www.numberanalytics.com/blog/advanced-shapiro-wilk-test-techniques>
- Leech, N. L., & Onwuegbuzie, A. J. (2007). A typology of mixed methods research designs. *Quality & Quantity*, 43(2), 265–275. <https://doi.org/10.1007/s11135-007-9105-3>
- Legg, K. (2013). An investigation into teachers' attitudes towards the teaching of modern foreign languages in the primary school. *Education 3-13*, 41(1), 55–62. <https://doi.org/10.1080/03004279.2012.710100>
- Lenth, R. (2024). *emmeans: Estimated marginal means, aka least-squares means*. R package version 1.10.0. <https://CRAN.R-project.org/package=emmeans>
- Leung, K. (2021). Assumptions of logistic regression, clearly explained. *Towards Data Science*. <https://towardsdatascience.com/assumptions-of-logistic-regression-clearly-explained-44d85a22b290/>
- Li, M., & DeKeyser, R. (2017). Perception practice, production practice and musical ability in L2 Mandarin tone-word learning. *Studies in Second Language Acquisition*, 39(4), 593–620. <https://doi.org/10.1017/S0272263116000358>
- Lichtman, K. (2016). Age and learning environment: Are children implicit second language learners? *Journal of Child Language*, 43(3), 707–730. <https://doi.org/10.1017/S0305000915000598>
- Lightbown, P. (2019). Perfecting practice. *The Modern Language Journal*, 103(3), 703–712. <https://doi.org/10.1111/modl.12588>
- Linck, J. A., & Cunnings, I. (2015). The utility and application of mixed-effects models in second language research. *Language Learning*, 65(Supplement 1), 185–207. <https://doi.org/10.1111/lang.12117>
- Lindgren, E. & Enever, J. (2017) Introduction: Mixed methods in early language learning research - examining complexity. In J. Enever & E. Lindgren (Eds.), *Early language learning: Complexity and mixed methods* (pp. 24–30). Multilingual Matters.
- Lindgren, E. & Muñoz, C. (2013). The influence of exposure, parents, and linguistic distance on young European learners' foreign language comprehension. *International Journal of Multilingualism*, 10(1), 105–129. <https://doi.org/10.1080/14790718.2012.679275>
- Lindstromberg, S. (2020). Intentional vocabulary learning. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 240–254). Routledge.

- Lomba-Portela, L., Domínguez-Lloria, S. & Pino-Juste, M. R. (2022). Resistances to educational change: Teachers' perceptions. *Education Sciences*, 12(359), 1-12. <https://doi.org/10.3390/educsci12050359>
- Lonsdale, D., & Le Bras, Y. (2009). *A frequency dictionary of French: Core vocabulary for learners*. Routledge.
- Lopriore, L., & Krikhaar, E. (2011). The school. In J. Enever (Ed.), *ELLiE Early language learning in Europe* (pp. 61–78). British Council. https://www.researchgate.net/publication/265383696_ELLiE_Early_language_learning_in_Europe/link/56681d4d08ae8905db8df8a6/download
- Lortie-Forgues, H., & Inglis, M. (2019). Rigorous large-scale educational RCTs are often uninformative: Should we be concerned? *Educational Researcher*, 48(3), 158–166. <https://doi.org/10.3102/0013189X19832850>
- Low, L., Brown, S., Johnstone, R., & Pirrie, A. (1995). *Foreign languages in Scottish primary schools – Evaluation of the Scottish pilot projects*. Final Report to Scottish Office. Scottish CILT.
- Low, L., Duffield, J., Brown, S., & Johnstone, R. (1993). *Evaluating foreign languages in primary schools*. Scottish CILT.
- Lüdtke, D. (2018). ggeffects: Tidy data frames of marginal effects from regression models. *Journal of Open Source Software*, 3(26), 772. <https://doi.org/10.21105/joss.00772>
- Macaro, E., & Masterman, L. (2006). Does intensive explicit grammar instruction make all the difference? *Language Teaching Research*, 10(3), 297–327. <https://doi.org/10.1191/1362168806lr197oa>
- Machida, T. (2016). Japanese elementary school teachers and English language anxiety. *TESOL Journal*, 7(1), 40–66. <https://doi.org/10.1002/tesj.189>
- Mackey, A. (2006). Feedback, noticing and instructed second language learning. *Applied Linguistics*, 27(3), 405–430. <https://doi.org/10.1093/applin/ami051>
- Macrory, G., & McLachlan, A. (2009). Bringing modern languages into the primary curriculum in England: Investigating effective practice in teacher education. *European Journal of Teacher Education*, 32(3), 259–270. <https://doi.org/10.1080/02619760902756012>
- Maie, R., Oikawa, G., Chen, Y., & Uchihara, T. (2025). Cumulative testing for L2 vocabulary learning: The impact of retrieval practice and proficiency. *TESOL Quarterly*, 0(0), 1–30. <https://doi.org/10.1002/tesq.3391>
- Marijanović, V., Panissal, N., & Billières, M. (2009). How do 9-11-year-old Croatians perceive sounds and read aloud in French? In M. Nikolov (Ed.), *Early learning of modern foreign languages: Processes and outcomes* (pp. 149–165). Multilingual Matters.

- Marinis, T., & Armon-Lotem, S. (2015). Sentence repetition. In S. Armon-Lotem, J. de Jong, & N. Meir (Eds.), *Assessing multilingual children: Disentangling bilingualism from language impairment* (pp. 95–121). Multilingual Matters.
- Marinova-Todd, S. H., Marshall, D. B., & Snow, C. E. (2000). Three misconceptions about age and L2 learning. *TESOL Quarterly*, 34(1), 9. <https://doi.org/10.2307/3588095>
- Marsden, E. (2006). Exploring input processing in the classroom: An experimental comparison of processing instruction and enriched input. *Language Learning*, 56(3), 507–566. <https://doi.org/10.1111/j.1467-9922.2006.00375.x>
- Marsden, E. (2007). Can educational experiments both test a theory and inform practice? *British Educational Research Journal*, 33(4), 565–588. <https://doi.org/10.1080/01411920701434094>
- Marsden, E., & Chen, H.-Y. (2011). The roles of structured input activities in processing instruction and the kinds of knowledge they promote: Roles of structured input activities in processing instruction. *Language Learning*, 61(4), 1058–1098. <https://doi.org/10.1111/j.1467-9922.2011.00661.x>
- Marsden, E., & David, A. (2008). Vocabulary use during conversation: A cross-sectional study of development from year 9 to year 13 among learners of Spanish and French. *The Language Learning Journal*, 36(2), 181–198. <https://doi.org/10.1080/09571730802390031>
- Marsden, E., Dudley, A., & Hawkes, R. (2023). Use of word lists in a high-stakes, low-exposure context: Topic-driven or frequency-informed. *The Modern Language Journal*, 107(3), 669–692. <https://doi.org/10.1111/modl.12866>
- Marsden, E., & Hawkes, R. (2023). Situating practice in a limited-exposure, foreign languages school curriculum. In Y. Suzuki (Ed.), *Practice and automatization in second language research* (pp. 89–118). Routledge.
- Marsden, E. & Hawkes, R. (In press). A narrative account of experiences at research-policy-practice interfaces in England - warts and all. *Studies in Second Language Learning and Teaching*. https://eprints.whiterose.ac.uk/id/eprint/231372/1/MarsdenHawkes_2025_post-print_SSLLT.pdf
- Marsden, E., Hawkes, R., Earnshaw, L., Hobson, V., & Team, NCELP. (2023). *NCELP final report*. https://eprints.whiterose.ac.uk/id/eprint/200637/1/NCELP_Final_Report_9_April_2023.pdf
- Marsden, E., & Torgerson, C. J. (2012). Single group, pre- and post-test research designs: Some methodological concerns. *Oxford Review of Education*, 38(5), 583–616. <https://doi.org/10.1080/03054985.2012.731208>

- Martin, C. (2000). *An analysis of national and international research on the provision of modern foreign languages in primary schools*. Qualifications and Curriculum Authority.
- Martin, C. (2012). Pupils' perceptions of foreign language learning in the primary school – findings from the key stage 2 language learning pathfinder evaluation. *Education 3-13*, 40(4), 343–362. <https://doi.org/10.1080/03004279.2012.691373>
- Masson, J. (2004). The legal context. In S. Fraser, V. Lewis, S. Ding, M. Kellett, & C. Robinson (Eds.), *Doing research with children and young people* (pp. 43–58). Sage Publications Ltd.
- McDonald, R. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum Associates.
- McDonough, K., & Chaikitmongkol, W. (2010). Collaborative syntactic priming activities and EFL learners' production of wh-questions. *The Canadian Modern Language Review*, 66(6), 817–841. <https://doi.org/10.3138/cmlr.66.6.817>
- McLachlan, A. (2009). Modern languages in the primary curriculum: Are we creating conditions for success? *The Language Learning Journal*, 37(2), 183–203. <https://doi.org/10.1080/09571730902928078>
- McNeish, D. (2018). Thanks coefficient Alpha, we'll take it from here. *Psychological Methods*, 23(3), 412–433. <https://doi.org/10.1037/met0000144>
- Mertens, D. M. (2015). *Research and evaluation in education and psychology* (4th ed.). Sage Publications Ltd.
- Mierzwa, E. (2019). Foreign language learning and teaching enjoyment: Teachers' perspectives. *Journal of Education, Culture and Society*, 10(2), 170–188. <https://doi.org/10.15503/jecs20192.170.188>
- Mihaljević Djigunović, J. (2009a). Impact of learning conditions on young FL learners' motivation. In M. Nikolov (Ed.), *Early learning of modern foreign languages processes and outcomes* (pp. 75–89). Multilingual Matters Ltd.
- Mihaljević Djigunović, J. (2009b). Individual differences in early language programmes. In M. Nikolov (Ed.), *The age factor and early language learning* (pp. 199–225). De Gruyter, Inc.
- Mihaljević Djigunović, J. (2012). Attitudes and motivation in early foreign language learning. *Centre for Educational Policy Studies Journal*, 2(3), 55–74. <https://doi.org/10.26529/cepsj.347>
- Mihaljević Djigunović, J. (2017). Developmental aspects of early English as a foreign language learning. In J. Enever & E. Lindgren (Eds.), *Early language learning: Complexity and mixed methods* (pp. 240–254). Multilingual Matters Ltd.
- Mihaljević Djigunović, J., & Lopriore, L. (2011). The learner: Do individual differences matter? In J. Enever (Ed.), *ELLiE Early language learning in Europe*. British Council. https://www.researchgate.net/publication/265383696_ELLiE_Early_language_learning_in_Europe/link/56681d4d08ae8905db8df8a6/download

- Mihaljević Djigunović, J., & Vilke, M. (2011). Eight years after: Wishful thinking vs the facts of life. In J. Moon & M. Nikolov (Eds.), *Research into teaching English to young learners* (pp. 67-86). University Press Pécs.
- Miles, M. B., Huberman, A. M., & Saldaña, J. (2020). *Qualitative data analysis: A methods sourcebook* (4th ed.). SAGE Publications Inc.
- Milton, J. (2006). Language lite? Learning French vocabulary in school. *Journal of French Language Studies*, 16(2), 187–205. <https://doi.org/10.1017/S0959269506002420>
- Milton, J. (2007). Lexical profiles, learning styles and the construct validity of lexical size tests. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 47–58). Cambridge University Press. <https://doi.org/10.1017/CBO9780511667268.005>
- Milton, J. (2009). Measuring second language vocabulary acquisition [electronic resource]. Multilingual Matters. <https://doi.org/10.21832/9781847692092>
- Milton, J. (2013). Measuring the contribution of vocabulary knowledge to proficiency in the four skills. *EuroSLA Monograph Series* 2, 0(0), 57–78. <https://eurosla.org/monographs/EM02/Milton.pdf>
- Miralpeix, I. (2020). L1 and L2 vocabulary size growth. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 189-206). Routledge.
- Mitchell, R. (1994). The ‘independent’ evaluation of bilingual primary education: A narrative account. In J. C. Alderson & A. Beretta (Eds.), *Evaluating second language education* (pp. 100–140). Cambridge University Press.
- Mitchell, R. (2000). Anniversary article. Applied linguistics and evidence-based classroom practice: The case of foreign language grammar pedagogy. *Applied Linguistics*, 21(3), 281–303. <https://doi.org/10.1093/applin/21.3.281>
- Mitchell, R., & Myles, F. (2019). Learning French in the UK setting: Policy, classroom engagement and attainable learning outcomes. *Apples – Journal of Applied Language Studies*, 13(1), 69–93. <https://doi.org/10.17011/apples/urn.201903011690>
- Moon, J. (2009). The teacher factor in early foreign language learning programmes: The case of Vietnam. In M. Nikolov (Ed.), *The age factor and early language learning* (pp. 312–326). De Gruyter, Inc.
- Morrow, V. (2008). Ethical dilemmas in research with children and young people about their social environments. *Children's Geographies*, 6(1), 49–61. <https://doi.org/10.1080/14733280701791918>
- Mosteller, F. & Boruch, R. (2002). Overview and new directions. In F. Mosteller & R. Boruch (Eds.), *Evidence matters: Randomized trials in education research* (pp. 2-14). Brookings Institution Press.

- Muenks, K., Wigfield, A., & Eccles, J. S. (2018). I can do this! The development and calibration of children's expectations for success and competence beliefs. *Developmental Review*, 48, 24–39. <https://doi.org/10.1016/j.dr.2018.04.001>
- Muijs, D., Barnes, A., Hunt, M., Powell, B., Arweck, E., & Lindsay, G. (2005). *Evaluation of the key stage 2 language learning pathfinders*. DfES Publications.
- Multon, K. D., Brown, S. D., & Lent, R. W. (1991). Relation of self-efficacy beliefs to academic outcomes: A meta-analytic investigation. *Journal of Counseling Psychology*, 38(1), 30–38. <https://doi.org/10.1037/0022-0167.38.1.30>
- Muñoz, C. (2006). The effects of age on foreign language learning: The BAF project. In C. Muñoz (Ed.), *Age and the rate of foreign language learning* (pp. 1–40). Multilingual Matters Ltd.
- Muñoz, C. (2009). Input and long-term effects of early learning in a formal setting. In M. Nikolov (Ed.), *The age factor and early language learning* (pp. 141–159). De Gruyter, Inc.
- Muñoz, C. (2014a). Exploring young learners' foreign language learning awareness. *Language Awareness*, 23(1–2), 24–40. <https://doi.org/10.1080/09658416.2013.863900>
- Muñoz, C. (2014b). Starting age and other influential factors: Insights from learner interviews. *Studies in Second Language Learning and Teaching*, 3(3), 465–484. <https://doi.org/10.14746/ssl.2014.4.3.5>
- Muñoz, C., & Lindgren, E. (2011). Out-of-school factors – the home. In J. Enever (Ed.), *ELLiE Early language learning in Europe* (pp. 103–122). British Council. https://www.researchgate.net/publication/265383696_ELLiE_Early_language_learning_in_Europe/link/56681d4d08ae8905db8df8a6/download
- Murphy, V. (2014). *Second language learning in the early school years: Trends and contexts*. Oxford University Press.
- Murphy, V. A., Macaro, E., Alba, S., & Cipolla, C. (2015). The influence of learning a second language in primary school on developing first language literacy skills. *Applied Psycholinguistics*, 36(5), 1133–1153. <https://doi.org/10.1017/S0142716414000095>
- Myles, F. (2005). The emergence of morpho-syntactic structure in French L2. In J. Dewaele (Ed.), *Focus on French as a foreign language: Multidisciplinary approaches* (pp. 86–106). Multilingual Matters.
- Myles, F. (2017). Learning foreign languages in primary schools: Is younger better? *Languages, Society and Policy*. <https://doi.org/https://doi.org/10.17863/CAM.9806>
- Myles, F., Hooper, J., & Mitchell, R. (1998). Rote or rule? Exploring the role of formulaic language in classroom foreign language learning. *Language Learning*, 48(323–263). <https://doi.org/10.1111/0023-8333.00045>

- Myles, F., & Mitchell, R. (2012). *Learning French from ages 5, 7, and 11: An investigation into starting ages, rates and routes of learning amongst early foreign language learners*. ESRC End of Award Report RES-062-23-1545. <http://www.rtpl.uk/wp-content/uploads/2017/03/EOA-Report-RES-062-23-1545.pdf>
- Nakata, T. (2020). Learning words with flash cards and word cards. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 304–319). Routledge.
- Nakata, T., & Suzuki, Y. (2019a). Effects of massing and spacing on the learning of semantically related and unrelated words. *Studies in Second Language Acquisition*, 41(2), 287–311. <https://doi.org/10.1017/S0272263118000219>
- Nakata, T., & Suzuki, Y. (2019b). Mixing grammar exercises facilitates long-term retention: Effects of blocking, interleaving, and increasing practice. *The Modern Language Journal*, 103(3), 629–647. <https://doi.org/10.1111/modl.12581>
- Nakata, T., Tada, S., Mclean, S., & Kim, Y. A. (2021). Effects of distributed retrieval practice over a semester: Cumulative tests as a way to facilitate second language vocabulary learning. *TESOL Quarterly*, 55(1), 248–270. <https://doi.org/10.1002/tesq.596>
- Nation, P. (2007). Fundamental issues in modelling and assessing vocabulary knowledge. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 1–10). Cambridge University Press.
- Nation, P. (2008). *Teaching vocabulary strategies and techniques*. Heinle, Cengage Learning.
- Nation, P. (2016). *Making and using word lists for language learning and testing*. John Benjamins Publishing Company.
- Nation, P. (2020). The different aspects of vocabulary knowledge. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 15–29). Routledge.
- Nation, P., & Sorrell, J. (2016). Corpus selection and design. In P. Nation (Ed.), *Making and using word lists for language learning* (pp. 95–105). John Benjamins Publishing Company.
- National Centre for Excellence in Language Pedagogy (NCELP). (2018). ncelp.org
- Newton, J. (2020). Approaches to learning vocabulary inside the classroom. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 255–270). Routledge.
- Nguyen, C. D. (2018). The construction of age-appropriate pedagogies for young learners of English in primary schools. *The Language Learning Journal*, 0(0), 1–14. <https://doi.org/10.1080/09571736.2018.1451912>
- Nikolov, M. (1999). ‘Why do you learn English?’ ‘Because the teacher is short.’ A study of Hungarian children’s foreign language learning motivation. *Language Teaching Research*, 3(1), 33–56. <https://doi.org/10.1177/136216889900300103>
- Nikolov, M. (2009). The age factor in context. In M. Nikolov (Ed.), *The age factor and early language learning* (pp. 1–38). De Gruyter, Inc.

- Nikolov, M., & Mihaljević Djigunović, J. (2006). Recent research on age, second language acquisition, and early foreign language learning. *Annual Review of Applied Linguistics*, 26, 234–260. <https://doi.org/10.1017/S0267190506000122>
- Nikolov, M., & Mihaljević Djigunović, J. (2011). All shades of every color: An overview of early teaching and learning of foreign languages. *Annual Review of Applied Linguistics*, 31, 95–119. <https://doi.org/10.1017/S0267190511000183>
- Norris, J. M. (2006). The why (and how) of assessing student learning outcomes in college foreign language programs. *The Modern Language Journal*, 90(iv), 576–583. https://doi.org/10.1111/j.1540-4781.2006.00466_2.x
- Norris, J. M. (2009) Understanding and improving language education through program evaluation: Introduction to the special issue. *Language Teaching Research* 13(1), 7-13. <https://doi.org/10.1177/1362168808095520>
- Norris, J. M. (2014). *Using evaluation for innovation and improvement in language and humanities education* [Keynote address]. Language Program Evaluation Symposium, Harvard University.
- Norris, J. M. (2016). Language program evaluation. *Modern Language Journal*, 100, 169-189. <https://doi.org/10.1111/modl.12307>
- Norris J. M. & Mills N. (2016). Introduction: Innovation and accountability in foreign language program evaluation. In J. M. Norris & N. Mills (Eds.) *AAUSC 2014 volume - Issues in language program direction: Innovation and accountability in language program evaluation* (pp. 1-14). Boston: Heinle Cengage.
- Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, 50(3), 417–528. <https://doi.org/10.1111/0023-8333.00136>
- Norris, J. M., & Ortega, L. (2001). Does type of instruction make a difference? Substantive findings from a meta-analytic review. *Language Learning*, 51(PART 1), 157–213. <https://doi.org/10.1111/j.1467-1770.2001.tb00017.x>
- Norris, J. M., & Watanabe, Y. (2013). Program evaluation. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Wiley-Blackwell. <https://doi.org/10.1002/9781405198431.wbeal0963>
- Northbrook, J. & Conklin, K. (2018). “What are you talking about?” An analysis of lexical bundles in Japanese junior high school textbooks. *International Journal of Corpus Linguistics*, 23(3), 311–334. <https://doi.org/10.1075/ijcl.16024.nor>
- Oakley, A. (2000). *Experiments in knowing: Gender and method in the social sciences*. The New Press.

- OECD. (2005). *Teachers Matter: Attracting, developing and retaining effective teachers*. OECD Publishing.
https://www.oecd.org/content/dam/oecd/en/publications/reports/2005/06/teachers-matter_g1gh5af3/9789264018044-en.pdf
- OECD. (2018). *Education at a glance: OECD indicators*. OECD Publishing. https://www.oecd.org/content/dam/oecd/en/publications/reports/2018/09/education-at-a-glance-2018_g1g921ab/eag-2018-en.pdf
- OFSTED. (2005). *Implementing languages entitlement in primary schools: An evaluation of progress in ten Pathfinder LEAs*. HMI 2476. [Implementing languages entitlement in primary schools](#)
- OFSTED. (2008a). *The changing landscape of languages: An evaluation of language learning 2004-2007*. [The changing landscape of languages](#)
- OFSTED. (2008b). *Primary languages in initial teacher training*. [Primary languages in initial teacher training](#)
- OFSTED. (2011). *Modern languages: Achievement and challenge 2007-10*. www.nationalarchives.gov.uk/doc/open-government-licence/
- Orosz, A. (2009). The growth of young learners' English vocabulary size. In M. Nikolov (Ed.), *Early learning of modern foreign languages processes and outcomes* (pp. 181–194). Multilingual Matters.
- Otwinowska, A., Foryś-Nogala, M., Kobosko, W., & Szewczyk, J. (2020). Learning orthographic cognates and non-cognates in the classroom: Does awareness of cross-linguistic similarity matter? *Language Learning*, 70(3), 685–731. <https://doi.org/10.1111/lang.12390>
- Pattison, E. M. (2014). 'It's important to put yourself in any lesson that you teach': Self-efficacy in action in the primary modern foreign languages classroom. *The Language Learning Journal*, 42(3), 334–345. <https://doi.org/10.1080/09571736.2013.841278>
- Peacock, M. (2009). The evaluation of foreign-language-teacher education programmes. *Language Teaching Research*, 13(3), 259–278. <https://doi.org/10.1177/1362168809104698>
- Pellicer-Sánchez, A. (2020). Learning single words vs. multiword items. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp.158-173). Routledge.
- Peng, J.-E., & MacIntyre, P. D. (2025). Directional or reciprocal? A random intercept cross-lagged panel analysis of the relationships between emotions, motivation, willingness to communicate, and L2 achievement. *Learning and Individual Differences*, 118, 102638. <https://doi.org/10.1016/j.lindif.2025.102638>

- Peters, E. (2014). The effects of repetition and time of post-test administration on EFL learners' form recall of single words and collocations. *Language Teaching Research*, 18(1), 75–94. <https://doi.org/10.1177/1362168813505384>
- Peters, E. (2020). Factors affecting the learning of single-word items. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 125–142). Routledge.
- Peters, E., Noreillie, A., Heylen, K., Bulté, B., & Desmet, P. (2019). The impact of instruction and out-of-school exposure to foreign language input on learners' vocabulary knowledge in two languages. *Language Learning*, 69(3), 747–782. <https://doi.org/10.1111/lang.12351>
- Peters, G.-J. Y. (2014). The alpha and the omega of scale reliability and validity: Why and how to abandon Cronbach's alpha and the route towards more comprehensive assessment of scale quality. *The European Health Psychologist*, 16(2), 56–69. <https://doi.org/10.31234/osf.io/h47fv>
- Phakiti, A. (2025). Commentary on the special issue on ability beliefs and learning a new language at school. *Studies in Second Language Learning and Teaching*, 15(2), 445–466. <https://doi.org/10.14746/ssllt.48252>
- Philp, J., Mackey, A., & Oliver, R. (2008). *Second language acquisition and the younger learner: Child's play?* John Benjamins.
- Pintrich, P. R., & Schrauben, B. (1992). Students' motivational beliefs and their cognitive engagement in classroom academic tasks. In D. Schunk & J. L. Meece (Eds.), *Student perceptions in the classroom* (pp. 149–183). Routledge.
- Plonsky, L. (2015). Statistical power, *p* values, descriptive statistics and effect sizes: A “back-to-basics” approach to advancing quantitative methods in L2 research. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 23–46). Routledge.
- Plonsky, L., & Oswald, F. L. (2014). How Big Is “Big”? Interpreting effect sizes in L2 research. *Language Learning*, 64(4), 878–912. <https://doi.org/10.1111/lang.12079>
- Porter, A. (2017). Verbal working memory and foreign language learning in English primary schools: Implications for teaching and learning. In J. Enever & E. Lindgren (Eds.), *Early language learning: Complexity and mixed methods* (pp. 88–107). Multilingual Matters.
- Porter, A. (2020). An early start to foreign language literacy in English primary school classrooms. *The Language Learning Journal*, 48(5), 656–671. <https://doi.org/10.1080/09571736.2019.1632918>
- Porter, A., & Graham, S. (2022). Research in primary languages: Contribution to teacher professional development. In K. McManus & M. S. Schmid (Eds.), *How special are early birds? Foreign language teaching and learning (Eurosla Studies 6)* (pp. 37–63). Language Science Press.

- Powell, B., Wray, D., Rixon, S., Medwell, J., Barnes, A., & Hunt, M. (2001). *Analysis and evaluation of the current situation relating to the teaching of modern foreign languages at key stage 2 in England*. Qualifications and Curriculum Authority.
- Puimège, E., & Peters, E. (2019). Learners' English vocabulary knowledge prior to formal instruction: The role of learner-related and word-related variables. *Language Learning*, 69(4), 943–977. <https://doi.org/10.1111/lang.12364>
- Punch, K., & Oancea, A. (2014). *Introduction to research methods in education* (2nd ed). Sage Publications.
- Punch, S. (2002). Research with children: The same or different from research with adults? *Childhood*, 9(3), 321-341. <https://doi.org/10.1177/0907568202009003005>
- Qian, D. D., & Lin, L. H. F. (2020). The relationship between vocabulary knowledge and language proficiency. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 66–80). Routledge.
- R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.r-project.org/>
- R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.r-project.org/>
- Raykov, T., & Marcoulides, G. A. (2019). Thanks coefficient Alpha, we still need you. *Educational and Psychological Measurement*, 79(1), 200–210. <https://doi.org/10.1177/0013164417725127>
- Rea-Dickins, P. (1994). Evaluation and English language teaching. *Language Teaching*, 27, 71–91. <https://doi.org/10.1017/S0261444800007679>
- Rea-Dickins, P. (1999). Some shifts, twists, turns and tactics in language programme evaluation. *Recherche et Pratiques Pédagogiques En Langues de Spécialité - Cahiers de l'APLIUT*, 18(3), 7–22. <https://doi.org/10.3406/apliu.1999.2292>
- Revelle, W. (2024). *psych: Procedures for psychological, psychometric, and personality research*. R package version 2.4.3. <https://CRAN.R-project.org/package=psych>
- Riazi, A. M., & Candlin, C. N. (2014). Mixed-methods research in language teaching and learning: Opportunities, issues and challenges. *Language Teaching*, 47(2), 135–173. <https://doi.org/10.1017/S0261444813000505>
- Richards, B., & Malvern, D. (2007). Validity and threats to the validity of vocabulary measurement. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 79–92). Cambridge University Press. <https://doi.org/10.1017/CBO9780511667268.007>

- Richards, H., Conway, C., Roskvist, A., & Harvey, S. (2013). Foreign language teachers' language proficiency and their language teaching practice. *The Language Learning Journal*, 41(2), 231–246. <https://doi.org/10.1080/09571736.2012.707676>
- Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, 17(3), 249–255. <https://doi.org/10.1111/j.1467-9280.2006.01693.x>
- Roehr-Brackin, K. (2020). Investigating explicit and implicit L2 knowledge and learning: Replications of Erlam (2005) and Roehr-Brackin and Tellier (2019). *Language Teaching*, 1–13. <https://doi.org/10.1017/S0261444820000415>
- Roehr-Brackin, K., Baranowska, K., Pavlekovic, R., & Scheffler, P. (2024). The role of individual learner differences in explicit language instruction. *The Modern Language Journal*, 108(4), 815–845. <https://doi.org/10.1111/modl.12963>
- Roehr-Brackin, K., & Tellier, A. (2019). The role of language-analytic ability in children's instructed second language learning. *Studies in Second Language Acquisition*, 41(5), 1111–1131. <https://doi.org/10.1017/S0272263119000214>
- Rogers, E.M. (2003). *Diffusion of innovations* (5th ed.). Free Press.
- Rohaani, E. J., Taconis, R., & Jochems, W. M. G. (2012). Analysing teacher knowledge for technology education in primary schools. *International Journal of Technology and Design Education*, 22(3), 271–280. <https://doi.org/10.1007/s10798-010-9147-z>
- Roothoof, H. (2017). Primary teachers' beliefs about teaching English to young learners. *Pulso Revista de educación*, (40), 2011–225. <https://doi.org/10.58265/pulso.5121>
- Ross, S. (1992). Program-defining evaluation in a decade of eclecticism. In J. C. Alderson & A. Beretta (Eds.), *Evaluating second language education* (pp. 167–196). Cambridge University Press.
- Rowe, J., Herrera, M., Hughes, B., & Cawley, M. (2012). Capacity building for primary languages through initial teacher education: Could specialist and non-specialist student teachers' complementary skills provide a winning combination? *The Language Learning Journal*, 40(2), 143–156. <https://doi.org/10.1080/09571736.2011.605904>
- Rule, S., & Marsden, E. (2006). The acquisition of functional categories in early French second language grammars: The use of finite and non-finite verbs in negative contexts. *Second Language Research*, 22(2), 188–218. <https://doi.org/10.1191/0267658306sr265oa>
- Ryan, R., & Deci, E. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, 25, 54–67. <https://doi.org/10.1006/ceps.1999.1020>

- Sagarra, N., & Alba, M. (2006). The key is in the keyword: L2 vocabulary learning methods with beginning learners of Spanish. *The Modern Language Journal*, 90(2), 228–243. <https://doi.org/10.1111/j.1540-4781.2006.00394.x>
- Saito, K. (2021). What characterizes comprehensible and native-like pronunciation among English-as-a-second-language speakers? Meta-analyses of phonological, rater, and instructional factors. *TESOL Quarterly*, 0(0), 1–35. <https://doi.org/10.1002/tesq.3027>
- Saito, K., & Plonsky, L. (2019). Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis. *Language Learning*, 69(3), 652–708. <https://doi.org/10.1111/lang.12345>
- Saito, K., Trofimovich, P., Abe, M., & In'nami, Y. (2020). Dunning-Kruger effect in second language speech learning: How does self perception align with other perception over time? *Learning and Individual Differences*, 79, 1–10. <https://doi.org/10.1016/j.lindif.2020.101849>
- Sakata, N. (2019). Profiling vocabulary for proficiency development: Effects of input and general frequencies on L2 learning. *System*, 87, 1–12. <https://doi.org/10.1016/j.system.2019.102167>
- Saldaña, J. (2021). *The coding manual for qualitative researchers* (4th ed.). SAGE. <http://prism.librarymanagementcloud.co.uk/chi-ac/items/540576>
- Salkind, N. J. (2014). *Statistics for people who (think they) hate statistics* (5th ed.). SAGE Publications, Inc.
- Šamo, R. (2009). The age factor and L2 reading strategies. In M. Nikolov (Ed.), *Early learning of modern foreign languages: Processes and outcomes* (pp. 121–131). Multilingual Matters.
- Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, 3(4), 207–217. <https://doi.org/10.1111/j.1467-9280.1992.tb00029.x>
- Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics*, 11(2), 129–158. <https://doi.org/10.1093/applin/11.2.129>
- Schmitt, N. (2008). Review article: Instructed second language vocabulary learning. *Language Teaching Research*, 12(3), 329–363. <https://doi.org/10.1177/1362168808089921>
- Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Palgrave Macmillan.
- Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning*, 64(4), 913–951. <https://doi.org/10.1111/lang.12077>

- Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*, 47(4), 484–503. <https://doi.org/10.1017/S0261444812000018>
- Schulz, J., Hamilton, C., Wonnacott, E., & Murphy, V. (2023). The impact of multi-word units in early foreign language learning and teaching contexts: A systematic review. *Review of Education*, 11(2). <https://doi.org/10.1002/rev3.3413>
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Company.
- Sharpe, K. (1995). The primacy of pedagogy in the early teaching of modern languages. *The Language Learning Journal*, 12(1), 40–42. <https://doi.org/10.1080/09571739585200411>
- Shintani, N. (2015). The effectiveness of processing instruction and production-based instruction on L2 grammar acquisition: A meta-analysis. *Applied Linguistics*, 36(3), 306–325. <https://doi.org/10.1093/applin/amu067>
- Shintani, N., Li, S., & Ellis, R. (2013). Comprehension-based versus production-based grammar instruction: A meta-analysis of comparative studies. *Language Learning*, 63(2), 296–329. <https://doi.org/10.1111/lang.12001>
- Shulman, L. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, 15(2), 4–14. <https://doi.org/10.1177/002205741319300302>
- Shulman, L. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review* 57(1), 1–22. <http://dx.doi.org/10.17763/haer.57.1.j463w79r56455411>
- Siddons, L. (2001). Practical reflections on the sound/spelling link. *Francophonie*, 23(Spring), 10–14. [Practical reflections on the sound/spelling link](#)
- Skehan, P. (1989). *Individual differences in second language learning*. Routledge. <https://doi.org/10.1017/S0272263100009979>
- Smithson, J. (2000). Using and analysing focus groups: Limitations and possibilities. *International Journal of Social Research Methodology*, 3(2), 103–119. <https://doi.org/10.1080/136455700405172>
- Spada, N. & Tomita, Y. (2010). Interactions between type of instruction and type of language feature: A meta-analysis. *Language Learning*, 60(2), 263–308. <https://doi.org/10.1111/j.1467-9922.2010.00562.x>
- Sparks, R. L., & Ganschow, L. (1991). Foreign language learning differences: Affective or native language aptitude differences? *The Modern Language Journal*, 75(1), 3–16. <https://www.jstor.org/stable/329830>
- Sparks, R. L., & Ganschow, L. (1993). Searching for the cognitive locus of foreign language learning difficulties: Linking first and second language learning. *The Modern Language Journal*, 77(3), 289–302. <https://www.jstor.org/stable/329098>

- Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *The Language Learning Journal*, 36(2), 139–152. <https://doi.org/10.1080/09571730802389975>
- Sturm, J. L. (2013). Explicit phonetics instruction in L2 French: A global analysis of improvement. *System*, 41(3), 654–662. <https://doi.org/10.1016/j.system.2013.07.015>
- Südkamp, A., Kaiser, J., & Möller, J. (2012) Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology*, 104(3), 743–762. <https://doi.org/10.1037/a0027627>
- Suzuki, Y., & DeKeyser, R. (2017). The interface of explicit and implicit knowledge in a second language: Insights from individual differences in cognitive aptitudes. *Language Learning*, 67(4), 747–790. <https://doi.org/10.1111/lang.12241>
- Suzuki, Y., Nakata, T., & DeKeyser, R. (2019a). Optimizing second language practice in the classroom: Perspectives from cognitive psychology. *Modern Language Journal*, 103(3), 551-561. <https://doi.org/10.1111/modl.12582>
- Suzuki, Y., Nakata, T., & DeKeyser, R. (2019b). The desirable difficulty framework as a theoretical foundation for optimizing and researching second language practice. *The Modern Language Journal*, 103(3), 713–720. <https://doi.org/10.1111/modl.12585>
- Szpotowicz, M. (2000). Young learners: How do they remember vocabulary? In J. Moon & M. Nikolov (Eds.), *Research into teaching English to young learners* (pp. 360–374). University Press Pécs.
- Szpotowicz, M. (2009). Factors influencing young learners' vocabulary acquisition. In Nikolov, M. (Ed.), *Early learning of modern foreign languages processes and outcomes* (pp.194–211). Multilingual Matters.
- Szpotowicz, M., & Lindgren, E. (2011). Language achievements: A longitudinal perspective. In J. Enever (Ed.), *ELLiE Early language learning in Europe* (pp. 125–142). British Council. https://www.researchgate.net/publication/265383696_ELLiE_Early_language_learning_in_Europe/link/56681d4d08ae8905db8df8a6/download
- Taylor, B., Pearson, P. D., Peterson, D.S., Rodriguez, M. C. (2005). The CIERA school change framework: An evidence-based approach to professional development and school reading improvement. *Reading Research Quarterly*, 40(1), 40–69. <https://doi:10.1598/RRQ.40.1.3>
- Taylor, F., & Marsden, E. (2014). Perceptions, attitudes, and choosing to study foreign languages in England: An experimental intervention. *Modern Language Journal*, 98(4), 902–920. <https://doi.org/10.1111/modl.12146> 0026-7902/14/902–920
- Teaching Schools Council. (2016). *Modern foreign languages pedagogy review: A review of modern foreign languages teaching practice in key stage 3 and key stage 4*. <https://tscouncil.org.uk/wp-content/uploads/2016/12/MFL-Pedagogy-Review-Report-2.pdf>

- Teddlie, C., & Tashakkori, A. (2009). *Foundations of mixed methods research*. SAGE Publications Inc.
- Tellier, A., & Roehr-Brackin, K. (2013). The development of language learning aptitude and metalinguistic awareness in primary-school children: A classroom study. *Essex Research Reports in Linguistics*, 62(1), 1-27. <https://doi.org/10.13140/RG.2.2.21365.58088>
- Teo, T., Lee, C. B., & Chai, C. S. (2008). Understanding pre-service teachers' computer attitudes: Applying and extending the technology acceptance model. *Journal of Computer Assisted Learning*, 24, 128–143. <https://doi.org/10.1111/j.1365-2729.2007.00247.x>
- The Nuffield Foundation. (2000). *Languages: The next generation. The final report and recommendations of the Nuffield languages inquiry*. <http://www.nuffield.org>
- Tierney, D., & Gallastegi, L. (2005). Where are we going with primary foreign languages? *The Language Learning Journal*, 31(1), 47–54. <https://doi.org/10.1080/09571730585200081>
- Tierney, D., & Gallastegi, L. (2011). The attitudes of the pupils towards modern languages in the primary school (MLPS) in Scotland. *Education 3-13*, 39(5), 483–498. <https://doi.org/10.1080/03004271003781538>
- Tinkham, T. (1997). The effects of semantic and thematic clustering on the learning of second language vocabulary. *Second Language Research*, 13(2), 138–163. <https://doi.org/10.1191/026765897672376469>
- Tinsley, T. (2019). *Language trends 2019: Language teaching in primary and secondary schools in England - survey report*. British Council. <https://www.britishcouncil.org/sites/default/files/language-trends-2019.pdf>
- Tinsley, T., & Board, K. (2012). *Language learning in primary and secondary schools in England. Findings from the 2012 Language trends survey - Research report*. CfBT Education trust. <https://edtlive.b-cdn.net/livenew/media/d0vereh5/language-learning-in-primary-and-secondary-schools-in-england.pdf>
- Tinsley, T., & Board, K. (2016). *Language trends 2015-16: The state of language learning in primary and secondary schools in England*. British Council. https://www.britishcouncil.org/sites/default/files/language_trends_survey_2016_0.pdf
- Tinsley, T., & Board, K. (2017). *Language trends 2016/17: Language teaching in primary and secondary schools in England - Survey Report*. British Council. https://www.britishcouncil.org/sites/default/files/language_trends_survey_2017_0.pdf
- Tinsley, T., & Comfort, T. (2012). *Lessons from abroad: International review of primary languages*. CfBT. <https://edtlive.b-cdn.net/livenew/media/qxzlk1hg/lessons-from-abroad-international-review-of-primary-languages.pdf>

- Tinsley, T., & Doležal, N. (2018). *Language trends 2018: Language teaching in primary and secondary schools in England - survey report*. British Council.
https://www.britishcouncil.org/sites/default/files/language_trends_2018_report.pdf
- Tonzar, C., Lotto, L., & Job, R. (2009). L2 vocabulary acquisition in children: Effects of learning method and cognate status. *Language Learning*, 59(3), 623–646.
<https://doi.org/10.1111/j.1467-9922.2009.00519.x>
- Torgerson, D. J., & Torgerson, C. J. (2008). *Designing randomised trials in health, education and the social sciences*. Palgrave Macmillan.
- Tracy, S. J. (2010). Qualitative quality: Eight “big-tent” criteria for excellent qualitative research. *Qualitative Inquiry*, 16(10), 837–851.
<https://doi.org/10.1177/1077800410383121>
- Tracy-Ventura, N., McManus, K., Norris, J., & Ortega, L. (2014). “Repeat as much as you can”: Elicited imitation as a measure of oral proficiency in L2 French. In P. Leclercq, A. Edmonds, & H. Hilton (Eds.), *Measuring L2 proficiency: Perspectives from SLA* (pp. 143–166). Multilingual Matters.
- Tragant, E. (2006). Language learning motivation and age. In C. Muñoz (Ed.), *Age and the rate of foreign language learning* (pp. 237–276). Multilingual Matters Ltd.
- Tragant Mestres, E., & Lundberg, G. (2011). The teacher’s role: What is its significance in early language learning? In J. Enever (Ed.), *ELLiE Early language learning in Europe* (pp. 81–100). British Council.
https://www.researchgate.net/publication/265383696_ELLiE_Early_language_learning_in_Europe/link/56681d4d08ae8905db8df8a6/download
- Trizano-Hermosilla, I., & Alvarado, J. M. (2016). Best alternatives to Cronbach’s alpha reliability in realistic conditions: Congeneric and asymmetrical measurements. *Frontiers in Psychology*, 7, 1–8. <https://doi.org/10.3389/fpsyg.2016.00769>
- Tschichold, C. (2012). French vocabulary in Encore Tricolore: Do pupils have a chance? *The Language Learning Journal*, 40(1), 7–19. <https://doi.org/10.1080/09571736.2012.658219>
- Tsui, A. B. M. (2003). *Understanding expertise in teaching: Case studies of ESL teachers*. Cambridge University Press.
- Tymms, P. (2017). Questionnaires. In R. Coe, M. J. Waring, L. V Hedges, & J. Arthur (Eds.), *Research methods and methodologies in education* (2nd ed.) (pp. 223–233). Sage Publications Ltd.
- Uchida, Y., & Sugimoto, J. (2019). Non-native English teachers’ confidence in their own pronunciation and attitudes towards teaching: A questionnaire survey in Japan. *International Journal of Applied Linguistics*, 30(1), 19–34.
<https://doi.org/10.1111/ijal.12253>

- Uchihara, T., Webb, S., Yanagisawa, A. (2019) The effects of repetition on incidental vocabulary learning: A meta-analysis of correlational studies. *Language Learning*, 69(3), 559-599. <https://doi.org/10.1111/lang.12343>
- United Nations Children's Fund - UNICEF. (2002). Evaluation technical notes. Children participating in research, monitoring and evaluation (M&E) — Ethics and your responsibility as a manager. <https://www.alnap.org/system/files/content/resource/files/main/technote1-ethics.pdf>
- University of St Andrews. (n.d.). *Analysing Likert scale-type data*. Retrieved October 16, 2025, from <https://www.st-andrews.ac.uk/media/ceed/students/mathssupport/Likert.pdf>
- Unsworth, S., Persson, L., Prins, T., & De Bot, K. (2015). An investigation of factors affecting early foreign language learning in the Netherlands. *Applied Linguistics*, 36(5), 527–548. <https://doi.org/10.1093/applin/amt052>
- Urhahne, D. & Wijnia, L. (2021) A review on the accuracy of teacher judgments. *Educational Research Review*, 32, 1-26. <https://doi.org/10.1016/j.edurev.2020.100374>
- Valdera Gil, F., & Crichton, H. (2018). Mother tongue plus two languages: Are Scottish primary teachers confident to deliver? *The Language Learning Journal*, 0(0), 1–15. <https://doi.org/10.1080/09571736.2018.1448430>
- van den Broek, G. S. E., Takashima, A., Segers, P. C. J., & Verhoeven, L. T. W. (2018). Contextual richness and word learning: Context enhances comprehension but retrieval enhances retention. *Language Learning*, 68(2), 546–585. <https://doi.org/10.1111/lang.12285>
- VanPatten, B. (2002). Processing instruction: An update. *Language Learning*, 52(4), 755–803. <https://doi.org/10.1111/1467-9922.00203>
- VanPatten, B. (2017). Processing instruction. In S. Loewen & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition* (pp. 166–180). Routledge.
- Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied Psycholinguistics*, 22(2), 217–234. <https://doi.org/10.1017/S0142716401002041>
- Vilkaitė-Lozdienė, L., & Schmitt, N. (2020). Frequency as a guide for vocabulary usefulness: High-, mid- and low-frequency words. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 81–96). Routledge.
- Wade, P., Marshall, H., & O'Donnell, S. (2009). *Primary modern foreign languages longitudinal survey of implementation of national entitlement to language learning at key stage 2: Final report*. DCSF.
- Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, 28(1), 46–65. <https://doi.org/10.1093/applin/aml048>

- Webb, S. (2020). Incidental vocabulary learning. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 225–239). Routledge.
- Webb, S., & Nation, P. (2017). *How vocabulary is learned*. Oxford University Press.
- Webb, S., Yanagisawa, A., & Uchihara, T. (2020). How effective are intentional vocabulary-learning activities? A meta-analysis. *The Modern Language Journal*, 104(4), 715–738. <https://doi.org/10.1111/modl.12671>
- Webb, T., & Sheeran, P. (2006). Does changing behavioral intentions engender behavior change? A meta-analysis of the experimental evidence. *Psychological Bulletin*, 132(2), 249–268. <https://doi.org/10.1037/0033-2909.132.2.249>
- Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary Educational Psychology*, 25(1), 68–81. <https://doi.org/10.1006/ceps.1999.1015>
- Wigfield, A., & Eccles, J. S. (2002). *Development of achievement motivation*. Academic Press.
- Williams, M., & Burden, R. (1997). Motivation in language learning: A social constructivist approach. *Cahiers de l'APLIUT*, 16(3), 19–27. <https://doi.org/10.3406/APLIU.1997.1201>
- Wingate, U. (2018). Lots of games and little challenge – a snapshot of modern foreign language teaching in English secondary schools. *The Language Learning Journal*, 46(4), 442–455. <https://doi.org/10.1080/09571736.2016.1161061>
- Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications. *ArXiv.Org.*, 08. <https://doi.org/10.48550/arXiv.1308.5499>
- Woodgate-Jones, A. (2008). Training confident primary modern foreign language teachers in England: An investigation into preservice teachers' perceptions of their subject knowledge. *Teaching and Teacher Education*, 24, 1–13. <https://doi.org/10.1016/j.tate.2006.10.008>
- Woolhouse, C., Bartle, P., Hunt, E., & Balmer, D. (2013). Language learning, cultural capital and teacher identity: Teachers negotiating the introduction of French into the primary curriculum. *The Language Learning Journal*, 41(1), 55–67. <https://doi.org/10.1080/09571736.2011.635798>
- Woore, R. (2007). 'Weisse Maus in meinem Haus': Using poems and learner strategies to help learners decode the sounds of the L2. *The Language Learning Journal*, 35(2), 175–188. <https://doi.org/10.1080/09571730701599211>
- Woore, R. (2009). Beginners' progress in decoding L2 French: Some longitudinal evidence from English modern foreign languages classrooms. *The Language Learning Journal*, 37(1), 3–18. <https://doi.org/10.1080/09571730902717398>
- Woore, R. (2010). Thinking aloud about L2 decoding: An exploration of the strategies used by beginner learners when pronouncing unfamiliar French words. *The Language Learning Journal*, 38(1), 3–17. <https://doi.org/10.1080/09571730903545210>

- Woore, R. (2014). Beginner learners' progress in decoding L2 French: Transfer effects in typologically similar L1-L2 writing systems. *Writing Systems Research*, 6(2), 167–189. <https://doi.org/10.1080/17586801.2013.838536>
- Woore, R. (2018). Learners' pronunciations of familiar and unfamiliar French words: What can they tell us about phonological decoding in an L2? *The Language Learning Journal*, 46(4), 456–469. <https://doi.org/10.1080/09571736.2016.1161062>
- Woore, R. (2022). What can second language acquisition research tell us about the phonics 'pillar'? *The Language Learning Journal*, 50(2), 172–185. <https://doi.org/10.1080/09571736.2022.2045683>
- Woore, R., Graham, S., Porter, A., Courtney, L., & Savory, C. (2018). *Foreign language education: Unlocking reading (FLEUR). A study into the teaching of reading to beginner learners of French in secondary school*. <https://ora.ox.ac.uk/objects/uuid:4b0cb239-72f0-49e4-8f32-3672625884f0>
- Wu, X. (2003). Intrinsic motivation and young language learners: The impact of the classroom environment. *System (Linköping)*, 31(4), 501–517. <https://doi.org/10.1016/j.system.2003.04.001>
- Xu, W., & Zammit, K. (2020). Applying thematic analysis to education: A hybrid approach to interpreting data in practitioner research. *International Journal of Qualitative Methods*, 19. <https://doi.org/10.1177/1609406920918810>
- Yang, W. W. (2009). Evaluation of teacher induction practices in a US university English language program: Towards useful evaluation. *Language Teaching Research*, 13(1), 77–98. <https://doi.org/10.1177/1362168808095524>
- Zhang, S. & Zhang, X. (2022). The relationship between vocabulary knowledge and L2 reading/listening comprehension: A meta-analysis. *Language Teaching Research*, 26(4), 696–725. <https://doi.org/10.1177/136216882091399>
- Zhao, T., & Murphy, V. A. (2017). Factors affecting the speed of word retrieval in children learning English as a foreign language. In M. del P. García Mayo (Ed.), *Learning foreign languages in primary school: Research insights* (pp. 29–43). Multilingual Matters.
- Zoblotsky, T. & Kaldon, C. (2018). *Roadblocks to implementing randomized controlled trials in educational research*. Sage Research Methods Cases. <https://doi.org/10.4135/9781526442246>
- Zohrabi, M. (2011). An introduction to course and/or program evaluation. *Journal of Pan-Pacific Association of Applied Linguistics*, 15(2), 59–70. [Introduction to course and/or program evaluation](#)

Appendix A.1

Detailed Discussion of Evaluation as a Research Method

This section provides a brief rationale for evaluation, mindful that designs should be chosen for suitability rather than any assumed hierarchy in effectiveness of methods (Gorard, 2017).

Evaluation as an approach to research: introduction

This section explores evaluation as a means of researching innovative pedagogy, acknowledging the difficulty of identifying and isolating factors influencing classroom learning through consideration of the potential contribution of randomised controlled trials (RCTs).

Historical context

Views have been mixed on evaluation in language education, with its explanatory, evidential potential recognised (Kiely, 2009; Rea-Dickins, 1999) alongside criticism of its association with “simplistic measures of learning outcomes and experimental comparisons” (Harris, 2009:56). However, synthesising language instruction research, Norris and Ortega (2000) documented a trend away from promotion of the “supremacy of quantitative data” (Rea-Dickins, 1994:77) towards increased interest in evaluation of instruction types impacting on classroom learning, with concomitant awareness of influential learner characteristics (e.g., age, language aptitude, L1 background).

Definition

Evaluation aims to determine effectiveness of interventions, programmes or policy (Norris, 2016; Norris & Watanabe, 2013), using varied sources and techniques to explore components (Zohrabi, 2011). Meeting certain standards, being “principled” and “systematic” (J. Alderson, 1992:274; Rea-Dickins, 1994:72, 1999:18), is essential (Beretta, 1992), with depth of analysis and scrutiny establishing credibility and accuracy (Bennett, 2003). Systematicity involves a “holistic” approach (Kiely, 2009:99), detailing aspects such as: theoretical background; delivery and development; teaching approaches; effectiveness; advantages and disadvantages (Norris & Watanabe, 2013).

Is evaluation research?

Whether evaluation constitutes research is contentious. Both can consider several factors while retaining a central focus, and develop empirical knowledge of how/why treatments succeed or fail (Norris & Mills, 2016). Research explores theoretical concepts, contributing to knowledge

(Norris, 2016), whereas evaluation's distinct feature is its ultimate focus on practice, normally resulting in real-world application of newfound knowledge (J. Alderson, 1992; Kiely & Rea-Dickins, 2005; Norris & Watanabe, 2013; Rea-Dickins, 1994), informing theory *and* practice (Marsden, 2007). This utility, (acknowledged by Beretta, 1992), combines empirical approaches with potential to engender improvement (Kiely, 2006; Norris, 2009, 2014, 2016; Norris & Watanabe, 2013; Peacock, 2009; Rea-Dickins, 1994, 1999; Yang, 2009).

Randomised controlled trials (RCTs)

RCTs are controlled experiments “outside a laboratory setting in the usual environment where social, economic or educational interactions occur” (Burtless, 2002:180). RCTs evidence the effectiveness of particular educational programmes or teaching approaches (Boruch et al., 2002; Haynes et al., 2012; Mosteller & Boruch, 2002; Torgerson & Torgerson, 2008), comparing outcomes of participants exposed to an intervention with those receiving different treatment or continuing as normal (Connolly et al., 2017; Haynes et al., 2012). Without comparison groups, improved outcomes cannot be solely attributed to interventions; in single group pre-/post-test designs, it is impossible to control for threats posed by history, maturation, test effects or regression (Cook & Campbell, 1979; Marsden & Torgerson, 2012).

Randomised controlled trials and internal validity

Randomised allocation improves internal validity (Cook & Campbell, 1979:56) - “the extent to which the effect measured is caused by the intervention” (Kiely & Rea-Dickins, 2005:24) - ensuring that results are trustworthy and interpretations justified by excluding other confounding variables (Marsden & Torgerson, 2012; Torgerson & Torgerson, 2008), and showing that the study measures what it claims (Shadish et al., 2002).

(i) Selection bias

To attribute differential learning outcome to the intervention, groups should be matched as closely as possible, usually achieved through random assignment (K. Punch & Oancea, 2014). RCTs try to ensure that potentially influential variables appear in both experimental and comparison groups. This minimises the risk of selection bias in non-randomised groups, which “creates an association that isn't true” (Torgerson & Torgerson, 2008:24). In educational contexts internal validity is threatened by possible imbalance in such variables as pupils' socio-economic status, gender or prior attainment (Haynes et al., 2012).

(ii) Temporal changes

Pupils may improve as they increase in maturity, regardless of interventions (Marsden & Torgerson, 2012; Torgerson & Torgerson, 2008). Randomised groups with similar characteristics should experience similar maturational changes, enabling outcome differential to be attributed to the intervention rather than usual teaching input and increased cognitive maturity (Marsden & Torgerson, 2012). This is particularly salient in longer interventions with greater time intervals between pre- and post-tests (Cook & Campbell, 1979; Marsden & Torgerson, 2012; Shadish et al., 2002). Groups should also experience history similarly.

(iii) Regression to the mean (RTM)

This statistical phenomenon can affect pre-/post-test designs without control groups (where differences in learners' pre-/post-intervention tests are attributed to innovative treatments) (Marsden & Torgerson, 2012; Torgerson & Torgerson, 2008). A group's test results usually cluster around the average outcome, but some register at extremes (high/low). If re-tested, extreme outcomes likely regress downwards or upwards towards the mean. When improvement is detected on re-testing, interventions designed to address lowest scores could be deemed successful, although these could have regressed upwards without intervention (Marsden & Torgerson, 2012; Torgerson & Torgerson, 2008). RTM occurs because of error in original outcomes, perhaps due to inaccurate marking/measurement processes or poor student performance (Torgerson & Torgerson, 2008).

Summary of evaluation as a research approach

Evaluation delves deep into educational programmes to understand effectiveness and how it is achieved. Given that greater rigour in language education research has been recommended for over two decades (Norris & Ortega, 2000), quantitative approaches (referring to amounts, quantity, frequency, degrees) are clearly necessary as they can provide data about effectiveness of interventions for learning outcomes (knowledge). Increased use of combined methods can improve validity and reliability, including qualitative data (involving speech, words, or images rather than numbers, K. Punch & Oancea, 2014). Data on attitudes, perceptions, or narratives from varied viewpoints enable understanding of individuals' or groups' experiences (Greig et al., 2013; King & Horrocks, 2010), particularly because individuals interpret the same events differently (Greig et al., 2013). RCTs, with quantitative *and* qualitative datasets as determined by RQs, appeared most useful for the purposes of the current study.

Appendix A.2

Individual Vocabulary Items: Discussion

Word class, concreteness and cognate status

Nouns constituted the largest proportion of lexical items across all vocabulary tests, and were among the highest-scoring items in each (Figures [10](#), [11](#), [12](#) and [13](#)). This appears to support suggestions made by Milton (2007) and Tschichold (2012) that nouns are easier to learn than verbs (which were never the highest-scoring). Scores on nouns were generally higher than: pronouns, with “*je*” (‘I’) and “*tu*” (‘you’) among the lowest-scoring items in recognition tests; conjunctions “*mais*” (‘but’) and “*aussi*” (‘also’); the adverb “*bien*” (‘good’/‘well’); the preposition “*dans*” (‘in’).

The highest-scoring items were concrete nouns ([Appendix E.8](#)), with some of these also cognates ([Appendix E.9](#)), which may have supported learnability (see Boers, 2020; Milton, 2007; E. Peters, 2020). However, concreteness did not consistently determine learnability, with other nouns such as “*enfant*” (‘child’) and “*fête*” (‘party’/‘celebration’ – in Standard only) - arguably similarly imageable - not as successfully conveyed.

Word length/syllable count

The length of “*non*” possibly contributed to its learnability, but other short, written words such as “*tu*” (‘you’), “*aussi*” (‘also’), “*dans*” (‘in’) and “*mais*” (‘but’) were less well recognised. This contradicted findings from Hashimoto and Egbert (2019), but was in line with Szpotowicz’s (2000) finding that number of syllables did not affect learnability. The longest word, “*aujourd’hui*” (‘today’) ([Appendix E.5](#)), proved challenging in written recognition, but other factors were probably involved (see below).

Distance between French and English in SSCs

“*Aujourd’hui*” had the highest number of phonemic and symbol-sound cross-linguistic differences ([Appendix E.10](#)). This perhaps explains its relatively low written recognition score in both approaches, supporting suggestions from S. Webb and Nation (2017) and E. Peters (2020), and aligning with findings from Szpotowicz (2000), that words which are difficult to pronounce (with more cross-linguistic differences relative to the L1) are harder to learn. However, in tests such as oral recognition, words with fewer such differences (“*regarder*” - ‘to look’, “*je*” - ‘I’, “*mais*” - ‘but’) were also among the lowest-scoring.

Frequency ranking and frequency/recency of encounter

Suggestions from Milton (2007) and findings from E. Peters (2014) that higher corpus-based frequency rankings may increase ease of learnability were not wholly supported by this study. In recognition tests, the highest-scoring items had both high (*“règle”* = 488; *“non”* = 75) and low (*“gâteau”* = 4845; *“stylo”* > 5000) corpus-informed frequency rankings, a finding replicated in written recall with the highest-scoring items *“chat”* (3138) and *“banane”* (> 5000) ([Appendix E.4](#)).

Frequency or recency of encounter in the classroom possibly had greater impact on learnability than frequency ranking, which aligns with E. Peters’ (2014) findings that additional encounters support form-meaning relationships; the importance of frequency of encounters with words has been repeatedly acknowledged (Barcroft, 2020; Nation, 2020; Schmitt, 2008; S. Webb & Nation, 2017). Words encountered more frequently and more recently ([Appendix E.6](#)) in one approach generally led to higher scores for those pupils (e.g., *“mais”* - ‘but’ - in Novel; *“dans”* - ‘in’ - and *“et”* - ‘and’ - in Standard).

However, frequency of encounter alone appears not to guarantee greater ease of learnability. For example, because of its high corpus-based frequency ranking [161] ([Appendix E.4](#)), the word *“ici”* (‘here’) was presented and revisited in Novel seven times, last in lesson 27, compared with two in Standard ([Appendix E.6](#)), yet was the lowest-scoring item in speaking.

Appendix A.3

Individual Phonics Items: Discussion

Discussion of individual SSCs

Pupils coped reasonably well with [i] and [u] in both tests, particularly in Standard. The similarity of the French [i] to its English equivalent was probably facilitative (see Hamada & Koda, 2008; Woore, 2014). However, the influence of English SSCs was not always positive, supporting Woore (2010, 2014). For example, the most common error in reading aloud “*char*” was the adoption of the English [ch]; similarly, pupils’ attempts at “*viser*” were hampered by their recognition of [er] as the English grapheme (e.g., as in the word ‘her’); and the ‘n’ in [on] in “*fond*” was often over-emphasised, as it would be in English, rather than reflecting the French nasal sound. The sight of the written word appeared to activate knowledge of pupils’ stored L1 SSCs which overrode the stored L2 pronunciation. These examples indicate that learners struggled when graphemes did not represent the phonemes they expected, given their knowledge of the correspondence in English (as found by Erler & Macaro, 2011). E. Peters (2020) suggested that learners may struggle to read words aloud when unfamiliar with phonemic structures different from those found in their L1; this was also found by Marijanović et al. (2009). This would appear to be the case here when learners attempted to transcribe unknown French words.

The SSCs [eu] and [au] were universally difficult to produce orally and to transcribe. Not only did learners apply their understanding of how to decode English words (as discussed above), but they also appeared to be trying to think of English words which could fit the sounds they had heard. For example, in the target word “*clou*”, a number of participants transcribed the target SSC [ou] as [oo] - a logical application of their L1 knowledge. However, the most common error was the use of [ue], which then produced the English word ‘clue’. Although occurring in transcription rather than RA, this is nevertheless perhaps supportive of Woore’s (2014) assertion (made in reference to pronunciation) that, although the similarities between French and English SSCs can be helpful in some respects, they can also increase the challenge when learners try to decode the written French form. There were other examples in both transcription and RA tests of pupils perhaps being distracted by the whole word, and producing English words, rather than trying to adopt French pronunciation: “*tiré*” (target SSC [é]) was most often decoded as ‘tyre’, and “*flou*” (target SSC [ou]) as ‘flow’ or ‘flew’. This could not be because the incorrect pronunciation had been memorised, as all words were unfamiliar.

Although pupils’ efforts were only scored for the target SSC and the rest of the utterance ignored, pupils were possibly sometimes negatively affected by not knowing other parts of the

word. This seemed a particular problem with “*ceux*”, where pupils’ struggle with the initial and final letter appeared to impact on their attempt at the target [eu] grapheme.

Appendix A.4

Elicited Imitation Grammar (Rescored Model)

Elicited Imitation Grammar (Rescored Model)

Test quality

Table 59.

Elicited Imitation Grammar (Rescored Model) Test Quality

		Elicited Post-test	imitation Delayed post-test
% cases complete prior to imputation		97	97
Reliability (EFA)	alpha	.78	.85
	omega	.82	.87
Facility		.38 <i>SD</i> .17	.41 <i>SD</i> .18
Discrimination		.40 <i>SD</i> .20	.47 <i>SD</i> .16
Normality of distribution link		Appendix M.25	

Descriptive statistics for the rescored Elicited Imitation Grammar task

Table 60.

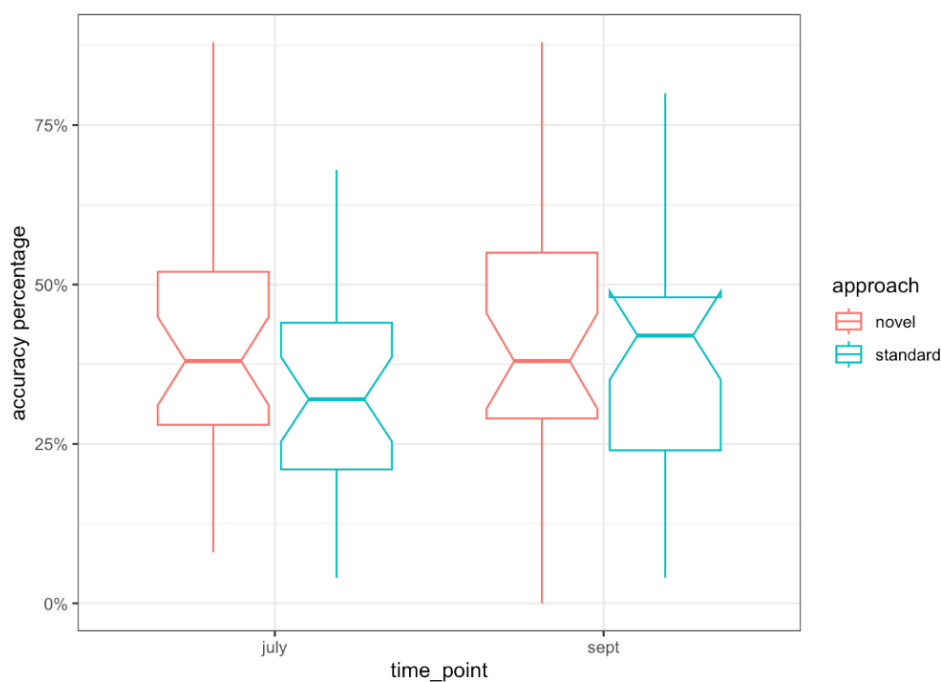
Descriptive Statistics for Rescored Elicited Imitation Grammar Task

Approach	Time- point (<i>k</i> = 12)	Descriptive statistics							
		<i>n</i>	<i>Mean</i> %	<i>SD</i>	<i>CI</i> %	Range /60		Effect size <i>d</i>	Effect size <i>CI</i>
Novel	Post-test	30	42	.20	.34 - .39	4-44	Within-group: Novel Post-test – Delayed post- test	.06	-.25 - .27
Standard	Post-test	30	34	.16	.28 - .40	2-34	Within-group: Standard Post-test – Delayed post- test	-.43	-.85 - -.08
Novel	Delayed post-test	30	43	.23	.34 - .51	0-44	Between- group: Novel – Standard Post-test	.44	-.06 - .98

Standard	Delayed post-test	30	40	.20	.32 - .47	2-40	Between- group: Novel – Standard Delayed post- test	.14	-.39 - .68

Figure 83.

Notched Box Plot Showing Elicited Imitation Grammar Scores by Time-point



Analyses using mixed-effects models

Model code

```
binomial_model_grammar_imitation_rescored_with_school_random = glmer(item_score ~ approach*time_point*scale(total_LAA_score) + (1|school/pupil_id) + (1|item_code), family = binomial, data = grammar_imitation_regression_rescored)
```

A generalised linear mixed effects binomial model was constructed for the EIG.

(i) Assumptions for the regression model

As the coefficient for the total_LAA_score_log was not significant in the Box-Tidwell test, it suggested that the linearity assumption held for the total_LAA_score.

Table 61.*Assumptions for the Regression Model*

<i>Predictors</i>	item score		
	<i>Odds Ratios</i>	<i>CI</i>	<i>p</i>
(Intercept)	.58	.36 – .92	.02
total LAA score	.89	.02 – 39.20	.95
total LAA score log	1.78	.04 – 77.33	.77
Random Effects			
σ^2	3.29		
τ_{00} pupil_id:school	.63		
τ_{00} item_code	.80		
τ_{00} school	.06		
ICC	.31		
N _{pupil_id}	60		
N _{school}	5		
N _{item_code}	25		
Observations	3000		
Marginal R^2 / Conditional R^2	.04 / .34		

Additionally, overdispersion in the model was not detected in the overdispersion test ($p = .98$).

(ii) Predictions based on mixed-effects models

The model summary ([Figure 84](#)) suggested an interaction between time-points (post-test and delayed post-test) and between baseline LAA score and rescored EIG scores, but not between approaches. Fixed effects (marginal R^2) accounted for 5% of variance in the EIG model; a combination of fixed and random effects (conditional R^2) accounted for 35% of total variance.

Figure 84.*Summary of Model (Elicited Imitation Grammar Rescored Model)*

<i>Predictors</i>	item score				
	<i>Odds Ratios</i>	<i>std. Error</i>	<i>CI</i>	<i>Statistic</i>	<i>p</i>
(Intercept)	.58	.14	.36 – .92	-2.31	.02
approach standard-novel	.74	.16	.48 – 1.13	-1.40	.16

time-point sept-july	1.20	.10	1.01 – 1.42	2.10	.04
total LAA score	1.58	.18	1.27 – 1.98	4.05	< .001
approach standard-novel × time-point sept-july	1.32	.23	.94 – 1.85	1.61	.11
approach standard-novel × total LAA score	.88	.19	.57 – 1.36	-.56	.58
time-point sept-july × total LAA score	1.07	.09	.90 – 1.26	.74	.46
(approach standard-novel × time-point sept-july) × total LAA score	.95	.16	.68 – 1.33	-.29	.77
Random Effects					
σ^2	3.29				
τ_{00} pupil_id:school	.61				
τ_{00} item_code	.81				
τ_{00} school	.06				
ICC	.31				
$N_{\text{pupil\_id}}$	60				
N_{school}	5				
$N_{\text{item\_code}}$	25				
Observations	3000				
Marginal R^2 / Conditional R^2	.05 / .35				

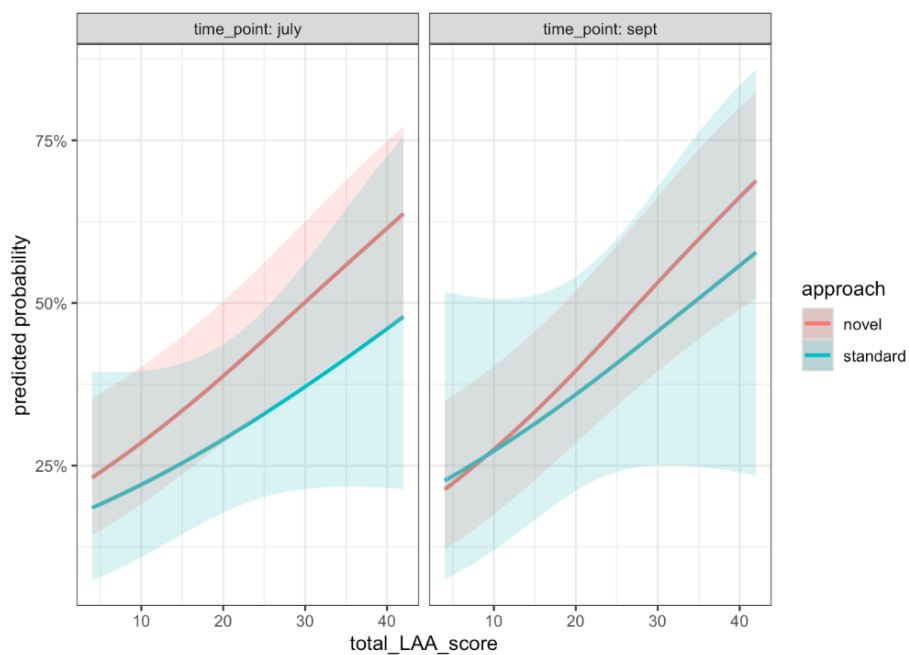
As with the original model (before re-scoring), pair-wise comparisons using emmeans (Lenth, 2024) indicated a statistically significant difference between the scores of pupils in Standard at delayed post-test compared with post-test ($SE = .09$, $z = -2.61$, $p = .009$), with the probability of a pupil in Standard scoring two points per item at delayed post-test increasing to 37% from 28% at post-test. In this rescored model, the probability of a pupil in Novel scoring one point per item was 40% at post-test and 41% at delayed post-test. The difference between Novel and Standard at post-test did not quite reach significance ($SE = .37$, $z = 1.88$, $p = .06$), with a 40% chance of scoring one point in Novel compared with a 30% probability in Standard; this had decreased further by the time of delayed post-test ($SE = .28$, $z = .71$, $p = .48$).

Ggpredict (Lüdtke, 2018) was used to explore further the significant interaction between baseline LAA score and EIG rescored item score (Figure 85). At both time-points pupils with lower baseline LAA scores appeared less likely to score one point (regardless of approach, although the likelihood was greater in Novel than Standard for all pupils except those pupils with

the lowest LAA scores - below ca. 9 – at delayed post-test, but this was not statistically significant).

Figure 85.

Elicited Imitation Grammar (Rescored Model) Predictions by Language-analytic Ability Score, Approach and Time-point



Return to [5.13.2](#).

Appendix B.1

Pre-test: Language-analytic Ability Test (from Kasproicz et al., 2019)

Name: _____

Section 1

Q1.

a) Replace the underlined noun or noun phrase with a **pronoun**. Write this in the box.

b) THEN put an (x) to show whether the pronoun is **singular** or **plural**.

Example:

Singular **Plural**

He

Ben loves cake.

x



Susan plays football with her friends.

Singular **Plural**

Normally, Simon hates to lose.

Last Saturday, Susan and Emily scored two goals.

Peter kicked the ball a long way.

Q2. Label each of the words below as either a **verb**, a **noun** or a **pronoun**. Write your answers in the boxes. You can write **V**, **N** or **P** if you prefer.

verb

noun

Example: The tiger **slept** under the **tree**.

The lioness was hunting. She **crawled** towards a **zebra** as it rested

in the **grass**.



The zebra and antelope **scattered**, because **they** were scared.

The lion was hungry, so **he** gave chase.

Q3. Here are some singular and plural sentences. Fill in the missing sentences.

Hint: All the sentences should be in the present tense.

Singular	Plural
Example: My cousin loves giraffes.	Example: My cousins love giraffes.
a. My best friend comes from Kenya.	a.
b.	b. Elephants eat roots, grasses, fruit and bark.
c. An African elephant has larger ears than an Asian elephant.	c.

Q4. Circle the **verb ending (suffix)** in each of the sentences below.
Hint: Draw your circles carefully!

Example: Samira **plays** tennis.

Ben loves fresh bread from the bakery.

Last weekend, Sara and Ben visited the bakery.

Sara talked to the baker.

Normally, Frank buys a cake.



Q5a. Change these verbs from the present tense into the **simple past tense**.

Present tense

Simple past tense

Example: **play**

→ **played**

walk

→

dance

→

live

→

Q5b. Change these verbs from the simple past tense into the **present tense**.

Simple past tense

Present tense

Example: **talked**

→ **talk**

used

→

called

→

enjoyed

→

Section 2

- These words are from a language called Esperanto.
Look how you say one animal and more than one animal.



one elephant → elefanto



more than one elephant → elefantoj



one cat → kato



more than one cat → katoj

Fill in the missing words. Use the examples above to help you.

One	More than one
 hundo	 1. _____
 2. _____	<u>leonoj</u> 
 3. _____	<u>delfenoj</u> 
 <u>porko</u>	 4. _____

2. Look how you talk about events happening NOW, IN THE PAST and IN THE FUTURE.

now ↓	I love → mi am <u>as</u>
past ←	I loved → mi am <u>is</u>
future →	I will love → mi <u>am<u>os</u></u>

Fill in the missing words below. Use the examples above and below to help you.

Example

How do you say...?

1 ↓ I hope = Mi <u>esperas</u>	I will hope = → Mi _____
2 → I will play = Mi <u>judo<u>s</u></u>	← I played = Mi _____
3 ← I needed = Mi <u>bezo<u>n</u>is</u>	↓ I need = Mi _____

Section 3

1. Somali is a language from Africa.

Look how you say “I” did something and “he” did something in the past.

Can you spot any patterns?

English meaning	Somali verb (“I”)	Somali verb (“he”)	English meaning
I went	tagay	tagtay	He went
I brought	keenay	keentay	He brought
I rose	kacay	kactay	He rose
I taught	baray	bartay	He taught

Fill in the missing words in the table below. Use the examples above to help you.

English meaning	Somali verb (“I”)	Somali verb (“He”)	English meaning
I saw	aragay	1.	He saw
I built	2.	dhistay	He built
I entered	3.	galtay	He entered
I ran away	cararay	4.	He ran away

1. This is a new language, *Oqaramana*.

ogrtsi means owl



Look at these examples.

Can you spot any patterns?



ogrtsi krit



ogrtsin krite



ogrtsi mrod



ogrtsin mrode

Now put a cross (X) next to the sentence that describes each picture. **Only choose one sentence for each picture.** Use the examples to help you.



- ☐ *ogrtsi riuk*
- ☐ *ogrtsin riuke*
- ☐ *ogrtsin riuk*
- ☐ *ogrtsi riuke*



- ☐ *ogrtsin gamne*
- ☐ *ogrtsi gamn*
- ☐ *ogrtsi gamne*
- ☐ *ogrtsin gamn*



- ☐ *ogrtsi dule*
- ☐ *ogrtsin dul*
- ☐ *ogrtsi dul*
- ☐ *ogrtsin dule*

Appendix B.2

Piloting Information - The Language-analytic Ability Test

The LAA test was first piloted with a Year 3 class of 31 pupils (aged 7-8) and returned a range of scores from 4-42 out of 44. After the first pilot, following feedback from class teachers, it was decided that having pupils complete the test over two sessions was more appropriate for younger learners. It was then repiloted with a class of 29 Year 2 pupils from a different school at the end of the school year so that they were as close as possible in age to the Year 3 participants at the start of September. The range of scores with this younger group of pupils was 3-34.

Appendix B.3

Language-analytic Ability Test Instructions

LAA QUIZ

Instructions: *As a general rule, please **do not give** any help beyond what is provided here. I would suggest you read the questions out loud for all pupils. Pupils should answer individually and not discuss their answers with each other. If a Teaching Assistant is supporting pupils, please ask them to help only in understanding what to do; please ask them not to provide any answers. (It's fine for them to scribe, but please get them to check intended spellings with pupils first).*

*I know it's really hard to see pupils struggle (and some will find this test difficult) but it's really important that this test shows a range of responses, so **if all pupils get full marks, it's not fit for purpose**. As a guide, a Year 3 group recently completing this test scored 4-42, and a Year 2 class scored 3-34. Please don't indicate to pupils whether something is right or wrong, please just encourage them. Please do reassure them that it's fine if they are not sure of an answer, just do the best they can and write what they think the answer is. The examples should help – if pupils still seem to be unsure, you can provide another example but please ensure that you don't give away anything that's included in the actual test.*

*If any pupils **really** struggle with handwriting, I have also sent the word document, and they could do it on a laptop.*

If they change their mind after writing an answer, it's fine, they can change it. Just make it clear to me (rub out or cross out neatly the wrong answer).

Please ensure they add their name.

Section 1

Q1 – Read a) and refer to example.

Read b). If singular / plural causes issues, refer to 'one / more than one'.

Q2 - Read and refer to example. If some pupils are struggling with the writing, they can put V, N or P.

Q3 - Read and refer to example. I've numbered the blank boxes 1-3 to help make sure pupils don't miss any out. Please ask them to make sure they answer all 3. It's fine to point out the headings of singular and plural. In #2, they are only expected to change the first part of the sentence (An elephant eats ...); in #3, they need to change both parts - African elephants have larger ears than Asian elephants.

Q4 - Read and refer to example.

Q5a) and b) Read and refer to example. These should be straightforward.

Section 2

1. Read the short intro and the examples in Esperanto. Point out the headings on the columns: One / More than one. I have numbered the blanks 1-4 so they hopefully don't miss any out.
2. Read the instructions and the table at the top of the page. Don't point out the 'stem' (i.e. 'am') just read the verbs with the endings with reference to the present, future, past. Fine to read the examples 1-3 in the table they have to complete.

Section 3

1. Read the introduction to Somali verbs. We want to see if pupils can spot/ apply the patterns themselves. I have numbered 1-4 for clarity.
2. Read the intro to Ogramana. Some of the pictures will need to be explained. You can say something like: *what do you think the owl in the first picture is doing? Yes, he's cooking. So choose the sentence that you think means 'an owl is cooking'.*

Pictures:

Knitting

Sleeping

Cooking

Eating

Playing music

Appendix C.1

Teacher Information and Consent Form

Information Page

‘Comme ci ou comme ça?’ An evaluation of two different approaches to the teaching of primary French.



Dear colleagues

I am currently carrying out a PhD research project: “*‘Comme ci ou comme ça?’* An evaluation of two approaches to the teaching of primary French.” I am a qualified teacher of foreign languages, with 25 years of classroom experience, and Senior Lecturer in teacher education at the University of Chichester. I would like to invite your school to take part in this research project. My research project is an evaluation of two different approaches to the teaching of primary French and is in collaboration with the National Centre for Excellence for Language Pedagogy.

Before agreeing to take part, please read this information sheet carefully and let me know if anything is unclear or you would like further information.

Please also read the information about General Data Protection Regulation (GDPR) that is provided at the end of this document.

Purpose of the study

The study is designed to evaluate the impact, in a limited-exposure context with non-specialist teachers, of a new approach to teaching French (which teaches high-frequency vocabulary,

systematic phonics and sequenced explicit core grammar) on year 3 pupils' learning outcomes and perceived motivation and on teachers' perceived confidence and enthusiasm, when compared with an approach that simulates the 'standard' pedagogy (which teaches topic-based vocabulary, non-systematic phonics, with little explicit grammar teaching). Both approaches are designed to facilitate progress in the language and motivate pupils and are fully compatible with the National Curriculum.

What would this mean for you?

Your school has been invited to take part as it has two parallel Year 3 classes which will enable pupils to be quasi-randomly allocated to either the standard or the new teaching approach, creating two new teaching groups (pupils will be allocated in 'matched pairs' to either group following a short language learning aptitude activity). Pupils will then follow this approach for the whole academic year. As both groups would be taught simultaneously, non-specialist class teachers would alternate, teaching one approach one week, and the alternative approach in the following week. Pupils will experience either the new or the standard approach. Year 3 French lessons would need to take place once a week for 40-45 minutes. This will ensure that pupils receive regular French lessons weekly for the duration of the project, with both approaches designed to facilitate progress in the language. All planning, teaching and assessment materials will be provided, along with supporting audio files. The duration of the project is one academic year, with the assumption that 30 sessions will be taught.

As a non-specialist teacher, you will have the opportunity to teach and evaluate two distinct approaches to teaching French, which will in turn support the development of your own linguistic and pedagogic competence and understanding.

Data would be collected in the following ways:

- Online questionnaires (Sept 2022, Feb 2023, July 2023) taking approximately 15 minutes each
- Short online lesson evaluations after each French lesson taking approximately 1 minute each
- Individual online or in person interviews (format to suit you) (Sept 2022, Feb 2023, July 2023, and once in academic year 2023-24) talking approximately 30 minutes each
- Lesson observations in visits from me just to ensure that the materials are being used as intended (approximately one or two per term and lesson audio recordings (approximately one in three lessons recorded)

You will also help collect data in the following ways:

- Administer hard copy child-friendly questionnaires (Sept 2022, Feb 2023, July 2023) taking approximately 15 minutes each
- Administer initial language learning aptitude test (Sept 2022) taking approximately 20 minutes and language tests (Feb 2023, July 2023, Sept 2023) – these would take place in a scheduled French lesson and would assess pupils’ performance in symbol-sound correspondences, vocabulary and grammar, using listening, speaking, reading, and writing.

Participation is voluntary

Participation in using the resources provided and providing data are optional. If you do decide to take part in providing data, you will be given a copy of this information sheet for your records and will be asked to complete a consent form. If you change your mind at any point during the study, you will be able to withdraw your participation in using these resources and providing data without having to provide a reason.

Anonymity and confidentiality

The data that you provide (e.g. audio recordings of interviews, notes from observations, post-lesson evaluations) will be stored by code number. Any information that identifies you will be stored separately from the data. You are free to withdraw from the study at any time during data collection and up to the end of September 2023.

Storing and using your data

Data will be stored on a password protected computer and in the University of York’s centrally managed filestore. The only person who will have access to the data with your name attached will be me. My supervisor will see data that has been pseudo-anonymised (your name removed). Data will be fully anonymised at the end of October 2023.

I am practising Open Science and anonymised data will be managed professionally and stored indefinitely with the University’s Research Data York service or free, open access IRIS repository (<https://www.iris-database.org>) and / or other long-term repository.

The data that I collect [transcripts of interviews / post-lesson evaluations] may be used in **anonymous** format in different ways, e.g. publications, presentations and online; it may also be used for future analysis. Please indicate on the consent form attached with a tick ☒ if you are happy for this anonymised data to be used in the ways listed.

You will be given the opportunity to comment on a written record of your interviews.

Please note: If we gather information that raises concerns about your safety or the safety of others, or about other concerns as perceived by the researcher, the researcher may pass on this information to another person.

Questions or concerns

If you have any questions about this participant information sheet or concerns about how your data is being processed, please feel free to contact me (Elaine Minett) by email (em1599@york.ac.uk), or the Chair of Ethics Committee via email education-research-admin@york.ac.uk. If you are still dissatisfied, please contact the University's Data Protection Officer at dataprotection@york.ac.uk

I hope that you will agree to take part. If you are happy to participate, please complete the form attached and email it back to me at em1599@york.ac.uk.

Please keep this information sheet for your own records.

Thank you for taking the time to read this information.

Yours sincerely

Elaine Minett

‘Comme ci ou comme ça?’ An evaluation of two different approaches to the teaching of primary French.

Consent Form

Please tick each box if you are happy to take part in this research.

Statement of consent	Tick each box
I confirm that I have read and understood the information given to me about the above-named research project and I understand that this will involve me taking part as described above.	
I understand that participation in this study is voluntary and that if I wish to withdraw, I can do so at any time during data collection and up to the end of September 2023.	
I understand that my data will not be identifiable and the anonymous data may be used in publications, presentations and online.	
I confirm that I have read the information about GDPR	

Name:

Signature:

Date:

Processing personal data

Under the General Data Protection Regulation (GDPR), the University has to identify a legal basis for processing personal data and, where appropriate, an additional condition for processing special category data.

In line with our charter which states that we advance learning and knowledge by teaching and research, the University processes personal data for research purposes under Article 6 (1)(e) of the GDPR:

Processing is necessary for the performance of a task carried out in the public interest

Special category data is processed under Article 9 (2) (j):

Processing is necessary for archiving purposes in the public interest, or scientific and historical research purposes or statistical purposes

Research will only be undertaken where ethical approval has been obtained, where there is a clear public interest and where appropriate safeguards have been put in place to protect data.

In line with ethical expectations and in order to comply with common law duty of confidentiality, we will seek your consent to participate where appropriate. This consent will not, however, be our legal basis for processing your data under the GDPR.

Protecting and storing personal data

Information that research participants provide will be treated confidentially and shared on a need-to-know basis only. The University is committed to the principle of data protection by design and default and will collect the minimum amount of data necessary for the project. In addition we will anonymise or pseudonymise data wherever possible.

We will put in place appropriate technical and organisational measures to protect your personal data and/or special category data (for example, data may be stored in secure filing cabinets and/or on a password protected computer).

Sharing of data

The default position is that personal data will only be accessible to members of the project team. In some cases, however, the research may be of a collaborative nature and hence the data will be made accessible to others from outside the University. Information specific to the project will include details of when this is the case, who the 3rd parties are, and what they will do with the data. It is possible that personal data may be shared anonymously with others for secondary research and/or teaching purposes.

Transfer of data internationally

The default position is that data will be stored on University devices and held within the European Economic Area in full compliance with data protection legislation.

However, data may be transferred to the project partners based outside the European Economic Area. Any international transfer will be undertaken in full compliance with the GDPR.

The University has access to cloud storage provided by Google which means that data can be located at any of Google's globally spread data centres. The University has data protection compliant arrangements in place with this provider. For further information see,

<https://www.york.ac.uk/it-services/google/policy/privacy/>

Your rights in relation to your data

Under the GDPR, you have a general right of access to your data, a right to rectification, erasure, restriction, objection or portability. You also have a right to withdrawal. Please note, not all rights apply where data is processed purely for research purposes. For information see,

<https://www.york.ac.uk/records-management/general-dataprotection-regulation/individual-rights/>

Right to complain

If you are unhappy with the way in which your personal data has been handled, you have a right to complain to the Information Commissioner's Office. For information on reporting a concern to the Information Commissioner's Office, see www.ico.org.uk/concerns

Appendix C.2

Child Assent Form

ASSENT FORM

Subject Number



What is research?

Research is finding out the answer to an important question, by carefully collecting information and looking at it very closely to see what we can find out. This research is being done to look at different ways of teaching French and how it can benefit pupils' learning.



You are being asked if you want to join in this research because we want to find out how you feel when you are being taught French. Please read this information carefully or ask someone to read it aloud. You will be given a copy to keep. If you have any questions, you should ask your teacher.



What will happen to me if I take part?

You will be asked to answer some questions in a questionnaire. There will be three questionnaires this year. You might also take part in a group interview with other pupils in the class. These questions will ask you how you feel about learning French.

If you want to join in, please write your name at the end of this form.

If you have any questions - please do ask them!

Do I have to take part?

Everyone will be learning French but you can decide if you want to fill in the questionnaires and take part in the interviews if you are asked. You can decide if we can use the information we find out about you. No one will be upset if you say "no". You can stop being in the research any time after it has started - just tell someone your teacher.



Will anything about the research upset me?

No, there is nothing in the research that should upset you. But if you become unhappy when you are taking part in the research we will stop and we will not ask you any more questions.

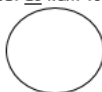
ASSENT FORM

Circle Yes or No

- ♦ Have you read about this research? Yes/No
- ♦ Has somebody else explained this research to you? Yes/No
- ♦ Do you understand what the research is about? Yes/No
- ♦ Have you asked any questions you want to? Yes/No
- ♦ Have you had time to think about taking part? Yes/No
- ♦ Are you happy to take part? Yes/No
- ♦ Are you happy to let us use the information we find out about you? Yes/No
- ♦ Are you happy to let us use the information in the future? Yes/No

If you **don't** want to take part, please tell your teacher.

If you don't enjoy writing yet, but **do** want to take part, please draw a smiley face:



If you enjoy writing and want to take part, please fill in the box below:

Your Name	Your Signature	Date

Thank you very much for your help with this research.

Appendix D.1

Data Management Plan

Plan Overview

A Data Management Plan created using DMPOnline

Title: 'Comme ci ou comme ça?' An evaluation of two different approaches to the teaching of primary French.

Creator: Elaine Minett

Principal Investigator: Elaine Minett, Emma Marsden

Data Manager: Elaine Minett, Emma Marsden

Project Administrator: Elaine Minett

Affiliation: University of York

Template: York DMP Template for Doctoral Student Research Projects

ORCID ID: 0000-0003-4974-373X

Project abstract:

A foreign language has been a statutory part of the National Curriculum since Sept 2014 (DfE, 2013), requiring pupils to make 'substantial progress', but very little (research-informed) guidance has been provided for primary school teachers on how to teach primary languages effectively in the limited-exposure context which typifies English provision. Ensuring that such 'substantial progress' takes place may be particularly challenging for non-specialist teachers who make up almost half of the primary Modern Foreign Language teacher population (Languages Trends, 2019). This study will evaluate the impact, in a limited-exposure context with non-specialist teachers, of a novel approach to teaching French (which teaches high-frequency vocabulary, systematic phonics and sequenced explicit grammar) on year 3 pupils' learning outcomes and perceived motivation and on teachers' confidence and enthusiasm, when compared with an approach that simulates the 'standard' pedagogy (which teaches topic-based vocabulary, non-systematic phonics, with little explicit grammar teaching).

The study will be carried out as a PhD study in six primary schools in the vicinity of the University of Chichester, where the PhD student works. These schools will either:

- have two-form entry in Year 3 (pupils aged 7-8), allowing both models of delivery to be implemented in the same year group. Pupils will be randomised to one approach or the other.

or, if funding is available to fund a visiting non-specialist teacher to work with half the Year 3 class, be a one-form entry in Year 3. In this case, the class teacher will teach one half of the class and the visiting teacher the other, with teachers alternating each week.

ID: B6285

Start date: 01-09-2022

Created using DMPOnline. Last modified 28 October 2021

1 of 9

End date: 01-03-2026

Last modified: 28-10-2021

'Comme ci ou comme ça?' An evaluation of two different approaches to the teaching of primary French.

1. Defining your data

1a. What data will you produce?

1a. What data will you produce?

Personal data - consisting of name, contact email address for teacher participants, age. I will only collect the personal data needed for the research.

Research data (which will be managed in accordance with legal and ethical requirements)

1. Questionnaires

All participating pupils will complete paper questionnaires at the start of the project, at the interim point (February 2023) and at the end of the project (July 2023) (initial one including consent forms) which will be analysed using SPSS. All teachers will complete short online questionnaires using Qualtrics (start of project, half-way through project and end of project) which will be analysed using SPSS. Consent forms will be included as part of the questionnaire online.

2. Short lesson evaluations

All teachers will complete a very **short evaluation** after each taught lesson to evaluate the impact of each teaching approach.

3. Interviews

All teachers will take part in individual **interviews** at the start of the project, part-way through the project, at the end of the project, and then once during the year following the end of the project to evaluate lasting impact. These may be in person or online.

Focus group pupil interviews will take part in group interviews part-way through and at the end of the study. These will be in person. I will record these interviews with participants using the audio recording facility on my laptop (which will be encrypted), then **transcribe** them into text. These will be analysed using NVIVO software.

4. Tests

All pupils will complete tests at the following points:

Start of project: Language aptitude test
Interim test - February 2023
Post-test - July 2023
Delayed post-test (approximately 6 weeks after the end of the project)

Tests will be of vocabulary, grammar and phonics.

Data will also include inter-rater reliability ratings where judgement is involved. This will be recorded in Excel.

5. Lesson recordings

Audio only, likely to be around 120, up to 360

6. Lesson observations

Observation notes taken on sample lesson observations of each model to ensure fidelity to treatment. These will be taken in Word.

1b. What formats and what software will you use?

All formats will be in uncompressed versions and will make use of widely used formats such as Office. Copies will be kept of documents in original formats if they are amended.

1. Hard copy child questionnaires; Qualtrics online questionnaires for teachers. Analysis using SPSS.
2. Qualtrics short evaluations emailed to teachers after each taught lesson.
3. Interviews recorded using the sound recording facility on the laptop and saved as NPS files; transcriptions created using Word. Analysis completed using NVIVO.
4. (i) Oral production of vocabulary recorded using sound recording facility on laptop.
(ii) Oral production of grammar-focus sentences recorded using sound recording facility on laptop. Hard copy of grammar-focus written responses.
(iii) Read-aloud tests recorded using sound recording facility on laptop. Paper copies of phoneme discrimination tests.
Excel spreadsheets of inter-rater reliability judgements.
5. Audio recordings will be completed using audio records, laptops or phones. A sample of these will be viewed and observation notes on the recordings completed in Word.
6. Observation notes taken in Word.

1c. How much data do you expect to generate?

1. Hard copy questionnaires - approximately 360 participants (60 x six schools) x 3 questionnaires, so 1080.
2. Teacher questionnaires - approximately 80KB per questionnaire x 3 in year = 240KB, x 10 teachers = 2,400KB.
3. Post-lesson teacher evaluations - 20KB per evaluation x 30 lessons = 600KB, x 10 teachers = 6,000KB.
4. Interview transcripts - pupils and teachers - estimated @ 150KB per transcript. Pupils - one focal group per model per school per collection point = 12 x 2 collections = 24. Teachers - four individual interviews x 10 teachers. Total: 5,200KB.
5. Audio recordings of lessons - estimated 20MB per recording. Number of recordings unknown. If 30 per teacher, then approximately 6,000 MB.
6. Audio recordings of interviews with pupils and teachers - estimated 25MB per recordings. Pupils x24 = 600MB; teachers x30 = 750MB.
7. Test data - collected by (or other online platform for data collection) and downloaded and then backed up on password protected folders. 60 pupils x 6 schools x 5 language tests x 3 phases of tests = 5400 files, of which 1/5 will include audio files (the speaking tests).

1d. Who owns the data you will generate?

According to my studentship agreement, the University owns all data I create, but I retain the copyright on publications based upon my data.

2. Looking after your data

2a. Where will you store your data?

I will store my data on the University's centrally managed network, in my personal filestore (the M: drive). I will use the VPN (Pulse downloaded onto personal laptop) to connect to my personal filestore and work on/add to my files.
Online questionnaires will be stored in the password-protected Qualtrics environment.
Hard copies of pupil questionnaires and test responses will be stored in a locked cabinet in my office at work. I will be the only person with access to the cabinet. On the rare occasion that I am unable to take them to the office directly after finishing in a school, I will take them home and store in a locked cupboard at home.

2b. How will you back-up your data?

My data will be stored on the University's centrally managed filestore. The filestore is regularly and automatically backed up by IT Services.

2c. Who else has a right to see or use this data during the project?

Only my supervisor should have access to my data during the project.

2d. How will you structure and name your folders?

Please see separate flow chart. Use will be made of capital letters, dashes or underscores rather than spaces or slashes.

2e. How will you name your files?

I will use the conventions:

[Interview-<participant number>] [YYYYMMDD], for example interview024_20220905 is an interview with participant 024 on 5 Sept 2022
[PFG-<pupil focus group number>] [YYYYMMDD], for example PFG1_20230210 is the transcript of the first pupil focus group, that took place on 10 February 2023

1b. What formats and what software will you use?

2f. How will you manage different versions of your files?

In most circumstances there will only ever be one version of each data file - each data collection point creates new data, which are stored in a new set of files.
If documents are modified, they will be saved as v_1 with major changes indicated as v_2 and minor changes as v_1_1.

2g. What additional information will be required to understand your data?

Supporting documentation / metadata will be provided in an accompanying Word document, to describe what data is contained within the datasheet, where it originated from, and how it was analysed in the original study.

This may include for example participant ID, institution ID. Data needs to be FAIR: findable, accessible, interoperable and re-usable. Findable - for example, when research materials are cited / where they can be found; accessible - available as part of Open Research practices; re-usable - will allow other researchers to use my data; interoperable - I will use most up to date software I can.

We plan to store our materials and data on an open repository such as IRIS or the OSF and assign them a DOI, to help their retrievability and accessibility.

3. Archiving your data

3a. What data should be kept or destroyed after the end of your project?

Data may be retained to validate research findings, in line with the University of York's Code of Practice on Research Integrity and commitment to transparency and open communication.
This could include, where prior consent has been obtained from participants:

- Questionnaire responses
- Interview transcripts
- Test results

The anonymised datasets of users' responses pseudonymised personal data are of long-term value for research purposes.

3b. For how long should data be kept after the end of your project?

In line with the University Research Data Management Policy, the data that will support my published research findings will be kept for 10 years from the date of last requested access.

3c. Where will the data you keep be archived at the end of the project?

I will provide a copy of the data supporting my thesis to the University's Research Data York service who will keep the selected data for the agreed retention period. Retained data will be registered with PURE.

The anonymised datasets, along with relevant metadata and documentation, will be uploaded to and stored in the free, open access IRIS repository (<https://www.iris-database.org>) and / or other long-term repository.

3d. When will you archive your data?

I will provide a copy of the data supporting my thesis to the University for long-term retention when I submit my thesis.

4. Sharing your data at the end of your project

4a. What data should or shouldn't be shared openly and why?

Teacher participants will be informed that their responses to short lesson evaluations, questionnaires and interviews will be anonymised and archived for data sharing with the University's Research Data York service. Explicit consent for this will be collected. Data will only be shared if participants have consented to data archiving and sharing.

Parents or carers of children will be informed that responses to questionnaires and interviews, as well as test data, will be anonymised and archived for data sharing with the University's Research Data York service. Explicit consent for this will be collected. Data will only be shared if participants have consented to data archiving and sharing.

Recordings of lessons will not be shared.

Raw audio recordings of interviews will not be shared but, where permission is given, anonymised transcripts could be shared.

If not all of my participants give informed consent for their anonymised data to be shared, I will exclude their results from the final dataset that I publish.

Identifiable data will not be shared outside the PhD study team.

4b. Who should have access to the final dataset(s) and under what conditions?

The data will be open access, provided the data are anonymised and provided participants have given consent for their data to be used in this way.

4c. How will you share your final dataset(s)?

A copy of the data supporting my thesis/publication will be retained by the University's Research Data York service University, and my supervisor will field any requests for access on my behalf. Data will be made available to the requestor where appropriate.

5. Implementing your plan

5a. Who is responsible for making sure this plan is followed?

I will take responsibility for carrying out the actions required by this plan and report them to my supervisor as appropriate.

5b. How often will this plan be reviewed and updated?

My supervisor and I will review this plan every 6 months and will agree updates if necessary.

5c. What actions have you identified from the rest of this plan?

- Contact IT support regarding encryption for laptop
- Test VPN on downloaded Pulse
- Learn how to anonymise my data so that it can be shared.
- Ensure that I request informed consent from my participants for archiving and sharing their data.

Find out how to add metadata to files.

5d. What policies are relevant to your project?

This project is covered by the University of York Research Data Management Policy.

5e. What further information do you need to carry out these actions?

See notes above.

-

Appendix D.2

Data Collection Logs

Spring term	Observation booked for Spring term	T questionnaire 2 completed	Pupil questionnaire booked	FG	T interview booked
Sch A T3	01.02.23 - DONE	✓	01.03.23 + me DONE	08.03.23	28.02.23 DONE
Sch A T4	N/A	T questionnaire 1	01.03.23 + me DONE	08.03.23	24.02.23 DONE
Sch B T1	16.3.23 - DONE	✓	28.02.23 and 01.03.23 DONE	09.03.23	17.04.23
Sch B T2	23.03.23 - DONE	✓	I'm not needed	09.03.23	17.04.23
Sch C T3	21.03.23 - DONE	✓	28.02.23 + me DONE	14.03.23	28.02.23 DONE
Sch E T3	21.02.23 - DONE	✓	28.02.23 + me DONE	07.03.23	28.02.23 DONE
Sch E T7	21.03.23 - DONE	✓	28.02.23 + me DONE	07.03.23	09.02.23 DONE
Sch F T8	06.02.23 - DONE	✓	2nd week after half-term	06.03.23	T 06.02.23 DONE
Sch F T9	27.03.23 - DONE	✓	2nd week after half-term	06.03.23	T 06.03.23 DONE

Session 30 to be taught by	Observation booked	T questionnaire 3 completed	Pupil questionnaire	FG	T interview booked	Pupil test 1	Pupil test 2	Sp test 1
w/b 19.06.23	19.04.23 - DONE	Yes	28.06.23 - DONE	07.07.23 - DONE	27.06.23 - DONE	28.06.23 - DONE	05.07.23 - DONE	05.07.23 - DONE (a)
w/b 26.06.23	10.05.23 - DONE	Yes	04.07.23 - DONE	17.07.23 - DONE	13.07.23 - DONE	04.07.23 - DONE	11.07.23 - DONE	17.07.23 - DONE
w/b 26.06.23	03.05.23 - DONE	Yes	04.07.23 - DONE	17.07.23 - DONE	13.07.23 - DONE	04.07.23 - DONE	11.07.23 - DONE	17.07.23 - DONE
w/b 19.06.23	02.05.23 - DONE	Yes	27.06.23 - DONE	04.07.23 - DONE	27.06.23 - DONE	27.06.23 - DONE	04.07.23 - DONE	04.07.23 - DONE
w/b 19.06.23	25.04.23 - DONE	Yes	26.06.23 - DONE	11.07.23	27.06.23 - DONE	28.06.23 - DONE	04.07.23	04.07.23 - DONE
w/b 19.06.23	23.05.23 - DONE		26.06.23 - DONE	11.07.23	29.06.23	28.06.23 - DONE	04.07.23	04.07.23 - DONE
w/b 03.07.23	05.06.23 - DONE	Yes	10.07.23 - DONE	10.07.23 & 14.07.23 - Done		10.07.23 - DONE	12.07.23 - DONE	10.07.23 - DONE
w/b 03.07.23	12.06.23 - DONE	Yes	10.07.23 - DONE	10.07.23 & 06.07.23 - DONE		10.07.23 - DONE	12.07.23 - DONE	10.07.23 - DONE

Sept data	Pupil test 1	Pupil test 2	Sp test 1
Sch A T3	04.09.23 DONE	04.09.23 DONE	07.09.23
Sch B T1	18.09.23 DONE	18.09.23 DONE	26.09.23
Sch B T2			
Sch C T3	05.09.23 DONE	05.09.23 DONE	11.09.23
Sch E T3	07.09.23 DONE	07.09.23 DONE	12.09.23
Sch E T7			
Sch F T8	06.09.23 DONE	06.09.23 DONE	08.09.23
Sch F T9			

	SchA T3	SchA T6	SchB T1	SchB T2	SchC T3	Sch E T3	SchE T7	SchF T8	SchF T9						
Audio recording	Yes - 1	Yes - 1	Yes - 1	Yes - 1	Yes - 1	Yes - 1	Yes - 1	Yes - 1	Yes - 1						
Date received	05.10.22	05.10.22	15.09.22	26.09.22	04.10.22	05.10.22	10.10.22	10.10.22	14.11.22						
Approach	S	N	N	N	N	N	S	S	N						
Week	3	3	2	3	3	3	3	5	5						
Audio recording	Yes - 2	Yes - 2	Yes - 2	Yes - 2 - IRR	Yes - 2	Yes - 2	Yes - 2	Yes - 2 - IRR	Yes - 2						
Date received	02.11.22	09.11.22	20.10.22	10.11.22	01.11.22	08.11.22	05.11.22	21.11.22	12.12.22						
Approach	N	N	S	S	N	N	N	N	N						
Week	6	7	6	7	6	7	6	6	10						
Audio recording	Yes - 3	Yes - 3	Yes - 3	Yes - 3	Yes - 3	Yes - 3	Yes - 3	Yes - 3	Yes - 3 - IRR whole lesson						
Date received	30.11.22	07.12.22	05.12.22	12.12.22	29.11.22	06.12.22	13.12.22	12.12.22	26.01.23						
Approach	S	S	N	S	S	N	N	S	S						
Week	9	10	9	11	9	10	11	10	13						
Audio recording	Yes - 4		Yes - 4	Yes - 4	Yes - 4	Yes - 4	Yes - 4	Yes - 4	Yes - 4						
Date received	10.01.23		12.01.23	19.01.23	10.01.23	10.01.23	24.01.23	26.01.23	20.02.23						
Approach	N		N	N	N	S	N	N	N						
Week	13		13	14	12	13	15	13	16						

IRR T9 standard
T3 (SchA) standard
T3 (SchE) novel

Appendix E.1

Mapping of the National Curriculum and Testing Content

Taken from the *Languages programmes of study: Key stage 2* ((DfE, 2013a:2-3). **Bold text shows which objectives are addressed through tests.**

Pupils should be taught to:

- **listen attentively to spoken language and show understanding by joining in and responding**
- explore the patterns and sounds of language through songs and rhymes and **link the spelling, sound and meaning of words**
- engage in conversations; ask and answer questions; express opinions and respond to those of others; seek clarification and help
- **speak in sentences, using familiar vocabulary, phrases and basic language structures**
- **develop accurate pronunciation and intonation so that others understand when they are reading aloud or using familiar words and phrases**
- present ideas and information orally to a range of audiences
- **read carefully and show understanding of words, phrases and simple writing**
- appreciate stories, songs, poems and rhymes in the language
- broaden their vocabulary and develop their ability to understand new words that are introduced into familiar written material, including through using a dictionary
- **write phrases from memory, and adapt these to create new sentences, to express ideas clearly**
- **describe people, places, things and actions orally and in writing**
- **understand basic grammar appropriate to the language being studied, including (where relevant): feminine, masculine and neuter forms and the conjugation of high-frequency verbs; key features and patterns of the language; how to apply these, for instance, to build sentences; and how these differ from or are similar to English.**

Appendix E.2

Piloting the Vocabulary Tests

(i) Introduction

It quickly became evident that piloting a scheme which aimed to reflect fairly the aims and content of both the Standard and the more Novel way of learning French was challenging, given that the Novel approach was by its very nature new, placing a greater emphasis on verb structures and high frequency vocabulary as well as SSCs. At the time of the creation of the pilot tests, the Novel scheme of work was not available in its entirety, and it was therefore impossible to finalise the vocabulary content. Sourcing a class of pupils who could participate in such a pilot when the Novel scheme was newly created was also difficult. For this reason, listening, reading and writing vocabulary tests were first piloted (at the same time as the phonics test) in June 2022 with three Year 3 pupils, then with the whole class of 26 pupils, who had followed a scheme of work which covered similar content to the Standard approach; the goal in this pilot was to check test format rather than content, in the knowledge that the latter would need to be revised once a suitable pilot class was available and the Novel scheme was completed. This round of piloting checked clarity of instructions and the time needed for completion (32-38 minutes), as well as providing some feedback on test content and the practicalities of administering tests to young learners.

The speaking pilot was conducted with a subsample of nine pupils from the same class, again using three versions of the test.

(ii) Findings from the pilot vocabulary tests

The initial design for all vocabulary tests consisted of three different counterbalanced test versions in order for all items to be tested in all knowledge types: meaning recognition, oral modality; meaning recognition, written modality; and written and spoken production. Had this design been adopted, the common words would have been divided into three sets, with each set featuring in a different testing approach in each test phase (mid-, post- and delayed post-test). As with the initial pilot, pupils in each school in each approach (Novel or Standard) would have been allocated to three groups of approximately equal language analytical aptitude and would have rotated through a different test type in each testing phase. By the end of the testing phase, all pupils would have been tested on all lexical items in all three test types. However, although this would have enabled conclusions to be drawn regarding whether a group had made gains, it would not have provided information relating to individual pupils. It would also have been

difficult to conclude whether any gains were due to the difficulty of particular tests, treatment effectiveness, or because the words in a specific set were easier than in another.

The initial pilot also revealed that, practically, it was highly challenging to administer three versions of the same test to pupils within one class, given the young age of the pupils, especially when one involved listening when other pupils were reading or writing. Tests were online and some pupils took a long time to log into their accounts, and there were issues with headphones needed for the listening test not working. It was therefore decided to allocate words only to one test type ([E.3](#)); this approach was used in the next pilot.

(iii) Repiloting the vocabulary tests

Initial changes were made to the structure and content of the tests to better reflect the newly completed Novel scheme of work and a single test consisting of listening, reading, speaking and writing quizzes was created. Some schools across the country had begun to implement the teaching of the Novel scheme of work in September 2022 and, through contact with the creator of the schemes, a school was sourced which was able to repilot the final version of the listening, reading and writing vocabulary tests in February 2023. (This version of the tests contained 11 listening, 10 reading and 11 writing items rather than the 12 planned for the final test, because the class teacher had requested the removal of words that had not yet been taught; the speaking contained 12 items). In this pilot, the 29 Year 3 pupils' listening scores ranged from 1-10 ($M = 6.5$), reading scores from 0-10 ($M = 6.48$), speaking scores from 3-9 (out of 24) ($M = 6.5$) and writing scores from 0-11 (out of 22) ($M = 4.75$); no individual items appeared to be particularly difficult. Although speaking and writing scores were lower than reading and listening, this was expected given the productive nature of the task and the young age of the learners. No further repiloting was deemed necessary.

Appendix E.3

Vocabulary Content Allocated to Test Type

LISTENING	SPEAKING	READING	WRITING
regarder	parler	midi	oui
cheval	orange	stylo	chat
salut	univers	non	bien
je	bonjour	jour	neuf
jeudi	dimanche	gâteau	banane
moto	et	manger	enfant
chien	ici	règle	mercredi
mais	crayon	lundi	mardi
mal	jeu	aujourd'hui	aussi
vendredi	samedi	dans	fête
deux	maison	tu	écouter
semaine	un	au revoir	gomme

Note. In all following vocabulary lists, the following heading terms are used:

Listening = vocabulary recognition in the oral modality

Speaking = vocabulary recall in the oral modality

Reading = vocabulary recognition in the written modality

Writing = vocabulary recognition in the written modality

Appendix E.4

Distribution of Corpus-based Frequency Ratings According to Test Type

	LISTENING		SPEAKING		READING		WRITING	
	regarder	425	parler	106	midi	2483	oui	284
	cheval	2220	une orange	3912	un stylo	n/a	chat	3138
	salut	2205	univers	2112	non	75	bien	47
	je	22	bonjour	1972	jour	78	neuf	787
	jeudi	1112	dimanche	1235	gâteau	4845	banane	n/a
	une moto	n/a	et	6	manger	1388	enfant	126
	un chien	1755	ici	167	une règle	488	mercredi	1168
	mais	30	un crayon	n/a	lundi	1091	mardi	1044
	mal	277	un jeu	291	aujourd'hui	233	aussi	44
	vendredi	1086	samedi	1356	dans	11	une fête	1490
	deux	41	la maison	325	tu	112	écouter	429
	la semaine	245	un	3	au revoir	1274	gomme	n/a
		9418		11485		12078		8557
		ave 856.18		ave 1,044.09		ave 1098		ave 855.7
0-1000	6		6		6		6	
1001-2000	3		3		3		3	
>2000	3		3		3		3	

Note. In all following vocabulary lists, the following heading terms are used:

Listening = vocabulary recognition in the oral modality

Speaking = vocabulary recall in the oral modality

Reading = vocabulary recognition in the written modality

Writing = vocabulary recognition in the written modality

Appendix E.5

Word Length in Letters According to Test Type

LISTENING	length		SPEAKING	length		READING	length		WRITING	length
regarder	8		parler	6		midi	4		oui	3
cheval	6		orange	6		un stylo	5		chat	4
salut	5		univers	7		non	3		bien	4
je	2		bonjour	7		jour	4		neuf	4
jeudi	5		dimanche	8		un gâteau	6		banane	6
moto	4		et	2		manger	6		enfant	6
chien	5		ici	3		une règle	5		mercredi	8
mais	4		un crayon	6		lundi	5		mardi	5
mal	3		un jeu	3		aujourd'hui	10		aussi	5
vendredi	8		samedi	6		dans	4		une fête	4
deux	4		la maison	6		tu	2		écouter	7
la semaine	7		un	2		au revoir	8		gomme	5
Total:	61			62			62			61
Mean:	5.08			5.17			5.17			5.08

Note. In all following vocabulary lists, the following heading terms are used:

Listening = vocabulary recognition in the oral modality

Speaking = vocabulary recall in the oral modality

Reading = vocabulary recognition in the written modality

Writing = vocabulary recognition in the written modality

Appendix E.6

Recency by Lesson When Last Encountered and Number of Times Encountered for Vocabulary in Each Test Type

LISTENING	last encountered		#of times encountered	
	Novel	Standard	Novel	Standard
regarder	30	12	7	5
cheval	29	30	9	11
salut	20	6	8	3
je	30	30	14	13
jeudi	24	26	5	5
une moto	14	12	5	5
un chien	30	19	10	8
mais	24	24	4	3
mal	29	26	6	4
vendredi	24	30	5	5
deux	26	26	8	12
la semaine	24	22	2	2
Total:	304	263	83	76
Mean	25.33	21.92	6.92	6.33
READING	last encountered		#of times encountered	
	Novel	Standard	Novel	Standard
midi	16	12	5	5
un stylo	26	14	12	4
non	26	23	8	5
jour	24	27	6	3
gâteau	26	27	2	5
manger	30	25	3	2
une règle	29	12	8	5
lundi	24	26	9	6
aujourd'hui	29	27	9	6
dans	16	26	4	9
tu	30	30	15	18
au revoir	20	11	8	3
Total:	296	260	89	71
Mean	24.67	21.67	7.42	5.92
SPEAKING	last encountered		# of times encountere	
	Novel	Standard	Novel	Standard
parler	30	12	8	5
une orange	28	30	11	11
univers	19	12	6	5
bonjour	19	11	8	4
dimanche	24	26	8	4
et	23	27	3	8
ici	27	18	7	2
un crayon	24	14	4	5
un jeu	26	12	12	4
samedi	30	30	7	5
la maison	30	26	7	2
un	26	30	7	23
Total:	306	248	88	78
Mean	25.5	20.67	7.33	6.5
WRITING	last encountered		# of times encounter	
	Novel	Standard	Novel	Standard
oui	20	23	6	5
un chat	30	19	11	9
bien	21	20	5	2
neuf	26	23	2	8
banane	28	30	11	10
enfant	6	17	2	4
mercredi	24	26	5	4
mardi	27	30	8	5
aussi	20	16	5	3
une fête	30	14	4	3
écouter	30	12	7	5
une gomme	25	10	4	3
Total:	287	240	70	61
Mean	23.92	20	5.83	5.08

Note. In all following vocabulary lists, the following heading terms are used:

Listening = vocabulary recognition in the oral modality

Speaking = vocabulary recall in the oral modality

Reading = vocabulary recognition in the written modality

Writing = vocabulary recognition in the written modality

Appendix E.7

Distribution of Word Classes According to Test Type

	LISTENING			SPEAKING			READING			WRITING	
	regarder	v		parler	v		midi	n		oui	other
	cheval	n		orange	n		un stylo	n		chat	n
	salut	other		univers	n		non	other		bien	adv
	je	pron		bonjour	other		jour	n		neuf	other: number
	jeudi	n: day		dimanche	n: day		gâteau	n		banane	n
	une moto	n		et	conj		manger	v		enfant	n
	un chien	n		ici	adv		une règle	n		mercredi	n: day
	mais	conj		un crayon	n		lundi	n: day		mardi	n: day
	mal	adv		un jeu	n		aujourd'hui	adv		aussi	conj
	vendredi	n: day		samedi	n: day		dans	prep		une fête	n
	deux	other: number		la maison	n		tu	pron		écouter	v
	semaine	n		un	other: number		au revoir	other		gomme	n
Key:											
n	Noun	4			5			5			5
n: day	Noun: Day of the week	2			2			1			2
adv	Adverb	1			1			1			1
v	Verb	1			1			1			1
pron	Pronoun	1			0			1			0
other: nur	Other: number	1			1			0			1
other	Other	1			1			2			1
conj	Conjunction	1			1			0			1
prep	Preposition	0			0			1			0

Note. In all following vocabulary lists, the following heading terms are used:

Listening = vocabulary recognition in the oral modality

Speaking = vocabulary recall in the oral modality

Reading = vocabulary recognition in the written modality

Writing = vocabulary recognition in the written modality

Appendix E.8

Concreteness of Words According to Test Type

LISTENING		SPEAKING		READING		WRITING
regarder		parler		midi		oui
cheval		une orange		un stylo		un chat
salut		univers		non		bien
je		bonjour		jour		neuf
jeudi		dimanche		gâteau		banane
une moto		et		manger		enfant
un chien		ici		une règle		mercredi
mais		un crayon		lundi		mardi
mal		un jeu		aujourd'hui		aussi
vendredi		samedi		dans		une fête
deux		la maison		tu		écouter
la semaine		un		au revoir		gomme
6		6		7		6

Note. In all following vocabulary lists, the following heading terms are used:

Listening = vocabulary recognition in the oral modality

Speaking = vocabulary recall in the oral modality

Reading = vocabulary recognition in the written modality

Writing = vocabulary recognition in the written modality

Appendix E.9

Cognate Status of Words According to Test Type

LISTENING		SPEAKING		READING		WRITING
regarder		parler		midi		oui
cheval		une orange		un stylo		un chat
salut		univers		non		bien
je		bonjour		jour		neuf
jeudi		dimanche		gâteau		banane
une moto		et		manger		enfant
chien		ici		une règle		mercredi
mais		un crayon		lundi		mardi
mal		un jeu		aujourd'hui		aussi
vendredi		samedi		dans		une fête
deux		la maison		tu		écouter
la semaine		un		au revoir		gomme
1		2		2		2

NB: it was decided not to classify *fête* or *gâteau* as cognates or near-cognates as, although these may be decipherable by older learners, they were likely unfamiliar to younger pupils.

Note. In all following vocabulary lists, the following heading terms are used:

Listening = vocabulary recognition in the oral modality

Speaking = vocabulary recall in the oral modality

Reading = vocabulary recognition in the written modality

Writing = vocabulary recognition in the written modality

Appendix E.10

Differences Between English and French Symbol-Sound Correspondences According to Test Type

LISTENING			SPEAKING			READING			WRITING	
regarder	1		parler	1		midi	1		oui	2
cheval	1		une orange	1		un stylo	1		un chat	2
salut	1		univers	2		non	1		bien	1
je	1		bonjour	1		jour	1		neuf	1
jeudi	3		dimanche	3		gâteau	1		banane	1
une moto	0		et	1		manger	2		enfant	2
chien	2		ici	2		une règle	0		mercredi	2
mais	1		un crayon	0		lundi	1		mardi	1
mal	0		un jeu	2		aujourd'hui	4		aussi	2
vendredi	3		samedi	2		dans	2		une fête	0
deux	2		la maison	0		tu	1		écouter	2
la semaine	1		un	1		au revoir	2		gomme	1
	16			16			17			17
<i>M</i> =	1.33			1.33			1.42			1.42
<i>SD</i> =	0.985			0.888			0.996			0.667
0 - very easy										
1 - easy										
2 - medium										
3 - difficult										
4 - very difficult										

This adopted a ‘distances removed’ approach related to symbol-sound correspondences – i.e., how many differences are there between the English and the French in terms of how it would be pronounced? (so “regarder” = 1; “cheval” = 1).

Note. In all following vocabulary lists, the following heading terms are used:

Listening = vocabulary recognition in the oral modality

Speaking = vocabulary recall in the oral modality

Reading = vocabulary recognition in the written modality

Writing = vocabulary recognition in the written modality

Appendix E.11

Syllable Count of Vocabulary According to Test Type

LISTENING			SPEAKING			READING			WRITING	
regarder	3		parler	2		midi	2		oui	1
cheval	2		une orange	2		un stylo	2		chat	1
salut	2		univers	3		non	1		bien	2
je	1		bonjour	2		jour	1		neuf	1
jeudi	3		dimanche	2		gâteau	2		banane	2
une moto	2		et	1		manger	2		enfant	2
chien	2		ici	2		une règle	2		mercredi	2
mais	1		un crayon	2		lundi	2		mardi	3
mal	1		un jeu	1		aujourd'hui	3		aussi	2
vendredi	2		samedi	2		dans	1		une fête	1
deux	1		la maison	2		tu	1		écouter	3
semaine	2		un	1		au revoir	3		gomme	1
TOTAL:	22			22			22			21
<i>M</i> =	1.83			1.83			1.83			1.75

Note. In all following vocabulary lists, the following heading terms are used:

Listening = vocabulary recognition in the oral modality

Speaking = vocabulary recall in the oral modality

Reading = vocabulary recognition in the written modality

Writing = vocabulary recognition in the written modality

Appendix E.12

Sample Vocabulary Listening Test (Recognition, Oral Modality)

Part A: Listening - What do these words mean?

You will hear 12 French words. **Choose** the English word or words that best match what you hear.

You will hear each word twice. Choose one correct answer only.

1/12

Listen.

Choose **the English word or words** that best match what you hear.

Your teacher can read aloud the words below.

- ☐ today
- ☐ Friday
- ☐ week
- ☐ Wednesday

2/12

Listen.

Choose **the English word or words** that best match what you hear.

Your teacher can read aloud the words below.

- ☐ yes
- ☐ hi / hello
- ☐ good / well
- ☐ and

Appendix E.13

Sample Vocabulary Reading Test (Recognition, Written Modality)

Reading11 1/12

Read the French word.

Choose **the English word** that best matches what you read.

tu

Your teacher can read aloud the words below.

- ☐ two (1)
- ☐ I (2)
- ☐ also (3)
- ☐ you (4)

Reading4 2/12

Read the French word.

Choose **the English word** that best matches what you read.

jour

Your teacher can read aloud the words below.

- ☐ Monday (1)
- ☐ Sunday (2)
- ☐ week (3)
- ☐ day (4)

Appendix E.14

Sample Vocabulary Writing Test (Recall, Written Modality)

Part C: Writing:

Read the 12 English words and numbers below and **write them in French**. Don't worry if you can't remember how to spell them, just do your best.

Please include accents if you know them. If you want to use a letter with an accent, you can click on the button you need in the bar at the top.

If you want to hear the English word read aloud, you can ask your teacher.

Writing4 1/12

Type this word **in French**:

cat

Writing11 2/12

Type this word **in French**:

rubber

Writing9 3/12

Type this word **in French**:

also

Appendix E.15

Sample Vocabulary Speaking Test (Recall, Oral Modality)

1 Quiz: Speaking	2 Part 1: Vocab Can you say the words in French for these numbers or English words?	3 to speak
4 orange	5 universe	6 hello / good morning

Appendix E.16

Teacher Test Instructions

For all quizzes, pupils must work independently without discussion. Teachers can support with understanding of what to do but not with the French.

QUIZ 1

Vocabulary:

*The Teacher version is the only one with sound. Sound has been removed from pupil version so they cannot click and listen at random. **Please ensure you have uploaded the PUPIL version of the quiz for pupils to work from.** Only use the teacher version yourselves.*

You can answer the questions yourself if you want a practice run, because anything you answer will be recorded on the teacher version not the pupil quiz anyway.

There are three parts to this quiz: listening, reading and writing.

Pupils do not need to answer if they do not know. Advise them to do their best and try to give an answer if they can – if they really can't, they can leave a blank. The only information they must give is at the start: school / group letter / name.

TELL PUPILS NOT TO CLICK ON THE ARROW IN THE BOTTOM RIGHT OF THE SCREEN UNTIL THEY ARE SURE THEY HAVE ANSWERED THE QUESTION AS THEY WISH TO ANSWER IT. They cannot go back once they have clicked.

Instructions for Part A: Listening

There are 12 items. Pupils listen to a recorded French word and choose the correct meaning from the multiple-choice options.

The teacher should read the instructions to pupils, and also the four English options.

You should only click to play the recording once – the word is recorded twice with a short pause between the two recordings, so pupils will hear it twice. *If for any reason you have to re-play the recording because of a loud noise for example, please make a note of this and let me know. **Do not** repeat it just because pupils ask to hear it again. **Please do not say it yourself.***

Instructions for Part B: Reading

There are 12 items. Pupils read the French word and choose the correct meaning from the multiple-choice options.

The teacher should read the instructions to pupils, and also the four English options. (The instructions are only recorded for the first item).

Please do NOT read the French words aloud, as it is a reading quiz.

Instructions for Part C: Writing

There are 12 items. Pupils read the English word and write it in French.

Teachers can read aloud the English instructions and words as needed. If pupils wish to use accents, these are on the accent bar at the top of the slide. There are two accents: é and ê if needed.

QUIZ 2

Phonics and grammar:

*The Teacher version is the only one with sound. Sound has been removed from pupil version so they cannot click and listen at random. **Please ensure you have uploaded the PUPIL version of the quiz for pupils to work from.** Only use the teacher version yourselves.*

You can answer the questions yourself if you want a practice run, because anything you answer will be recorded on the teacher version not the pupil quiz anyway.

There are two parts to this quiz: phonics and grammar.

Instructions for Part D: phonics:

Pupils will first look at a written word with gap(s). They then listen to a recorded word and decide which letter / letters are missing.

You should only click to play the recording of the word ONCE** – the word is recorded twice with a short pause between the two recordings, so pupils will hear it twice. If for any reason you have to re-play the recording because of a loud noise for example, please make a note of this and let me know. **Do not** repeat it just because pupils ask to hear it again. **Please do not say it yourself.

There is one letter with an accent: é. If pupils think they need to use this, it is on a bar at the top of the slide and they can click on it when they're ready to use it.

There are two examples, you can use these to model what the pupils have to do (i.e., type in 'ai' for example 1, and type in 'a' for example 2).

From number 1 onwards, the instructions are not recorded, just to save time. You can read them to the pupils if you wish.

Pupils need to click on the arrow in the bottom right of the slide to move to the next question. They cannot go back.

*Pupils do not need to answer if they do not know. Advise them to do their best and try to give an answer if they can – if they really can't, they can leave a blank. The only information they **must** give is at the start: school / group letter / name.*

Instructions for Part E: Grammar

This should be straightforward. It is multiple choice. The instructions are only recorded at the start, after that pupils can work through at their own pace.

Please do not read aloud the French sentences, it should be a reading activity. You can support pupils in the following way as an example:

So what word is missing here? Can you see the blank? Do you think it's the first option, or the second one or the last one? Click the one that you think would fill in the gap.

Appendix E.17

Language-analytic Aptitude Scores of Pupil Subsample for Oral Production Tests: Post-Tests

	<i>School</i>	<i>A</i>	<i>School</i>	<i>B</i>	<i>School</i>	<i>C</i>	<i>School</i>	<i>E</i>	<i>School</i>	<i>F</i>
	Novel	Standard	Novel	Standard	Novel	Standard	Novel	Standard	Novel	Standard
<i>Pupil Pair 1</i>	36	35	38	39	39	30	41	40	34	27
<i>Pupil Pair 2</i>	31	32	34	33	28	25	31	31	13	14
<i>Pupil Pair 3</i>	22	25	26	25	17	24	30	25	10	18
<i>Pupil Pair 4</i>	24	21	20	20	16	19	20	24	8	9
<i>Pupil Pair 5</i>	17	16	11	11	13	17	19	19	7	5
<i>Pupil Pair 6</i>	10	11	5	6	11	12	8	8		
<i>Total LAA scores</i>	140	140	134	134	124	127	149	147	72	73

Overall total (Novel approach): 619 ($M = 10.67$)

Overall total (Standard approach): 621 ($M = 10.71$)

Language-analytic Aptitude Scores of Pupil Subsample for Oral Production Tests: Delayed Post-Tests

	<i>School</i>	<i>A</i>	<i>School</i>	<i>B</i>	<i>School</i>	<i>C</i>	<i>School</i>	<i>E</i>	<i>School</i>	<i>F</i>
	Novel	Standard	Novel	Standard	Novel	Standard	Novel	Standard	Novel	Standard
<i>Pupil Pair 1</i>	36	35	38	39	39	30	41	40	34	27
<i>Pupil Pair 2</i>	31	32	34	33	28	25	31	31	13	14
<i>Pupil Pair 3</i>	22	25	26	25	17	24	30	25	10	18
<i>Pupil Pair 4</i>	24	21	20	20	16	19	20	24	8	9
<i>Pupil Pair 5</i>	17	16	11	11	13	17	19	19	7	5
<i>Pupil Pair 6</i>	10	11	7	7	11	12	8	8		
<i>Total LAA scores</i>	140	140	136	135	124	127	149	147	72	73

Overall total (Novel approach): 621 ($M = 10.71$)

Overall total (Standard approach): 622 ($M = 10.72$)

Appendix E.18

Host words for Symbol-Sound Correspondences in the Phonics Transcription Tests and Position in Post-tests and Delayed Post-tests

Grapheme	Host Word	Post-test Position	Delayed Post-test Position
[on]	don	1	4
[ou]	clou	2	7
[ch]	chauve	3	5
[er/et/ez/é (1)]	viré	4	9
[i]	fiche	5	2
[au]	tableau	6	8
[SFC]	jettes/t/d/x	7	6
[eu]	leu	8	7
[er/et/ez/é (2)]	briser	9	1
[u]	valu	10	3

Appendix E.19

Host words for Symbol-Sound Correspondences in the Phonics Read Aloud Tests and Position in Post-tests and Delayed Post-tests

Grapheme	Host Word	Post-test Position	Delayed Post-test Position
[on]	fond	3	4
[ou]	flou	1	5
[ch]	char	4	3
[er/et/ez/é (1)]	tiré	5	8
[i]	mixte	6	2
[au]	couteau	7	9
[SFC]	pattes	8	1
[eu]	ceux	2	6
[er/et/ez/é (2)]	viser	9	10
[u]	fallu	10	7

Appendix E.20

Piloting the Phonics Tests

(i) Introduction

Similar issues were encountered in sourcing a class of pupils to pilot the phonics test pilot as in vocabulary, as many schools teach from schemes where little focus is given to SSCs. Two pilots were therefore conducted for RA and three for transcription. Both tests were first piloted in July 2022 with Year 3 pupils, who had followed a scheme of work which covered similar content to the Standard approach. The pupils had not done explicit phonics-focused work, although the teacher reported that they had drawn pupils' attention to SSCs as they occurred in the scheme. At this point three versions of the same transcription test and RA test were piloted (following the counterbalanced model explained in [E.2](#)).

Transcription

Pupils had to listen to a recorded French word, played twice, and write the missing grapheme (including some digraphs), with the number of missing letters indicated. It was accepted that, as with the vocabulary, the goal in this pilot was to check format and practicalities of administration rather than test content.

Read Aloud

Pupils were asked to read aloud an unknown French word embedded in a PowerPoint presentation.

(ii) Findings from the pilot phonics tests

Transcription

Similar technical issues were encountered as in the vocabulary tests, such as delays caused by logging in to devices, and practical issues relating to individual headphone use. Year 3 pupils found the transcription test (all versions consisted of nine items) very challenging: in the first of the test versions, completed by 11 pupils, only seven points were scored (out of 99); in the second test version, allocated to nine pupils, only two points were scored (out of 81); in the third test version, completed by seven pupils, only four points were scored (out of 63 possible). This floor effect made it difficult to determine whether difficulties were due to test format or content; it was also impossible to know whether pupils who had received more sustained explicit instruction would cope any better with the transcription test.

Read Aloud

Individual RA tests were initially piloted with only three pupils, partly due to time constraints and also because it was evident that a further pilot would be needed for transcription. Scores ranged from 6-12/18 on the three test versions (each consisting of 9 items and scored on a 0-1-2 basis, where 1 represented reasonable communication of phoneme without complete accuracy). Pupils had been selected by the class teacher according to their academic achievement in English (rather than French), making it difficult to determine whether the test versions differed significantly in terms of challenge or if the disparity in results was due to individual aptitude.

(iii) Repiloting the phonics tests

Transcription

Contact was made with a school in the north of England through the Novel scheme's creator, where a non-specialist French teacher had been teaching Year 3 pupils since September 2022 following the Novel scheme of work ([E.2](#)). This teacher agreed to pilot materials online in February 2023, with adjustments to ensure that pupils were not being tested on material not yet covered. The difficulties encountered with online log-in confirmed that it was the right decision to provide a single test to be completed by all pupils, rather than administering three versions of the same test simultaneously. Pupils still struggled with transcription, with scores ranging from 0-3/10, even after receiving explicit instruction on phonics in the Novel scheme. As it was possible that it was the nature of the form recall which was causing problems, a further pilot was organised which compared pupils' results on form recall with a multiple-choice recognition option.

The final pilot took place at the start of June 2023 in a local school which did not follow the Novel scheme, but where the teacher was a French specialist and the allocation for French was 60 minutes each week. A split-class pilot was conducted, whereby half the class completed the form recall transcription pilot (writing the graphemes themselves) and half completed a multiple-choice recognition pilot (consisting of four options, plus "I don't know"). This was administered online as a whole class and included a confidence rating. The French teacher gave a general indication of pupils' French ability, and they were then randomly allocated by the researcher to recall or recognition. Eleven pupils completed form recall, and 14 completed the multiple-choice option. On form recall, scores ranged from 0-6 out of 10 ($M = 3.00$), a better range than in the previous pilot. On recognition, scores also ranged from 0-6 out of 10 ($M = 3.79$). This seemed to suggest a very slight advantage on multiple-choice recognition, although it was not possible to know how accurately pupils had been grouped according to existing ability. The argument in favour of multiple-choice options was that pupils may have felt that they were at least able to respond, but it introduced the possibility of just guessing. Additionally, the

multiple-choice format sometimes appeared to be a disadvantage. For example, the [u] in “*valu*” was most often correctly identified in the February pilot, and in this pilot, where pupils had to identify it for themselves, seven of 11 did so correctly. In the multiple-choice option, only two of 14 identified it correctly, appearing to be distracted by the [ou] option.

Read Aloud

This was repiloted online in February 2023 with six pupils of varying abilities in French from the school which had been following the Novel scheme. Pupils were tested on 10 unknown French words, with scoring on a 0-1-2 basis. Scores ranged from 4-15/20.

Appendix E.21

Sample Phonics Transcription Test

phonics9

1/10

Read the word.

b r i s _ _

Listen to the word.

Type the two missing letters in the box below.

Appendix E.22

Sample Phonics Read Aloud Test

Part 2: Now you are going to see some new words in French – can you read these words out loud in French?

15

flou

16

ceux

17

fond

18

Appendix E.23

Items for Written Grammar Recognition and Position in Post-tests and Delayed Post-tests

Item with Multiple Choice Options	Post-test Position	Delayed Post-test Position
gr1:1 Tu _____ un chat ? as / ai / a	2	9
gr1:2 _____ est amusant. (amusant = funny) Je / Tu / Le chien	5	10
gr1:3 _____ ai une gomme. J' / Tu / Il	10	15
gr1:4 _____ as un stylo. Marie / Tu / J'	12	4
gr1:5 Le lion _____ triste. (triste = sad) suis / est / es	1	5
gr1:6 J' _____ une règle. a / as / ai	6	3
gr2:1 Tu _____ le cheval. écoute / écoutes / écouter	16	18
gr2:2 Elle _____ une orange. manges / manger / mange	7	11
gr2:3 _____ parles français. Tu / Je / Elle	9	13
gr2:4 Tu _____ le football? regardes / regarder / regarde	4	1
gr2:5 _____ regarde la moto. Tu / Nous / Je	11	17
gr2:6 _____ écoutes bien. J' / Tu / Elle	14	7
gr3:1 Le stylo est _____. brun / brune / bruns	15	8
gr3:2 La maison est _____. petit (small) / petite (small) / petits (small)	18	12
gr3:3 _____ est bleu. Les gommages / Le crayon / La maison	17	2

gr3:4 La banane est _____.	8	14
verts / vert / verte		
gr3:5 _____ est anglais (anglais = English)	13	16
Tom / Sarah / Sarah et Louise		
gr3:6 _____ est orange.	3	6
Le chat / La règle / It could be either 'Le chat' OR 'La règle'		

Note. The correct answers are shown in bold.

Appendix E.24

Items for Spoken Grammar Recall and Position in Post-tests and Delayed Post-tests

Item and translation	Post-test Position	Delayed Post-test Position
gr1:1 I <u>have</u> a dog.		
J' _____ un chien.	1	5
gr1:2 The pen <u>is</u> orange.		
Le stylo _____ orange.	5	1
gr1:3 You <u>have</u> a horse?		
Tu _____ un cheval?	6	2
gr1:4 The pencil <u>is</u> blue.		
Le crayon _____ bleu.	10	6
gr2:1 <u>You</u> listen to the child.		
_____ écoutes l'enfant.	2	7
gr2:2 <u>I</u> prefer cats.		
_____ préfère les chats.	3	8
gr2:3 <u>He/She</u> speaks English.		
_____ parle anglais.	7	3
gr2:4 <u>I</u> have a house.		
_____ ai une maison.	8	13
gr2:5 <u>You</u> have a motorbike?		
_____ as une moto?	12	12
gr3:1 She <u>eats</u> a cake.		
Elle _____ un gâteau.	4	4
gr3:2 I <u>eat</u> a banana.		
Je _____ une banane.	9	11
gr3:3 He <u>looks at</u> the pen.		
Il _____ le stylo.	11	14
gr5:1 ___?___ is big.		
_____?_____ est grand.	13	10
gr5:2 ___?___ is brown.		
_____?_____ est brune.	14	9

Appendix E.25

Items for Elicited Imitation Grammar and Elicited Imitation Speaking Tests and Position in Post-tests and Delayed Post-tests

Item	Translation	Post-test Position	Delayed Post-test Position
<i>J'ai un chien brun. Et toi?</i>	I have a brown dog. And you?	1	7
<i>Le cheval regarde le chat.</i>	The horse looks at the cat.	2	4
<i>Le chien mange une orange ?</i>	Is the dog eating an orange?	3	1
<i>Le stylo est bleu et brun.</i>	The pen is blue and brown.	4	11
<i>La banane est petite et jaune.</i>	The banana is small and yellow.	5	3
<i>Tu as deux chiens? J'ai un chat.</i>	You have two dogs? I have a cat.	6	13
<i>Le chat écoute le cheval.</i>	The cat is listening to the horse.	7	15
<i>Tu as un animal? Moi, non.</i>	You have a pet? I don't.	8	8
<i>Elle a une règle. J'ai une gomme.</i>	She has a ruler. I have a rubber.	9	2
<i>C'est vendredi? Non, jeudi.</i>	Is it Friday? No, Thursday.	10	5
<i>Elle parle français. C'est super.</i>	She speaks French. It's great.	11	6
<i>C'est mercredi aujourd'hui.</i>	It's Wednesday today.	12	10
<i>Elle mange un gâteau super.</i>	She is eating a super cake.	13	14
<i>Il regarde la gomme dans le sac.</i>	He is looking at the rubber in the bag.	14	12
<i>Je mange une banane verte et jaune.</i>	I'm eating a green and yellow banana.	15	9

Note. Targeted items appear in bold text.

Appendix E.26

Piloting the Grammar Tests

As with vocabulary and phonics, a counterbalanced design was initially adopted, with three different versions of each grammar test.

(i) Introduction

Written grammar recognition

A class of 28 Year 4 pupils participated in the pilot in September 2022 (as for vocabulary – [E.2](#)), having followed a standard scheme of French since the previous September in which grammar was given relatively little explicit emphasis. Pupils were asked to read short, written sentences in French, each with one or two words missing (indicated by a single blank). They then chose the correct answer from three options, all semantically correct. The gaps could only be completed by pupils demonstrating their understanding of the focus grammatical structure, which in the first round of piloting included “*avoir*”/“*être*” verbs (‘to be’/‘to have’), regular ‘-er’ verb endings, singular and plural nouns, gender and adjectival agreement.

Spoken grammar recall

The oral modality test was chosen to avoid too heavy a literacy focus for these young learners. Pupils were individually shown a series of short English sentences embedded in a PowerPoint presentation, with the incomplete French translation below. The word containing the target structure was represented by a blank ([E.28](#)). Pupils were asked to supply the French word orally.

Elicited Imitation Grammar (EIG)

The pilot test consisted of three counterbalanced versions of the EIG. Before starting the test, the researcher gave the following instructions: “*You are going to hear some sentences. Try to say what you hear. The pictures will help you understand what’s being said. What you say should have the same meaning as what you hear. Don’t worry if you can’t remember everything you hear but try to say as much as you can and remember to speak clearly. Speak when the pictures disappear*”. Pupils listened to the 15 utterances once through headphones (unless it was noisy in the room, in which case the utterance was repeated). Utterances were pre-recorded into a PowerPoint presentation to ensure that all learners had the same experience, and three examples were included. Courtney et al. (2017) and Graham et al. (2017) used 18-19 items with slightly older learners, but it was difficult to create sentences when working with a limited set of vocabulary and core grammatical structures which had to feature in both schemes – this also

made it difficult to control for imageability as much as might have been ideal. Length of sentences and range of structures used were controlled for. Utterances consisted of seven-ten syllables, as Graham et al. (2017) found floor/ceiling effects below/above this. A short delay of five seconds was included before pupils were asked to attempt to repeat the utterance. Oral praise was given regardless of success of utterance.

(ii) Findings from the pilot grammar tests

Written grammar recognition

Administering three test versions to the whole class again proved problematical and time-consuming. Scores were very low, with the cause of this floor effect again unclear. Several reasons were considered: these pupils had experienced little sustained French grammar instruction in Year 3 and were being tested after the summer break; progress in such young learners is likely to be slow; and these tests moved away from rote-learned phrases, requiring recognition of individual sentence components.

Spoken grammar recall

A similar floor effect was seen in the individual spoken recall tests, piloted with six pupils of varying abilities. Scores ranged from 0-8/24 ($M = 2.20$).

Elicited Imitation Grammar

Due to time constraints, this was piloted with just four Year 4 pupils in July 2022. Two versions of the test were trialled. Pupils scored 8-25/58 on one version, and 25/58 on the other. Although pupils responded positively to the treasure hunt context, they found it challenging to reproduce the sentences, especially after the delay. Graham et al. (2017) had also found that a delay made sentence repetition impossible with slightly older learners. It was necessary to make the use of headphones optional for a few individuals, as not all pupils agreed to wear them.

(iii) Repiloting the grammar tests

Written grammar recognition

The schools approached were reluctant to pilot a grammar test, even when following the Novel scheme, as they were concerned that pupils were only at the very beginning of their learning journey. Eventually a school was sourced in the southwest of England through a contact of the Novel scheme's creator. The class of 26 Year 3 pupils piloted the written grammar recognition test online in June 2023. This test consisted of a single version, with six items targeting each of "avoir"/"être", '-er' verb endings and adjectival agreement. Scores ranged from 4-15/18 ($M = 7.69$).

Spoken grammar recall

This was not repiloted.

Elicited Imitation Grammar

The EIG was repiloted online in March 2023 with the school in the north of England which had been teaching the Novel scheme. It was reduced to a single version of the test and was conducted with six pupils of varying abilities. The 15 utterances focused on “*avoir*”/“*être*”, ‘-er’ verb endings and adjectival agreement. The length of utterances was reduced to six-eight syllables and the number of examples from three to two. A delay was still included: when the recorded utterance stopped playing, the picture began to spin to act as a distractor; when the picture stopped spinning (after three seconds), pupils were able to speak.

This pilot was conducted online and due to technical difficulties, it was not possible to test all sentences with all pupils. Where pupils *were* tested on all items, scores ranged from 6-32/50.

Appendix E.27

Sample Written Grammar Recognition Test

Part E: Grammar

Read the sentences carefully. Each sentence has one or two words missing. **Click on the correct word or words** from the choices below.

gr2:4 Tu _____ le football?

☐ regardes

☐ regarder

☐ regarde

gr3:3 _____ est bleu.

☐ Les gommes

☐ Le crayon

☐ La maison

gr1:6 J' _____ une règle.

☐ a

☐ as

☐ ai

Appendix E.28

Sample Spoken Grammar Recall Test

Part 3: Grammar

Answer as many questions as you can. It doesn't matter if you don't know an answer. You may see some pictures in between the questions!

27

Section 1: 2/6

I have a dog.

J' _____ un chien.

28



29

00:03

Section 2: 3/6

You listen to the child.

_____ écoutes l'enfant.

30

Section 1: 3/6



31

00:03

Section 2: 4/6

I prefer cats.

_____ préfère les chats.

32

Section 1: 4/6



33

00:03

Section 2: 5/6

She eats a cake.

Elle _____ un gâteau.

34



35

00:03

Section 1: 5/6

The pen is orange.

Le stylo _____ orange.

36

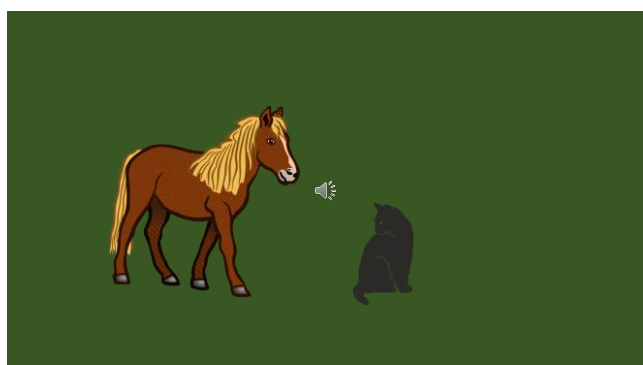
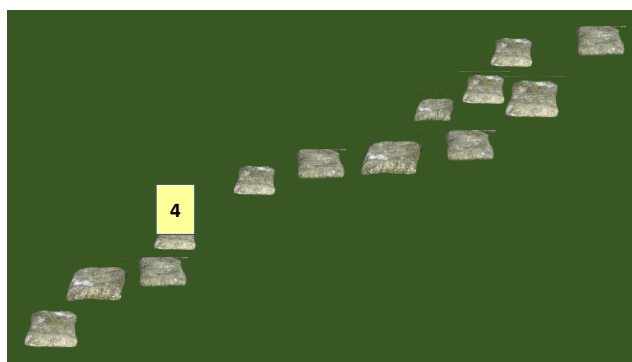
Appendix E.29

Sample Slides from Elicited Imitation Grammar Test

Part 5: EI
Can you cross the
stepping stones
to reach the
magic door?



Listen carefully
and repeat what
you hear. Don't
start speaking
until the picture
stops spinning!



Appendix E.30

Piloting the Speaking Tests

(i) Introduction

Pilots of the open-ended speaking test were affected by time constraints in both the July 2022 and March 2023 rounds, particularly as the latter was conducted online and impacted by technical issues.

Picture Response

The PR task was piloted with a small number of pupils who had just started Year 4 in September 2022 and included four slides embedded in a PowerPoint presentation which consisted of single images.

Elicited Imitation Speaking

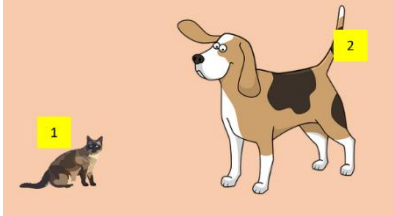
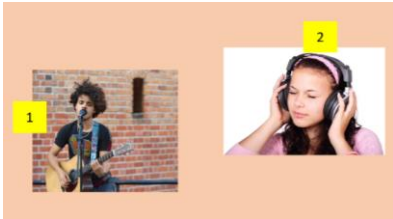
As the EIS test consisted of a rescoring of the same task as the EIG, this was not piloted separately (see [E.26](#)).

(ii) Findings from the Picture Response task

Pupils' responses to pictures were limited to individual lexical items. However, given the age of these young learners and the relatively short exposure to learning French, this was expected. The PR task was not repiloted.

Appendix E.31

Picture Response: Pictures and Position in Post-tests and Delayed Post-tests

Picture #	Post-test Position	Delayed Post-test Position
1	1	2
		
2	2	3
		
3	3	4
		
4	4	1
		

Introduction by the researcher:

Try to say something about each slide. You could say something or ask me something. Talk about both pictures if you can - you could even compare them.

Appendix E.32

Picture Response Scoring Guidelines

The following guidance was followed:

- *Pronouns, nouns*

For each correctly produced pronoun, verb or noun with determiner, 2 points were available (e.g., “*le chat*” for ‘the cat’ scored 1+1; “*la chat*” (incorrect gender) scored 0+1; “*le chaT*” (correct noun incorrectly pronounced scored 1+.50).

- *Verbs*

Correctly conjugated verbs scored 2 points; verbs (incorrectly) in the infinitive scored 1. Where pupils incorrectly translated the English gerund ‘he **is** singing’ using “*est*” for ‘is’, .50 was deducted (so, “*il est chanter*” scored 2+.50 - 2 for “*il*”, 1 for use of infinitive, minus .50 for use of “*est*”; “*il est chante*” scored 2+1.5).

- *Adjectives*

A correctly modified adjective scored 1 point (incorrectly modified, but appropriate meaning scored .50), with an additional point available for correct position where appropriate (e.g., “*bleue madeleine*” for ‘blue cake’ scored 1+1: correct adjective, correct noun, but no determiner; “*madeleine bleue*” scored 1+1, +1 additional point for the correct order). For pronunciation errors, .50 was deducted even if two were made (so, “*grand*” for the masculine singular scored 1; “*grande*” and “*grANde*” both scored .50). As Picture 4 (at post-test) featured the colour blue for multiple items, “*bleu(e)*” was credited only once. However, “*est*” (‘is’) was credited twice as appropriate, as it demonstrated ability to use a noun and phrase accurately.

- *Conjunctions*

For conjunctions, 1 point was available – the only one actually used was “*et*” (‘and’). If used twice within the same clause, it was only credited once (and was reduced to .50 if the [t] was sounded incorrectly).

Appendix E.33

Scoring Guidelines for Elicited Imitation Speaking Task

Adapted from: Tracy-Ventura et al. (2014). “Repeat as much as you can”: Elicited imitation as a measure of oral proficiency in L2 French. In P. Leclercq, A. Edmonds & H. Hilton (Eds.), *Measuring L2 proficiency: Perspectives from SLA*. Clevedon, UK: Multilingual Matters.

SCORING GUIDELINES FOR FRENCH ELICITED IMITATION TASK

SCORE 0

Criteria	Examples
No attempt	
Whole utterance is garbled (unintelligible)	
- Minimal repetition, then item abandoned: - Under half of the idea units are conveyed clearly. - May contain extraneous words that were not in the original stimulus	Le flançais une super [ang]. (<i>target form: Elle parle français. C’est super!</i>).

SCORE 1

Criteria	Examples
When only about half of idea units are represented in the string but a lot of important information in the original stimulus is left out	Le mange et banane et jaune et viert. (<i>target form : Je mange une banane verte et jaune.</i>) Le chat une quitte le cheval. (<i>target form : Le chat écoute le cheval.</i>)
When barely half of lexical words get repeated and meaningful content is unrelated (or opposed) to stimulus, frequently with hesitation markers	
When the component parts of the sentence are conveyed but in the	Le cheval écoute le chatte. (<i>target form : Le chat écoute le cheval</i>)

incorrect order which affects communication of meaning	
• Or when string does not in itself constitute a self-standing sentence with some (targetlike or non-targetlike) meaning. (This may happen more often with shorter items, where, if only 2 of 3 content words are repeated and no grammatical relation between them is attempted, then score 1) • Also when half of a long stimulus is left out, and the sentence produced is incomplete	Le stylo une bleu and blown (rhyming with brown) (<i>target form: le stylo est bleu et brun</i>). This conveys two of the four idea units but does not constitute a self-standing sentence with some meaning – no grammatical relations between the content words)

SCORE 2

Criteria	Examples
• When content of string preserves at least more than half of the idea units in the original stimulus; string is meaningful, and the meaning is close or related to original, but it departs from it in some slight changes in content, which makes content inexact, incomplete, or ambiguous • Cases of extra marking or more marked morphology should be considered as meaning change. For example, a present tense repeated as past or as future should be scored as meaning change (score 2). • Similarly, singular/plural differences between stimulus and repeated string change the meaning, not only the grammar (score 2). Omission of an article may lead to confusion between singular and plural forms. E.g., 'J'ai un chat' in the original stimulus could be conveyed as 'J'ai une chat' or 'j'ai le chat' and this would not be an issue. However, if a pupil said 'j'ai chat' it would be unclear whether the 'chat' was singular or plural, and would introduce ambiguity. (An exception to this is 'animal' which is clearly singular without the article).	Le cheval regarde un chien (<i>target form : Le cheval regarde un chat</i>). Le stylo est bleu et brown. (<i>le stylo est bleu et brun</i>). C'est mercredi (<i>target form : c'est mercredi aujourd'hui</i>) or c'est mercredi unjundhui (where aujourd'hui is unclear)

However, where the utterance includes a singular form of the verb which clarifies the number of items, then this is acceptable. E.g., les stylo bleu could not be accepted as a singular noun, whereas les stylo est bleu could be. Changes of person (il for elle) change the meaning, so score 2	

SCORE 3

Criteria	Examples
- Original, complete meaning is preserved as in the stimulus. Pronunciation needs to be comprehensible but not necessarily fully authentic. Strings which are quite ungrammatical can get a 3 score, as long as exact meaning is preserved. Synonymous substitutions are acceptable. - Changes in grammar or sentence structure that don't affect meaning should be scored as 3. - Ambiguous changes in grammar that COULD be interpreted as meaning changes from a NS perspective should be scored as 2. That is, as a general principle in case of doubt about whether meaning has changed or not, score 2.	Le cheval le regarde le chatte (<i>target form: le cheval regarde le chat</i>)

SCORE 4

Criteria	Examples
Exact reproduction of meaning. String matches stimulus such that both form	

and meaning are completely comprehensible • Order of adjectives can be different without meaning being affected • Missing liaison acceptable • Omission of c' in 'c'est' is tolerable.	E.g.: Je mange une banane jaune et verte instead of <i>Je mange une banane verte et jaune.</i> Elle parle en français. Est super. (<i>target form : Elle parle français. C'est super.</i>)

Appendix E.34

Vocabulary Items for Vocabulary Recognition in the Oral Modality (Listening) Tests and Position in Post-tests and Delayed Post-tests

Vocabulary Item	Translation	Post-test Position	Delayed Post-test Position
<i>regarder</i>	to look (at)	1	3
<i>cheval</i>	horse	2	9
<i>salut</i>	hi, bye	3	2
<i>je</i>	I	4	7
<i>deux</i>	two	5	6
<i>chien</i>	dog	6	5
<i>jeudi</i>	Thursday	7	12
<i>mais</i>	but	8	10
<i>semaine</i>	week	9	11
<i>mal</i>	bad(ly)	10	8
<i>moto</i>	motorbike	11	4
<i>vendredi</i>	Friday	12	1

Appendix E.35

Vocabulary Items for Vocabulary Recognition in the Written Modality Tests and Position in Post-tests and Delayed Post-tests

Vocabulary Item	Translation	Post-test Position	Delayed Post-test Position
<i>midi</i>	midday	1	3
<i>stylo</i>	pen	2	6
<i>non</i>	no	3	8
<i>jour</i>	day	4	2
<i>gâteau</i>	cake	5	9
<i>manger</i>	to eat	6	7
<i>règle</i>	ruler	7	10
<i>lundi</i>	Monday	8	4
<i>aujourd'hui</i>	today	9	12
<i>dans</i>	in	10	11
<i>tu</i>	you	11	1
<i>au revoir</i>	goodbye	12	5

Appendix E.36

Vocabulary Items for Vocabulary Recall in the Oral Modality Tests and Position in Post-tests and Delayed Post-tests

Vocabulary Item	Translation	Post-test Position	Delayed Post-test Position
to speak	<i>parler</i>	1	6
orange	<i>orange</i>	2	8
universe	<i>univers</i>	3	3
hello	<i>bonjour</i>	4	1
Sunday	<i>dimanche</i>	5	12
and	<i>et</i>	6	10
here	<i>ici</i>	7	7
pencil	<i>crayon</i>	8	2
game	<i>jeu</i>	9	5
Saturday	<i>samedi</i>	10	4
house	<i>maison</i>	11	9
one	<i>un</i>	12	11

Appendix E.37

Vocabulary Items for Vocabulary Recall Tests in the Written Modality and Position in Post-tests and Delayed Post-tests

Vocabulary Item	Translation	Post-test Position	Delayed Post-test Position
nine	<i>neuf</i>	1	5
yes	<i>oui</i>	2	9
Wednesday	<i>mercredi</i>	3	8
cat	<i>chat</i>	4	1
good/well	<i>bien</i>	5	10
banana	<i>banane</i>	6	12
child	<i>enfant</i>	7	11
Tuesday	<i>mardi</i>	8	4
also	<i>aussi</i>	9	3
to listen	<i>écouter</i>	10	6
rubber	<i>gomme</i>	11	2
party/celebration	<i>fête</i>	12	7

Appendix F.1

Teachers' Post-lesson Evaluation

What is your response to this lesson? *Select **one answer** for each statement.*

	Very strongly agree	Strongly agree	Agree	Disagree	Strongly disagree	Very strongly disagree
1. I enjoyed this lesson	☐	☐	☐	☐	☐	☐
2. I was tense during this lesson	☐	☐	☐	☐	☐	☐
3. Students understood the material during this lesson.	☐	☐	☐	☐	☐	☐
4. Students were motivated in this lesson.	☐	☐	☐	☐	☐	☐
5. Students were focused in this lesson.	☐	☐	☐	☐	☐	☐
6. I was satisfied with what the students knew by the end of the lesson.	☐	☐	☐	☐	☐	☐
7. I was satisfied that the French I taught was useful .	☐	☐	☐	☐	☐	☐
8. The activities gave me sufficient evidence about the students' actual learning .	☐	☐	☐	☐	☐	☐

Appendix F.2

Piloting Teachers' Post-lesson Evaluations

The post-lesson evaluation was piloted with three primary school teachers of French, who completed it over four weeks (13 returns). No issues arose and no changes were made to content or format.

Appendix F.3

Primary French Teacher Questionnaire I Final

Q3 Your age (*please select **one category***):

☐ 20-30 (1)

☐ 31-40 (2)

☐ 41-50 (3)

☐ 51-60 (4)

☐ 61-70 (5)

Q4 How long have you been teaching in primary schools? *Please select **one category**.*

☐ Less than one year (1)

☐ More than one and less than five years (2)

☐ More than five and less than ten years (3)

☐ More than ten and less than twenty years (4)

☐ More than twenty years (5)

Q5 Do you currently teach phonics in English? *Please select **one answer**.*

☐ Yes (1)

☐ No (2)

Q6 Have you taught phonics in English before? *Please select **one answer**.*

☐ Yes (1)

☐ No (2)

Q7 How confident do you feel teaching phonics to pupils in English? **Drag the slider** to indicate your response: 1 being not at all confident, 6 being very confident.

Not at all confident

Very confident

1

2

3

4

5

6

Level of confidence ()



Q8 Please consider your **general** teaching experience. To what extent do you agree with the following statements? *Please select **one answer** for each statement.*

	Very strongly agree (1)	Strongly agree (2)	Agree (3)	Disagree (4)	Strongly disagree (5)	Very strongly disagree (6)	N/A (7)
I generally enjoy teaching. (1)	☐	☐	☐	☐	☐	☐	☐
Teaching is so enjoyable that I like preparing the lessons for the class. (2)	☐	☐	☐	☐	☐	☐	☐
I often have reasons to be happy when teaching. (3)	☐	☐	☐	☐	☐	☐	☐
I generally teach with enthusiasm. (4)	☐	☐	☐	☐	☐	☐	☐
I generally feel tense when teaching. (5)	☐	☐	☐	☐	☐	☐	☐
I often worry that my teaching isn't going well. (6)	☐	☐	☐	☐	☐	☐	☐
Preparing to teach often causes me to worry. (7)	☐	☐	☐	☐	☐	☐	☐
I feel uneasy when I think about teaching. (8)	☐	☐	☐	☐	☐	☐	☐

Q9 Now please focus on French. What is your own level of linguistic competence in French?

*Please select **one** answer.*

- ☐ Zero / almost zero (1)
- ☐ Pre-GCSE / very basic (2)
- ☐ GCSE / O-level (basic, low-intermediate) (3)
- ☐ A-level of equivalent (intermediate) (4)

- ☐ Degree (low-advanced / advanced) (5)
- ☐ Native (or native-like) speaker (6)

Q10 Have you taught French before? *Please select **one answer**.*

- ☐ Yes (1)
- ☐ No (2)

Q11 For how many years have you taught French? *Please give a rough total, even if the experience has been disjointed over the years.*

Q12 Now think about how you feel about **the teaching of French**. To what extent do you agree with the following statements? *Please select **one answer** for each statement.*

	Very strongly agree (1)	Strongly agree (2)	Agree (3)	Disagree (4)	Strongly disagree (5)	Very strongly agree (6)	N/A (7)
I generally enjoy teaching French. (1)	☐	☐	☐	☐	☐	☐	☐
I look forward to teaching French. (2)	☐	☐	☐	☐	☐	☐	☐
Teaching French is so enjoyable that I like preparing the lessons for the class. (3)	☐	☐	☐	☐	☐	☐	☐
I often have reasons to be happy when teaching French. (4)	☐	☐	☐	☐	☐	☐	☐
I generally teach French with enthusiasm. (5)	☐	☐	☐	☐	☐	☐	☐
I generally feel tense when teaching French. (6)	☐	☐	☐	☐	☐	☐	☐
I often worry that my French teaching isn't going well. (7)	☐	☐	☐	☐	☐	☐	☐
Preparing to teach French often gives me cause to worry. (8)	☐	☐	☐	☐	☐	☐	☐
I feel uneasy when I think about teaching French. (9)	☐	☐	☐	☐	☐	☐	☐

Appendix F.4

Teacher Questionnaire 2 & 3

☐ Q3

💡 ★ ...

In this section please think about your experiences teaching **the more standard (topic-driven) approach**. To what extent do you agree with the following statements? Please select **one answer** for each approach.

	Very strongly agree	Strongly agree	Agree	Disagree	Strongly disagree	Very strongly disagree
I enjoy teaching French through this approach	☐	☐	☐	☐	☐	☐
I look forward to teaching French through this approach	☐	☐	☐	☐	☐	☐
Teaching French through this approach is so enjoyable that I like preparing the lessons for the class	☐	☐	☐	☐	☐	☐
I often have reasons to be happy when teaching French through this approach	☐	☐	☐	☐	☐	☐
When teaching French through this approach, I generally teach with enthusiasm	☐	☐	☐	☐	☐	☐
	Very strongly agree	Strongly agree	Agree	Disagree	Strongly disagree	Very strongly disagree
I generally feel tense when teaching French through this approach	☐	☐	☐	☐	☐	☐
I often worry that my French teaching isn't going so well with this approach	☐	☐	☐	☐	☐	☐
Preparing to teach French through this approach often gives me cause to worry	☐	☐	☐	☐	☐	☐
I feel uneasy when I think about teaching French through this approach	☐	☐	☐	☐	☐	☐

☐ Q4

💡 ★ ...

In this section please think about your experiences teaching **the more novel (language-driven) approach**. To what extent do you agree with the following statements? Please select **one answer** for each approach.

	Very strongly agree	Strongly agree	Agree	Disagree	Strongly disagree	Very strongly disagree
I enjoy teaching French through this approach	☐	☐	☐	☐	☐	☐
I look forward to teaching French through this approach	☐	☐	☐	☐	☐	☐
Teaching French through this approach is so enjoyable that I like preparing the lessons for the class	☐	☐	☐	☐	☐	☐
I often have reasons to be happy when teaching French through this approach	☐	☐	☐	☐	☐	☐
When teaching French through this approach, I generally teach with enthusiasm	☐	☐	☐	☐	☐	☐
	Very strongly agree	Strongly agree	Agree	Disagree	Strongly disagree	Very strongly disagree
I generally feel tense when teaching French through this approach	☐	☐	☐	☐	☐	☐
I am often worried that my teaching isn't going so well with this approach	☐	☐	☐	☐	☐	☐
Preparing to teach this approach often gives me cause to worry	☐	☐	☐	☐	☐	☐
I feel uneasy when I think about teaching this approach	☐	☐	☐	☐	☐	☐

Appendix F.5

Number of Post-Lesson Evaluations Completed by Teachers by Approach

<i>Teacher (School – Sch)</i>	<i>Approach</i>	<i>Approach</i>	<i>Additional Information</i>
	<i>Novel</i>	<i>Standard</i>	
T1	15	15	
T2	11	11	
T3 (SchA)	25	24	Taught all lessons in terms 2 & 3
T3 (SchC)	29	28	Taught all lessons
T3 (SchE)	16	16	Taught half of lessons and also covered on occasions for absent teacher
T4	5	5	Left the school after term 1
T7	14	14	
T8	10	12	
T9	14	12	
Total	139	137	

Appendix G.1

Primary French Questionnaire: Pupils 1

Info_1

Click on the down arrow and **choose** the school letter that your teacher gives you.

Info_2

Your name: Please **type** your first name.

Info_3

Your group letter: **Click** on the down arrow and **choose** the **letter** that your teacher gives you.

Example_1

PRIMARY FRENCH

Thank you for completing this questionnaire. Your answers will help us to understand what it's like to learn French. Please remember these things when you answer:

Your answers are really important to us.

There are no right or wrong answers, so just be honest and tell us what you really think.

Just choose one answer each time. Look carefully at the faces or the answers and choose the best one for you.

Let's try a couple together:

Example 1: Do you like eating ice-cream? **Tick ✓ one answer next to the face that best shows how you feel.**

- ☐ Yes, a lot 
- ☐ Yes, quite a lot 
- ☐ A bit 
- ☐ No 

Example_2

Do you wear a uniform to school? **Tick ✓ one answer.**

- ☐ Yes (2)
- ☐ No (1)

Example_3

Do you think you could count backwards from 10 to 1? **Tick ✓ one answer.**

- ☐ I'm sure I could 
- ☐ I think I could 
- ☐ I might be able to 
- ☐ I definitely couldn't 

Intro

The only people who will see your answers will be the researchers, so you can tell us exactly what you think! Your answers will tell us how you feel about learning French this year, which will help us to find the best ways to teach you. Sometimes, the questions we ask might seem

the same, but it's to help us find out what you think in different ways. If you're not sure what a question is asking you, just put your hand up and ask for help.

Section 1: Your languages

Background_Q1

Apart from English, do you speak any other languages? **Tick ✓ one answer.**

☐ Yes

☐ No

Background_Q1b

If you said 'yes', please tell us which languages you speak at home or with any members of your family. **Write them here.**

Background_Q2

Apart from your lessons in school, have you ever learnt French? (for example, in a club after school or at the weekends?) **Tick ✓ one answer.**

☐ Yes

☐ No

Background_Q3

Did you learn French in Key Stage 1? **Tick ✓ one answer.**

☐ Yes

☐ No

Background_Q3b

Please tell us which years you learnt French in. **Choose each year when you did French.**

You can tick ✓ as many answers as you need.

☐ Reception

☐ Year 1

☐ Year 2

Section 2: Your thoughts about learning French this year

Q4

Are you looking forward to learning French this year? **Tick ✓ one answer.**

☐ Yes

☐ No

☐ I don't know

Q1_SE

How well do you think you will be able to learn French in Year 3? **Tick ✓ one answer.**

- ☐ Very well 
 - ☐ Quite well 
 - ☐ OK 
 - ☐ Not very well 
-

Q2_SE

Please tell us why you think this. **Write your answer here.**

Q3_SE

Now think about some of the things you will be learning to do in French. Do you think you could do these things? You will see circles of different sizes. **Tick ✓ the circle that best matches how you feel.**

Could you put the words in the right order in a sentence?

- ☐ I'm sure I could 
 - ☐ I think I could 
 - ☐ I might not be able to 
 - ☐ I definitely couldn't 
-

Q4_SE

Could you say a sentence that you want to say?

- ☐ I'm sure I could
- ☐ I think I could
- ☐ I might not be able to
- ☐ I definitely couldn't



Q5_SE

Could you spell the French words that you will learn?

- ☐ I'm sure I could
- ☐ I think I could
- ☐ I might not be able to
- ☐ I definitely couldn't



Q6_SE

Could you read words out loud in French?

- ☐ I'm sure I could
- ☐ I think I could
- ☐ I might not be able to
- ☐ I definitely couldn't



Q7_SE

Could you pronounce words clearly in French?

- ☐ I'm sure I could
- ☐ I think I could
- ☐ I might not be able to
- ☐ I definitely couldn't



Q8_SE

Could you remember the meaning of French words?

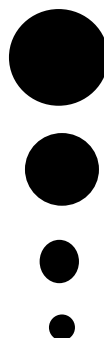
- ☐ I'm sure I could
- ☐ I think I could
- ☐ I might not be able to
- ☐ I definitely couldn't



Q9_SE

Could you know the French words for some English words?

- ☐ I'm sure I could
- ☐ I think I could
- ☐ I might not be able to
- ☐ I definitely couldn't



Thank you for completing this questionnaire.

Appendix G.2

Pupil Questionnaire 2/3

Section 1: Your thoughts about learning French

Are you enjoying learning French this year? Choose one answer above the face that best shows how you feel.

☐ Yes, a lot



☐ Yes, quite a lot



https://york.eu.qualtrics.com/QESetSection/Block/Ajax/GetSurveyPrintPreview?ContextSurveyID=SV_3h8k8z2FmJdPxc&ContextLibraryID=UR... 3/10

19/09/2025, 16:05

Qualtrics Survey Software

☐ A bit



☐ No



0:00 / 0:18

Do you look forward to French lessons? Choose one answer above the face that best shows how you feel.

☐ Yes, a lot



☐ Yes, quite a lot



☐ A bit



☐ No



0:00 / 0:19

Would you like to have more French lessons? Choose one answer above the face that best shows how you feel.

☐ Yes, a lot



☐ Yes, quite a lot



☐ A bit



☐ No



Block 2

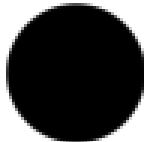
0:00 / 0:36

Section 2: What do you think you could do in French?

Now think about some of the things you have been learning to do in French. If I asked you to do these things, do you think you could do them? **Choose the answer above the circle that best matches how you feel.**

Could you put the words in the right order in a sentence?

☐ I'm sure I could



☐ I think I could



☐ I might not be able to



☐ I definitely couldn't



Could you say a sentence that you want to say?

☐ I'm sure I could



☐ I think I could



☐ I might not be able to



☐ I definitely couldn't



0:00 / 0:13

Could you spell the French words that you have learnt?

☐ I'm sure I could



☐ I think I could



☐ I might not be able to



☐ I definitely couldn't



https://york.eu.qualtrics.com/Q/EdtSection/Blocks/Vjao/GetSurveyPrintPreview?ContextSurveyID=SV_3h8M82FmJuPxc&ContextLibraryID=URL...

19/08/2025, 16:05

Qualtrics Survey Software

0:00 / 0:13

Could you read words out loud in French?

☐ I'm sure I could



☐ I think I could



☐ I might not be able to



☐ I definitely couldn't



Could you pronounce words clearly in French?

☐ I'm sure I could



☐ I think I could



☐ I might not be able to



☐ I definitely couldn't



https://york.eu.qualtrics.com/QED/section/Blocks/JavaScript/GetSurveyPrintPreview?ContextSurveyID=SV_3t8k82FmJzPoc&ContextLibraryID=URL... 7/16

19/08/2025, 16:05

Qualtrics Survey Software

0:00 / 0:11

Could you remember the meaning of French words?

☐ I'm sure I could



☐ I think I could



☐ I might not be able to



☐ I definitely couldn't



0:00 / 0:12

Could you know the French words for some English words?

☐ I'm sure I could



☐ I think I could



☐ I might not be able to



☐ I definitely couldn't



Section 3. Your French teachers.

Now think about your French teacher or teachers and answer these questions.

Choose one answer.

Do you think your French teachers like teaching you French?

☐ Yes, a lot



☐ Yes, quite a lot



☐ Not much



☐ No, not at all



0:00 / 0:14

Do you think your French teacher or teachers try to make you excited about learning French? **Choose one answer.**

☐ Yes, a lot



https://yorku.qualtrics.com/Q/EdtSection/Blocks/Ajax/GetSurveyPrintPreview?ContextSurveyID=SV_3tBk82FnJzPxc&ContextLibraryID=URL... 8/11

19/08/2025, 16:05

Qualtrics Survey Software

☐ Yes, quite a lot



☐ Not much



☐ No, not at all



Powered by Qualtrics

Appendix G.3

Pupil Questionnaire 3 – Open-ended Questions

What things have you enjoyed the most about French lessons in Year 3? **Write your answers here.**

0:00 / 0:09

Are there things you didn't like about the French lessons in Year 3? If there are, **write your answers here.**

Appendix G.4

Piloting Pupil Questionnaires

PQs were piloted with 52 Year 3 pupils from three local schools, using cognitive interviewing with 12 pupils in one school (Bielick, 2017:653). Piloting presented opportunities to observe response distribution, albeit on a smaller scale than that advocated (Bielick, 2017; Dörnyei & Taguchi, 2010). Pupil feedback led to minimal redrafting; changes were minor so repiloting was not deemed necessary. As children do not always recognise issues (Bell, 2007), primary school teachers and expert colleagues also reviewed questionnaires.

Appendix G.5

Original Questions from Carreira (2006) Used as Measure of Pupil Enjoyment

I really enjoy learning English.

I always look forward to the day when we have English class.

I would like to try to use the English we have learned.

I hope that we have more English lessons.

Appendix G.6

Piloting Pupil Group Interviews

The group interview schedule was piloted in June 2022 with two groups of six Year 3 pupils in a local school who had been learning Spanish since September 2021. Following this pilot, the group interview duration was decreased from 30 minutes to 20-25, deemed more appropriate given the young age of the participants. Manipulatives such as Spanish flags to indicate they wanted to contribute, and a feelings thermometer on which they placed their name cards, were found to be useful during piloting; these were retained for the study.

Appendix G.7

Pupils' Group Interview Schedule

State purpose:

We're here today to talk about your French lessons, and how you feel about learning French.

Reminder of confidentiality:

We need to remember that everyone in this group should feel they can say anything they want to say about the French lessons, and no one here will repeat what they say about the French lessons to other people in school. Do we all understand that?

(If pupils were to make any disclosure which raised concerns about their safety or the safety of others, I would point out that I may pass on this information to another person).

It's really important that everyone has a turn to say what they want, so I've put some flags from France and other French-speaking countries here on the table. If you've got something to say and you haven't had a chance, you can just pick one of those up and hold it in the air. Is that OK? Try to make sure you don't speak at the same time as anyone else, that way we can make sure we can listen carefully to what everyone is saying.

Now, you know that what you say is very important. To help me remember afterwards what you've said, I'm going to record our conversations, but instead of using your real name when I ask you to say something, I am going to use your French name. Are we happy with that? So, I've made you a card with your French name on. If you don't like that name, you can change it now. Here we go. Shall we have a practice saying our new name?

While we're having our chat, I might just stop you occasionally and check I've understood what you said. This doesn't mean it's wrong – there aren't any right and wrong answers, I just want to know what you think, OK? It just means I want to be really sure I've understood because what you say is very important to me. Is that OK? And if you don't understand, just tell me and I will say it again in different words. If you decide you want to change anything you've said, that's fine, you can tell me any time before the end of this school year.

Reminder of voluntary participation:

Remember that you don't have to take part if you don't want to. If you change your mind, that's OK. We can stop any time you want to, even if you just need a little break.

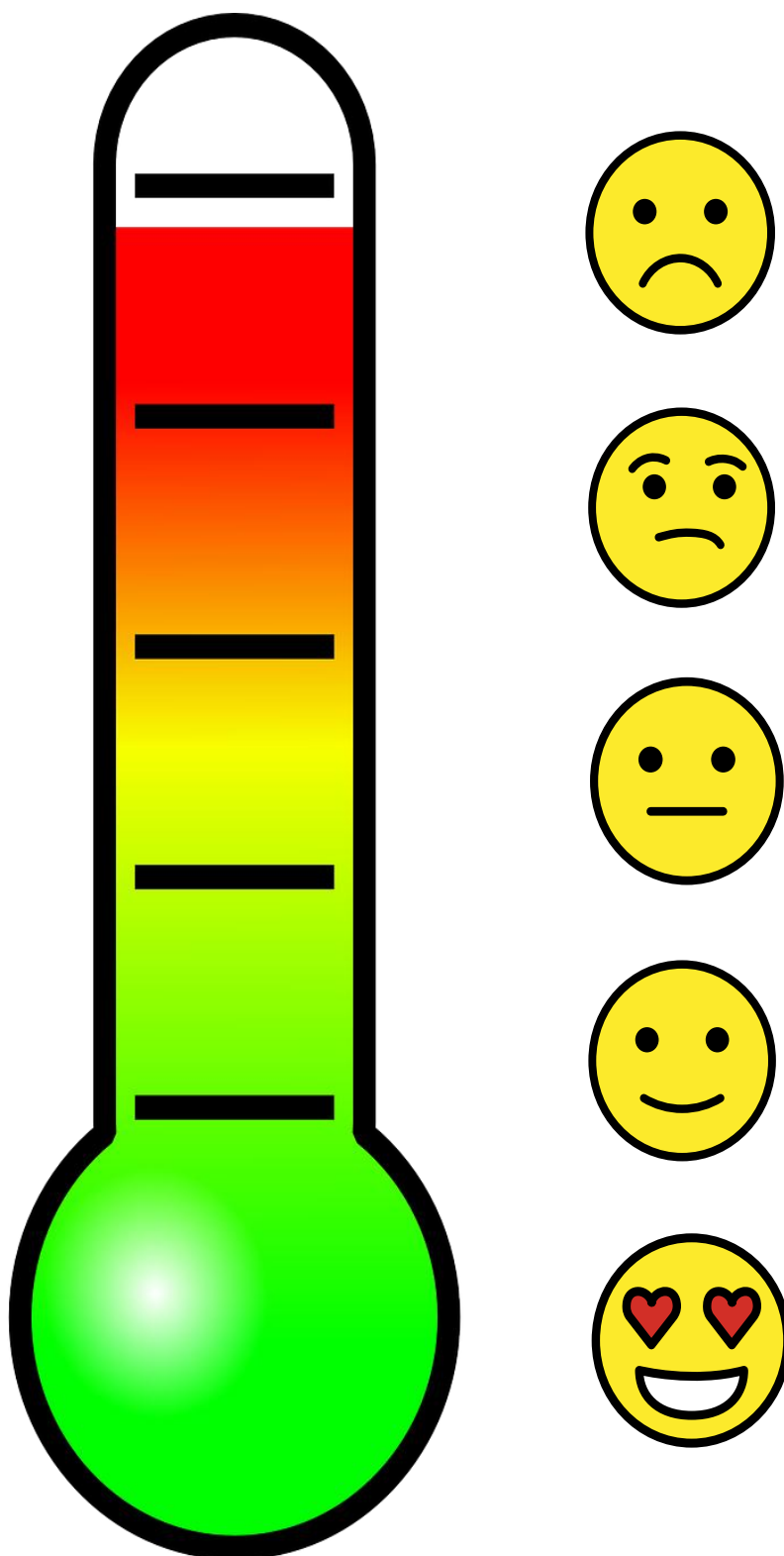
Do you have any questions before we start?

Shall we listen to ourselves first on the recording and see what we sound like?

1. Let's think back to some of your French lessons. How do you feel about learning French? Can you explain how you feel? Do you enjoy it? (use feelings thermometer)
2. Can you explain your answer? What do you like about learning French? Anything you don't like?
3. Do you have any favourite activities? Prompt verbally first. (Use worksheets or printouts of PPT slides as prompts if needed)
4. What do you think you have learnt in French? (prompt: words? sounds? spellings? how to say things? songs? poems?)
5. What things help you to learn French? (see if this leads to reference to approach-specific pedagogy – i.e., songs, phonics, explicit grammar teaching).
6. Do you think you're good at learning French? What makes you say that?
7. What things do you think you'll be able to do in French next year?
8. Do you think your French teachers like teaching you French? What gives you that feeling? / How do you know?
9. Do your French teachers try to make you excited about learning French? What makes you say that?

Appendix G.8

Feelings Thermometer Activity



Appendix G.9

Self-efficacy Questions from Pupil Questionnaires

Q1 SE Could you put the words in the right order in a sentence?

Q2 SE Could you say a sentence that you want to say?

Q3 SE Could you spell the French words that you will learn?

Q4 SE Could you read words out loud in French?

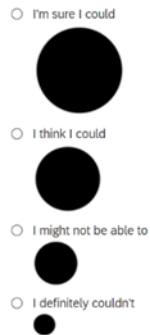
Q5 SE Could you pronounce words clearly in French?

Q6 SE Could you remember the meaning of French words?

Q7 SE Could you know the French words for some English words?

Appendix G.10

Response Options and Scoring System for Self-efficacy Questions



Scoring system:

I'm sure I could = 4 points

I think I could = 3 points

I might not be able to = 2 points

I definitely couldn't = 1 point

Total score possible = 28

Appendix G.11

Response Options for Enjoyment Questions in Pupil Questionnaire 2 (PQ2) and Pupil Questionnaire 3 (PQ3)

☐ Enjoy1



Section 1: Your thoughts about learning French

Are you enjoying learning French this year? Choose one answer above the face that best shows how you feel.

☐ Yes, a lot



☐ Yes, quite a lot



☐ A bit



☐ No



Yes, a lot = 4 points

Yes, quite a lot = 3 points

A bit = 2 points

No = 1 point

Appendix H.1

Piloting Teacher Interviews

Interviews were piloted in June 2022 with four teachers, two teachers piloting Interview1 and two piloting Interview2. No issues arose.

Appendix H.2

Teacher Interview Schedule 1

State purpose:

The purpose of this interview is to give you an opportunity to talk about your experiences teaching French so far, if you've taught it already, and to think ahead to both French approaches that you'll be teaching this year, the more standard topic-driven approach, and the more novel language-driven approach.

Because I want to find out how you feel about teaching French, please remember that there are no right or wrong answers, please just make your answers as honest and as detailed as you can.

Reminder of confidentiality:

I want to remind you that everything you say is said in confidence and will not be shared with anyone outside this interview. Your identity will be anonymised in the reporting of the project, and I will not use any responses which can only be traced back to you.

Data will be stored securely and the only person who will have access to the data with your name attached will be me. My supervisor will see data that has been pseudo-anonymised (your name removed).

This interview is being recorded. I will then give you a copy of the transcript and you will have the opportunity to make any changes that you wish to make.

Reminder of voluntary participation:

Please remember that your participation is voluntary. You have the right not to answer and can stop at any time if you wish, without having to explain your decision.

Do you have any questions before we start?

Pick up on responses to Teacher questionnaire 1 – So you have taught French before / you haven't taught French before ...

1. Have you taught any other subjects which were unfamiliar? How did you approach them? To what extent do you think you might draw on this experience while teaching French this year?

OR:

2. When you taught French for the first time, how did you approach that? To what extent do you think you might draw on this experience when teaching French this year?
3. Have you had any CPD to support your teaching of French?
 Prompt: focusing on linguistic competence or pedagogical competence? Or both?
 (Can you tell me about the content? What form did that take? How did you find it?
 Was this provided as part of your initial teacher training?)
4. How do you feel about teaching French this coming year?
5. If you don't feel confident, can you explain your answer?
 Prompt: is this related to grammatical understanding / vocabulary / confidence in French phonics? To pedagogy?
 If you feel confident, can you explain your answer?
 Prompt: is this related to grammatical understanding / vocabulary / confidence in French phonics? To pedagogy?
6. How do you feel about teaching each approach? Do you feel the same way or differently about each approach? Can you explain that answer?
 Prompt: How does the thought of teaching the language-driven approach make you feel? Any particular reasons for this?
 How does the thought of teaching the topic-driven approach make you feel? Any reasons for this?
7. How confident do you feel about making explicit connections between French and English to support pupils? Do you feel this is addressed equally in both approaches?
8. Are you worried about making mistakes when teaching French? What kind of mistakes?
9. How do you feel about explaining vocabulary / correcting pupils' pronunciation (when they read aloud) / supporting pupils when they write things down / your grammar knowledge to support pupils when they try to create simple sentences?
10. Which kinds of words do you think are the most important ones for children in Year 3 to learn in French?
11. Which aspect(s) of teaching French concern you the most? What aspects of the teaching are you hoping that the resources will most support you with? What do you most want out of a curriculum map and accompanying resources?

12. What do you most want to see at the end of a French lesson? If you were to plan a curriculum, what aspects would you think are the most important?
13. Is there anything else you would like to ask? Prompt: About either teaching approach?

Appendix H.3

Teacher Interview Schedule 2 & 3

State purpose (as for Interview1, [Appendix H.2](#))

Do you have any questions before we start?

Link to responses to Teacher Questionnaire.

1. Let's think back to your teaching of French over this half-term. How do you feel it's going? In each approach? (Use as check of fidelity to treatment too - so, to what extent have you followed the materials given?)
2. How does teaching the language-driven approach make you feel? Can you think of reasons for this?
3. How does teaching the topic-driven approach make you feel? Can you think of reasons for this?

Follow-up if unclear: Do you find you prefer teaching one approach over the other?

Can you explain your answer?

4. What do you enjoy about teaching each approach? (Ensure they discuss both approaches)
5. Do you feel more enthusiastic when teaching one approach than the other? Can you explain your answer?

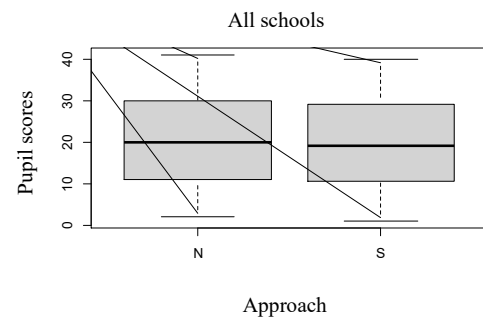
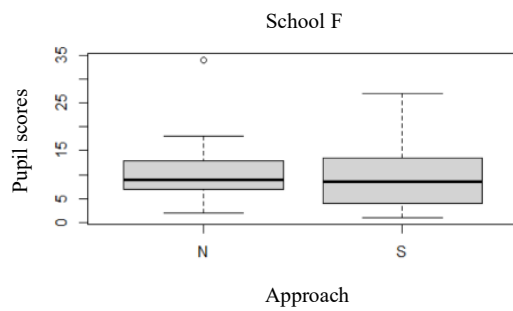
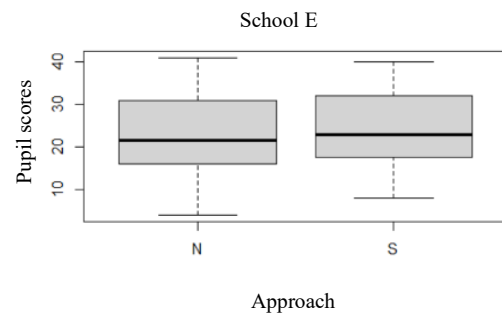
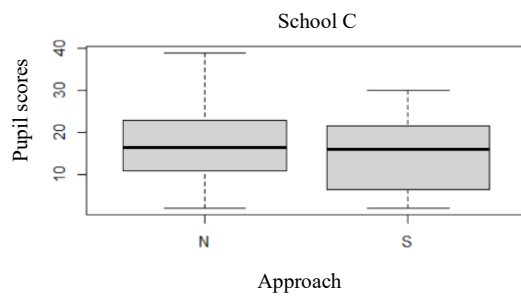
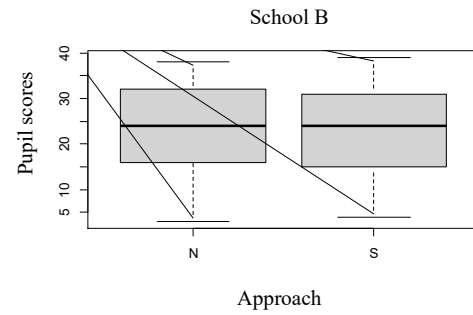
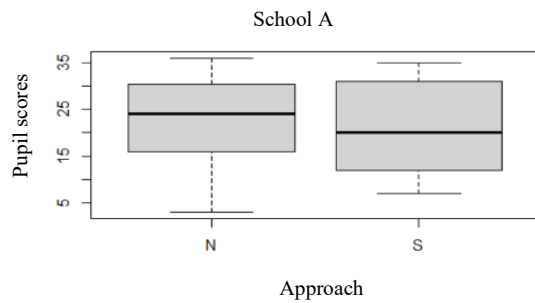
Prompt if needed: Does either approach make you feel more tense or anxious? (Generally, and when speaking?)

6. Let's think back to when you have been planning the French lessons. Have you noticed any differences in your preparation for each approach?
7. Do you feel more confident when teaching one approach or the other? is this related to grammatical understanding / vocabulary / confidence in French phonics? Also allow them to say if this is related to their pedagogic knowledge or feelings of confidence in terms of delivery (not just subject knowledge).
8. Does one approach make you feel more secure than the other? Can you explain your answer?

9. Have you noticed you make explicit connections between French and English when teaching to support learners? Is this more apparent in one approach than the other?
10. Do you think there's been any impact from each approach on your subject knowledge in French?
11. Do you worry about making mistakes when teaching French? What kind of mistakes? In one approach more than the other?
12. How do you feel about explaining vocabulary / correcting pupils' pronunciation (when they read aloud) / supporting pupils when they write things down / your grammar knowledge to support pupils when they try to create simple sentences? – in each approach?
13. Which kinds of words do you think are the most important ones for children in Year 3 to learn in French?
14. Do you feel able to provide activities which enable pupils to practise symbol-sound correspondences? Is this more evident in one approach than the other?
15. Have you noticed any impact of the teaching approach on your relationship with pupils?
16. Which aspects seem to go down well / less well with pupils from each approach? How do pupils respond to each approach? Any impact on their motivation? Or attention?
17. Do pupils make more progress in one approach than the other? Do you think that the pupils have learnt a) more and b) different things in one approach compared with the other?
18. Which aspect(s) of teaching French concern you the most?
19. Is there anything else you would like to tell me or ask me about either teaching approach?

Appendix I.1

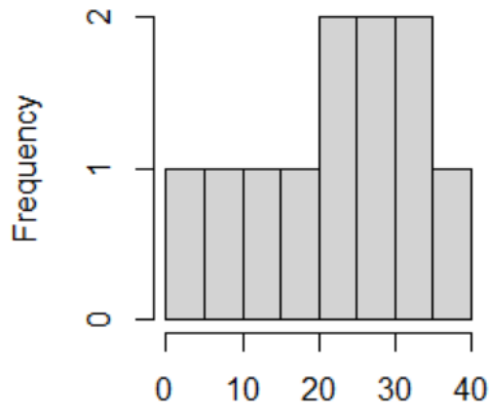
Box Plots Showing Data Distribution for Language-analytic Ability Test Scores According to Approach in Schools A, B, C, E, F and All Schools



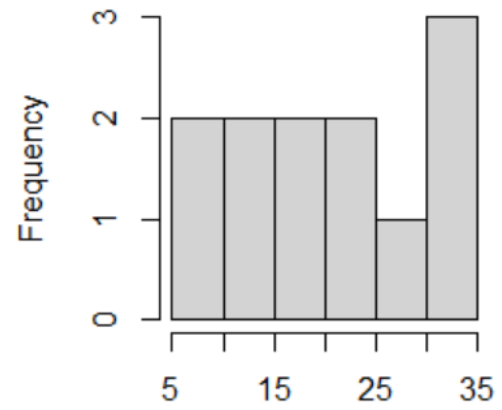
Appendix I.2

*Histograms Showing Data Distribution for Language-analytic Ability Test Scores
According to Approach in Schools A, B, C, E, F and All Schools*

School A

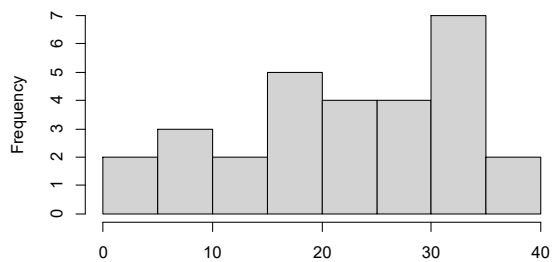


Novel scores

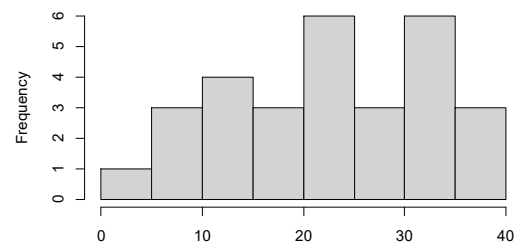


Standard scores

School B

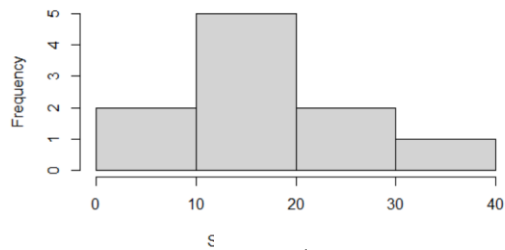


Novel scores

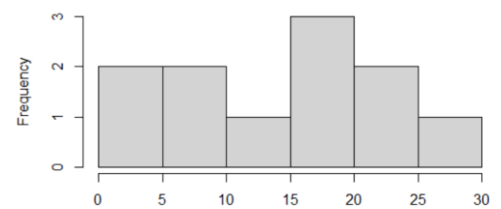


Standard scores

School C

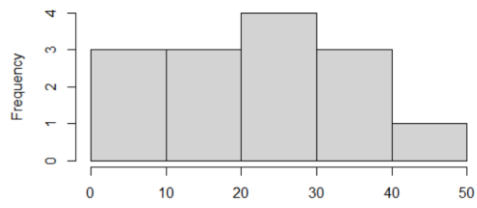


Novel scores

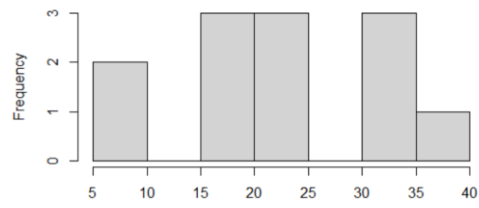


Standard scores

School E

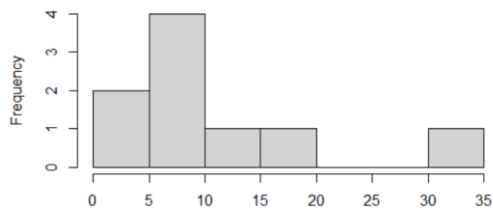


Novel scores

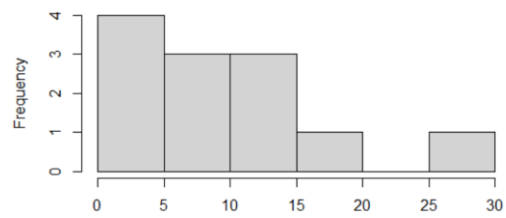


Standard scores

School F

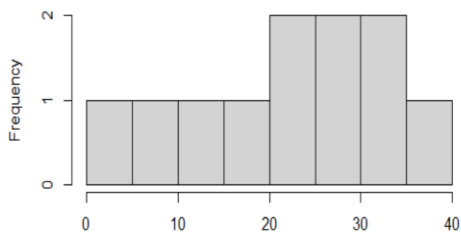


Novel scores

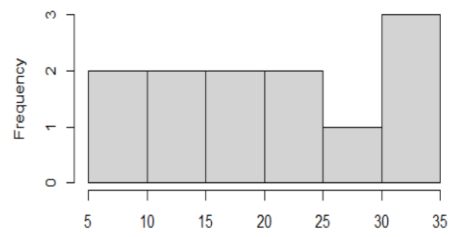


Standard scores

All schools



Novel scores



Standard scores

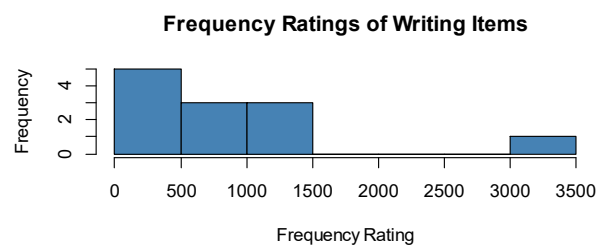
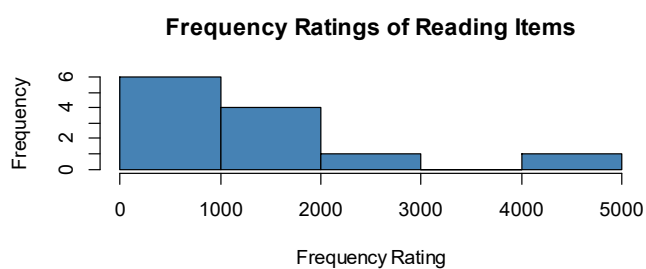
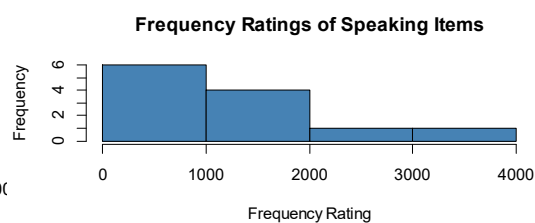
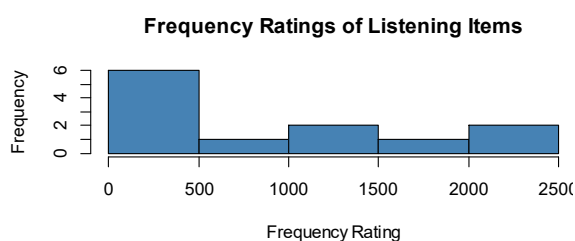
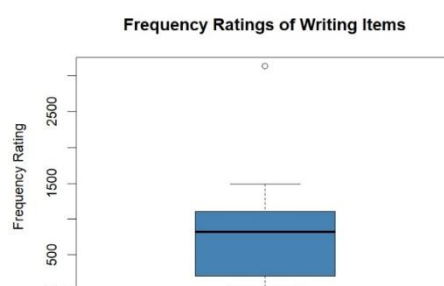
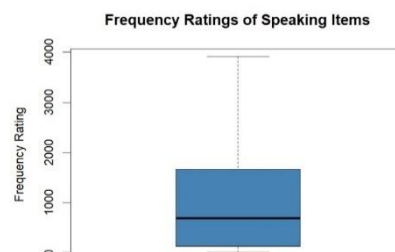
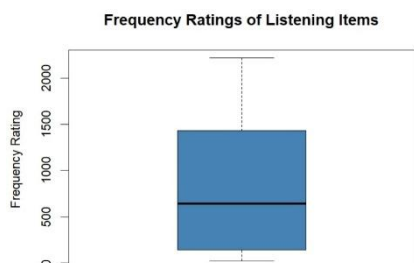
Appendix I.3

Skew, Kurtosis and Shapiro-Wilk Scores for Language-analytic Ability Test Scores by Approach

	School	A	School	B	School	C	School	E	School	F	All	schools
	Novel	Standard	Novel	Standard	Novel	Standard	Novel	Standard	Novel	Standard	Novel	Standard
Shapiro-Wilk W <i>p</i> -value	94 .51	93 .37	93 .09	96 .25	97 .93	95 .68	97 .86	97 .95	83 .05	93 .37	96 .02	96 .02
Skew	-.52	.09	-.26	-.28	.57	.16	-.04	-.03	1.44	.89		
Kurtosis	2.22	1.56	1.72	1.98	2.88	1.79	2.14	2.08	4.27	3.27		

Appendix I.4

Comparing Corpus-Based Frequency Ratings in Vocabulary Test Content in Receptive and Productive Modes, and Oral and Written Modalities – Normality of Distribution



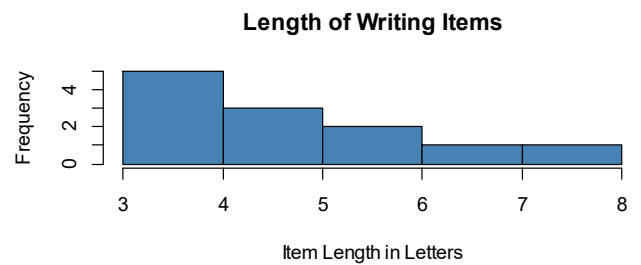
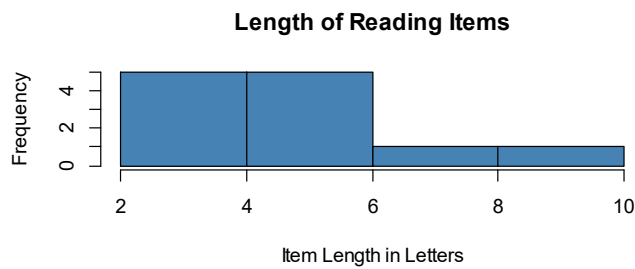
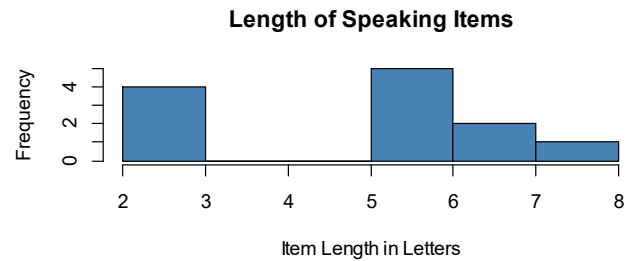
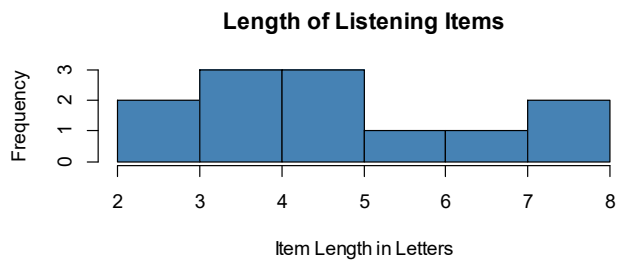
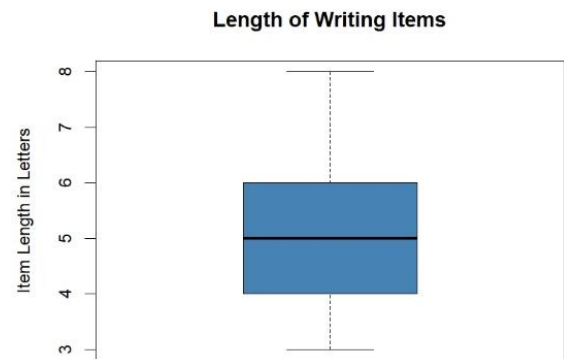
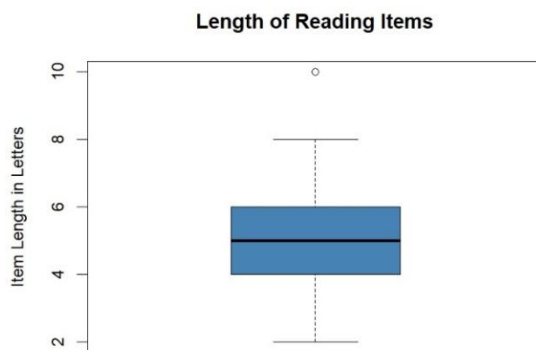
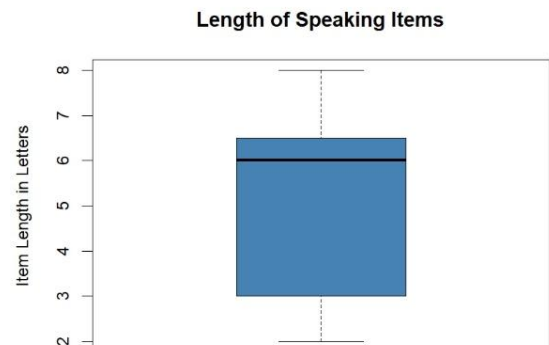
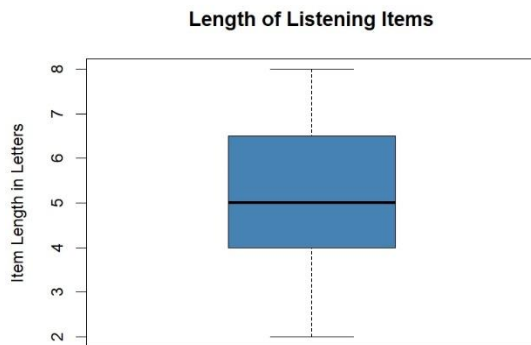
Note that where the list contained a lexical item with a corpus-based frequency rating above 5000, this appeared as N/A in the test list. In the statistical test, this was replaced by the mean rating for that test type.

Corpus-Based Frequency Rating

Test type	Skew	Kurtosis	Shapiro-Wilk
Receptive, oral	.41	-5.62	W = 86, $p = .06$
Productive, oral	.85	.08	W = 84, $p = .03$
Receptive, written	1.20	.63	W = 76, $p < .05$
Productive, written	1.08	.56	W = 82, $p = .02$

Appendix I.5

Comparing Word Length in Vocabulary Test Content in Receptive and Productive Modes, and Oral and Written Modalities – Normality of Distribution

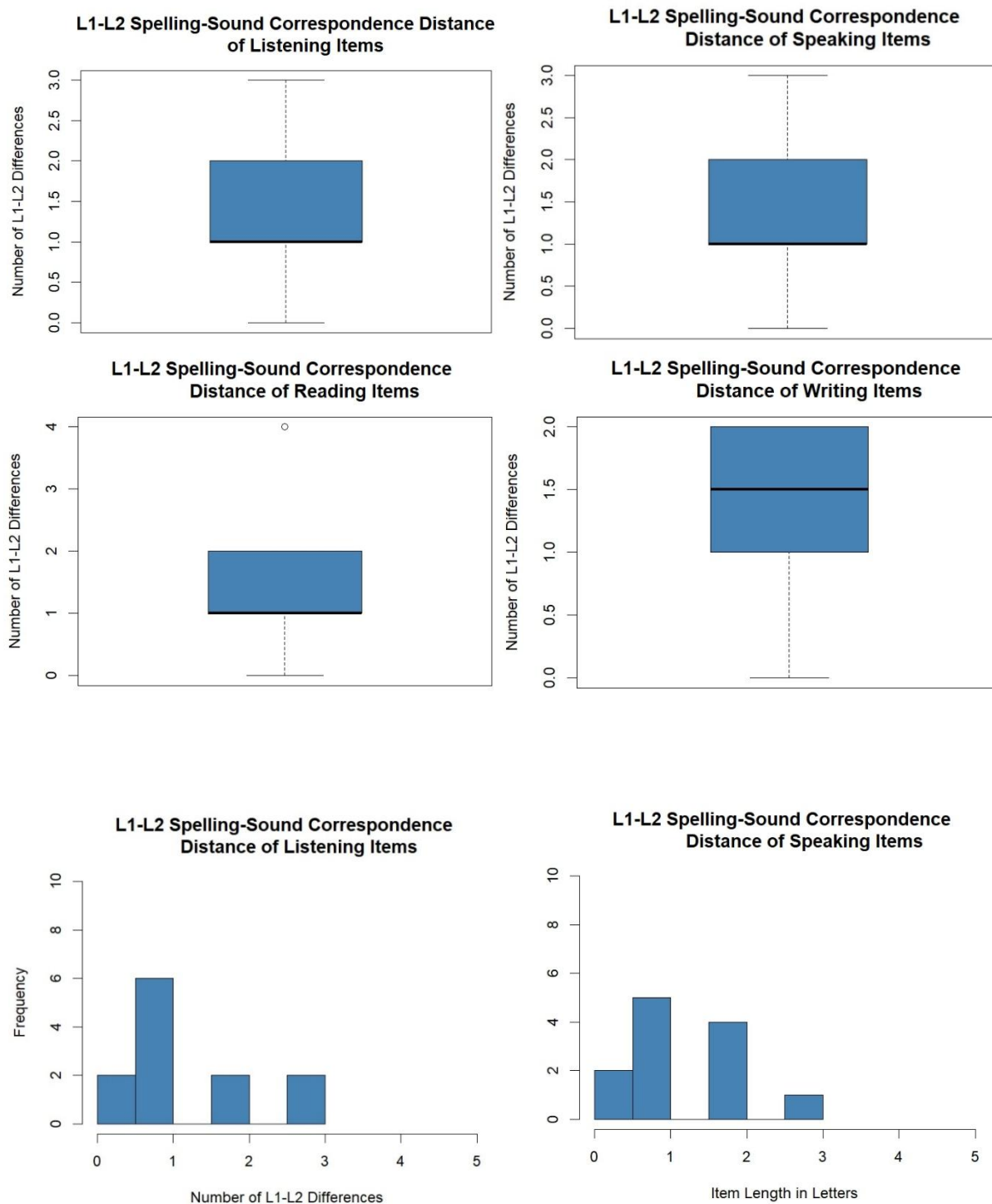


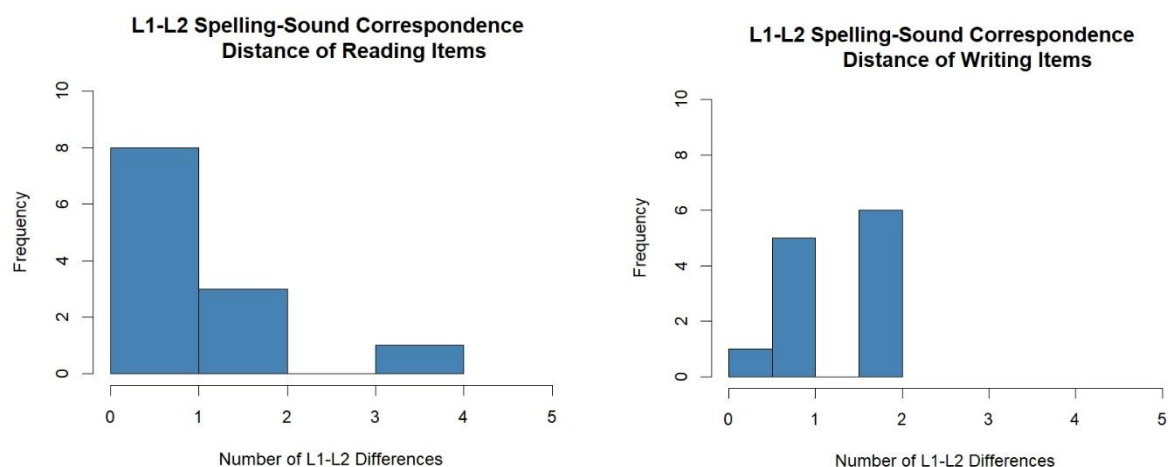
Word Length

Test type	Skew	Kurtosis	Shapiro-Wilk
Receptive, oral	.15	-.48	W = 94, $p = .55$
Productive, oral	-.33	-.61	W = 85, $p = .04$
Receptive, written	.58	-.08	W = 93, $p = .35$
Productive, written	.42	-.35	W = 93, $p = .37$

Appendix I.6

Comparing Symbol-Sound Correspondence Distance Between the English and the French in Vocabulary Test Content in Receptive and Productive Modes, and Oral and Written Modalities – Normality of Distribution



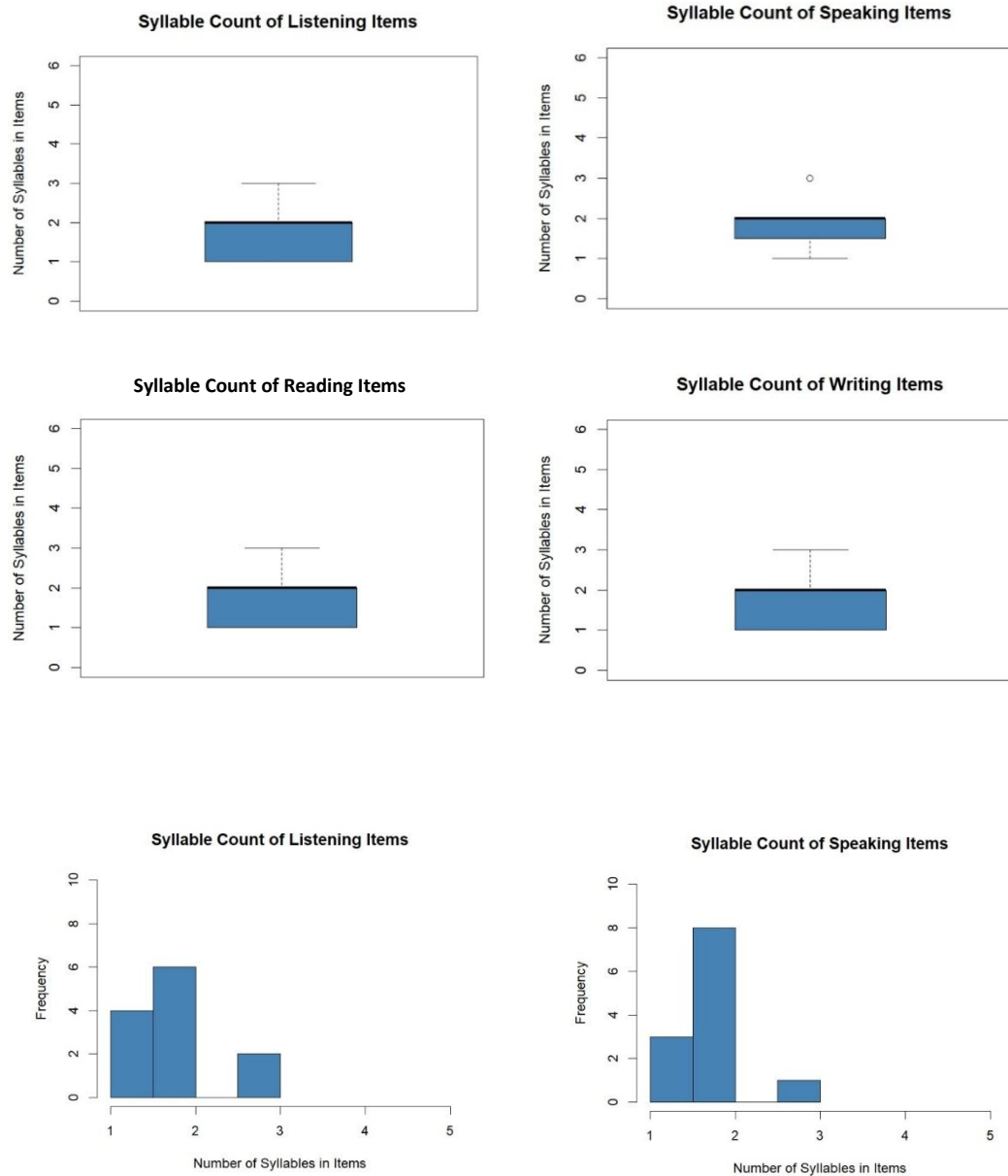


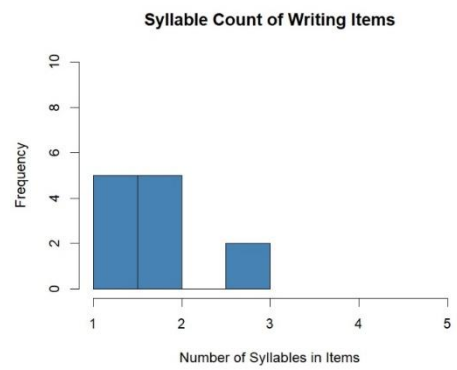
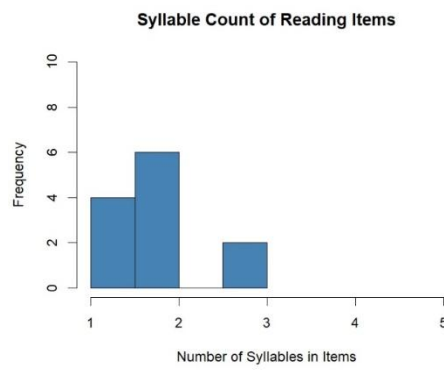
Symbol-Sound Correspondence Distance Between the French and the English

Test type	Skew	Kurtosis	Shapiro-Wilk
Receptive, oral	.33	-.42	W = 86, $p = .05$
Productive, oral	.08	-.41	W = 90, $p = .16$
Receptive, written	.96	.47	W = 78, $p = .005$
Productive, written	-.44	-.39	W = 77, $p = .004$

Appendix I.7

Comparing Syllable Count in Vocabulary Test Content in Receptive and Productive Modes, and Oral and Written Modalities – Normality of Distribution



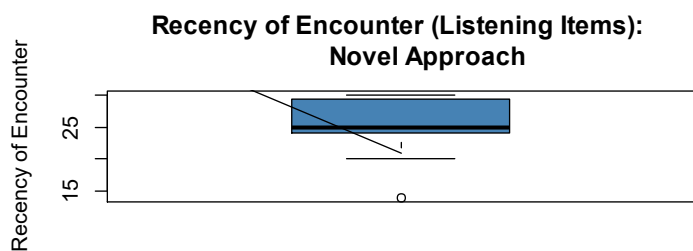


Syllable Count

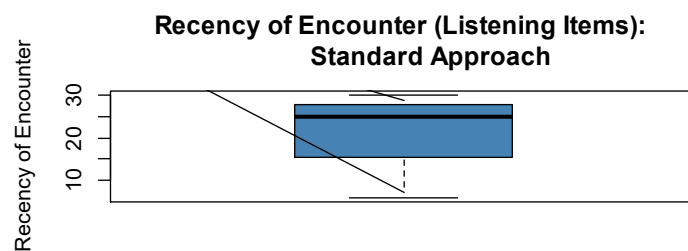
Test type	Skew	Kurtosis	Shapiro-Wilk
Receptive, oral	.16	-.50	$W = 82, p = .02$
Productive, oral	-.04	-.21	$W = 75, p = .003$
Receptive, written	.16	-.50	$W = 82, p = .02$
Productive, written	.29	-.54	$W = 81, p = .01$

Appendix I.8

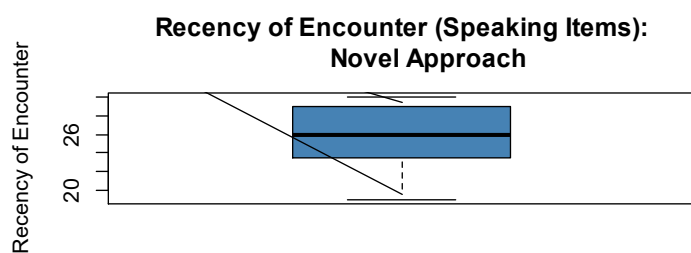
Comparing Recency of Encounter in Vocabulary Test Content in Receptive and Productive Modes, and Oral and Written Modalities – Normality of Distribution



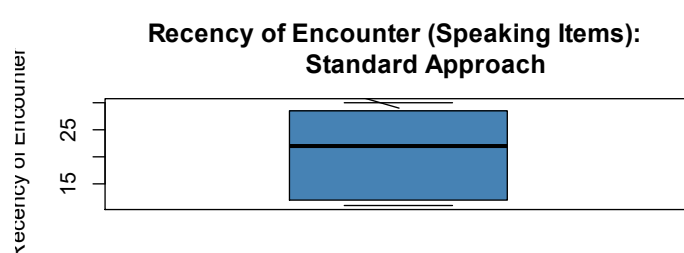
Teaching Approach



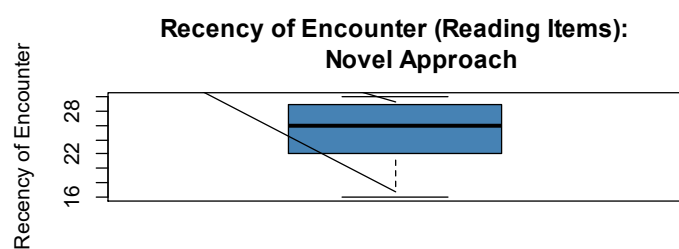
Teaching Approach



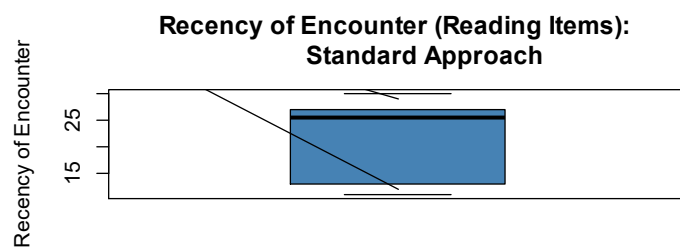
Teaching Approach



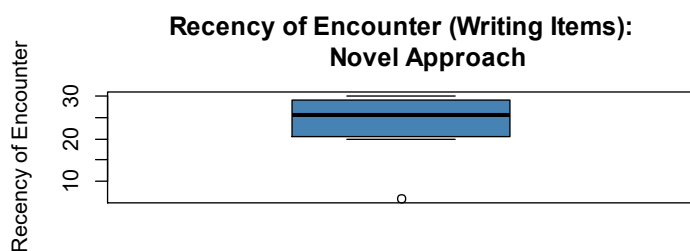
Teaching Approach



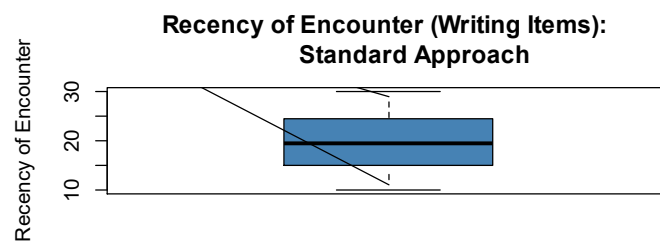
Teaching Approach



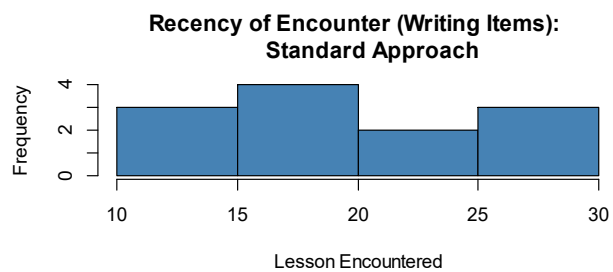
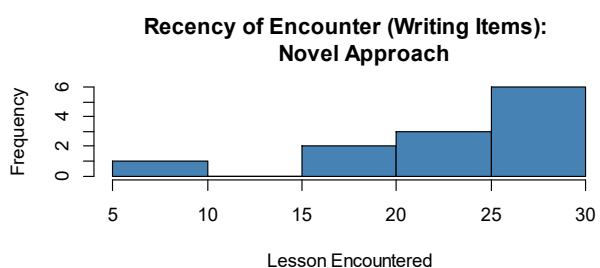
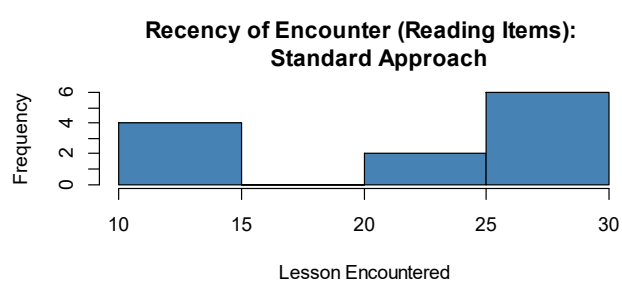
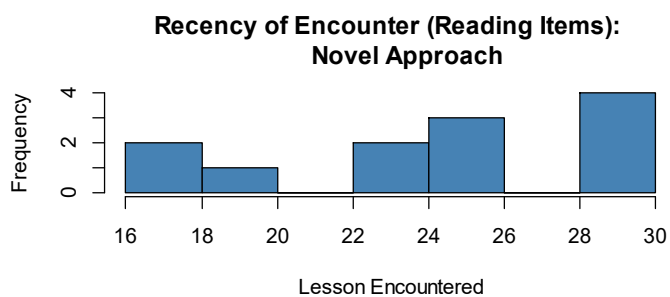
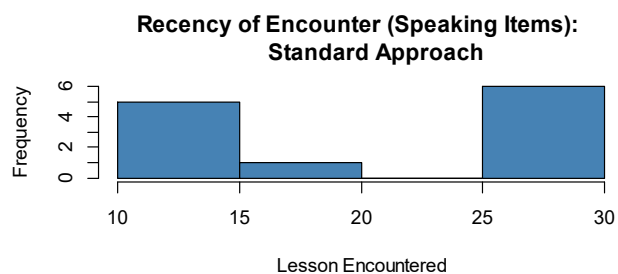
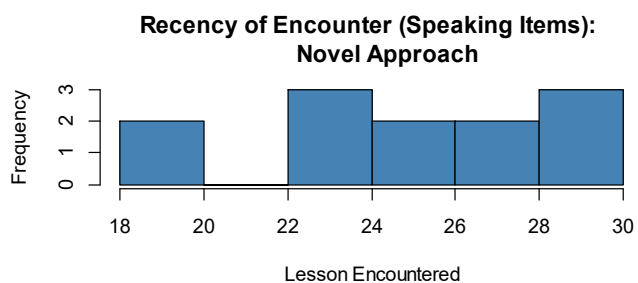
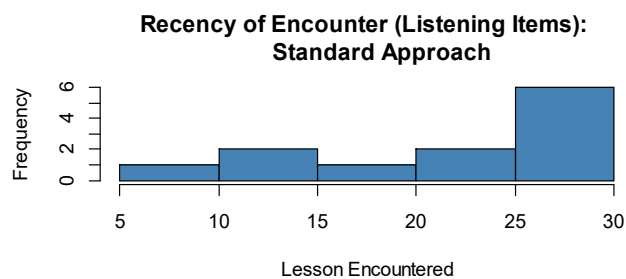
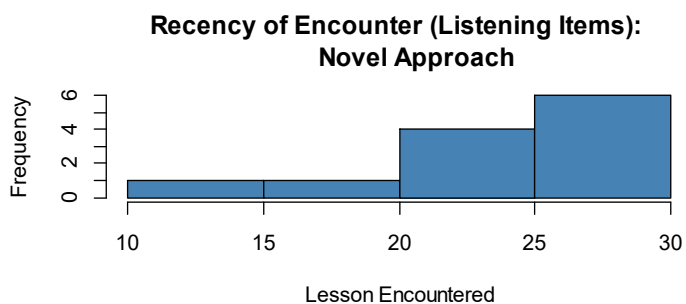
Teaching Approach



Teaching Approach



Teaching Approach

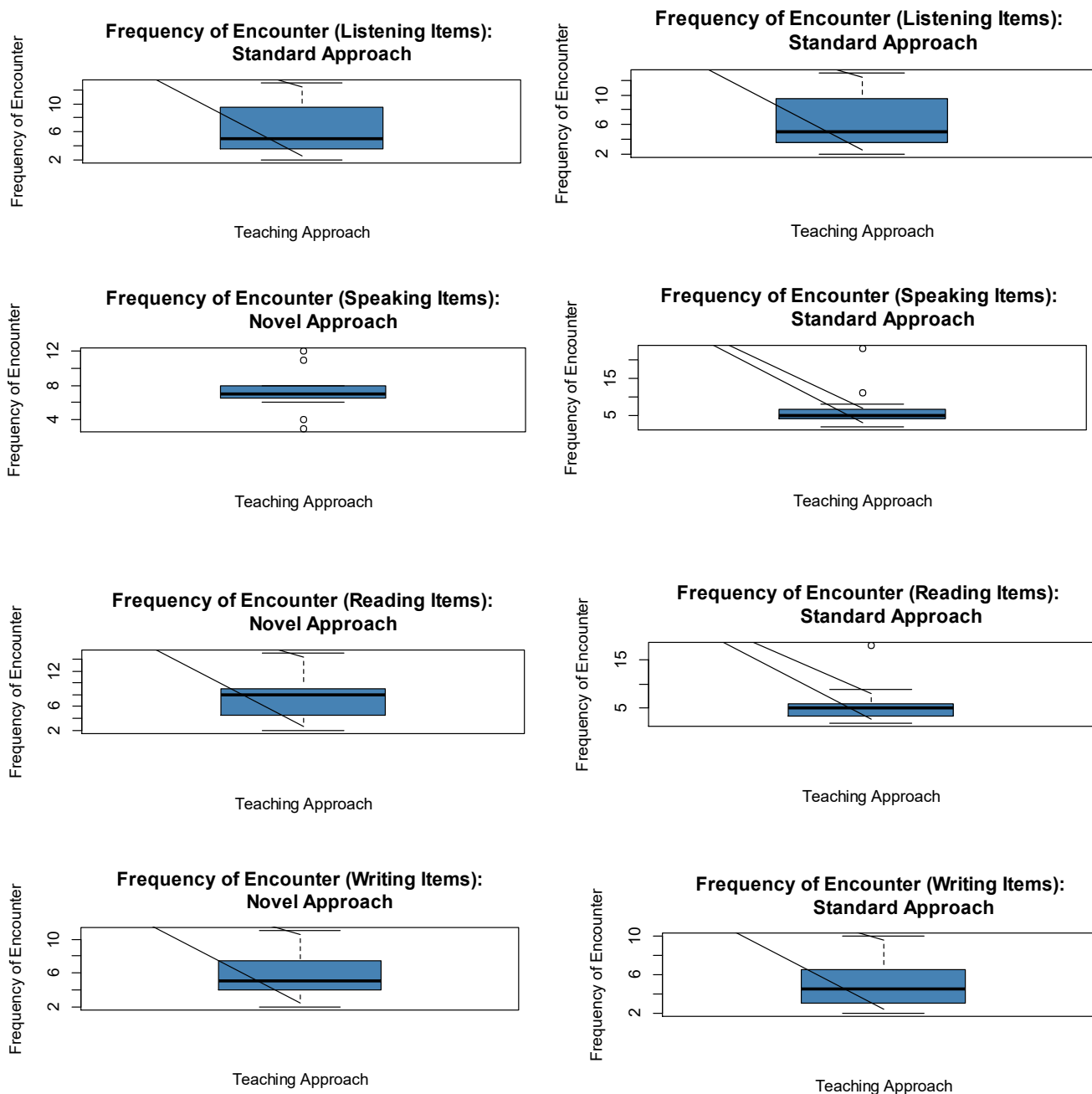


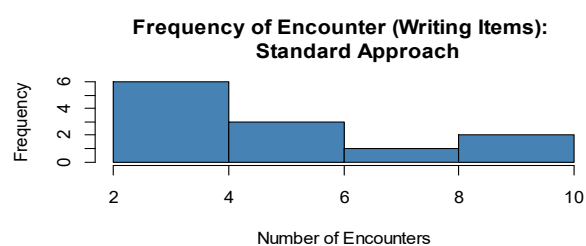
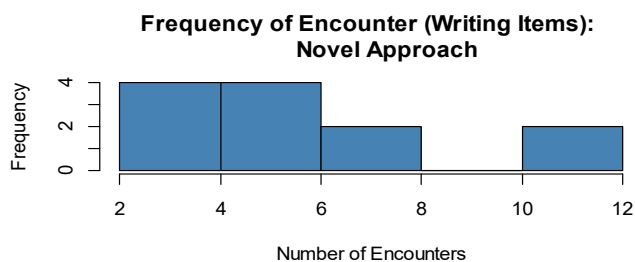
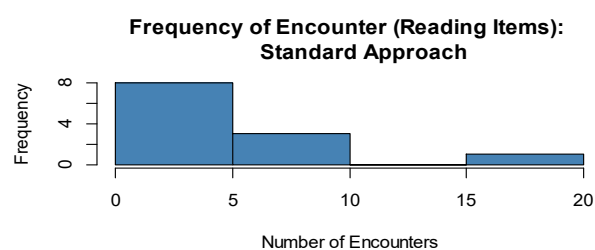
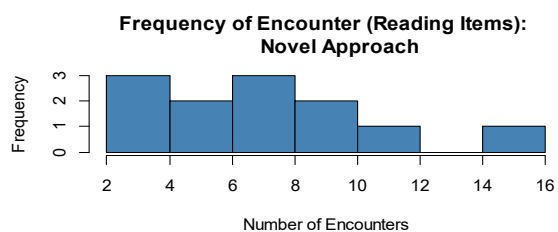
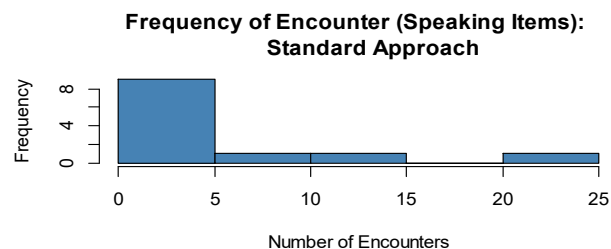
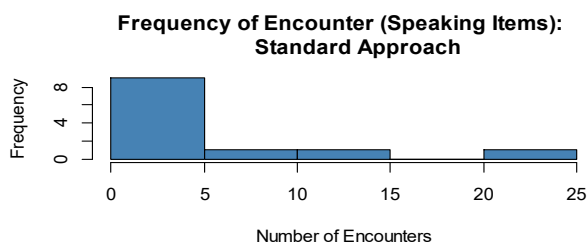
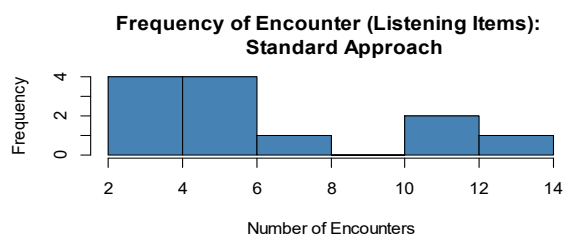
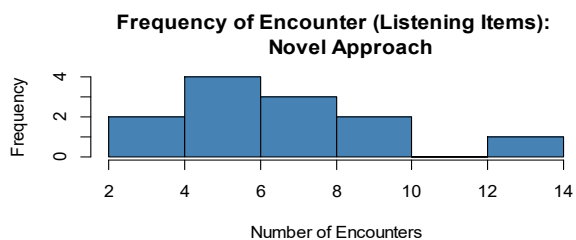
Recency of Encounter

Test type	Approach	Skew	Kurtosis	Shapiro-Wilk
Receptive, oral	Novel	.70	-.03	$W = 85, p = .04$
	Standard	-.68	-1.05	$W = 87, p = .07$
Productive, oral	Novel	-.31	-.48	$W = 91, p = .20$
	Standard	-.02	-.80	$W = 81, p = .01$
Receptive, written	Novel	-.49	-.44	$W = 87, p = .07$
	Standard	-.40	-.68	$W = 80, p = .01$
Productive, written	Novel	-1.08	.53	$W = 81, p = .01$
	Standard	.10	-.56	$W = 96, p = .76$

Appendix I.9

Comparing Frequency of Encounter in Vocabulary Test Content in Receptive and Productive Modes, and Oral and Written Modalities – Normality of Distribution





Frequency of Encounter

Test type	Approach	Skew	Kurtosis	Shapiro-Wilk
Receptive, oral	Novel	.47	-.11	$W = 89, p = .70$
	Standard	.50	-.51	$W = 86, p = .05$
Productive, oral	Novel	.13	-.23	$W = 93, p = .34$
	Standard	1.50	1.13	$W = 68, p < .001$
Receptive, written	Novel	.30	-.31	$W = 96, p = .75$
	Standard	1.47	1.13	$W = 71, p = .001$
Productive, written	Novel	.40	-.40	$W = 91, p = .21$
	Standard	.10	-.56	$W = 88, p = .08$

Appendix J.1

Checking Assumptions for Mixed-effects Models

Binomial models

Field et al. (2012) recommended checking linearity, independence of errors and multicollinearity. In mixed-effects models ([4.4.4.2](#)) cases of data are not always independent; multicollinearity was also not relevant to these models as they only had one continuous predictor. The assumption of linearity between predictor and logit was checked using the Box-Tidwell test (Leung, 2021) – see [Chapter 5, Results I](#). If the coefficient for the total_LAA_score_log is not significant, it suggests that the linearity assumption holds for total_LAA_score. Overdispersion was also checked to avoid inflating calculations of marginal R^2 values (see [4.4.4.6](#)).

Ordinal models

Proportionality of odds was checked by using the Brant Test (Brant, 1990), to determine whether observed deviations were larger than what could be explained by chance. The main check was the Omnibus measure (values $> .05$ are preferable), with other p -values reported if $\leq .05$, in which case results were interpreted with caution. Where assumptions were not met, this is clearly reported in [Chapter 5, Results I](#) and [Chapter 6, Results II](#), and caution was advised when presenting/interpreting these results. Descriptive statistics including effect sizes and visuals also supported interpretation of the data.

Appendix K.1

Example Lesson Plan: Novel

Wk 3 Timing:	- Describing me & others – Asking questions - Silent Final Consonant (SFC) - s Resources needed: - PPT - Knowledge organisers - Mini WBs / paper or French book for slide 17. - Possible photocopy of slide 16 if needed for some pupils (school to copy)	Teacher support	Support / challenge
Slide 1	for your ref only		
1 min Slide 2: Recap the changes to adjectives covered last lesson	1. Read the grammar explanation about adjective agreement to remind pupils.		
8 mins Slide 3: practise written comprehension of masculine and feminine	Lire! 1. Explain that there are two names for each register entry, a boy and a girl. 2. Pupils have to work out which is the correct name by looking at the form of présent or absent. 3. Do the example (E) (or as many as needed) with the whole class. 4. Click to reveal the answers. Pupils check their own.		Pupils can just write the initial letter of the correct name to save writing the whole name if it's easier. Or some pupils might benefit from having a

adjective agreement			copy of slide 3 they can write on. Allow paired discussion as support.
5 mins Slide 4: practise applying knowledge of SFC (Silent Final Consonant) to known and unknown words	Parler! Time now to practise speaking. We're going to work in pairs. Take it in turns to try to read all the words on the board. You've got 60 seconds – do you think you can do it? What do we need to remember if the word ends in 't'? That's right, we don't pronounce it. Your partner is going to check if you manage to keep silent when it comes to the 't'. If you say it by mistake, you'll need to start again. We haven't met all these French words yet, but you already know how to pronounce the vowels, so I think you'll manage to pronounce these words correctly. 1. Start the timer. Pupil A tries to read aloud all of the words without pronouncing the –t on the end, within the 60 seconds allowed. 2. Pupil B listens. If Pupil A says a –t by mistake, Pupil B can direct him/her to start again from the beginning. 3. They swap roles and repeat the task.	It may be helpful to agree an order for reading the words aloud. Eg top row left to right, then middle, then bottom.	Some pupils might find it helpful to recap the vowel sounds first. You can have children working in 3s if they are not confident, so that two together can act as Partner A and one as Partner B.
5 mins Slide 5 (9 slides) reintroduce the word 'Salut', which has a SFC, and revisit 'Bonjour' practise spoken	1. Click to bring up the teacher saying 'Bonjour'. Accompany this with a circular gesture with left palm facing the class, if desired. 2. Click to bring up the teacher saying 'Salut'. Accompany this with a circular gesture with right palm facing the class, if using gestures. 3. Click to bring up the SFC icon 'shh' and point to the 't' at the end of Salut. 1. Click to bring up either an adult/teacher or a child/friend.	Note: The names have been chosen with	

production of the greetings 'salut' and 'bonjour' with the appropriate level of formality.	2. Pupils say an appropriate greeting. 3. Click for a call out to introduce the two words Madame and Monsieur, on the first and second slides. 4. Click to bring up the person's matching response. 5. Repeat with items 2-8.	familiar SSC so for additional practice in pronunciation, pupils could also say the names.	
10 mins Slides 14-15	Very quickly recap m/f ending on absent / present using Slide 14. Then focus on slide 15 - écouter: 1. Tell pupils that they will just hear 'am' or 'is' in French. They will not hear 'je' or 'il' or 'elle' – instead there will be a beep. 2. Click to listen first time for 'suis' or 'est'. 3. Pupils write 'I' or 'she/he'. 4. Listen again and focus on the adjective form. Is it a boy or a girl? Write B/G. 5. They also write P or A for present or absent. (If necessary they could listen again for this additional detail).	Transcript: 1. [Je] suis présente. 2. [il] est absent. 3. [elle] est absente. 4. [Je] suis présent. 5. [il] est présent. 6. [elle] est présente. 7. [il] est absent. 8. [Je] suis absent. Non, Madame ! [Je] suis présent !	This is not intended as a listening <i>test</i>. Encourage pupils to try the first time on their own, then they can compare answers with a partner before listening again. Refer to knowledge organiser as needed.
5 mins Slide 16: practise aural comprehension of this week's new vocabulary; written comprehension	Écouter – sections A & B 1. Click to listen. Students circle the correct word (if they have the sheet printed). If not, they can use letters A, B,C. Click for answers. 2. Students refer to their listening table, and write the corresponding numbers into the gaps in Section B. 3. Click to reveal answers. Pupils check their own. Section C – in pairs, take it in turns to say words aloud. Section D (if time) – pupils can choose any of the new words.	Transcript: E. présente 1. elle est 2. je suis 3. ici 4. présent 5. là	Ensure you model using the example. Some pupils may find it easier to have a photocopy of this slide to write on.
5-10 mins Slide 17:	Écrire! Section A: Students write the missing verbs on mini-WBs / paper / into French books.	If time is short, you could translate the	Refer to knowledge organiser as needed.

practise written production of this week's two verb forms, and written comprehension of this week's new vocabulary.	Section B: Then they translate the sentences into English. Section C: Finally they write down how they will respond to the register when their own name is called, and about anyone who is not there.	sentences into English together as a class.	
National Curriculum references:	- read carefully and show understanding of words, phrases and simple writing - develop accurate pronunciation and intonation so that others understand when they are reading aloud or using familiar words and phrases - listen attentively to spoken language and show understanding by joining in and responding - write phrases from memory		

Appendix K.2

Example Lesson Plan: Standard

Wk 3 Timing:	Numbers 1-12 Resources needed: - PPT - French booklets	Teacher support	Support / challenge
5 mins Slide 1: introduce numbers 1-10	1. Click on the opening line to play audio – tu as combien de doigts? (how many fingers do you have?) 2. Clicking brings on each number, then clicking again brings on the audio. 3. Get pupils to count on their fingers as shown, listening and repeating. 4. Practise each number several times, using different voices.	Pronunciation comes on when you click on the word.	
2 mins Slide 2: practise numbers 1-10	Let's see if we can remember them ourselves. 1. Count the numbers with pupils. They put their hands up like in the slide and repeat after the teacher. 2. Do this two or three times.		Confident pupils may be able to volunteer the numbers before the teacher.
4 mins Slide 3: familiarise pupils with the written form of numbers	How many fingers are there? We're going to practise again, but this time we're going to look at how the numbers are written, as well as how they sound. 1. Click and listen, repeat together. 2. Isolate and practise the diphthongs shown in pink	Pronunciation on click.	Challenge – which two numbers have the same sound in them? (deux / neuf; six / dix)
4 mins Slide 4:	Teacher sings: 1, 2, 3, 4, to the tune of Frère Jacques Students repeat in song: 1, 2, 3, 4 T: 5, 6, / S: 5,6,		

practise the numbers in a song	T:7, 8, 9 / S: 7,8,9 T:10, 10, 10 /S 10, 10 , 10! Practise a few times.		
3 mins Slide 5: recall and produce numbers in a song	Repeat, this time without the words as prompts. Teacher leads.		A confident pupil could take on the role of teacher. p.7 of the French booklets has a list of numbers in case needed.
4 mins Slide 6: recognise the numbers when spoken	Let's see now if we can recognise the numbers when they're not in order. When I click the blue box, you'll hear a number. Show me on your fingers which number this is. Click on each blue box, from left to right in order, and you will hear the number which matches the next picture that comes up. Then click anywhere else on the screen to bring up the picture so pupils can check that they were holding up the right number of fingers. Click again and the picture disappears. The numbers are in the following order: Trois Dix Six Un Huit Cinq Neuf Sept Quatre Deux	Click on each blue box, from left to right in order, and you will hear the number which matches the next picture that comes up. Then click anywhere else on the screen to bring up the picture so pupils can check that they were holding up the right number of fingers. Click again and the picture disappears. It's easier to just play the audio once and repeat yourself as needed.	This is not meant as a test. Encourage pupils to try this independently first, then they can discuss in pairs.
4 mins	Now let's listen to a real French rhyme that children learn in France and other French-speaking countries. See if you can spot some of the numbers.	The word 'une' here is pronounced as two	

Slide 7: listen to a rhyme using numbers	1. Click on Text ‘Un deux’ and let it play through – you can click each image to come up at the right time, to support understanding, while the audio continues playing. 2. Repeat a few times and allow children to join in if / when they feel comfortable Translation: 1,2 1,2,3 That makes one hand. 1,2,3 4,5 That makes five fingers. 6,7,8,9,10 That makes two hands Who say hello to each other.	syllables, simply because of the rhyme (it’s normally one).	
3 mins Slide 8: say numbers out of order	Will you be able to say the numbers when they’re not in order? Let’s see. 1. Pupils should say (chorally) which number of finger (according to slide 1) is being pointed at. The arrow only flashes once and then disappears so pupils have to stay focused to see which finger is being pointed at.	The instruction has audio, but here pupils need to just respond with the correct number, so there is no audio for each number on this slide.	It might be helpful to recap all numbers in order before starting. Some pupils might find it helpful to have a prompt each time if they are really struggling (e.g., if it’s #4 being elicited, you could count un, deux, trois), but get them to try independently first. Allow paired discussion before taking answers.

1 min Slide 9: introduce 11 & 12	1. Click to bring on the written form and the digits. 2. listen and repeat a few times in different voices.	The audio plays once both numbers are on screen.	
3 mins Slide 10: recognise numbers in the written form	1. Pupils say the number as it appears (on click) – first in French, then in English. 2. Do this a couple of times to build confidence.		You can either do this as a whole class or have individuals lead (volunteers).
3 mins Slide 11: use the numbers in a context	Now let's see if we can say together in French how many candles there are on each cake. 1. Repeat the known numbers and recap eleven and twelve 2. Click on the text box to hear numbers 11 and 12 pronounced again.		
3 mins Slide 12: say the numbers in random order	As you click, the cakes disappear one by one and pupils have to say the number in French.		Allow paired discussion before taking answers.
6 mins (6 slides) Slides 13-18: recognise numbers within a sentence; introduce 'il y a'	Slide 13 : 1. Click on the slide title: Il y a combien de chiens? Picture of one dog appears. 2. Let's find out how many animals there are. Which animals do you think this slide is talking about? 3. Click to reveal the phrase. How many dogs are there? There is one dog = il y a un chien. 4. Click to reveal the second dog. How many dogs are there now? Il y a deux chiens = there are two dogs. 5. Listen and repeat phrases. As pupils become more confident, they can volunteer answers and you click to check.	Il y a = there is / are (it doesn't change in the plural). The 'ch' in French is pronounced in the same way as the 'sh' in English.	More confident pupils will enjoy predicting answers. Some will manage to say the whole phrase.

		'ien' – chien rhymes with 'bien' which pupils have met already.	
National Curriculum references:	- listen attentively to spoken language and show understanding by joining in and responding - explore the patterns and sounds of language through songs and rhymes and link the spelling, sound and meaning of words - develop accurate pronunciation and intonation so that others understand when they are reading aloud or using familiar words and phrases - appreciate stories, songs, poems and rhymes in the language - read carefully and show understanding of words		

Appendix K.3

Observed Lessons and Monitored Audio-recorded Lessons per Teacher

	Teacher 1	Teacher 2	Teacher 3 (School A/C/E)	Teacher 4	Teacher 7	Teacher 8	Teacher 9	Total
Number of Novel lessons observed	2	2	1 + 2 + 2	1	1	1	1	13
Number of Standard lessons observed	1	1	2 + 1 + 1	0	2	2	2	12
Number of Novel audio- recorded lessons monitored	5	4	5 + 5 + 5	2	5	4	7	42
Number of Standard audio- recorded lessons monitored	5	6	5 + 5 + 5	1	5	6	3	41

Appendix K.4

Observation Chart

Teacher ID:

School:

Observation #:

Date:

Approach:

Week:

Class info:

	5 mins	10 mins	15 mins	20 mins	25 mins	30 mins	35 mins	40 mins	45 mins
Consistent with approach intent									
Deviation from Lesson Plan									
Reminders given re behaviour / concentration									
Timing - behind + ahead									
Pronunciation issues/Teacher errors									
Additional comments									

VERY MINOR (and can therefore be disregarded) – slight deviation from the Lesson Plan activity which does not change the focus of the activity, and which does not introduce any element which is more consistent with the other approach. May involve missing out or changing a small element of an activity or practice which is unlikely to impact significantly on pupil progress (for example doing something as a whole class activity instead of individually when not suggested by the plan).

MINOR – slight deviation which, although it may introduce an element which is consistent with the other approach, reinforces or is consistent with what has already been taught in other parts of this lesson or in previous lessons in this approach. May also involve minor omission (e.g., missing out a significant part of practice, missing out part of a focus – e.g., one or two slides from a series of 4+).

DEVIATION – teacher draws attention to/includes something which was not part of the plan OR has not been part of previous lessons OR which is more consistent with the other approach. More serious omission.

Appendix K.5

Calculations for Lessons Delivered as Intended

Novel:

55 lessons were monitored = 495 possible measures, if nine available per lesson.

483 measures were observed (out of possible total of 495).

483 divided by 55 = an average of 8.78 measures per lesson.

In total, 150 Novel lessons were available. Multiplying 150 by the average number of measures per lesson gives a possible total of 1317 measures.

We know that in Novel, 92.34% of monitored measures were delivered as intended.

If we apply this to the total number of measures available, we could expect that 92.34% of 1317 (= 1216.12 measures) would indicate delivery as intended.

Dividing the total measures possible (1317) by the total number of lessons (150) gives the average number of measures per lesson (8.78).

Dividing the likely number of measures delivered as intended (1216.12) by the average number of measures per lesson (8.78) tells us the number of lessons that would likely be delivered as intended in Novel approach (138.51 out of 150).

Standard:

53 lessons were monitored = 477 possible measures, if 9 available per lesson.

460 measures were observed (out of possible total of 477).

460 divided by 53 = an average of 8.68 measures per lesson.

In total, 150 Standard lessons were available. Multiplying 150 by the average number of measures per lesson gives a possible total of 1302 measures.

We know that in Standard, 90.87% of monitored measures were delivered as intended.

If we apply this to the total number of measures available, we could expect that 90.87% of 1302 (= 1183.13 measures) would indicate delivery as intended.

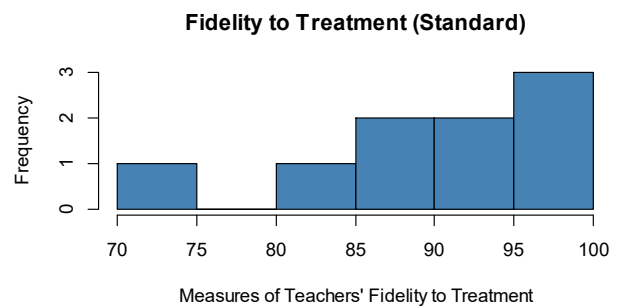
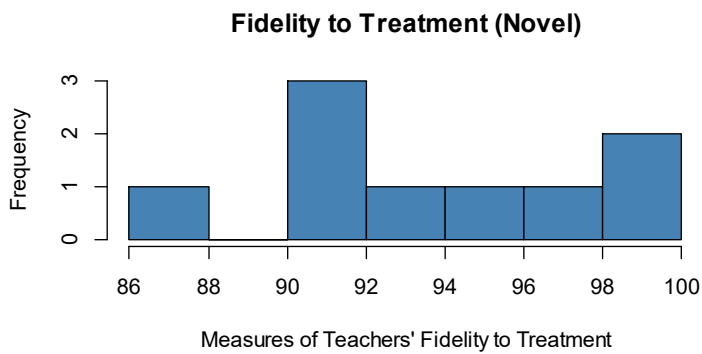
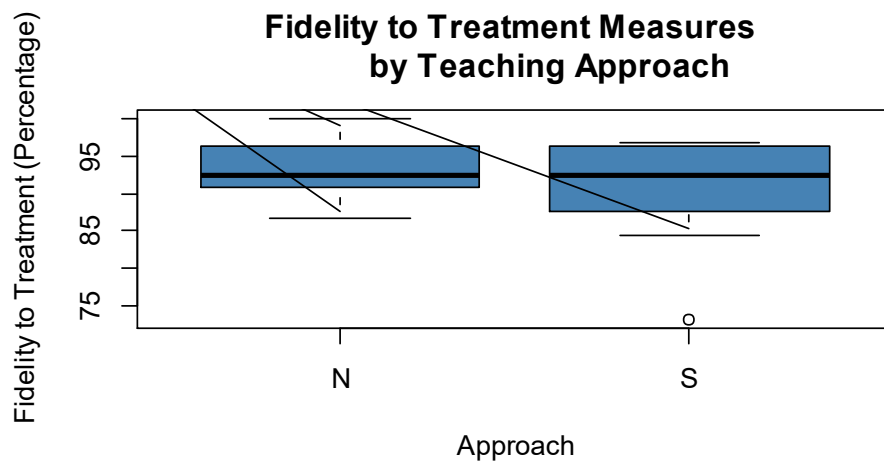
Dividing the total measures possible (1302) by the total number of lessons (150) gives the average number of measures per lesson (8.68).

Dividing the likely number of measures delivered as intended (1183.13) by the average number of measures per lesson (8.68) tells us the number of lessons that would likely be delivered as intended in Standard approach (136.31 out of 150).

Appendix K.6

Fidelity to Treatment: Descriptive Statistics

	<i>Mean</i>	<i>Standard Deviation</i>
Novel approach	93.6	4.3
Standard approach	90.0	7.6



Appendix K.7

Descriptive Deviations Chart

Novel					Standard				
		Lesson	Minor Deviations	Deviations			Lesson	Minor Deviations	Deviations
School A	Teacher 3	4		aT	School A	Teacher 3	3	e	
		6		aT, aT			9		
		13					16		
		20	a				17		
		26	b				22	aT	
		29	a				23		
	Teacher 4	3					28		
		7		aC, aC		Teacher 4	10	a	
		9							
School B	Teacher 1	2			School B	Teacher 1	6		
		7		aT			16		
		9					18	a	
		13	a, c				22	a, a	
		19	aT				24	a	
		27	a, c				26		
		29				Teacher 2	3	f, g	
	Teacher 2	6		aT			7	f, a	
		14	b				11	a, g	
		18	a, c				17	aT, a, a	
		20	c				23	a	
		24	a				27		aT
		30	a				29	a	
School C	Teacher 3	3			School C	Teacher 3	9		
		6					15	e	
		7	a	aT			18		
		12	aT				19		
		22					26		
		23					30	aT	
		28							
School E	Teacher 3	3	c		School E	Teacher 3	6	h, a	
		7					13	e	
		10					22	aT	
		16					26		
		18					28		
		19					30		
		23				Teacher 7	3	e	
	Teacher 7	6	aT				10		
		11					18		
		15					21	a	
		20					23		
		26		aT			27		
		28					29		
School F	Teacher 8	6			School F	Teacher 8	3		
		13					5	d, b	
		15					10	c	
		25					16	c	
		29					18	a, c	
	Teacher 9	5		aT			20		a
		10					26		
		16					28		
		18	d			Teacher 9	4	c, i, d, d,	
		24		aT			13	a, e, c, g	

		26	a	
		28		
		30		aT

		21	c, c, e	
		25		
		29	a	

Minor Deviation: a slight deviation which, although it may introduce an element which is usually consistent with the other approach, reinforces or is consistent with what has already been taught in other parts of this lesson or in previous lessons in this approach. It may also involve minor omission (e.g., missing out a significant part of practice, missing out part of a focus (e.g., one or two slides from a series of 4+).

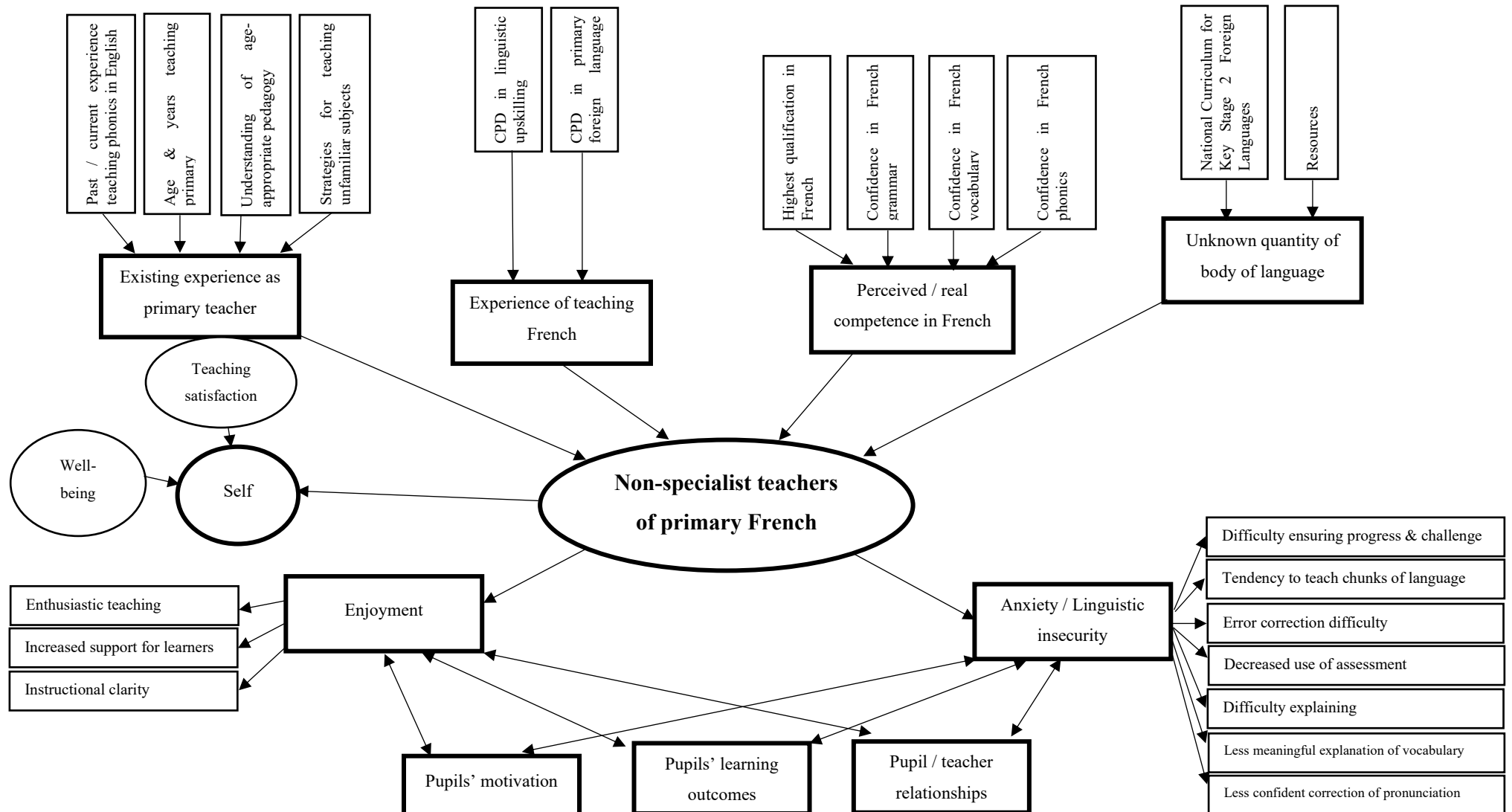
Deviation: Teacher draws significant attention to / includes something which was not part of the lesson plan OR has not been part of previous lessons OR which is more consistent with the other approach. A more serious omission.

Categories:

- a) Omission of / reduction in practice / activity
- aT) Omission of / reduction in practice / activity due to time
- aC) Omission of /reduction in practice / activity due to circumstances beyond teachers' control
- b) Omission of feedback
- c) Change in nature of activity
- d) Lack of clarity in explanation of pronunciation
- e) Unintended focus on pronunciation / grapheme-phoneme correspondences
- f) Lesson rushed
- g) Lesson finished early
- h) Unintended language analysis
- i) Lack of clarity / content in explanation of activity

Appendix L.1

A First-Draft Conceptual Framework Showing Identified Influences on Enjoyment and Anxiety of Non-specialist Teachers of Primary French



Appendix L.2

Sample Interview 1 Emotions Grid

Transcript	Paraphrase (& context)	Emotion identified	Cause of emotion
T7: Erm, OK. I, I felt quite kind of in the dark, erm, when I first had to teach it because when I qualified as a teacher, erm, we weren't teaching French in primary. So it was kind of put on me and I felt a bit, er, well, very nervous about it.	Felt daunted when asked to teach French for the first time	Nervousness / insecurity	Lack of prior experience / training / preparation
T7: Erm, so, I used a lot of that essentially and, erm, where possible to begin with, I tried to use, erm, the sound clips and things that are, that were available on there, as much as possible, just because of my own confidence really. Erm, I was very nervous – still am really about pronouncing things incorrectly, or, you know, saying something wrong. That was always my biggest worry. So, to begin with, when I first started teaching French, definitely, I would try to rely on videos or sound clips, and PowerPoints, where it minimised the amount of, erm, talking that I'd have to do.	Refers to resources such as sound clips which supported them with pronunciation or helped them avoid having to speak in front of the class	Insecurity Nervousness	Lack of confidence in ability to pronounce correctly – inadequate subject knowledge Lack of confidence speaking in front of the class / inadequate subject knowledge
Y7: I think it definitely helped to build my confidence, erm, to some extent in that I was seeing, err, the French being spoken on	Getting support from videos increased their confidence and made them worry less	Reassurance Increased sense of security	Support from resources

videos so I could build my confidence over the years a bit more about how to pronounce words, or maybe just worry less about if I pronounce this slightly wrong. [laughs quietly]			Greater teaching experience over the years (NB: this would link to Driscoll, Jones and Macrory, 2004, which found that class teachers developed confidence with language teaching experience)

Appendix L.3

Example of First-stage Coding Taken from Pupil Questionnaire 1 Open-Ended Question

Pupil starting point enjoyment

Nodes\\first stage coding

Name	Description
anticipates difficulty	Thinks they will find French hard
confidence	
excitement	
good at other languages, existing knowledge of other languages - transferable skills	Either has L1 in a language other than English or has learnt a different FL and feels this will be relevant to French
interest	
lack of prior knowledge or language learning experience	Has not learnt French or any other language before
likes French	
looking forward to it	
not good at French or other languages	Has experienced failure when learning French or other language
novelty	Attracted to idea of doing something new
positive about languages	

Name	Description
positive about learning	Likes idea of learning generally, not French specifically, likes learning 'new things'
previous contact with the country	Has visited France or French-speaking country
prior knowledge	Has learnt or picked up some French previously
ref to family and friends	Has relatives or friends learning French (or other language); has some contact with French friends
scared	
unsure	Doesn't know how they feel about learning French

Appendix L.4

Example of Code Merging Taken from Pupil Questionnaire 1 Open-Ended Question

Pupil interviews end

Nodes\\Second stage coding N

Name	Description
confidence	
can work without support	
confidence - gets things right	
OK to make mistakes	Feels that making mistakes is part of the learning process
pride, sense of achievement	
using it outside school	i.e., feels secure enough to want to teach another
wants to show off their French	Takes opportunities to volunteer words in French
curiosity	Shows interest in how the language works or in the culture or people
dislikes	
affected by behaviour	Refers to the negative impact of other pupils' poor behaviour in French lessons

Name	Description
affected by fatigue	Refers to being tired in French or finding lessons tiring
boring	Explicitly refers to French as boring
dislikes - classroom factors	For example, heat in the classroom
dislikes forgetting	Not being able to remember French words has a negative impact on their enjoyment
dislikes not knowing words	
dislikes reading aloud	
dislikes repeating words	Dislikes copying words said by the teacher or by the French native speaker
dislikes revisiting	Dislikes going back over language
dislikes spending too long on French	Refers to lessons lasting too long, or particular activities being long
dislikes time-sensitive activities	Dislikes the timed activities
dislikes vocab sheets	This refers to the vocabulary grids
fear	Feeling scared about an aspect of French
feels ambivalent	Feels 'in the middle'; non-specific positive and negative expressed
general progress	

Name	Description
progress - actions	Refers to having made progress in doing actions for French words
progress - communication	
progress - games	Can play French games better than at start
progress - knowledge about France	
progress - pronunciation	
progress - reading stories	
progress - topic - animals	
progress - topic - carnival	
progress - topic - days of week	
progress - topic - food	
progress - words and sentences	

Name	Description
integrative motivation	Being able to communicate with/understand people from France, use the language on holidays there, make friends while there;
lacks confidence	
likes	
complete satisfaction	Likes everything about French
enjoys it	
fun	
interest in culture	
likes - animals	Likes learning about animals
likes - pacperson	Reference to specific activity
likes a bit	
likes a lot	
likes activities	General reference to activities
likes games	
likes grammar activity	
likes guessing the hiding place activity	Reference to specific activity

Name	Description
likes learning new language(s)	
likes learning words	
likes maze activity	Reference to specific activity
likes Menton festival	
likes pronouncing words, sounds	
likes spelling	Learning how to spell words
likes teacher	
likes the sound of words	
likes vocab grids	
loves	Language used conveys stronger emotion than liking (e.g., it's fabulous, I love it)
links to visiting the country	Refers to actual past visits or potential future ones, or just the general concept of visiting France
no contact outside school	Cannot practise French outside school
perceptions of future progress	Describes what they think they will be able to do in Year 4
future - alphabet	
future - animals	

Name	Description
future - better pronunciation	
future - communicate	
future - general improvement	Doesn't refer to anything specific ('will get/be better')
future - know more words	
future - numbers	
future - overall progress	
future - writing	
personal connections	
ref to difficulty	
teacher - enjoyment	Whether they think the teacher enjoys teaching them
teacher - classroom management, pupil behaviour	Pupil perceptions of how teacher is impacted by class behaviour
teacher - enjoyment of pupil progress or effort	Refers to how the teacher seems to like seeing pupils make progress/put effort into French
teacher - fun activities or games	Teacher appears to derive enjoyment from fun activities

Name	Description
teacher - lack of enjoyment	Mentions the teacher does not appear to enjoy French
teacher - positive relationships	Refers to impact of positive relationships with pupils on teacher
teacher encouragement	Refers to how teacher encourages them
teacher fatigue	Refers to teacher seeming tired
teacher knowledge	Any reference to teacher subject knowledge
tries hard	Pupils refers to own effort learning French
what helps	Pupils describes strategies/aspects of approach which they find helpful
support from peers	
support - general	
support from family	
support from Knowledge Organiser	
support from PowerPoint	
support from repeating	
support from revisiting	

Name	Description
support from sounding out words	
support from teacher	
support from video	
support from writing	
visual support helps	

Appendix L.5

Examples of Coding Types Taken from Teacher Questionnaire 2 (Mid-project)

- **descriptive coding (identifying basic topics), used as a starting point**

Examples:

“Lack of experience teaching French”

“Past experience learning French”

“Planning/preparation”

- **concept coding, going beyond what is immediately apparent (particularly relevant if aiming for generalisability)**

Examples:

- “doing things right myself”

- “doing things the right way”

- “doing right by learners”

- “feeling out of depth”

- **sub-coding, amplifying original entries**

Example:

“Pronunciation:”

- ‘insecure’, ‘incompetent’

- ‘reliant’

- ‘worried’, ‘nervous’, ‘anxious’

- **simultaneous coding, using multiple codes for the same section of text (cf. Braun & Clark, 2006)**

Example:

Teacher 1

I: OK. And do you feel equally confident in both approaches explaining vocab?

T1: I think so, yes.

I: Yeah. OK. What about correcting children's pronunciation? How do you feel about doing that now?

T1: Yeah, not as confident as I would like to be.

I: OK.

T1: Erm, I think it's, it's OK if it's something that you've just literally heard the audio say and you've said it.

I: Yeah.

T1: But again, if you have to go back to it, that's sometimes when I think, "Oh, I hope I've done it right".
I: Yeah. And do-do you think that's the same in both? You feel equally confident in both?
T1: Yeah. Yeah, I do. Yeah.

Coded as:

"Feeling insecure, incompetent"

"(Lack of) confidence correcting pupils' pronunciation"

"No difference in confidence in either approach"

"Pronunciation"

"Confidence"

- **emotion coding, an affective method helpful when participants related experiences**

Examples:

"Worry"

"Enthusiasm"

"Pleasure"

"Tension"

- **In Vivo coding, using participants' own wording**

Example:

"Feeling out of my depth now"

Appendix L.6

Interview Analysis – Taken from the Literature Review on Specialist Teaching and Used to Inform Coding

- Teachers feel like they are learning alongside pupils (Myles, 2017)
- Influence of teacher on young learners (Enever, 2011; Mihaljević Djigunović & Vilke, 2000; Muñoz, 2014a; Nikolov & Mihaljević Djigunović, 2006)
- Linguistic competence issues of non-specialist teachers:
 - o Oral fluency, pronunciation and intonation (Cable et al., 2010; DCSF, 2005; Moon, 2009; Muijs et al., 2005; OFSTED, 2011)
 - o Grammatical accuracy (Cable et al., 2010; Muijs et al., 2005)
 - o Difficulty ensuring progress (Cable et al., 2010; Driscoll, 1999; Graham et al., 2017; McLachlan, 2009; Muijs et al., 2005)
 - o Teaching chunks rather than promoting grammatical / syntactical understanding (Cable et al., 2010; Nguyen, 2018)
 - o Supporting/challenging pupils (Cable et al., 2010; Driscoll, 1999; Muijs et al., 2005)
 - o Decreased likelihood of error correction (Cable et al., 2010; Muijs et al., 2005; H. Richards et al., 2013)
- Focus on speaking and listening rather than reading and writing (Driscoll, 1999)
- Overuse of games (Muijs et al., 2005)
- Pedagogy seen as different from other areas (Finch et al., 2018)
- Teacher confidence:
 - o Affected by linguistic proficiency (Barnes, 2006; Finch et al., 2018; Muijs et al., 2005; Woodgate-Jones, 2008)
 - o Impacts on lesson pace (Muijs et al., 2005)
 - o Impacts on teacher and pupil motivation (Barton et al., 2009; Crichton & Templeton, 2010)
 - o Can lead to teacher reluctance (Tinsley & Board, 2016)
- [Negative] impact on pupil outcomes:
 - o Writing (Board & Tinsley, 2014, 2015; Cable et al., 2010; Tinsley & Board, 2012)
 - o Speaking – poor model of pronunciation and intonation provided (Larson-Hall, 2008; Muñoz, 2009; Tinsley & Board, 2017).
 - o Vocabulary acquisition (Graham et al., 2014; Szpotowicz, 2009; Tsui, 2003; Unsworth et al., 2015)
 - o Grammar (Board & Tinsley, 2014, 2015; Tinsley & Board, 2012; Unsworth et al., 2015)
- Continuing Professional Development:
 - o In linguistic upskilling (Board & Tinsley, 2015; Cable et al., 2010; Finch et al., 2018; Legg, 2013; OFSTED, 2005)
 - o Lack of accessible training (Board & Tinsley, 2015; Cable et al., 2010; Holmes & Myles, 2019; Tinsley & Board, 2016; Tinsley & Doležal, 2018)
 - o Lack of training in specific areas (SSCs, assessment, planning for progression, grammatical understanding, relationship between FL and L1 literacy) (Holmes &

- Myles, 2019; Moon, 2009; Muijs et al., 2005; OFSTED, 2005; Powell et al., 2001)
 - Positive impact of Continuing Professional Development/training (Rowe et al., 2012)
 - Positive impact of experience teaching the foreign language (Cable et al., 2010; Cooke & Faez, 2018; Driscoll, 1999; Driscoll et al., 2004; Muijs et al., 2005; Rowe et al., 2012)
- Resources:
 - Need for video/CDs (DCSF, 2005; Driscoll, Jones, Martin, et al., 2004; OFSTED, 2005)
 - Need for accurate model of spoken language (OFSTED, 2011; Tinsley & Board, 2016)
- Linguistic insecurity – anxiety, tension, nervous, worried, uneasy
- Non-specialist teachers’ physical responses
- Non-specialist teachers’ emotional responses
- Effect on delivery:
 - Impact on ability to explain accurately (T. Farrell & Richards, 2007; H. Richards et al., 2013; Tsui, 2003)
 - Self-efficacy – importance of perception of ability (Dörnyei & Csizér, 1998; Horwitz, 1996; Pattison, 2014; Sharpe, 1995)
 - Only low level of foreign language required for Y3 (T. Farrell & Richards, 2007)
 - Non-specialist teachers’ reliance on more experienced/competent colleagues (Du Plessis et al., 2015)
 - Less positive teacher/pupil relationships (Du Plessis, 2015)
 - Anxiety arising from (i) grammar (Macrory and McLachlan, 2009) and (ii) speaking (Amengual Pizarro, 2019; Horwitz, 1996; Machida, 2016)
- Provision of schemes of work and materials seen as supportive/essential (Driscoll, Jones, Martin, et al., 2004; Kasprowicz, 2025; Muijs et al., 2005).

Appendix L.7

Example of Advanced Coding Taken from Teacher Interview 2 – Developing and Naming Themes

Teacher mid-project

Nodes\\third stage\\negative emotions

Name	Description
doing it right myself	Teacher worries about making mistakes generally, or being corrected by a pupil who knows more than they do
doing it the right way	Teacher worries about ensuring that the lesson is delivered the way it was intended (being involved in the project adds pressure)
doing right by them	Teacher worries about letting learners down in some way
managing the class	Teacher expresses concerns about not managing the class effectively
out of my depth now	Teacher feels insecure
tension	Expresses tension related to teaching or preparing French
worry	Teaching or preparing French causes them to worry

Appendix L.8

Thoughts on Teacher Interview 1 – Example of Researcher Memo

Name: Thoughts on teacher pre-project info

Created On: 15/08/2024 17:10:08

Created By: EM

Modified On: 20/08/2024 15:00:28

Modified By: EM

Size: 3 KB

15.08.24 thinking that 'ready for a change' is a possible theme emerging from how teachers feel about teaching the Novel approach. Even for teachers who haven't taught much/any French before in primary school, there is a sense of frustration that we haven't got things right in this country, there's a need to try something different.

I: Do you feel the same, do you feel differently?

T3: Well, I've only ever really taught French via topic.

I: Yeah.

T3: Erm, you know, it's always been, you know, transport or ... you know ...

I: ((overlaps)) Yeah, yeah.

T3: ... animals or whatever and that has always seemed like a good way of doing it.

I: Yeah.

T3: I must admit there have been times when I have taught that way where I'm thinking, 'I want them to be able to say so... [corrects], a sentence', which would require a verb, and I haven't, that hasn't been part of the scheme [laughs]. I do sometimes do that, but erm, I think I, I wonder whether with the new approach the sentence level will, will progress more quickly.

I: Ok.

T3: That's what I'm wondering...

There is also the appeal of trying something new generally:

T9: I feel quite excited about the Novel approach ...

I: OK.

T9: ... just because, yeah, I like, it's, it's a shiny new thing! I like ...
[laughter]

I: I love that!

- T9: It's a very shallow reason.
- I: Yeah, yeah.
- T9: Pedagogically, why it's, that kind, it's, that's interesting, because I, I, you know, I'm constantly changing my tables around in the classroom, for example – this was in a big horseshoe with two bits in the middle of it - just to see what works for different situations.
- I: Yeah.
- T9: So, from that point of view ...
- I: [interjects] Yeah.
- T9: That's especially interesting, yeah.

T9 also refers to this novel approach as a chance to see things differently:

Erm, so, it's quite, it's quite difficult because you get used to being introduced to language in a fairly standard way all of the time, so the opportunity to break out from that ...

However, the very newness means that it is daunting.

- I: Do you feel the same about both, or do you feel differently?
- T3: Erm ... I think I feel perhaps a, if anything, a little bit more daunted about the Novel approach, and I think that's probably because I've only been used to the topic-based approach.
- T7: Erm, yeah, I mean I guess the topic-driven one is something that I've come across before so I guess maybe I would feel more comfortable with that but yeah, I feel quite open-minded actually about them both, I would say.

19.08.24 Started going through the transcripts again with a view to really analysing emotions experienced at different times by teachers. (see word doc: Teachers' emotions). Naming emotions is helpful. Also, thinking about what the cause might be. Could potentially consider whether this is something related to things outside/within the teacher's own control.

20.08.24 Thinking about a theme emerging relating to 'comfort zone'.

Linked Item

Nodes\\Second stage coding\\appeal of each approach\\novel is new and different

Appendix L.9

Example of Concept Table from Group Interview Data Showing Impact of Approach on Pupil Confidence at Mid-point: Novel

Confidence Category	Supporting Participant Quotes	Analytic Ability	School
Very low	<i>I'm bad at it ... Because I can only remember one word.</i>	Low	B
	[What do you think you'll be able to do in Year 4?] <i>Nothing.</i>	Low	E
Low	<i>I only know a few words ... not much at all.</i>	Medium	B
	[Do you think you're good at learning French?] <i>I'm not sure about that.</i>	Low	E
Medium	<i>I think I understand ... French a little bit more, because sometimes I just get the words mixed up, so ... now that I've done a little bit of more French, I think I can guess at them.</i>	Medium	E
High	<i>I can pronounce words much better than I used to, but also, I can say some sentences.</i>	Low	B
Very high	<i>I think I'm amazing at it.</i>	High	A
	<i>I can think I'm a hundred per cent.</i>	Medium	B

Appendix L.10

Analytic Memo (NVivo, 21.03.24)

Does one approach make pupils feel more inclined to envision a time when they might use it?

This might suggest the idea of higher-frequency language being more useful.

Appendix M.1

Normality of Distribution for Vocabulary Recognition in the Oral Modality

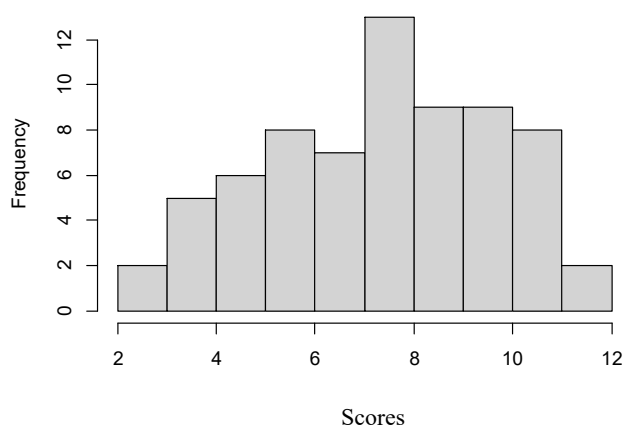
Normality of raw data distribution was checked using visual inspections as recommended by Larson-Hall (2016), as well as through the Shapiro-Wilk test. In the vocabulary recognition post-test in the oral modality, data were treated as normally distributed (Novel: $W = .97, p = .06$; Standard: $W = .96, p = .03$) (see below in [Appendix M.2](#)). Although the p -value was $< .05$ in the Shapiro-Wilk test, it was acknowledged that this test can generate significant results from small deviations from normality in larger data sets (Field et al., 2012) and visual inspection of the data supported a judgement of normal distribution; skew and kurtosis measures were acceptable. A similar approach was adopted in the vocabulary recognition delayed post-test in the oral modality (Novel: $W = .97, p = .07$; Standard: $W = .95, p = .004$) (see [Appendix M.3](#)).

Return to [Table 19](#).

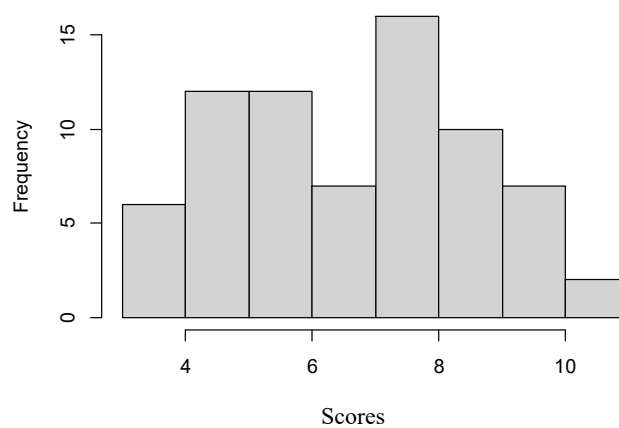
Appendix M.2

Normality of Distribution Checks of Post-test Receptive Tests: Vocabulary Recognition in the Oral Modality (raw data)

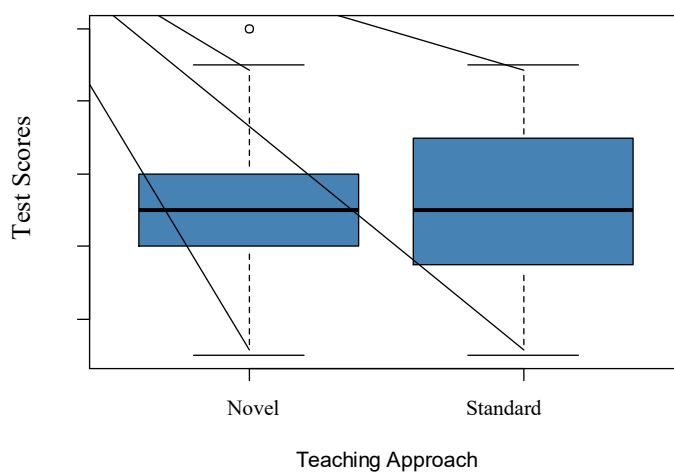
Post-test Vocabulary Recognition in the Oral Modality Scores: Novel



Post-test Vocabulary Recognition in the Oral Modality Scores: Standard



Post-test Vocabulary Recognition in the Oral Modality Scores by Teaching Approach

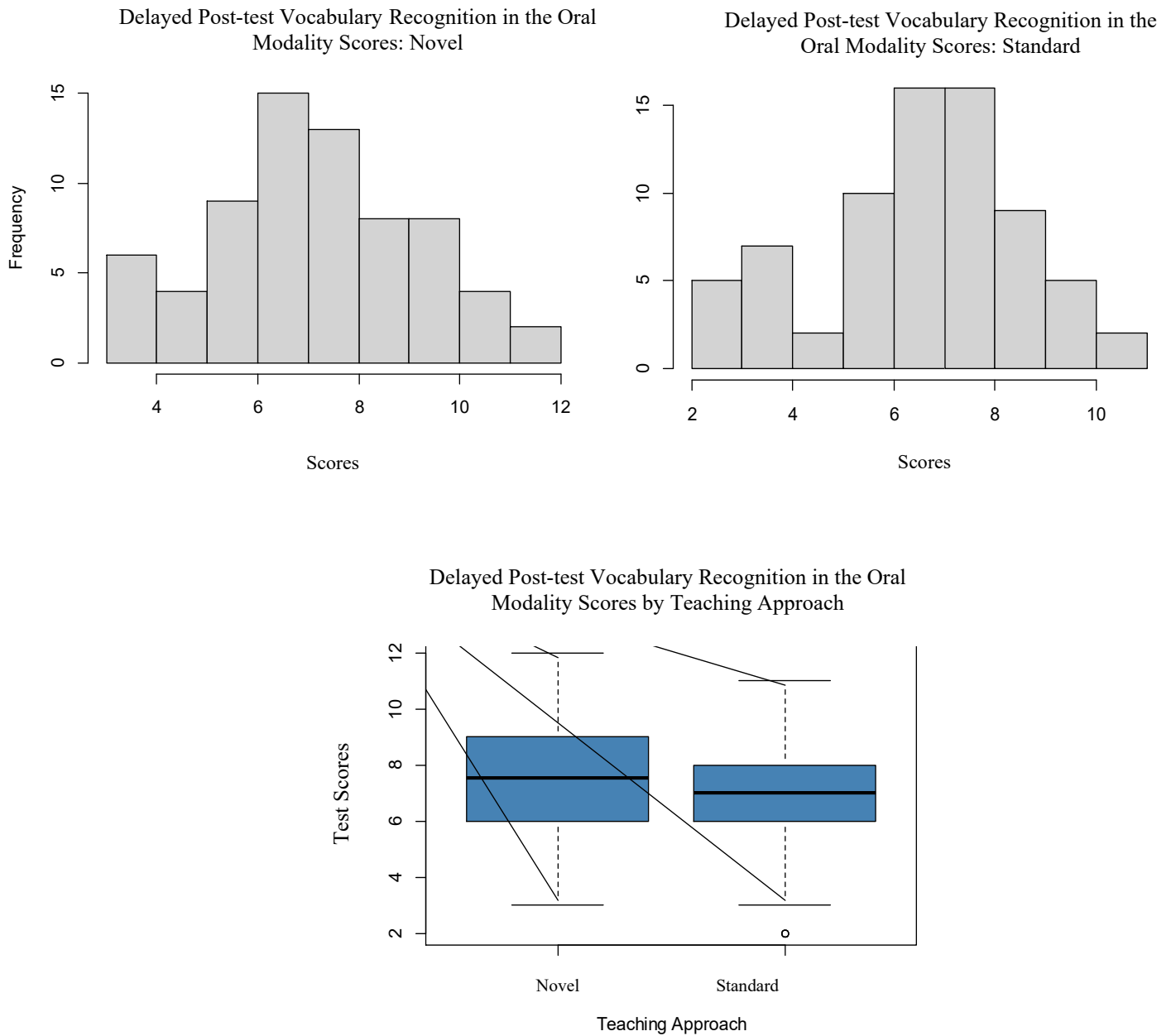


Approach	Vocabulary Recognition in the Oral Modality: Post-test	
	Skew	Kurtosis
Novel	-.46	-.66
Standard	-.13	-.76

Return to [Table 19](#).

Appendix M.3

Normality of Distribution Checks of Delayed Post-test Receptive Tests: Vocabulary Recognition in the Oral Modality (raw data)



Approach	Vocabulary Recognition in the Oral Modality: Delayed Post-test	
	Skew	Kurtosis
Novel	-.24	-.29
Standard	-.92	-.16

Return to [Table 19](#).

Appendix M.4

Normality of Distribution for Vocabulary Recognition in the Written Modality

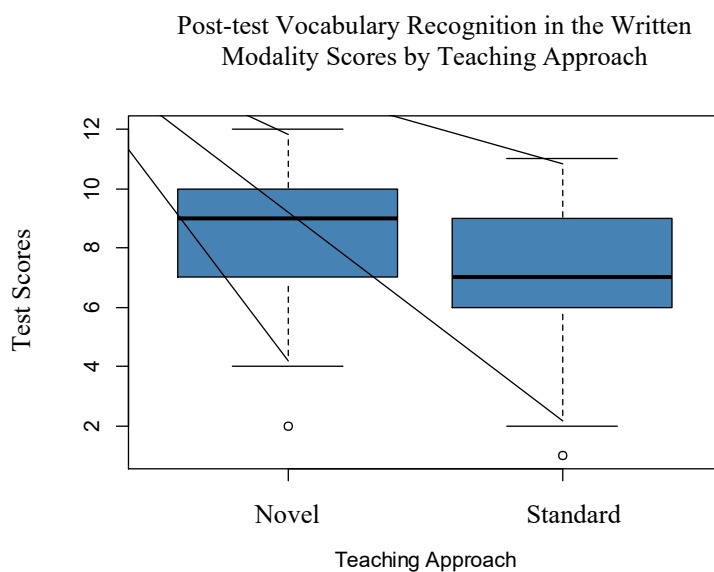
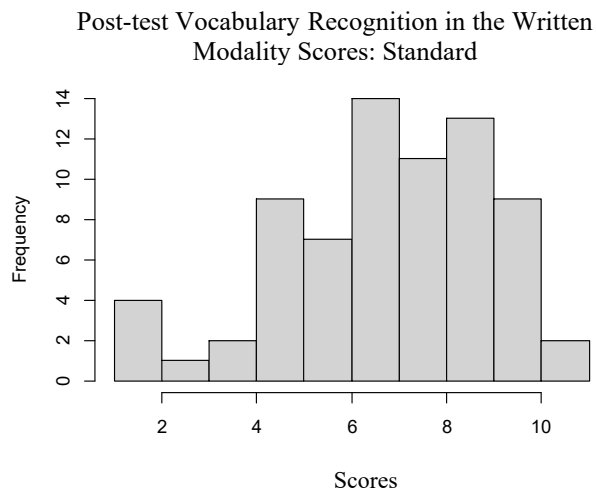
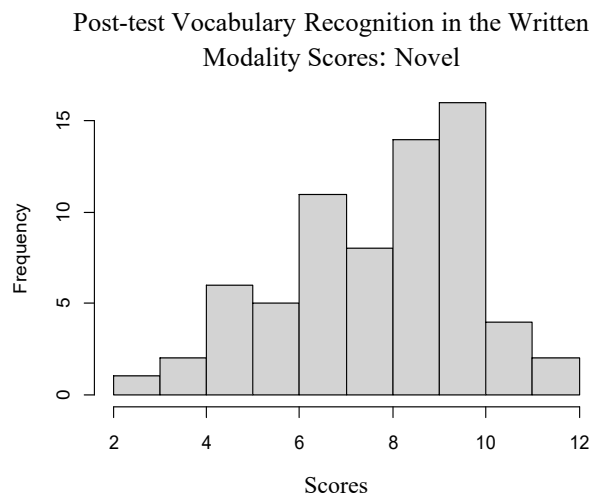
In the vocabulary recognition post-test in the written modality, the decision not to treat data as normally distributed was supported both by visual inspections of histograms and boxplots (see [Appendix M.5](#)) and by the outcome of the Shapiro-Wilk test (Novel: $W = .95, p = .01$; Standard: $W = .94, p = .003$).

At delayed post-test, the Shapiro-Wilk test (Novel: $W = .97, p = .06$; Standard: $W = .98, p = .21$) indicated that data were normally distributed and this was supported by visual inspections of the data (see [Appendix M.6](#)).

Return to [Table 19](#).

Appendix M.5

Normality of Distribution Checks of Post-test Vocabulary Recognition in the Written Modality Tests (raw data)



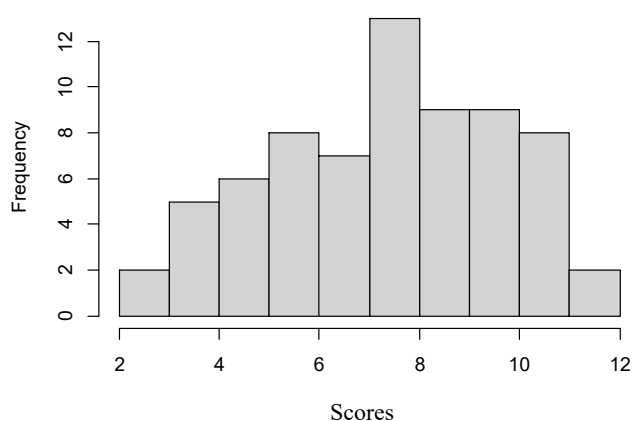
Approach	Vocabulary Recognition in the Written Modality: Post-test	
	Skew	Kurtosis
Novel	-.93	-.17
Standard	-1.12	< .001

Return to [Table 19](#).

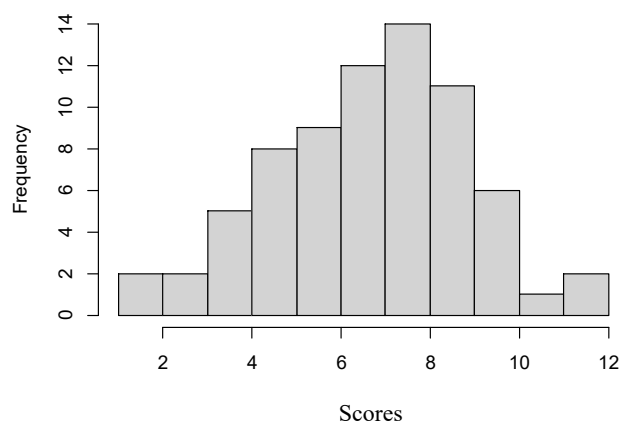
Appendix M.6

Normality of Distribution Checks of Delayed Post-test Vocabulary Recognition in the Written Modality (raw data)

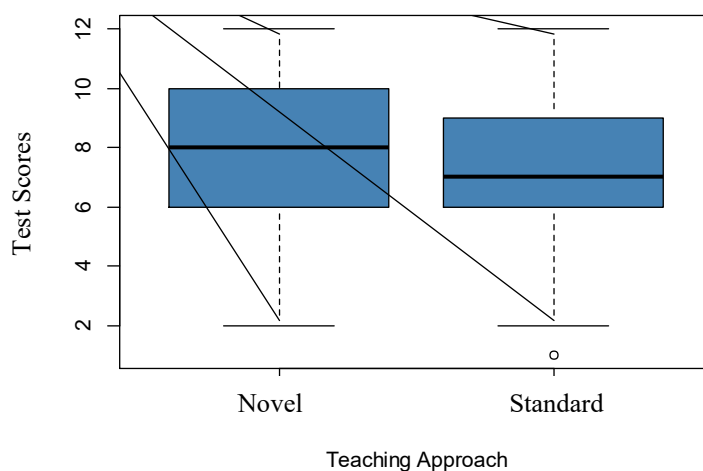
Delayed Post-test Vocabulary Recognition in the Written Modality Scores: Novel



Delayed Post-test Vocabulary Recognition in the Written Modality Scores: Standard



Delayed Post-test Vocabulary Recognition in the Written Modality Scores by Teaching Approach



Approach	Vocabulary Recognition in the Written Modality: Delayed Post-test	
	Skew	Kurtosis
Novel	-.46	-.66
Standard	-.46	-.08

Return to [Table 19](#).

Appendix M.7

Normality of Distribution for Vocabulary Recall in the Written Modality

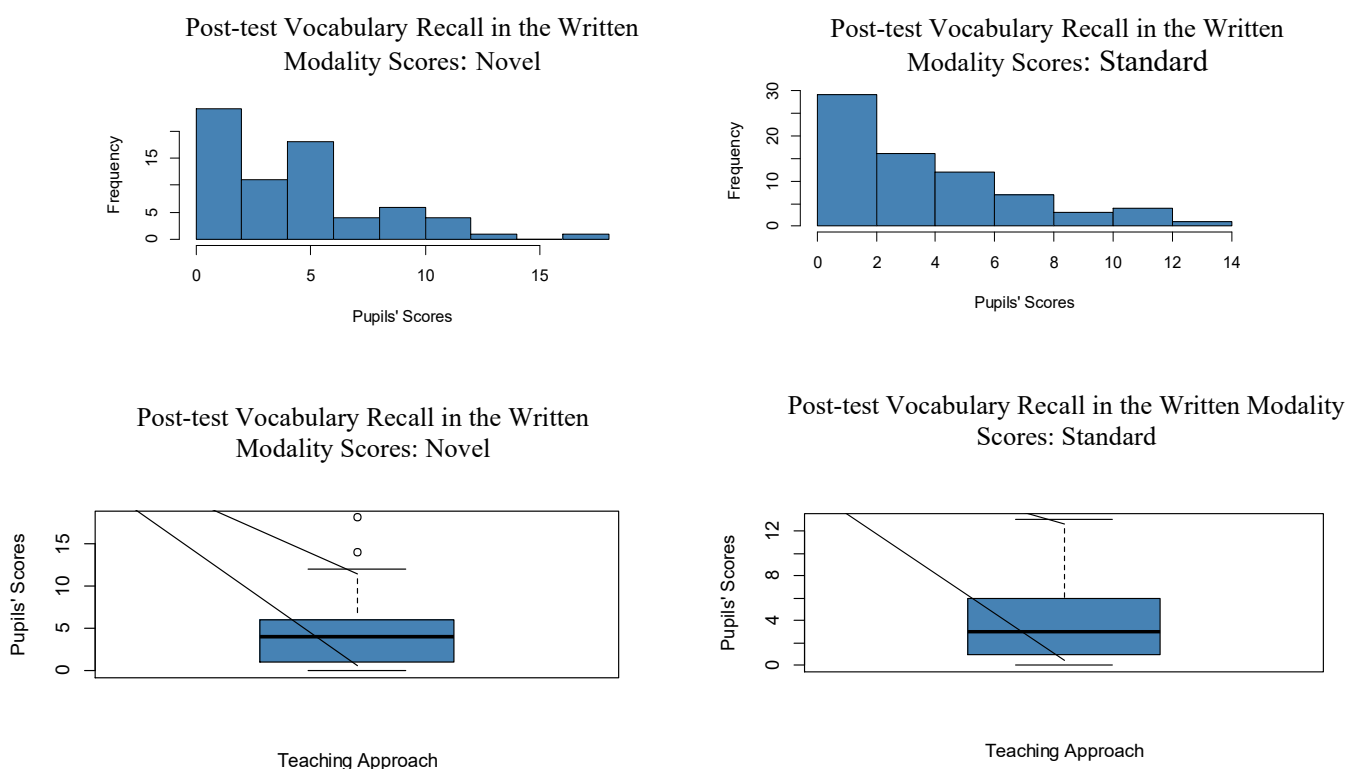
In the written modality post-test, data were considered to deviate from normal distribution, according to both the Shapiro-Wilk test (Novel: $W = .92, p < .001$; Standard: $W = .91, p < .001$) and visual inspection (see [Appendix M.8](#)).

In delayed post-tests, data from the vocabulary recall in the written modality were not normally distributed, a conclusion drawn from the Shapiro-Wilk test (Novel: $W = .81, p < .001$; Standard: $W = .88, p < .001$) and visual inspection (see [Appendix M.9](#)).

Return to [Table 19](#).

Appendix M.8

Normality of Distribution Checks of Post-test Receptive Tests: Vocabulary Recall in the Written Modality (raw data)



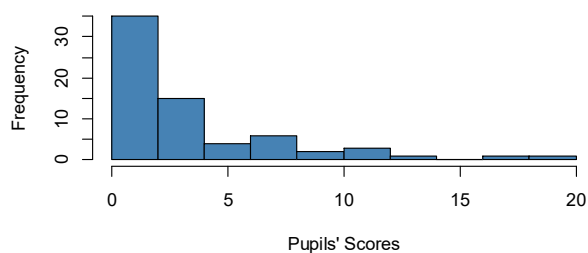
Approach	Vocabulary Recall in the Written Modality: Post-test	
	Skew	Kurtosis
Novel	1.61	.64
Standard	1.40	-.25

Return to [Table 19](#).

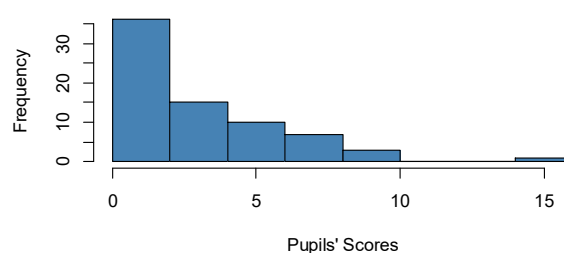
Appendix M.9

Normality of Distribution Checks of Delayed Post-test Receptive Tests: Vocabulary Recall in the Written Modality (raw data)

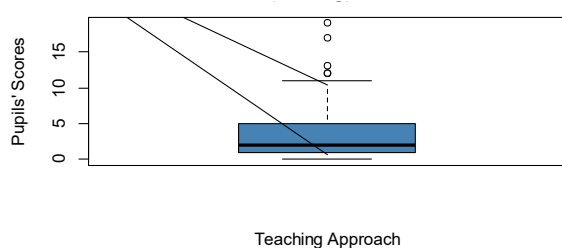
Delayed Post-test Vocabulary Recall in the Written Modality Scores: Novel



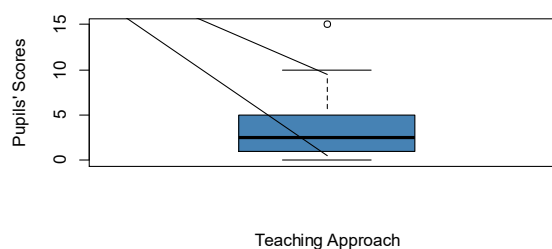
Delayed Post-test Vocabulary Recall in the Written Modality Scores: Standard



Delayed Post-test Vocabulary Recall in the Written Modality Scores: Novel



Delayed Post-test Vocabulary Recall in the Written Modality Scores: Standard



Approach	Vocabulary Recall in the Written Modality: Delayed Post-test	
	Skew	Kurtosis
Novel	2.81	2.19
Standard	2.19	1.69

Return to [Table 19](#).

Appendix M.10

Normality of Distribution for Vocabulary Recall in the Oral Modality

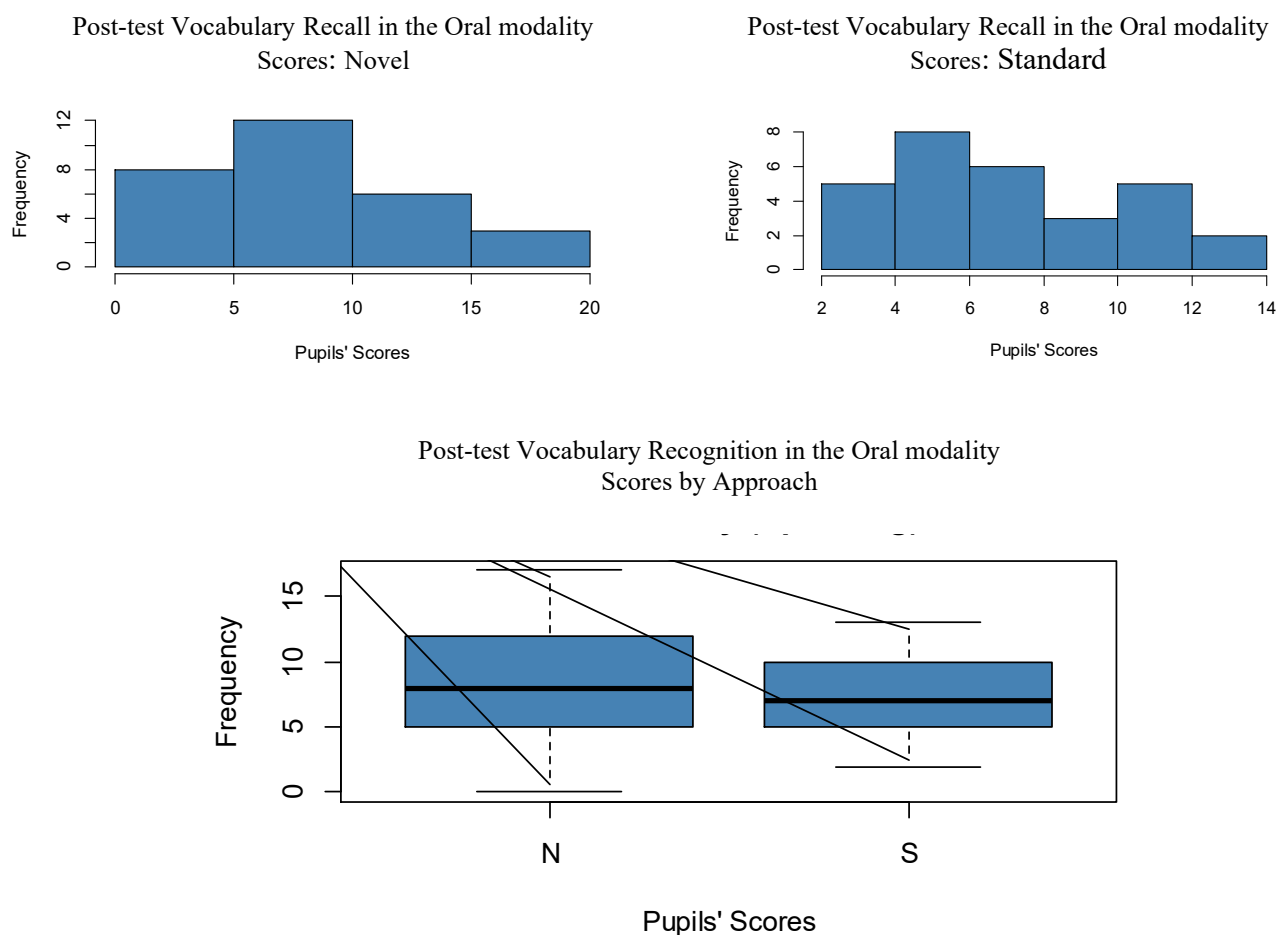
In the vocabulary recall speaking post-test, data appeared to be normally distributed (Novel: $W = .96, p = .42$; Standard: $W = .94, p = .13$) (see [Appendix M.11](#)).

In the delayed post-test, although the data from the Standard scores were normally distributed ($W = .95, p = .16$), the decision not to treat data as normally distributed was supported both by visual inspections of histograms and boxplots and by the outcome of the Shapiro-Wilk test in the Novel approach ($W = .94, p = .008$) ([Appendix M.12](#)).

Return to [Table 19](#).

Appendix M.11

Normality of Distribution Checks of Post-test Receptive Tests: Vocabulary Recall in the Oral Modality (raw data)

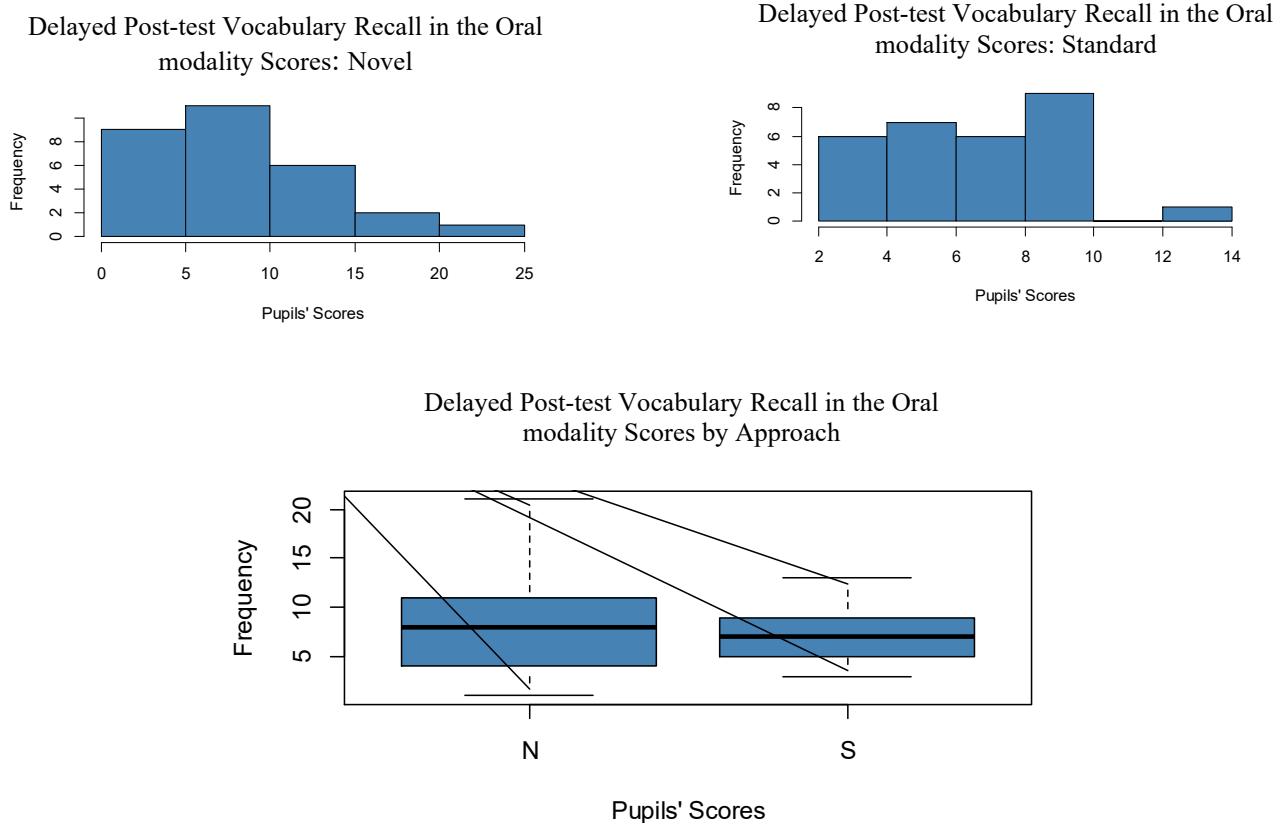


Approach	Vocabulary Recall in the Oral Modality: Post-test	
	Skew	Kurtosis
Novel	.19	-.56
Standard	.26	-.68

Return to [Table 19](#).

Appendix M.12

Normality of Distribution Checks of Delayed Post-test Receptive Tests: Vocabulary Recall in the Oral Modality (raw data)



Approach	Vocabulary Recall in the Oral Modality: Delayed Post-test	
	Skew	Kurtosis
Novel	.87	.001
Standard	.14	-.58

Return to [Table 19](#).

Appendix M.13

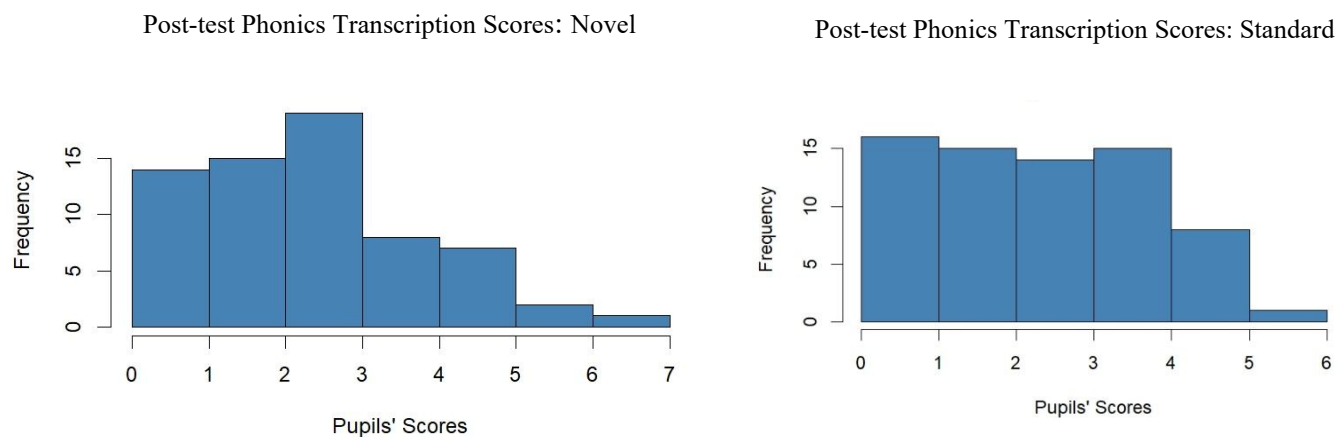
Normality of Distribution for Phonics Transcription Tests

According to both the Shapiro-Wilk test (Novel: $W = .93, p < .001$; Standard: $W = .94, p = .003$) and visual inspection, data in the phonics transcription test appeared to deviate from normal distribution at post-test (see [Appendix M.14](#)). In delayed post-tests, although data from the Novel phonics transcription tests were normally distributed according to visual inspection of histograms, the Shapiro-Wilk test for Novel ($W = .94, p = .003$) and the data from Standard were not ($W = .92, p < .001$) ([Appendix M.15](#)).

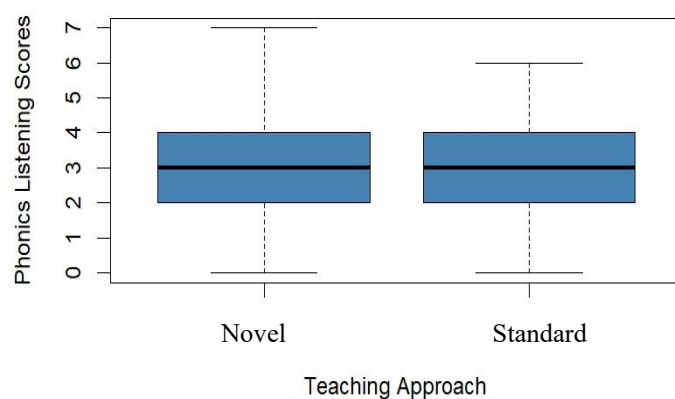
Return to [Table 25](#).

Appendix M.14

Normality of Distribution Checks of Post-test Phonics Transcription Tests



Post-test Phonics Transcription Scores by Approach



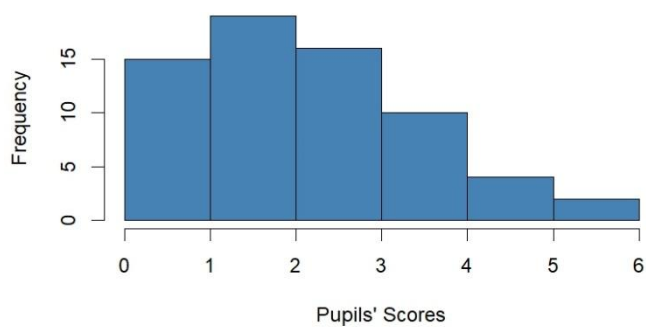
Approach	Phonics Transcription: Post-test	
	Skew	Kurtosis
Novel	.90	-.22
Standard	.07	-.84

Return to [Table 25](#).

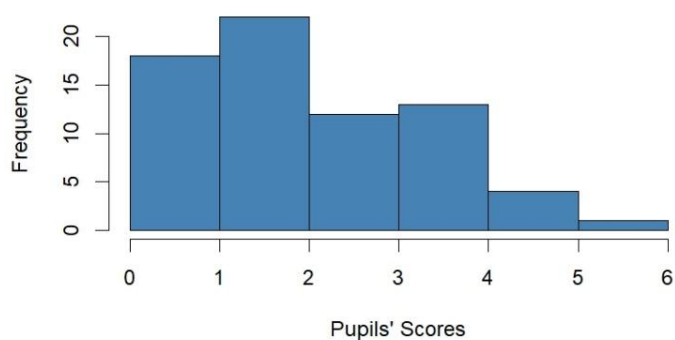
Appendix M.15

Normality of Distribution Checks of Delayed Post-test Phonics Transcription Tests

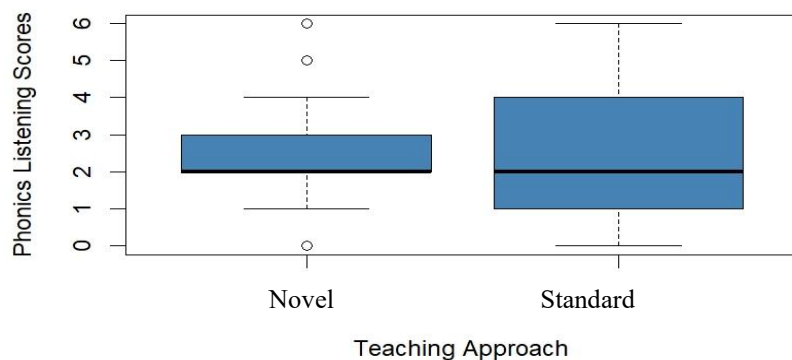
Delayed Post-test Phonics Transcription Scores: Novel



Delayed Post-test Phonics Transcription Scores: Standard



Delayed Post-test Phonics Transcription Scores by Approach



Approach	Phonics Transcription: Delayed Post-test	
	Skew	Kurtosis
Novel	.25	-.38
Standard	.52	-.53

Return to [Table 25](#).

Appendix M.16

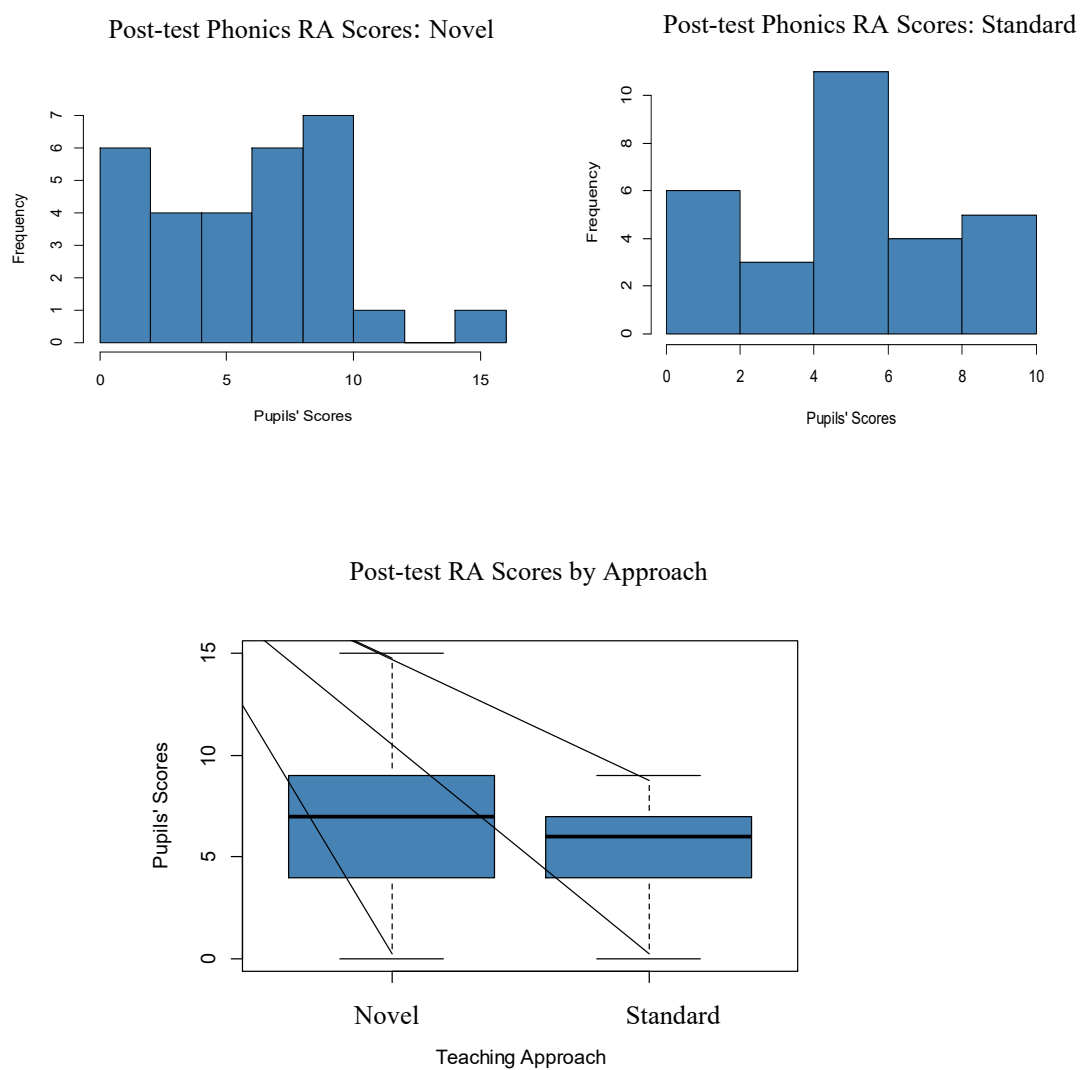
Normality of Distribution for Phonics Read Aloud Tests

The Shapiro-Wilk tests (Novel: $W = .97, p = .56$; Standard: $W = .93, p = .06$) indicated that data for the phonics Read Aloud test were normally distributed at post-test, and this observation was supported by visual inspection of the Standard approach data, although less so of Novel. Skew and kurtosis values were acceptable (see [Appendix M.17](#)). In delayed post-tests, visual inspection of histograms appeared to indicate deviation from normal distribution, although the Shapiro-Wilk tests did not reach significance (Novel: $W = .95, p = .15$; Standard: $W = .95, p = .14$). Skew and kurtosis values were again acceptable ([Appendix M.18](#)).

Return to [Table 25](#).

Appendix M.17

Normality of Distribution Checks of Post-test Phonics Read Aloud Tests



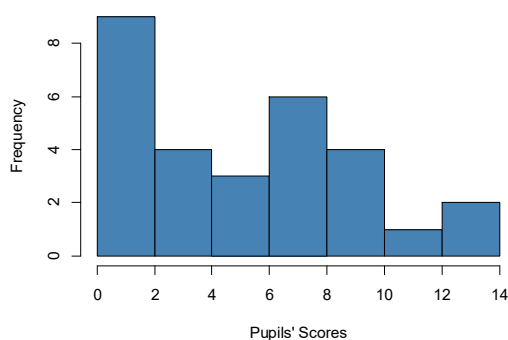
Approach	Phonics Read Aloud: Post-test	
	Skew	Kurtosis
Novel	.14	-.40
Standard	-.35	-.54

Return to [Table 25](#).

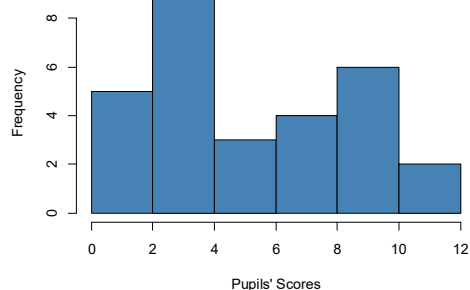
Appendix M.18

Normality of Distribution Checks of Delayed Post-test Phonics Read Aloud Tests

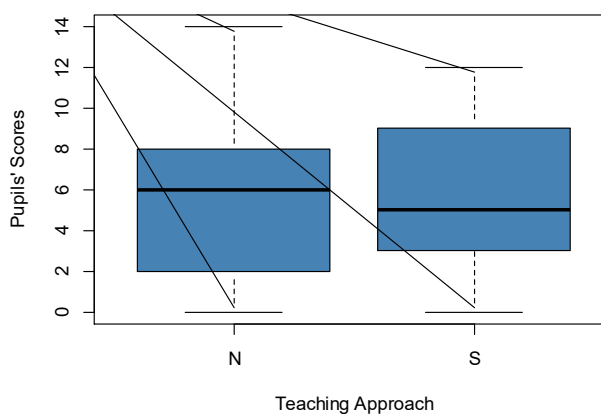
Delayed Post-test Phonics RA Scores: Novel



Delayed Post-test Phonics RA Scores: Standard



Delayed Post-test RA Scores by Approach



Approach	Phonics Read Aloud: Delayed Post-test	
	Skew	Kurtosis
Novel	.41	-.52
Standard	-.06	-.71

Return to [Table 25](#).

Appendix M.19

Normality of Distribution Checks for Written Grammar Recognition Tests

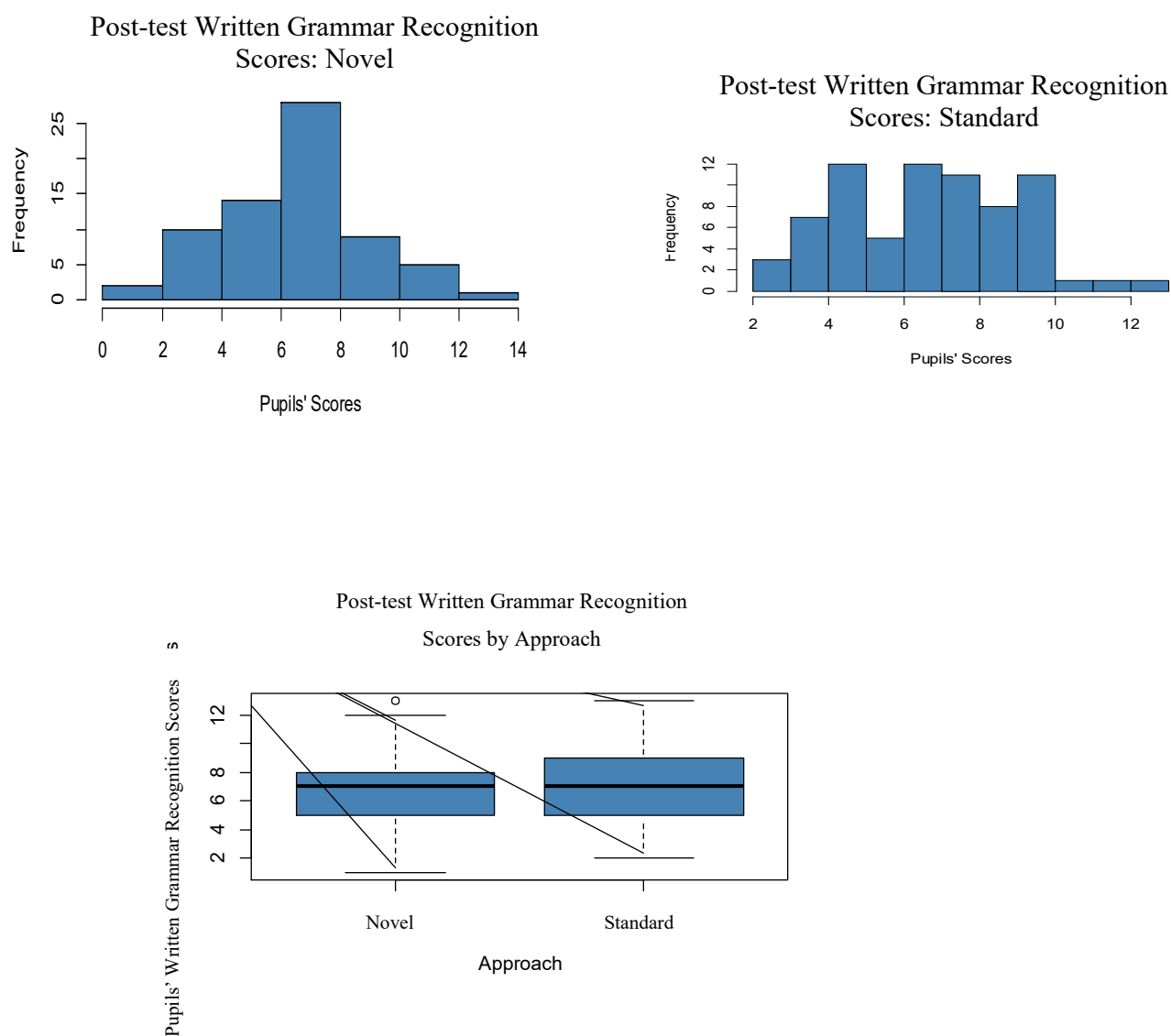
At post-test, visual inspections of data, as recommended by Larson-Hall (2016), indicated that the raw data distribution was normal for both Novel and Standard approaches (see below). This was supported by the Shapiro-Wilk test results (Novel: $W = .97$, $p = .18$; Standard: $W = .97$, $p = .07$). Skew and kurtosis measures were acceptable (see [Appendix M.20](#)).

In the delayed post-test, the Novel approach data were normally distributed (Novel: $W = .97$, $p = .09$), whereas in Standard the p -value was slightly below .05 (Standard: $W = .96$, $p = .04$), and the histogram was irregular (see [Appendix M.21](#)).

Return to [Table 29](#).

Appendix M.20

Normality of Distribution Checks of Post-test: Written Grammar Recognition (raw data)

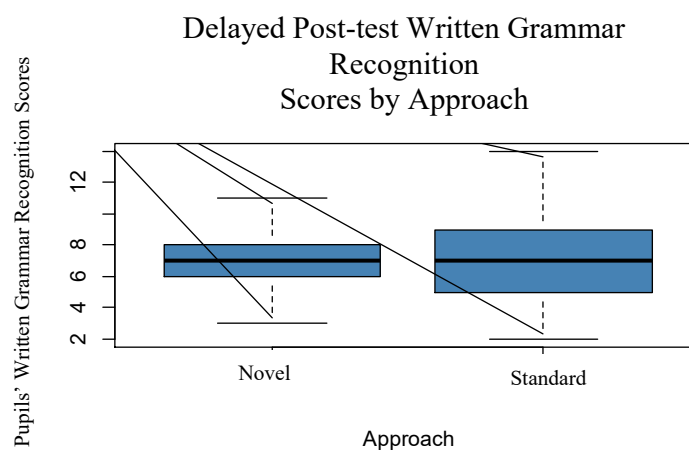
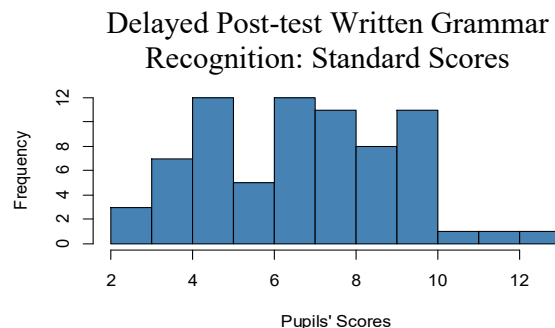
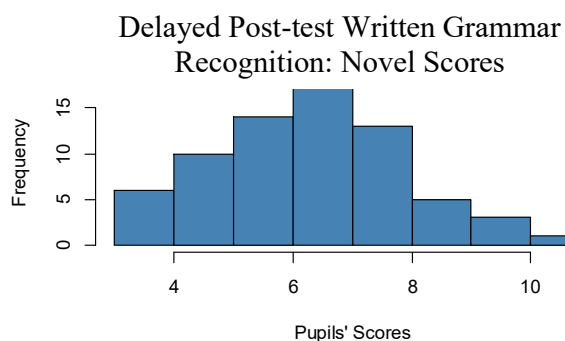


Approach	Written Grammar	Recognition: Post-test
	Skew	Kurtosis
Novel	.24	-.06
Standard	.15	-.59

Return to [Table 29](#).

Appendix M.21

Normality of Distribution Checks of Delayed Post-test: Written Grammar Recognition (raw data)



Approach	Written Grammar Recognition: Delayed Post-test	
	Skew	Kurtosis
Novel	.29	-.23
Standard	.56	-.17

Return to [Table 29](#).

Appendix M.22

Normality of Distribution for Spoken Grammar Recall Tests

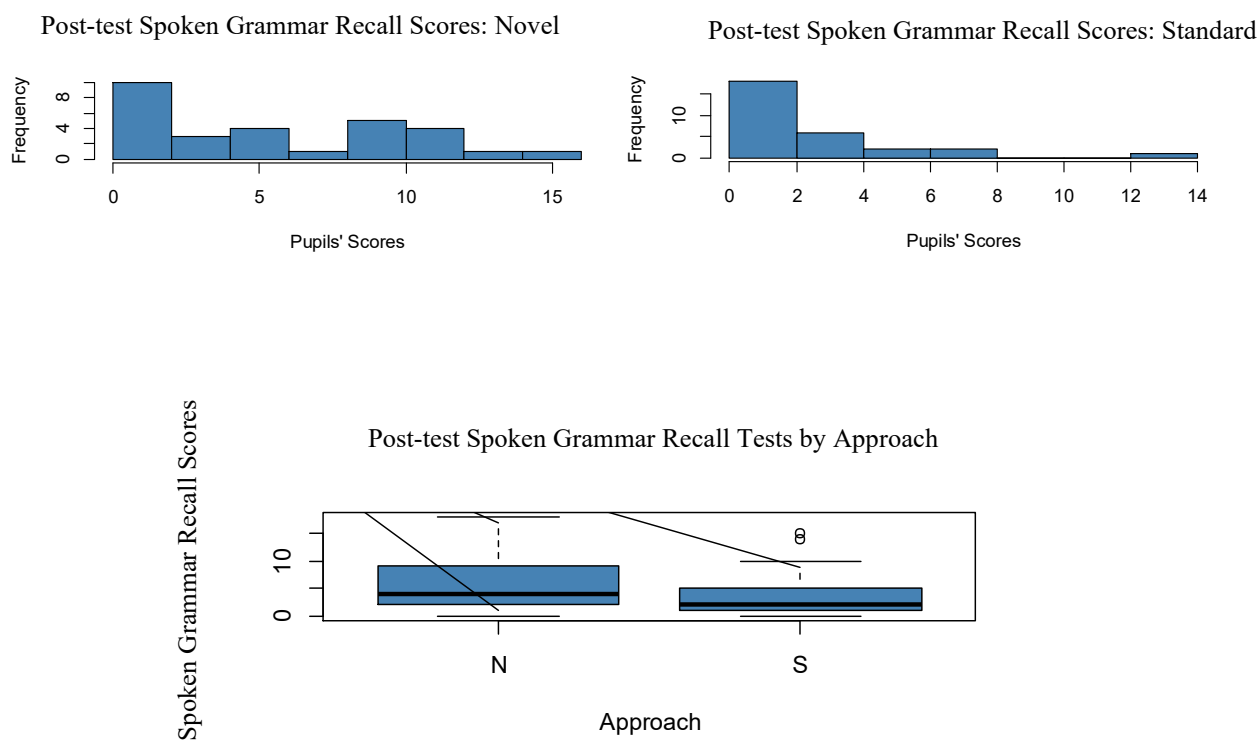
The p -values in the Shapiro-Wilk tests (Novel: $W = .93, p = .04$; Standard: $W = .79, p < .001$) indicated that data were not normally distributed, and this was supported by visual inspection of histograms and bar charts (see [Appendix M.23](#)).

It was evident from inspections of visual data (see [Appendix M.24](#)) and the p -values returned by the Shapiro-Wilk tests, that data were also not normally distributed at delayed post-test (Novel: $W = .90, p = .01$; Standard: $W = .79, p < .001$). Data were slightly skewed and the kurtosis value showed leptokurtic distribution.

Return to [Table 29](#).

Appendix M.23

Normality of Distribution Checks of Post-test: Spoken Grammar Recall (raw data)

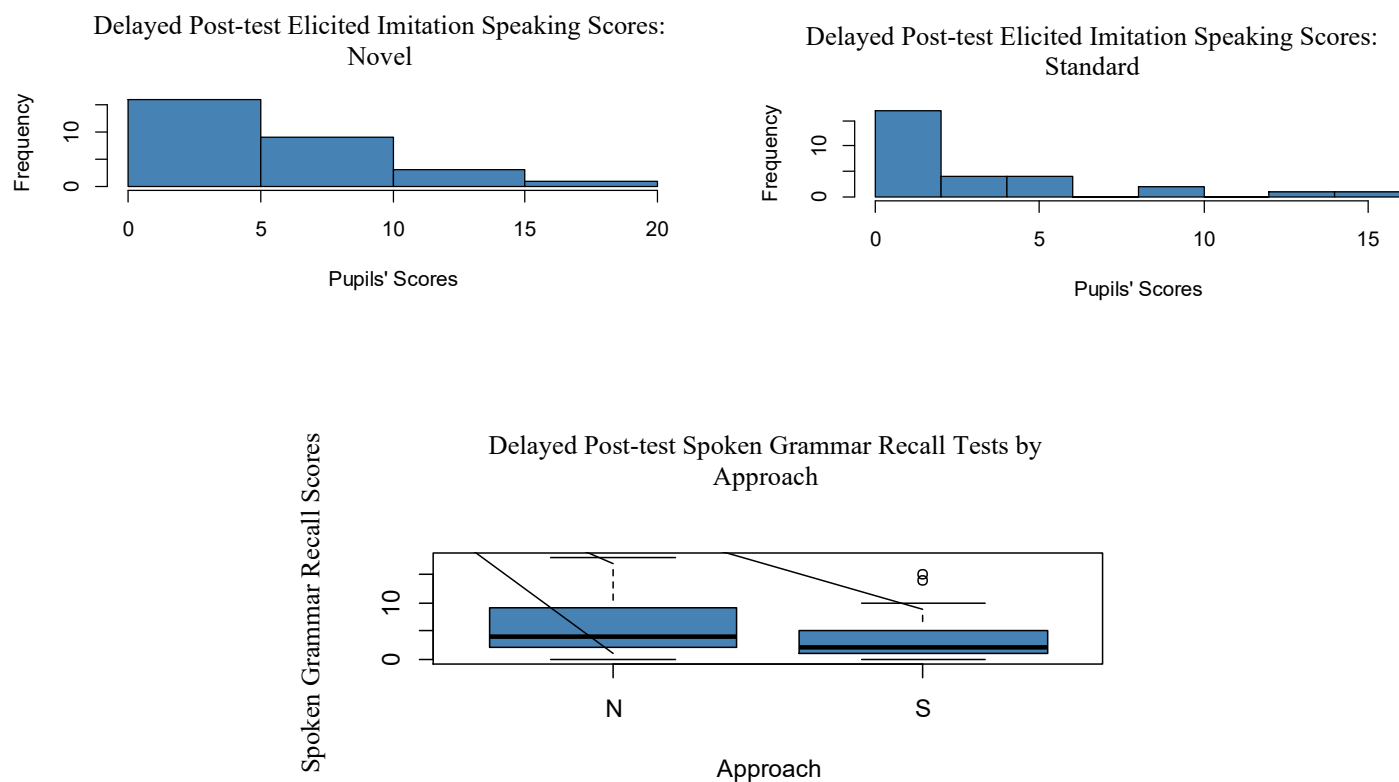


Approach	Spoken Grammar	Recognition: Post-test
	Skew	Kurtosis
Novel	.34	-.71
Standard	2.08	2.38

Return to [Table 29](#).

Appendix M.24

Normality of Distribution Checks of Delayed Post-test: Spoken Grammar Recall (raw data)



Approach	Spoken Grammar	Recognition: Delayed Post-test
	Skew	Kurtosis
Novel	.88	-.18
Standard	1.70	8.04

Return to [Table 29](#).

Appendix M.25

Normality of Distribution for Elicited Imitation Grammar Tests

At post-test, skew and kurtosis scores for the Elicited Imitation Grammar test were acceptable (see [Appendix M.26](#)). Inspections of visual data, combined with the p -values in the Shapiro-Wilk tests (Novel: $W = .96, p = .41$; Standard: $W = .96, p = .27$) indicated that data did not deviate significantly from normal distribution.

Results at delayed post-test were similar, again indicating that data were normally distributed, both from the Shapiro-Wilk tests (Novel: $W = .97, p = .61$; Standard: $W = .96, p = .29$) and the visual data (see [Appendix M.27](#)). Skew and kurtosis were also acceptable.

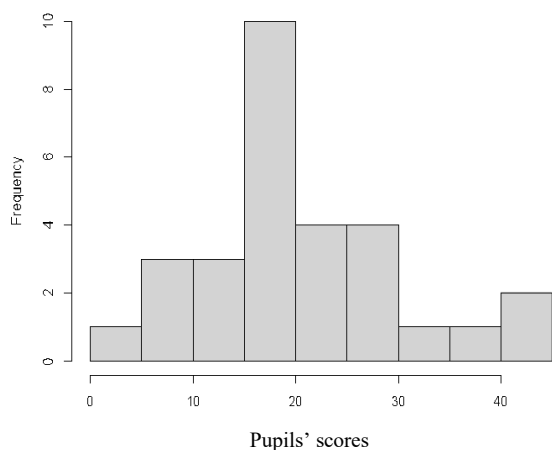
Return to [Table 29](#) (Elicited Imitation model – test quality).

Return to [Table 58](#) (Elicited Imitation rescored model – test quality).

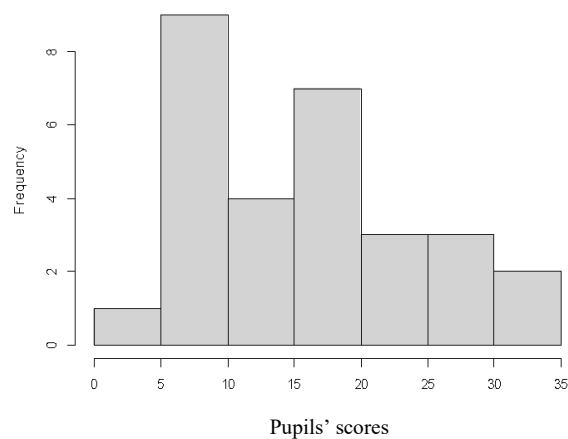
Appendix M.26

Normality of Distribution Checks of Post-test: Elicited Imitation Grammar (raw data)

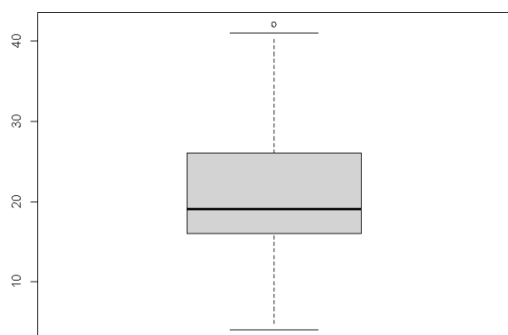
Post-test Elicited Imitation Grammar Scores: Novel



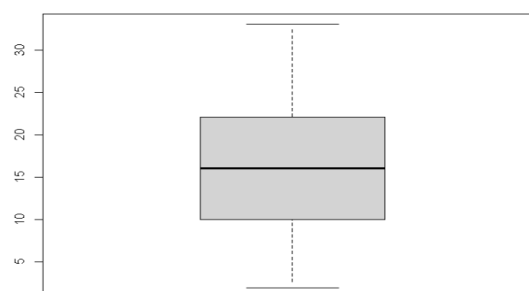
Post-test Elicited Imitation Grammar Scores: Standard



Post-test Elicited Imitation Grammar Scores: Novel



Post-test Elicited Imitation Grammar Scores: Standard



Approach	Elicited Imitation	Grammar: Post-test
	Skew	Kurtosis
Novel	.58	-.14
Standard	.41	-.86

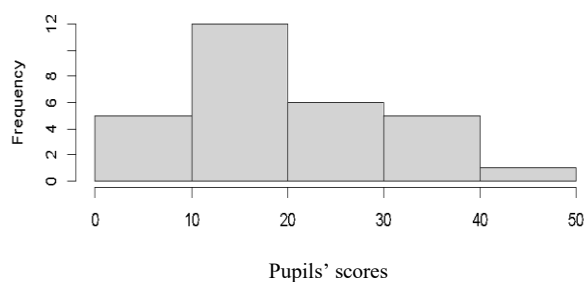
Return to [Table 29](#) (Elicited Imitation model – test quality).

Return to [Table 58](#) (Elicited Imitation rescored model – test quality).

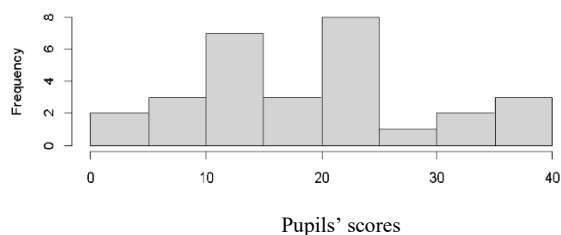
Appendix M.27

Normality of Distribution Checks of Delayed Post-test: Elicited Imitation Grammar (raw data)

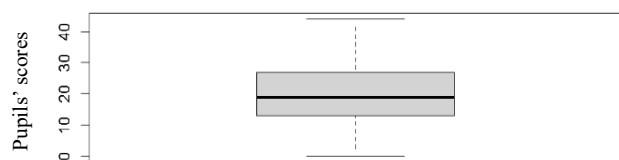
Delayed Post-test Elicited Imitation Grammar Scores:
Novel



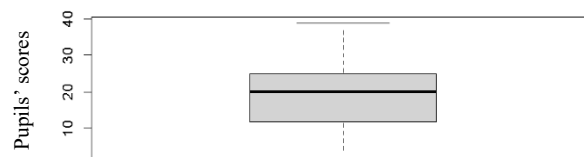
Delayed Post-test Elicited Imitation Grammar Scores:
Standard



Delayed Post-test Elicited Imitation Grammar Scores:
Novel



Delayed Post-test Elicited Imitation Grammar Scores:
Standard



Approach	Elicited Imitation	Grammar: Delayed Post-test
	Skew	Kurtosis
Novel	.26	-.47
Standard	.32	-.48

Return to [Table 29](#) (Elicited Imitation model – test quality).

Return to [Table 58](#) (Elicited Imitation rescored model – test quality).

Appendix M.28

Normality of Distribution for Picture Response Test

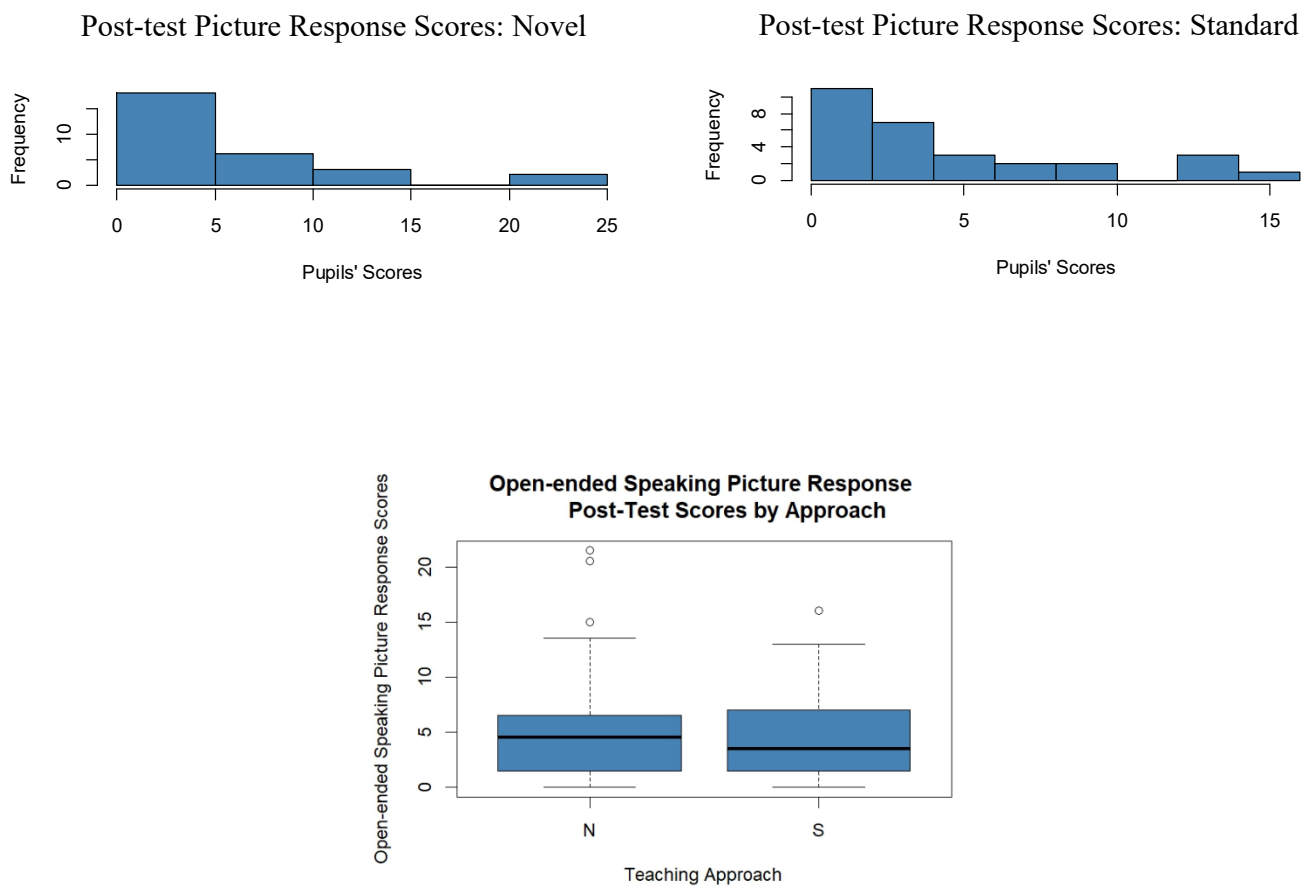
Normality of raw data distribution was checked using visual inspections as recommended by Larson-Hall (2016), as well as through the Shapiro-Wilk test. In the open-ended speaking Picture Response post-test, data appeared not to be normally distributed (Novel: $W = .82, p < .001$; Standard: $W = .88, p = .003$) and this was supported by the appearance of the boxplot and histograms (see [Appendix M.29](#)). Data were slightly skewed; kurtosis measures were acceptable.

In the delayed post-test, the decision not to treat data as normally distributed was supported both by visual inspections of histograms and boxplots (see [Appendix M.30](#)) and by the outcome of the Shapiro-Wilk test (Novel: $W = .87, p = .002$; Standard: $W = .87, p = .001$) and skew measures.

Return to [5.15.1](#).

Appendix M.29

Normality of Distribution Checks of Post-test: Picture Response (raw data)

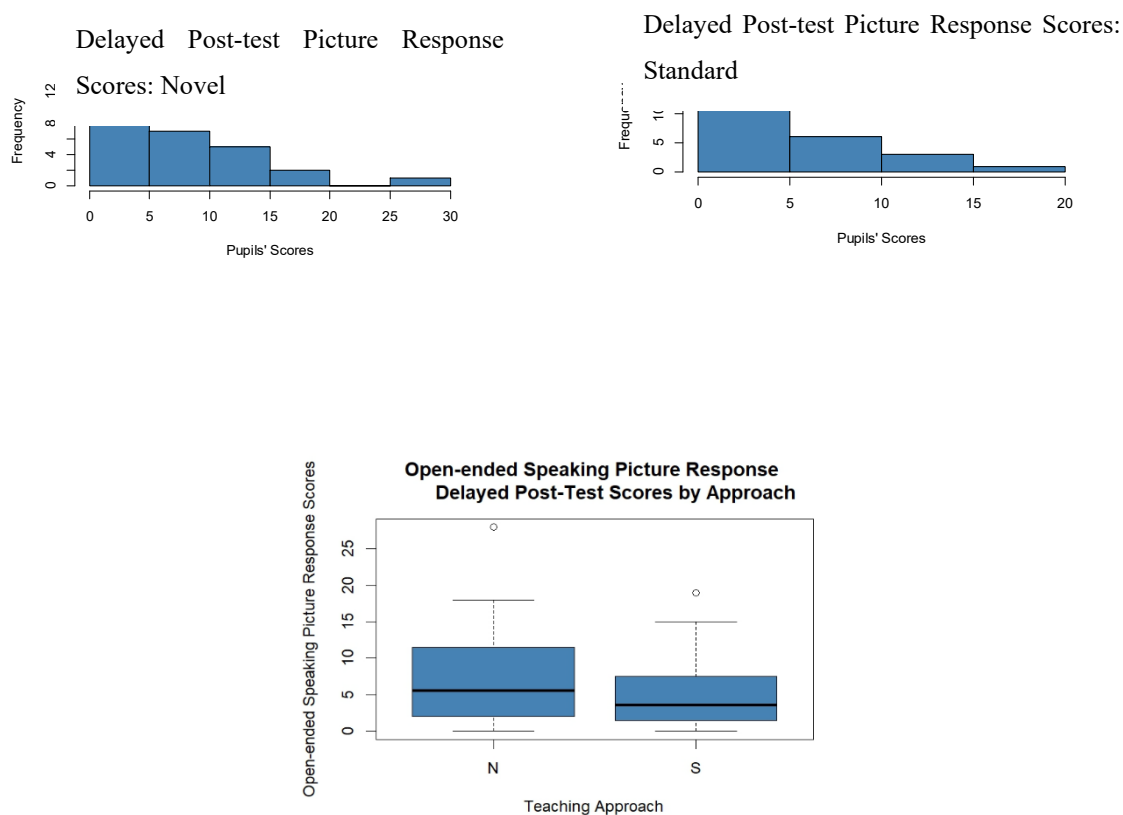


Approach	Open-ended Speaking	Picture Response: Post-test
	Skew	Kurtosis
Novel	1.61	.69
Standard	1.09	-.12

Return to [5.15.1](#).

Appendix M.30

Normality of Distribution Checks of Delayed Post-test: Picture Response (raw data)



Approach	Open-ended Speaking	Picture Response: Delayed Post-test
	Skew	Kurtosis
Novel	1.36	.67
Standard	1.29	.22

Return to [5.15.1](#).

Appendix M.31

Normality of Distribution for Elicited Imitation Speaking Tests

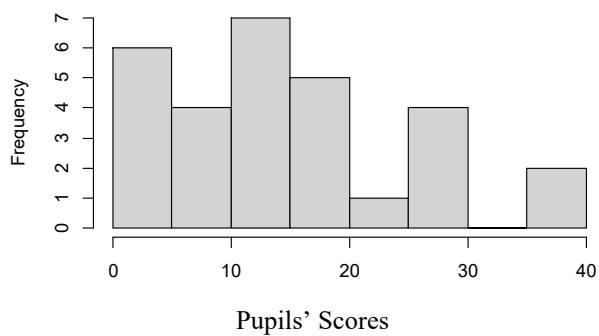
A combination of visual checks and the Shapiro-Wilk test (Novel: $W = .97, p = .04$; Standard: $W = .93, p = .05$) indicated that data were not normally distributed, at least for Standard scores, at post-test (see [Appendix M.32](#)). In the delayed post-test (see [Appendix M.33](#)), the Shapiro-Wilk test results (Novel: $W = .93, p = .06$; Standard: $W = .92, p = .03$) again indicated that data were not normally distributed for Standard.

Return to [Table 35](#).

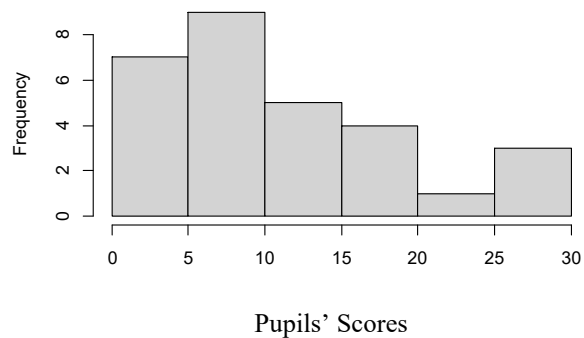
Appendix M.32

Normality of Distribution Checks of Post-test: Elicited Imitation Speaking (raw data)

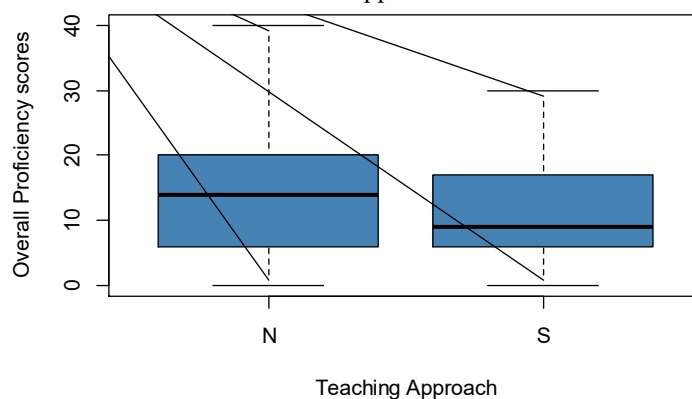
Post-test Elicited Imitation Speaking Scores: Novel



Post-test Elicited Imitation Speaking Scores: Standard



Post-test Elicited Imitation Speaking Scores by Teaching Approach



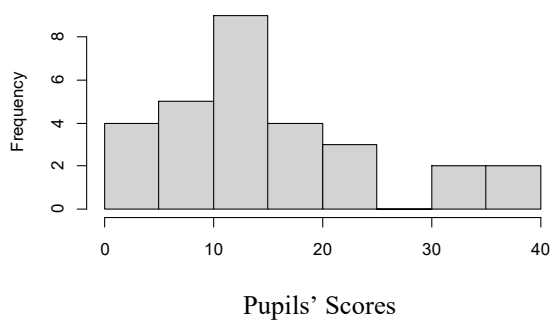
Approach	Elicited Imitation Speaking: Post-test	
	Skew	Kurtosis
Novel	.71	-.33
Standard	.77	-.39

Return to [Table 35](#).

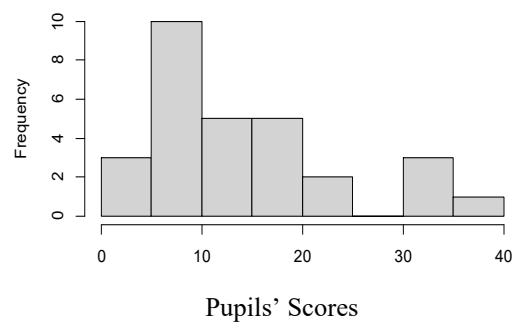
Appendix M.33

Normality of Distribution Checks of Delayed Post-test: Elicited Imitation Speaking (raw data)

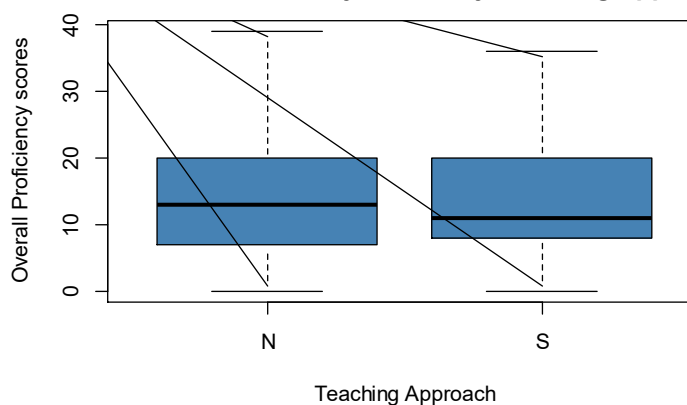
Delayed Post-test Elicited Imitation Speaking Scores:
Novel



Delayed Post-test Elicited Imitation Speaking Scores:
Standard



Delayed Post-test Elicited Imitation Speaking Scores by
Teaching Approach



Approach	Elicited Imitation	Speaking: Delayed Post-test
	Skew	Kurtosis
Novel	.83	-.16
Standard	.86	-.26

Return to [Table 35](#).

Appendix M.34

Confirmatory Factor Analysis Results for Elicited Imitation Grammar

	PT	DPT
Alpha:	.76	.84
Omega:	.79	.004*

* This has been checked but cannot be explained.

Confirmatory Factor Analysis Results for Elicited Imitation Speaking

	PT	DPT
Alpha:	.86	.86
Omega:	.88	.88

Appendix N.1

Model Code: Vocabulary Recognition in the Speaking and Writing Modalities

```
binomial_model_receptive_with_school_random = glmer(item_score ~ approach*time_point*task*scale(total_LAA_score) + (1|school/pupil_id) + (1|item), data = vocab_receptive_regression, family = binomial)
```

This was calculated using the `glmer` function from the `lme4` package [Bates et al., 2015] in R version 4.2.3 [R Core Team, 2023]), with `family = binomial (link = "logit")` because the data were binary, with answers correct or incorrect

Appendix N.2

Linearity Check: Vocabulary Recognition in the Speaking and Writing Modalities

<i>Predictors</i>	item_score		
	<i>Odds Ratios</i>	<i>CI</i>	<i>p</i>
(Intercept)	1.97	1.19 – 3.24	.008
total LAA score	2.56	.58 – 11.20	.21
total LAA score log	.67	.15 – 2.93	.56
Random Effects			
σ^2	3.29		
τ_{00} pupil_id:school	.31		
τ_{00} item	1.25		
τ_{00} school	.05		
ICC	.33		
N _{pupil_id}	141		
N _{school}	5		
N _{item}	24		
Observations	6768		
Marginal R^2 / Conditional R^2	.06 / .37		

Appendix N.3

Model Summary: Vocabulary Recognition in the Speaking and Writing Modalities

<i>Predictors</i>	item_score				
	<i>Odds Ratios</i>	<i>std. Error</i>	<i>CI</i>	<i>Statistic</i>	<i>p</i>
(Intercept)	1.99	.51	1.21 – 3.29	2.69	.007
approach standard-novel	.79	.09	.64 – .98	-2.10	.04
time-point sept-july	.97	.06	.87 – 1.09	-.49	.62
task reading-listening	1.16	.54	.47 – 2.88	.33	.74
total LAA score	1.73	.11	1.53 – 1.95	8.86	< .001
approach standard-novel × time-point sept-july	.95	.11	.76 – 1.20	-.43	.66
approach standard-novel × task reading-listening	.77	.09	.61 – .97	-2.24	.03
time-point sept-july × task reading-listening	.85	.10	.68 – 1.07	-1.36	.17
approach standard-novel × total LAA score	1.14	.13	.91 – 1.41	1.15	.25
time-point sept-july × total LAA score	1.04	.06	.92 – 1.16	.58	.56
task reading-listening × total LAA score	1.25	.08	1.11 – 1.41	3.71	< .001
approach standard-novel × time-point sept-july × task reading-listening	1.45	.34	.92 – 2.30	1.60	.11
(approach standard-novel × time-point sept-july) × total LAA score	.99	.12	.78 – 1.25	-.08	.93
(approach standard-novel × task reading-listening) × total LAA score	.98	.12	.77 – 1.24	-.18	.85
(time-point sept-july × task reading-listening) × total LAA score	1.05	.12	.83 – 1.33	.42	.67
(approach standard-novel × time-point sept-july × task reading-listening) × total LAA score	.95	.23	.59 – 1.51	-.23	.82

task reading-listening) ×
total LAA score

Random Effects

σ^2	3.29
τ_{00} pupil_id:school	.30
τ_{00} item	1.26
τ_{00} school	.05
ICC	.33
N _{pupil_id}	141
N _{school}	5
N _{item}	24
Observations	6768
Marginal R^2 / Conditional R^2	.07 / .37

Appendix N.4

Pairwise Comparisons Using Emmeans for Vocabulary Recognition in the Spoken and Written Modalities: Scores by Time-point and Approach

task = listening:

contrast	odds.ratio	SE	df	null	z.ratio	p.value
novel / standard	1.11	.14	Inf	1	.81	.42

task = reading:

contrast	odds.ratio	SE	df	null	z.ratio	p.value
novel / standard	1.44	.18	Inf	1	2.88	.004

Results are averaged over the levels of time_point. Tests are performed on the log odds ratio scale.

task = listening:

approach	prob	SE	df	asympt.LCL	asympt.UCL
novel	.66	.08	Inf	.49	.79
standard	.64	.08	Inf	.47	.78

task = reading:

approach	prob	SE	df	asympt.LCL	asympt.UCL
novel	.72	.07	Inf	.56	.84
standard	.64	.08	Inf	.47	.78

Results are averaged over the levels of time_point. Confidence level used: .95.

Intervals are back-transformed from the logit scale.

Appendix N.5

Vocabulary Recall in the Written Modality Model Code

```
ordinal_model_writing_with_school_random = clmm(item_score ~ approach*time_point*scale(total_LAA_score) + (1|school/pupil_id) + (1|item), data = vocab_writing_regression)
```

Appendix N.6

Vocabulary Recall in the Oral Modality Model Code

```
ordinal_model_speaking_with_school_random = clmm(item_score ~ approach*  
time_point*scale(total_LAA_score) + (1|school/pupil_id) + (1|item), dat  
a = vocab_speaking_regression)
```

Appendix N.7

Brant Test Results for Vocabulary Recall in the Written Modality

	pr(>chi)
Omnibus	.13
approach.standard-novel	.02
time_point.sept-july	.99
scale(total_LAA_score)	.05
approach.standard-novel:time_point.sept-july	.78
approach.standard-novel:scale(total_LAA_score)	.003
time_point.sept-july:scale(total_LAA_score)	.65
approach.standard-novel:time_point.sept-july:scale(total_LAA_score)	.80

Appendix N.8

Summary of Model (Writing Vocabulary)

<i>Predictors</i>	item score				
	<i>Odds Ratios</i>	<i>std. Error</i>	<i>CI</i>	<i>Statistic</i>	<i>p</i>
0 1	8.57	4.21	3.27 – 22.46	4.37	< .001
1 2	27.10	13.41	10.28 – 71.48	6.67	< .001
approach [.standard-novel]	.72	.15	.48 – 1.09	-1.55	.12
time-point [.sept-july]	.70	.07	.57 – .86	-3.41	.001
total LAA score	2.32	.28	1.82 – 2.94	6.88	< .001
approach [.standard-novel] × time point [.sept-july]	1.09	.23	.72 – 1.65	.41	.68
approach [.standard-novel] × total LAA score	1.20	.26	.79 – 1.83	.85	.39
time-point [.sept-july] × total LAA score	.94	.10	.77 – 1.16	-.58	.56
(approach [.standard-novel] × time- point [.sept-july]) × total LAA score	.90	.19	.60 – 1.37	-.48	.63
Random Effects					
σ^2	3.29				
τ_{00} pupil_id:school	1.19				
τ_{00} item	2.25				
τ_{00} school	.19				
ICC	.52				
N _{pupil_id}	141				
N _{school}	5				
N _{item}	12				
Observations	3384				
Marginal R^2 / Conditional R^2	.10 / .57				

Appendix N.9

Pairwise Comparisons Using Emmeans for Vocabulary Recall in the Written Modality Scores by Time-point and Approach

approach = novel:

contrast	estimate	SE	df	z.ratio	p.value
july - sept	.40	.15	Inf	2.68	.007

approach = standard:

contrast	estimate	SE	df	z.ratio	p.value
july - sept	.32	.15	Inf	2.12	.03

time_point = july, approach = novel:

item_score	prob	SE	df	asympt.LCL	asympt.UCL
1	.86	.06	Inf	.73	.98
2	.09	.04	Inf	.02	.17
3	.05	.02	Inf	.003	.10

time_point = sept, approach = novel:

item_score	prob	SE	df	asympt.LCL	asympt.UCL
1	.90	.05	Inf	.81	.99
2	.07	.03	Inf	.009	.12
3	.03	.02	Inf	.001	.07

time_point = july, approach = standard:

item_score	prob	SE	df	asympt.LCL	asympt.UCL
1	.90	.05	Inf	.80	.99
2	.07	.03	Inf	.01	.13
3	.04	.02	Inf	.001	.07

time_point = sept, approach = standard:

item_score	prob	SE	df	asympt.LCL	asympt.UCL
1	.92	.04	Inf	.85	.99
2	.05	.02	Inf	.005	.10
3	.03	.01	Inf	.0005	.05

Confidence level used: .95

Note. 1 indicates a score of 0, 2 represents a score of 1, and 3 is a score of 2.

Appendix N.10

Brant Test Results for Vocabulary Recall in the Oral Modality

	pr(>chi)
Omnibus	.58
approach.standard-novel	.05
time_point.sept-july	.69
scale(total_LAA_score)	.35
approach.standard-novel:time_point.sept-july	.48
approach.standard-novel:scale(total_LAA_score)	1.00
time_point.sept-july:scale(total_LAA_score)	1.00
approach.standard-novel:time_point.sept-july:scale(total_LAA_score)	.57

Appendix N.11

Summary of Model (Speaking Vocabulary)

<i>Predictors</i>	item score				
	<i>Odds Ratios</i>	<i>std. Error</i>	<i>CI</i>	<i>Statistic</i>	<i>p</i>
0 1	2.19	1.16	.78 – 6.17	1.48	.14
1 2	6.06	3.22	2.14 – 17.15	3.40	.001
approach [.standard-novel]	.81	.20	.50 – 1.32	-.84	.40
time-point [.sept-july]	.95	.12	.73 – 1.22	-.42	.68
total LAA score	2.09	.28	1.61 – 2.71	5.58	< .001
approach [.standard-novel] × time point [.sept-july]	.83	.22	.50 – 1.38	-.73	.47
approach [.standard-novel] × total LAA score	.77	.19	.47 – 1.26	-1.03	.30
time-point [.sept-july] × total LAA score	1.00	.13	.77 – 1.30	-.00	1.00
(approach [.standard-novel] × time point [.sept-july]) × total LAA score	1.13	.30	.67 – 1.91	.47	.64
Random Effects					
σ^2	3.29				
τ_{00} pupil_id:school	.64				
τ_{00} item	2.71				
τ_{00} school	.18				
ICC	.52				
$N_{\text{pupil\_id}}$	60				
N_{school}	5				
N_{item}	12				
Observations	1440				
Marginal R^2 / Conditional R^2	.08 / .56				

Appendix N.12

Pairwise Comparisons Using Emmeans for Vocabulary Recall in the Oral Modality by Time-point and Approach

time_point = july:

contrast	estimate	SE	df	z.ratio	p.value
novel - standard	.11	.28	Inf	.40	.69

time_point = sept:

contrast	estimate	SE	df	z.ratio	p.value
novel - standard	.30	.28	Inf	1.09	.28

Appendix N.13

Pairwise Comparisons Using Emmeans for Vocabulary Recall in the Oral Modality by Approach and Time-point

approach = novel:

contrast	estimate	SE	df	z.ratio	p.value
july - sept	-.04	.19	Inf	-.22	.83

approach = standard:

contrast	estimate	SE	df	z.ratio	p.value
july - sept	.15	.18	Inf	.83	.41

time_point = july, approach = novel:

item_score	prob	SE	df	asympt.LCL	asympt.UCL
1	.67	.12	Inf	.43	.91
2	.18	.05	Inf	.08	.28
3	.15	.07	Inf	.01	.29

time_point = sept, approach = novel:

item_score	prob	SE	df	asympt.LCL	asympt.UCL
1	.66	.12	Inf	.42	.90
2	.18	.05	Inf	.08	.29
3	.16	.07	Inf	.01	.30

time_point = july, approach = standard:

item_score	prob	SE	df	asympt.LCL	asympt.UCL
1	.69	.12	Inf	.46	.92
2	.17	.05	Inf	.07	.27
3	.14	.07	Inf	.009	.27

time_point = sept, approach = standard:

item_score	prob	SE	df	asympt.LCL	asympt.UCL
1	.72	.11	Inf	.51	.94
2	.16	.05	Inf	.05	.26
3	.12	.06	Inf	.006	.24

Confidence level used: .95

Appendix N.14

Phonics Read Aloud Model Code

```
ordinal_model_phonics_speaking_with_school_random = clmm(item_score ~ a  
pproach*time_point*scale(total_LAA_score) + (1|school/pupil_id) + (1|it  
em_code), data = phonics_speaking_regression)
```

Appendix N.15

Brant Test Results for Phonics Read Aloud Tests

	pr(> chi)
Omnibus	.12
approach.standard-novel	.009
time_point.sept-july	.77
scale(total_LAA_score)	.20
approach.standard-novel:time_point.sept-july	.19
approach.standard-novel:scale(total_LAA_score)	.86
time_point.sept-july:scale(total_LAA_score)	.97
approach.standard-novel:time_point.sept-july:scale(total_LAA_score)	.98

Appendix N.16

Summary of Model (Phonics Read Aloud)

<i>Predictors</i>	item_score				
	<i>Odds Ratios</i>	<i>std. Error</i>	<i>CI</i>	<i>Statistic</i>	<i>p</i>
0 1	3.99	2.50	1.17 – 13.63	2.20	.03
1 2	8.65	5.45	2.52 – 29.75	3.43	.001
approach [.standard-novel]	.83	.22	.49 – 1.41	-.67	.50
time-point [.sept-july]	.77	.12	.57 – 1.03	-1.77	.08
total LAA score	1.72	.24	1.32 – 2.25	3.97	< .001
approach [.standard-novel] × time- point [.sept-july]	1.60	.48	.88 – 2.88	1.55	.12
approach [.standard-novel] × total LAA score	.86	.23	.51 – 1.45	-.58	.56
time-point [.sept-july] × total LAA score	1.41	.21	1.05 – 1.90	2.31	.02
(approach [.standard-novel] × time point [.sept-july]) × total LAA score	1.14	.34	.63 – 2.04	.43	.67
Random Effects					
σ^2	3.29				
τ_{00} pupil_id:school	.72				
τ_{00} item_code	3.64				
τ_{00} school	.01				
ICC	.57				
N _{pupil_id}	60				
N _{school}	5				
N _{item_code}	10				
Observations	1200				
Marginal R^2 / Conditional R^2	.05 / .59				

Appendix N.17

Pairwise Comparisons Using Emmeans for Phonics Speaking Read Aloud Scores by Time-point and Approach

```
emmeans(ordinal_model_phonics_speaking_with_school_random, ~
item_score|time_point|approach, mode="prob") # Estimated means

pairs(emmeans(ordinal_model_phonics_speaking_with_school_random, ~
time_point|approach, type="response"), adjust = "bonf") # Pairwise comparisons
```

time_point = july, approach = novel:

item_score	prob	SE	df	asympt.LCL	asympt.UCL
1	.74	.13	Inf	.49	.98
2	.12	.05	Inf	.03	.22
3	.14	.08	Inf	-.01	.29

time_point = sept, approach = novel:

item_score	prob	SE	df	asympt.LCL	asympt.UCL
1	.82	.09	Inf	.64	1.01
2	.09	.04	Inf	.004	.17
3	.09	.05	Inf	-.02	.19

time_point = july, approach = standard:

item_score	prob	SE	df	asympt.LCL	asympt.UCL
1	.81	.10	Inf	.62	.006
2	.09	.04	Inf	.007	.18
3	.10	.06	Inf	-.01	.21

time_point = sept, approach = standard:

item_score	prob	SE	df	asympt.LCL	asympt.UCL
1	.82	.10	Inf	.62	1.01
2	.09	.04	Inf	.006	.17
3	.09	.06	Inf	-.02	.20

Confidence level used: .95

NOTE: Results may be misleading due to involvement in interactions

approach = novel:

contrast	estimate	SE	df	z.ratio	p.value
july - sept	.50	.22	Inf	2.31	.02

approach = standard:

contrast	estimate	SE	df	z.ratio	p.value
july - sept	.03	.21	Inf	.16	.87

Appendix N.18

Written Grammar Recognition Model Code

```
binomial_model_grammar_reading_with_school_random = glmer(item_score ~  
approach*time_point*scale(total_LAA_score) + (1|pupil_id) + (1|item_cod  
e), data = grammar_reading_regression, family = binomial)
```

Appendix N.19

Linearity Check for Written Grammar Recognition Test

<i>Predictors</i>	item score		
	<i>Odds Ratios</i>	<i>CI</i>	<i>p</i>
(Intercept)	61	.47 – .79	< .001
total LAA score	.82	.34 – 2.02	.67
total LAA score log	.40	.57 – 3.42	.46
Random Effects			
σ^2	3.29		
τ_{00} pupil_id:school	.04		
τ_{00} item_code	.31		
τ_{00} school	.00		
ICC	.10		
N _{pupil_id}	140		
N _{school}	5		
N _{item_code}	18		
Observations	5040		
Marginal R^2 / Conditional R^2	.006 / .10		

Appendix N.20

Summary of Written Grammar Recognition Model

<i>Predictors</i>	item score				
	<i>Odds Ratios</i>	<i>std. Error</i>	<i>CI</i>	<i>Statistic</i>	<i>p</i>
(Intercept)	.61	.08	.47 – .80	-3.65	< .001
approach standard-novel	1.06	.07	.93 – 1.22	.89	.34
time-point sept-july	.97	.06	.86 – 1.09	-.46	.65
total LAA score	1.16	.04	1.08 – 1.24	4.19	< .001
approach standard-novel × time-point sept-july	1.03	.12	.81 – 1.30	.24	.81
approach standard-novel × total LAA score	1.05	.07	.92 – 1.21	.75	.45
time-point sept-july × total LAA score	1.00	.06	.89 – 1.13	.05	.96
(approach standard-novel × time-point sept-july) × total LAA score	1.35	.16	1.07 – 1.71	2.48	.01
Random Effects					
σ^2	3.29				
τ_{00} pupil_id	.04				
τ_{00} item_code	.31				
ICC	.10				
N _{pupil_id}	140				
N _{item_code}	18				
Observations	5040				
Marginal R^2 / Conditional R^2	.01 / .10				

Appendix N.21

Pairwise Comparisons for Written Grammar Recognition by Time-point

time_point = july:

contrast	odds.ratio	SE	df	null	z.ratio	p.value
novel / standard	.95	.09	Inf	1	-.51	.61

time_point = sept:

contrast	odds.ratio	SE	df	null	z.ratio	p.value
novel / standard	.93	.09	Inf	1	-.82	.41

Tests are performed on the log odds ratio scale

Appendix N.22

Pairwise Comparisons for Written Grammar Recognition by Approach

approach = novel:

contrast	odds.ratio	SE	df	null	z.ratio	p.value
july / sept	1.04	.09	Inf	1	.49	.63

approach = standard:

contrast	odds.ratio	SE	df	null	z.ratio	p.value
july / sept	1.01	.09	Inf	1	.16	.87

Tests are performed on the log odds ratio scale

approach = novel:

## time_point	prob	SE	df	asympt.LCL	asympt.UCL
## july	.38	.03	Inf	.31	.45
## sept	.37	.03	Inf	.30	.44

##

approach = standard:

## time_point	prob	SE	df	asympt.LCL	asympt.UCL
## july	.39	.03	Inf	.32	.46
## sept	.38	.03	Inf	.32	.45

##

Confidence level used: .95

Intervals are back-transformed from the logit scale

Appendix N.23

Spoken Grammar Recognition Model Code

```
ordinal_model_grammar_speaking_with_school_random = clmm(item_score ~ a  
pproach*time_point*scale(total_LAA_score) + (1|school/pupil_id) + (1|it  
em_code), data = grammar_speaking_regression)
```

Appendix N.24

Brant Test Results for Spoken Grammar Recall Test

	pr(>chi)
Omnibus	.52
approach.standard-novel	.56
time_point.sept-july	.26
scale(total_LAA_score)	.68
approach.standard-novel:time_point.sept-july	.40
approach.standard-novel:scale(total_LAA_score)	.25
time_point.sept-july:scale(total_LAA_score)	.45
approach.standard-novel:time_point.sept-july:scale(total_LAA_score)	.91
H0: Proportional odds assumption holds	

Appendix N.25

Summary of Model (Spoken Grammar Recall)

<i>Predictors</i>	item score				
	<i>Odds Ratios</i>	<i>std. Error</i>	<i>CI</i>	<i>Statistic</i>	<i>p</i>
0 1	8.93	4.03	3.69 – 21.62	4.86	< .001
1 2	20.10	9.16	8.23 – 49.13	6.58	< .001
approach [.standard-novel]	.39	.10	.23 – .63	-3.76	< .001
time-point [.sept-july]	1.11	.17	.83 – 1.50	.72	.47
total LAA score	1.88	.25	1.45 – 2.43	4.78	< .001
approach [.standard-novel] × time- point [.sept-july]	1.73	.52	.95 – 3.13	1.80	.07
approach [.standard-novel] × total LAA score	1.01	.26	.62 – 1.67	.05	.96
time-point [.sept-july] × total LAA score	1.08	.16	.81 – 1.44	.49	.62
(approach [.standard-novel] × time- point [.sept-july]) × total LAA score	.96	.28	.54 – 1.71	-.14	.89
Random Effects					
σ^2	3.29				
τ_{00} pupil_id:school	.56				
τ_{00} item_code	1.39				
τ_{00} school	.41				
ICC	.42				
N _{pupil_id}	60				
N _{school}	5				

N _{item_code}	14
Observations	1680
Marginal R^2 / Conditional R^2	.10 / .48

Appendix N.26

Pairwise Comparisons for Speaking Grammar Recall by Time-point and Approach

```
pairs(emmeans(ordinal_model_grammar_speaking_with_school_random, ~
approach|time_point, type="response"), adjust = "bonf") # Pairwise comparisons
```

NOTE: Results may be misleading due to involvement in interactions

time_point = july:

contrast	estimate	SE	df	z.ratio	p.value
novel - standard	1.22	.30	Inf	4.08	< .0001

time_point = sept:

contrast	estimate	SE	df	z.ratio	p.value
novel - standard	.68	.29	Inf	2.34	.02

```
emmeans(ordinal_model_grammar_speaking_with_school_random, ~
item_score|approach|time_point, mode="prob") # Estimated means
```

approach = novel, time_point = july:

item_score	prob	SE	df	asympt.LCL	asympt.UCL
1	.84	.06	Inf	.71	.96
2	.08	.03	Inf	.02	.14
3	.08	.04	Inf	.01	.15

approach = standard, time_point = july:

item_score	prob	SE	df	asympt.LCL	asympt.UCL
1	.95	.03	Inf	.90	1.00
2	.03	.01	Inf	.003	.06
3	.02	.01	Inf	.001	.05

approach = novel, time_point = sept:

item_score	prob	SE	df	asympt.LCL	asympt.UCL
1	.86	.06	Inf	.74	.97
2	.07	.03	Inf	.02	.13
3	.07	.03	Inf	.009	.13

approach = standard, time_point = sept:

item_score	prob	SE	df	asympt.LCL	asympt.UCL
1	.92	.03	Inf	.85	.99
2	.04	.02	Inf	.006	.08
3	.04	.02	Inf	.003	.07

Confidence level used: .95

Appendix N.27

Elicited Imitation Grammar Model Code

```
ordinal_model_grammar_imitation_with_school_random = clmm(item_score ~  
  approach*time_point*scale(total_LAA_score) + (1|school/pupil_id) + (1|  
  item_code), data = grammar_imitation_regression)
```

Appendix N.28

Brant Test Results for Elicited Imitation Grammar Test

	pr(>chi)
Omnibus	< .001
approach.standard-novel	.68
time_point.sept-july	.08
scale(total_LAA_score)	< .001
approach.standard-novel:time_point.sept-july	.33
approach.standard-novel:scale(total_LAA_score)	.12
time_point.sept-july:scale(total_LAA_score)	.04
approach.standard-novel:time_point.sept-july:scale(total_LAA_score)	.31

Appendix N.29

Summary of Model (Elicited Imitation Grammar)

<i>Predictors</i>	item score				
	<i>Odds Ratios</i>	<i>std. Error</i>	<i>CI</i>	<i>Statistic</i>	<i>p</i>
0 1	1.64	.39	1.03 – 2.61	2.08	.04
1 2	2.03	.48	1.27 – 3.23	2.97	.003
approach [.standard-novel]	.74	.16	.49 – 1.13	-1.39	.16
time-point [.sept-july]	1.22	.10	1.04 – 1.44	2.39	.02
total LAA score	1.61	.18	1.29 – 2.00	4.28	<.001
approach [.standard-novel] × time point [.sept-july]	1.34	.23	.96 – 1.86	1.73	.08
approach [.standard-novel] × total LAA score	.92	.20	.60 – 1.40	-.40	.69
time-point [.sept-july] × total LAA score	1.08	.09	.92 – 1.28	.92	.36
(approach [.standard-novel] × time- point [.sept-july]) × total LAA score	.92	.16	.66 – 1.28	-.51	.61
Random Effects					
σ^2	3.29				
τ_{00} pupil_id:school	.58				
τ_{00} item_code	.81				
τ_{00} school	.06				
ICC	.31				
N _{pupil_id}	60				
N _{school}	5				
N _{item_code}	25				
Observations	3000				
Marginal R^2 / Conditional R^2	.05 / .34				

Appendix N.30

Pairwise Comparisons for Elicited Imitation Grammar by Time-point and Approach

approach = novel:

contrast	estimate	SE	df	z.ratio	p.value
july - sept	-.05	.12	Inf	-.47	.64

approach = standard:

contrast	estimate	SE	df	z.ratio	p.value
july - sept	-.35	.12	Inf	-2.89	.004

time_point = july, approach = novel:

item_score	prob	SE	df	asympt.LCL	asympt.UCL
1	.59	.06	Inf	.47	.72
2	.05	.006	Inf	.04	.06
3	.36	.06	Inf	.24	.48

time_point = sept, approach = novel:

item_score	prob	SE	df	asympt.LCL	asympt.UCL
1	.58	.07	Inf	.45	.71
2	.05	.005	Inf	.04	.06
3	.37	.06	Inf	.25	.49

time_point = july, approach = standard:

item_score	prob	SE	df	asympt.LCL	asympt.UCL
1	.69	.06	Inf	.58	.81
2	.04	.006	Inf	.03	.06
3	.26	.05	Inf	.16	.37

time_point = sept, approach = standard:

item_score	prob	SE	df	asympt.LCL	asympt.UCL
1	.62	.06	Inf	.49	.74
2	.05	.006	Inf	.04	.06
3	.34	.06	Inf	.22	.45

Confidence level used: .95

Appendix N.31

Elicited Imitation Speaking Model Code

```
model_speaking_imitation_long = clmm(item_score ~ approach*time_point*s  
cale(total_LAA_score) + (1|school/pupil_id) + (1|item), data = speaking  
_imitation_long)
```

Appendix N.32

Brant Test Results for Elicited Imitation Speaking Test

	pr(>chi)
Omnibus	.06
approach.standard-novel	.12
time_point.sept-july	.23
scale(total_LAA_score)	.10
approach.standard-novel:time_point.sept-july	.55
approach.standard-novel:scale(total_LAA_score)	.22
time_point.sept-july:scale(total_LAA_score)	.57
approach.standard-novel:time_point.sept-july:scale(total_LAA_score)	.02

Appendix N.33

Summary of Model (Elicited Imitation Speaking)

<i>Predictors</i>	item score				
	<i>Odds Ratios</i>	<i>std. Error</i>	<i>CI</i>	<i>Statistic</i>	<i>p</i>
0 1	1.34	.53	.62 – 2.89	.74	.46
1 2	4.43	1.75	2.04 – 9.60	3.77	< .001
2 3	22.00	8.84	10.01 – 48.38	7.69	< .001
3 4	46.72	19.06	21.00 – 103.93	9.42	< .001
approach [.standard-novel]	.75	.22	.42 – 1.35	-.96	.34
time-point [.sept-july]	1.36	.14	1.12 – 1.67	3.02	.003
total LAA score	2.34	.36	1.72 – 3.17	5.46	< .001
approach [.standard-novel] × time- point [.sept-july]	1.36	.28	.91 – 2.04	1.51	.130
approach [.standard-novel] × total LAA score	1.07	.32	.59 – 1.92	.21	.83
time-point [.sept-july] × total LAA score	.88	.09	.72 – 1.08	-1.21	.23
(approach [.standard-novel] × time- point [.sept-july]) × total LAA score	1.13	.23	.76 – 1.69	.59	.55
Random Effects					
σ^2	3.29				
τ_{00} pupil id:school	1.16				
τ_{00} item	1.68				
τ_{00} school	.10				
ICC	.47				
N _{pupil id}	60				
N _{school}	5				
N _{item}	15				
Observations	1800				
Marginal R^2 / Conditional R^2	.11 / .53				

Appendix N.34

Pairwise Comparisons for Elicited Imitation Speaking by Time-point and Approach

approach = novel:

contrast	estimate	SE	df	z.ratio	p.value
july - sept	-.16	.14	Inf	-1.08	.28

approach = standard:

contrast	estimate	SE	df	z.ratio	p.value
july - sept	-.47	.15	Inf	-3.17	.002

approach = novel, time_point = july:

item_score	prob	SE	df	asympt.LCL	asympt.UCL
1	.56	.11	Inf	.35	.76
2	.25	.04	Inf	.17	.33
3	.15	.05	Inf	.05	.24
4	.02	.01	Inf	.005	.04
5	.02	.01	Inf	.004	.04

approach = standard, time_point = july:

item_score	prob	SE	df	asympt.LCL	asympt.UCL
1	.66	.10	Inf	.47	.85
2	.20	.05	Inf	.11	.30
3	.10	.04	Inf	.03	.18
4	.02	.007	Inf	.003	.03
5	.01	.006	Inf	.002	.03

approach = novel, time_point = sept:

item_score	prob	SE	df	asympt.LCL	asympt.UCL
1	.52	.11	Inf	.31	.73
2	.26	.04	Inf	.19	.33
3	.17	.05	Inf	.06	.27
4	.03	.01	Inf	.006	.05
5	.03	.01	Inf	.004	.05

approach = standard, time_point = sept:

item_score	prob	SE	df	asympt.LCL	asympt.UCL
1	.55	.11	Inf	.34	.762
2	.25	.04	Inf	.17	.33
3	.15	.05	Inf	.05	.25
4	.02	.01	Inf	.005	.04
5	.02	.01	Inf	.003	.04

Note. Item score 1 represents a score of 0, 2 a score of 1, 3 a score of 2, 4 a score of 3, and 5 the maximum possible score of 4.

Confidence level used: .95

Appendix O.1

Sample Vocabulary Grid Activity

Écris en anglais : Can you get at least 15 points?

 x3	le chat	la table	quoi ?
	répéter	chanter	ou
 x2	avec	tous les jours	facile
	ton	ta	utiliser
 x1	l'amie	mon	ma
	le déjeuner	le weekend	l'ami
	passer	manger	voici



Écris en français : Can you get at least 15 points?

 x3	(the) photo	(the) table	or
	(the) pet, animal	(the) dog	(the) cat
 x2	he has	she has	what
	with	every day	I have
 x1	to sing, singing	to use, using	easy
	(the) boy	of	to repeat, repeating
	(the) party, celebration	(the) girl	(the) fruit



Appendix O.2

Sample Materials from Novel Approach on Menton's Fête du Citron

Follow up 2: Le Carnaval de Menton [1/10]

Décris l'image.



La banane est .



parler

Follow up 2: Le Carnaval de Menton [2/10]

Décris l'image.



Le garçon est .



parler

Follow up 2: Le Carnaval de Menton [3/10]

Décris l'image.



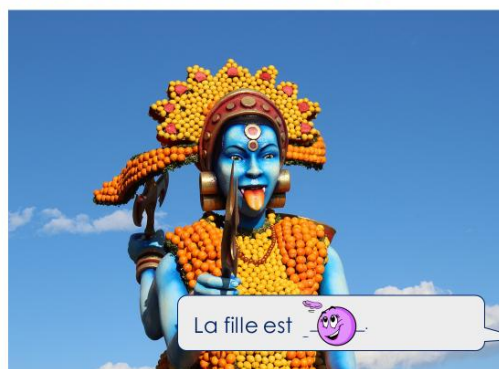
La tête est .



parler

Follow up 2: Le Carnaval de Menton [4/10]

Décris l'image.



La fille est .



parler

Follow up 2: Le Carnaval de Menton [5/10]

Décris l'image.



Le numéro huit est et il porte un .



parler

Follow up 2: Le Carnaval de Menton [6/10]

Décris l'image.



Le Datta est .



parler

Note. This resource is based on Y3_4_rouge_t3_w28_lesson (by Guillou, Marie-Odile; Hawkes, Rachel), licensed under CC BY-NC-SA 4.0.

Appendix P.1

Normality of Distribution (Self-efficacy Data)

PQ1:

Skew and kurtosis measures were acceptable for both Novel and Standard ratings at PQ1. However, the results of the Shapiro-Wilk tests for PQ1 data (Novel: $W = .97, p = .04$; Standard: $W = .97, p = .11$) indicated that data were not normally distributed in Novel (see [Appendix P.2](#)).

PQ2:

In PQ2, the Shapiro-Wilk test results (Novel: $W = .96, p = .02$; Standard: $W = .98, p = .31$) again indicated that data were not normally distributed in the Novel approach data, and this was supported by visual inspection of the data (see [Appendix P.3](#)).

PQ3:

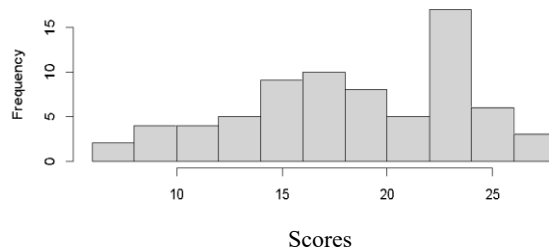
Visual inspection of the data (see [Appendix P.4](#)) suggested that the ratings in Standard approach at PQ3 were not normally distributed, and this was supported by the Shapiro-Wilk test results (Novel: $W = .98, p = .48$; Standard: $W = .97, p = .05$).

Return to [Table 39](#).

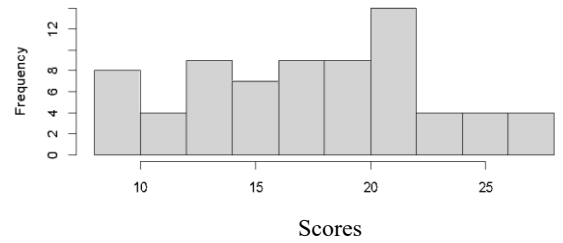
Appendix P.2

Normality of Distribution for PQ1 Self-Efficacy Data

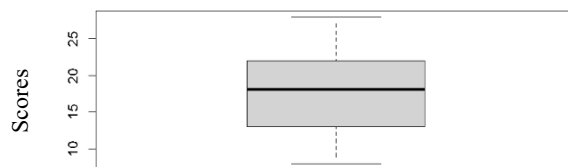
PQ1 Self-efficacy Scores: Novel



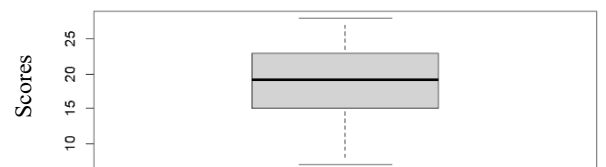
PQ1 Self-efficacy Scores: Standard



PQ1 Self-efficacy Scores: Novel



PQ1 Self-efficacy Scores: Standard



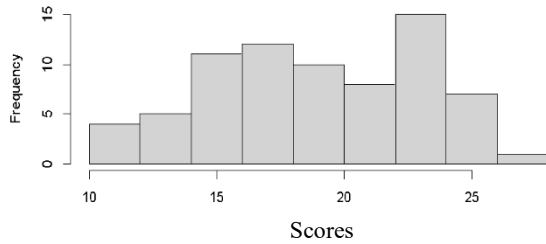
Approach	Self-efficacy Ratings: PQ1	
	Skew	Kurtosis
Novel	.53	.78
Standard	.12	.80

Return to [Table 39](#).

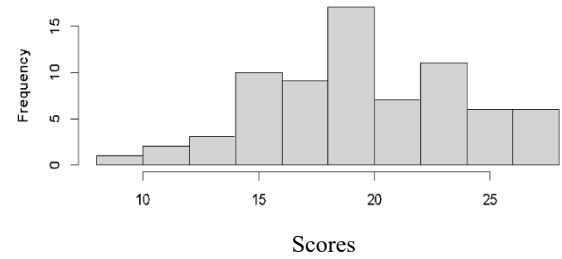
Appendix P.3

Normality of Distribution for PQ2 Self-Efficacy Data

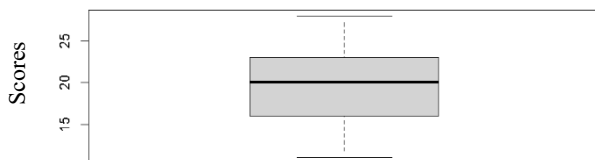
PQ2 Self-efficacy Scores: Novel



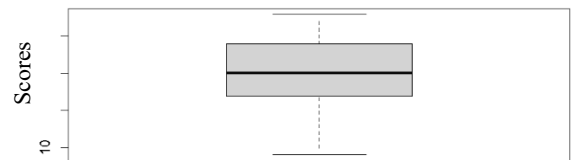
PQ2 Self-efficacy Scores: Standard



PQ2 Self-efficacy Scores: Novel



PQ2 Self-efficacy Scores: Standard



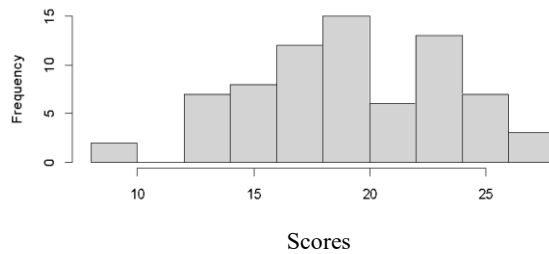
Approach	Self-efficacy Ratings: PQ2	
	Skew	Kurtosis
Novel	-.32	-.94
Standard	-.14	-.46

Return to [Table 39](#).

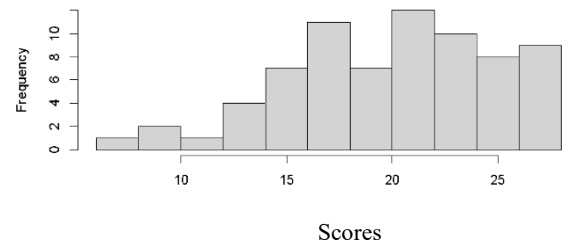
Appendix P.4

Normality of Distribution for PQ3 Self-Efficacy Data

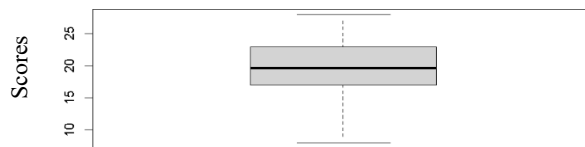
PQ2 Self-efficacy Scores: Novel



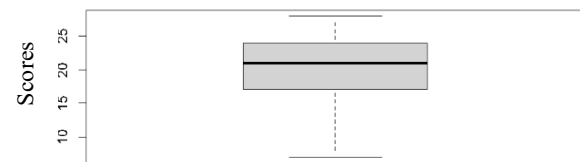
PQ2 Self-efficacy Scores: Standard



PQ2 Self-efficacy Scores: Novel



PQ2 Self-efficacy Scores: Standard



Approach	Self-efficacy Ratings: PQ3	
	Skew	Kurtosis
Novel	-.32	-.31
Standard	-.71	-.36

Return to [Table 39](#).

Appendix P.5

Normality of Distribution (Enjoyment Data)

PQ2:

While skew and kurtosis measures were acceptable for Novel enjoyment ratings at PQ2, the results of the Shapiro-Wilk tests (Novel: $W = .93, p < .001$; Standard: $W = .88, p < .001$) and visual inspection of histograms indicated that data were not normally distributed (see [Appendix P.6](#)).

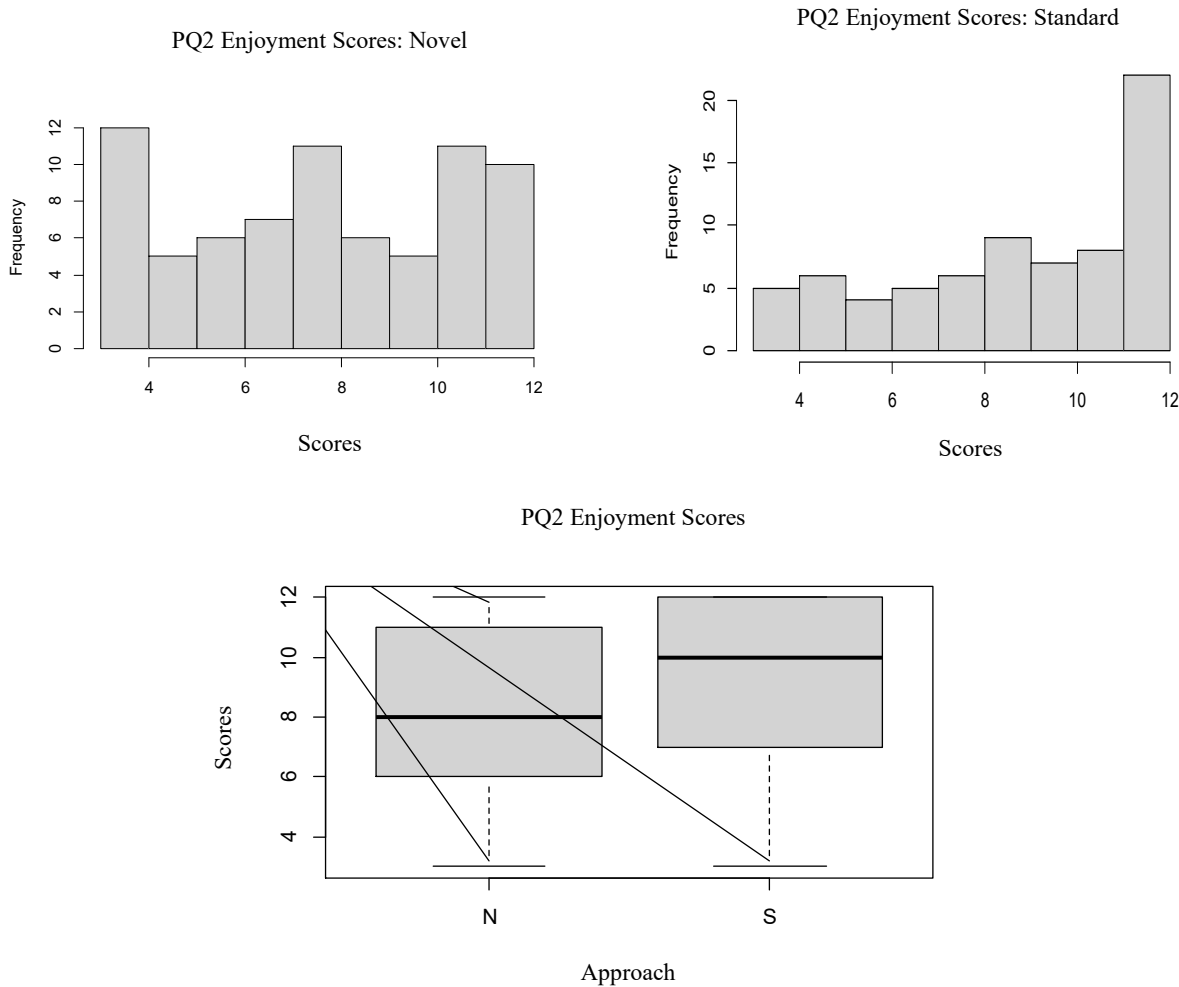
PQ3:

As with PQ2, visual inspection of the PQ3 enjoyment ratings (see [Appendix P.7](#)) indicated that the data were not normally distributed; this was supported by the results of the Shapiro-Wilk tests (Novel: $W = .93, p < .001$; Standard: $W = .92, p < .001$).

Return to [Table 41](#).

Appendix P.6

Normality of Distribution for PQ2 Enjoyment Data (raw ratings)

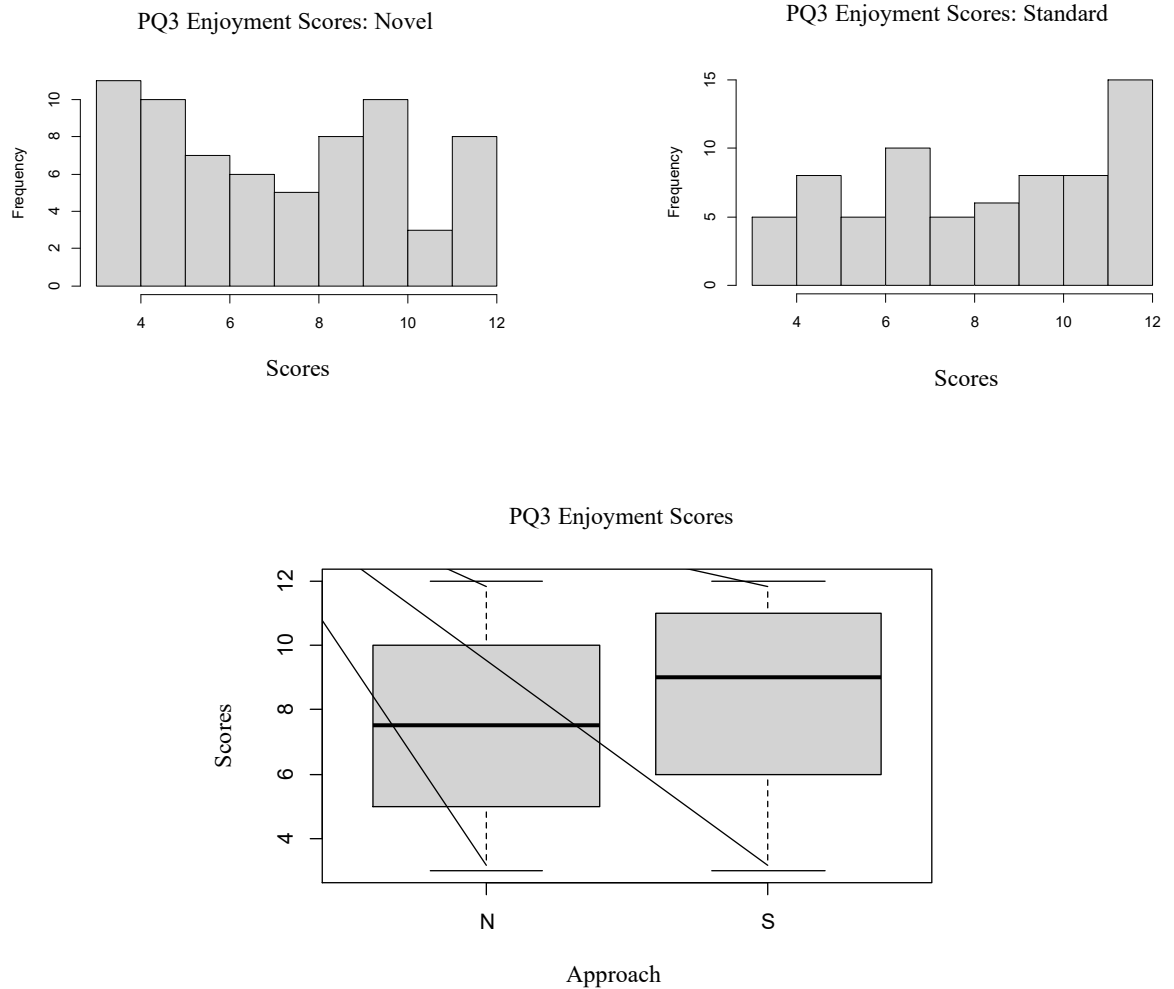


Approach	Enjoyment Ratings PQ2	
	Skew	Kurtosis
Novel	-.21	1.85
Standard	-.61	2.16

Return to [Table 41](#).

Appendix P.7

Normality of Distribution for PQ3 Enjoyment Data (raw ratings)



Approach	Enjoyment Ratings PQ3	
	Skew	Kurtosis
Novel	.05	1.77
Standard	-.29	1.87

Return to [Table 41](#).

Appendix Q.1

Self-efficacy Model Code

```
model_self_efficacy = clmm(response ~ approach*time_point*enjoyment_bas  
eline + (1|school/pupil_id) + (1|question), data = self_efficacy_data_1  
ong)
```

Appendix Q.2

Brant Test Results for Self-efficacy Ratings

Omnibus	< .001
approach.standard-novel	.37
time_point.t2-t1	< .001
time_point.t3-t1	.005
enjoyment_baseline.Unsure-No	.04
enjoyment_baseline.Yes-No	.10
approach.standard-novel:time_point.t2-t1	.95
approach.standard-novel:time_point.t3-t1	.96
approach.standard-novel:enjoyment_baseline.Unsure-No	.54
approach.standard-novel:enjoyment_baseline.Yes-No	.68
time_point.t2-t1:enjoyment_baseline.Unsure-No	.03
time_point.t3-t1:enjoyment_baseline.Unsure-No	.52
time_point.t2-t1:enjoyment_baseline.Yes-No	.02
time_point.t3-t1:enjoyment_baseline.Yes-No	.72
approach.standard-novel:time_point.t2-t1:enjoyment_baseline.Unsure-No	.06
approach.standard-novel:time_point.t3-t1:enjoyment_baseline.Unsure-No	.002
approach.standard-novel:time_point.t2-t1:enjoyment_baseline.Yes-No	.70
approach.standard-novel:time_point.t3-t1:enjoyment_baseline.Yes-No	.09

Appendix Q.3

Summary of Self-efficacy Model

<i>Predictors</i>	response				
	<i>Odds Ratios</i>	<i>std. Error</i>	<i>CI</i>	<i>Statistic</i>	<i>p</i>
1 2	.15	.04	.09 – .24	-7.80	< .001
2 3	.63	.15	.39 – 1.02	-1.90	.06
3 4	4.34	1.07	2.68 – 7.02	5.98	< .001
approach [.standard-novel]	1.12	.33	.63 – 1.99	.38	.70
time-point [.t2-t1]	1.71	.24	1.29 – 2.26	3.73	< .001
time-point [.t3-t1]	1.52	.22	1.14 – 2.02	2.85	.004
positive anticipation baseline [.Unsure-No]	1.22	.52	.52 – 2.82	.46	.65
positive anticipation baseline [.Yes-No]	2.34	1.00	1.01 – 5.39	1.99	.05
approach [.standard-novel] × time point [.t2-t1]	1.65	.47	.94 – 2.89	1.75	.08
approach [.standard-novel] × time point [.t3-t1]	2.13	.62	1.20 – 3.78	2.58	.01
approach [.standard-novel] × positive anticipation baseline [.Unsure-No]	1.13	.97	.21 – 6.09	.14	.89
approach [.standard-novel] × positive anticipation baseline [.Yes-No]	.74	.62	.14 – 3.81	-.36	.72
time-point [.t2-t1] × positive anticipation baseline [.Unsure-No]	.73	.30	.33 – 1.66	-.74	.46
time-point [.t3-t1] × positive anticipation baseline [.Unsure-No]	1.36	.58	.59 – 3.13	.73	.47

time-point [.t2-t1] × positive anticipation baseline [.Yes-No]	.78	.32	.35 – 1.72	-.62	.54
time-point [.t3-t1] × positive anticipation baseline [.Yes-No]	1.62	.67	.72 – 3.66	1.16	.25
(approach [.standard-novel] × time point [.t2-t1]) × positive anticipation baseline [.Unsure-No]	1.46	1.21	.29 – 7.40	.45	.65
(approach [.standard-novel] × time point [.t3-t1]) × positive anticipation baseline [.Unsure-No]	.61	.52	.12 – 3.22	-.58	.56
(approach [.standard-novel] × time point [.t2-t1]) × positive anticipation baseline [.Yes-No]	1.02	.82	.21 – 4.97	.02	.98
(approach [.standard-novel] × time point [.t3-t1]) × positive anticipation baseline [.Yes-No]	.54	.45	.11 – 2.76	-.74	.46
Random Effects					
σ^2	3.29				
τ_{00} pupil_id:school	.92				
τ_{00} question	.09				
τ_{00} school	.10				
ICC	.25				
N _{pupil_id}	146				
N _{school}	5				
N _{question}	7				
Observations	3066				
Marginal R^2 / Conditional R^2	.05 / .29				

Appendix Q.4

Pairwise Comparisons for Self-efficacy Ratings by Time-point

time_point = t1, approach = novel, enjoyment_baseline = No:

response	prob	SE	df	asympt.LCL	asympt.UCL
1	.19	.10	Inf	-.00	.38
2	.31	.06	Inf	.19	.43
3	.37	.09	Inf	.20	.55
4	.13	.07	Inf	-.01	.26

time_point = t2, approach = novel, enjoyment_baseline = No:

response	prob	SE	df	asympt.LCL	asympt.UCL
1	.12	.07	Inf	-.01	.25
2	.25	.08	Inf	.09	.41
3	.43	.05	Inf	.33	.53
4	.20	.10	Inf	-.00	.39

time_point = t3, approach = novel, enjoyment_baseline = No:

response	prob	SE	df	asympt.LCL	asympt.UCL
1	.26	.13	Inf	.01	.51
2	.34	.03	Inf	.28	.41
3	.31	.11	Inf	.10	.52
4	.09	.05	Inf	-.02	.19

time_point = t1, approach = standard, enjoyment_baseline = No:

response	prob	SE	df	asympt.LCL	asympt.UCL
1	.25	.14	Inf	-.02	.52
2	.34	.04	Inf	.25	.43
3	.32	.12	Inf	.09	.55
4	.09	.06	Inf	-.03	.22

time_point = t2, approach = standard, enjoyment_baseline = No:

response	prob	SE	df	asympt.LCL	asympt.UCL
1	.12	.07	Inf	-.03	.26
2	.25	.09	Inf	.07	.43

3	.43	.05	Inf	.34	.53
4	.20	.12	Inf	-.02	.43

time_point = t3, approach = standard, enjoyment_baseline = No:

response	prob	SE	df	asympt.LCL	asympt.UCL
1	.14	.09	Inf	-.03	.30
2	.27	.09	Inf	.10	.44
3	.42	.07	Inf	.28	.56
4	.17	.10	Inf	-.03	.38

time_point = t1, approach = novel, enjoyment_baseline = Unsure:

response	prob	SE	df	asympt.LCL	asympt.UCL
1	.17	.04	Inf	.09	.26
2	.30	.03	Inf	.24	.37
3	.39	.04	Inf	.31	.46
4	.14	.04	Inf	.07	.21

time_point = t2, approach = novel, enjoyment_baseline = Unsure:

response	prob	SE	df	asympt.LCL	asympt.UCL
1	.17	.04	Inf	.09	.25
2	.30	.03	Inf	.23	.36
3	.39	.04	Inf	.32	.47
4	.14	.04	Inf	.07	.21

time_point = t3, approach = novel, enjoyment_baseline = Unsure:

response	prob	SE	df	asympt.LCL	asympt.UCL
1	.15	.04	Inf	.08	.23
2	.29	.03	Inf	.22	.35
3	.41	.03	Inf	.34	.47
4	.16	.04	Inf	.08	.23

time_point = t1, approach = standard, enjoyment_baseline = Unsure:

response	prob	SE	df	asympt.LCL	asympt.UCL
1	.20	.05	Inf	.10	.30
2	.32	.03	Inf	.26	.38

3	.36	.05	Inf	.27	.46
4	.12	.03	Inf	.05	.19

time_point = t2, approach = standard, enjoyment_baseline = Unsure:

response	prob	SE	df	asympt.LCL	asympt.UCL
1	.10	.03	Inf	.04	.16
2	.23	.04	Inf	.14	.31
3	.44	.02	Inf	.41	.48
4	.23	.06	Inf	.12	.34

time_point = t3, approach = standard, enjoyment_baseline = Unsure:

response	prob	SE	df	asympt.LCL	asympt.UCL
1	.10	.03	Inf	.04	.16
2	.23	.04	Inf	.14	.31
3	.44	.02	Inf	.41	.48
4	.23	.06	Inf	.12	.34

time_point = t1, approach = novel, enjoyment_baseline = Yes:

response	prob	SE	df	asympt.LCL	asympt.UCL
1	.09	.02	Inf	.05	.14
2	.21	.03	Inf	.15	.28
3	.45	.01	Inf	.42	.47
4	.25	.05	Inf	.15	.34

time_point = t2, approach = novel, enjoyment_baseline = Yes:

response	prob	SE	df	asympt.LCL	asympt.UCL
1	.07	.02	Inf	.04	.11
2	.18	.03	Inf	.12	.24
3	.45	.01	Inf	.42	.47
4	.30	.06	Inf	.19	.41

time_point = t3, approach = novel, enjoyment_baseline = Yes:

response	prob	SE	df	asympt.LCL	asympt.UCL
1	.07	.02	Inf	.03	.10
2	.17	.03	Inf	.11	.23

3	.44	.01	Inf	.42	.47
4	.32	.06	Inf	.21	.44

time_point = t1, approach = standard, enjoyment_baseline = Yes:

response	prob	SE	df	asympt.LCL	asympt.UCL
1	.14	.03	Inf	.08	.20
2	.27	.03	Inf	.21	.33
3	.42	.03	Inf	.37	.47
4	.18	.04	Inf	.10	.25

time_point = t2, approach = standard, enjoyment_baseline = Yes:

response	prob	SE	df	asympt.LCL	asympt.UCL
1	.08	.02	Inf	.04	.11
2	.19	.03	Inf	.12	.25
3	.45	.01	Inf	.43	.47
4	.29	.05	Inf	.19	.40

time_point = t3, approach = standard, enjoyment_baseline = Yes:

response	prob	SE	df	asympt.LCL	asympt.UCL
1	.06	.02	Inf	.03	.09
2	.16	.03	Inf	.10	.22
3	.44	.02	Inf	.41	.47
4	.34	.06	Inf	.23	.45

Confidence level used: .95

Appendix Q.5

Pairwise Comparisons for Self-Efficacy Ratings by Approach

time_point = t1, enjoyment_baseline = No:

contrast	estimate	SE	df	z.ratio	p.value
novel - standard	.33	.93	Inf	.35	.72

time_point = t2, enjoyment_baseline = No:

contrast	estimate	SE	df	z.ratio	p.value
novel - standard	-.04	.91	Inf	-.05	.96

time_point = t3, enjoyment_baseline = No:

contrast	estimate	SE	df	z.ratio	p.value
novel - standard	-.80	.93	Inf	-.86	.39

time_point = t1, enjoyment_baseline = Unsure:

contrast	estimate	SE	df	z.ratio	p.value
novel - standard	.16	.34	Inf	.49	.63

time_point = t2, enjoyment_baseline = Unsure:

contrast	estimate	SE	df	z.ratio	p.value
novel - standard	-.58	.34	Inf	-1.73	.08

time_point = t3, enjoyment_baseline = Unsure:

contrast	estimate	SE	df	z.ratio	p.value
novel - standard	-.47	.34	Inf	-1.38	.17

time_point = t1, enjoyment_baseline = Yes:

contrast	estimate	SE	df	z.ratio	p.value
novel - standard	.43	.26	Inf	1.64	.10

time_point = t2, enjoyment_baseline = Yes:

contrast	estimate	SE	df	z.ratio	p.value
novel - standard	.04	.26	Inf	.15	.88

time_point = t3, enjoyment_baseline = Yes:

contrast	estimate	SE	df	z.ratio	p.value
novel - standard	-.08	.26	Inf	-.33	.74

Appendix Q.6

Enjoyment Model Code

```
model_enjoyment = clmm(response ~ approach*time_point*enjoyment_baseline + (1|school/pupil_id) + (1|question), data = na.omit(enjoyment_data_long))
```


Appendix Q.7

Brant Test Results for Enjoyment Ratings

Omnibus	< .001
approach.standard-novel	.05
time_point.t3-t2	.23
enjoyment_baseline.Unsure-No	< .001
enjoyment_baseline.Yes-No	< .001
approach.standard-novel:time_point.t3-t2	.09
approach.standard-novel:enjoyment_baseline.Unsure-No	.65
approach.standard-novel:enjoyment_baseline.Yes-No	.19
time_point.t3-t2:enjoyment_baseline.Unsure-No	.24
time_point.t3-t2:enjoyment_baseline.Yes-No	.32
approach.standard-novel:time_point.t3-t2:enjoyment_baseline.Unsure-No	.14
approach.standard-novel:time_point.t3-t2:enjoyment_baseline.Yes-No	.04

Appendix Q.8

Summary of Enjoyment Model

<i>Predictors</i>	response				
	<i>Odds Ratios</i>	<i>std. Error</i>	<i>CI</i>	<i>Statistic</i>	<i>p</i>
1 2	.13	.05	.06 – .29	-4.93	< . .001
2 3	.85	.35	.38 – 1.90	-.40	.69
3 4	5.46	2.27	2.42 – 12.32	4.09	< . .001
approach [.standard-novel]	1.40	.79	.46 – 4.26	.59	.56
time-point [.t3-t2]	.78	.20	.47 – 1.28	-.98	.33
positive anticipation baseline [.Unsure-No]	3.16	2.62	.62 – 16.08	1.39	.17
positive anticipation baseline [.Yes-No]	8.67	7.09	1.74 – 43.10	2.64	.008
approach [.standard-novel] × time point [.t3-t2]	1.05	.54	.38 – 2.88	.10	.92
approach [.standard-novel] × positive anticipation baseline [.Unsure-No]	6.89	11.46	.26 – 179.25	1.16	.25
approach [.standard-novel] × positive anticipation baseline [.Yes-No]	4.70	7.60	.20 – 111.68	.96	.34
time-point [.t3-t2] × positive anticipation baseline [.Unsure-No]	.53	.40	.12 – 2.32	-.84	.40
time-point [.t3-t2] × positive anticipation baseline [.Yes-No]	.50	.37	.12 – 2.13	-.93	.35
(approach [.standard-novel] × time point [.t3-t2]) × positive anticipation baseline [.Unsure-No]	.27	.41	.01 – 5.17	-.86	.39
(approach [.standard-novel] × time	.42	.62	.02 – 7.47	-.59	.55

point [.t3-t2)) ×
 positive anticipation baseline
 [.Yes-No]

Random Effects

σ^2	3.29
τ_{00} pupil_id:school	3.19
τ_{00} school	.08
τ_{00} question	.19
ICC	.51
N pupil_id	146
N school	5
N question	3
Observations	876
Marginal R^2 / Conditional R^2	.09 / .56

Appendix Q.9

Pairwise Comparisons for Enjoyment Ratings of Pupils Not Looking Forward to French by Approach and Time-point

time_point = t2, approach = novel, enjoyment_baseline = No:

response	prob	SE	df	asympt.LCL	asympt.UCL
1	.19	.17	Inf	-.14	.52
2	.42	.10	Inf	.22	.61
3	.30	.17	Inf	-.04	.64
4	.09	.09	Inf	-.09	.27

time_point = t3, approach = novel, enjoyment_baseline = No:

response	prob	SE	df	asympt.LCL	asympt.UCL
1	.22	.20	Inf	-.17	.61
2	.43	.07	Inf	.30	.57
3	.27	.18	Inf	-.08	.63
4	.08	.08	Inf	-.08	.24

time_point = t2, approach = standard, enjoyment_baseline = No:

response	prob	SE	df	asympt.LCL	asympt.UCL
1	.44	.34	Inf	-.23	1.11
2	.40	.15	Inf	.10	.70
3	.13	.15	Inf	-.16	.43
4	.03	.04	Inf	-.05	.11

time_point = t3, approach = standard, enjoyment_baseline = No:

response	prob	SE	df	asympt.LCL	asympt.UCL
1	.30	.28	Inf	-.25	.86
2	.44	.04	Inf	.37	.51
3	.21	.19	Inf	-.17	.58
4	.05	.06	Inf	-.08	.18

Appendix Q.10

Pairwise Comparisons for Enjoyment Ratings of Unsure Pupils by Approach and Time-point

time_point = t2, approach = novel, enjoyment_baseline = Unsure:

response	prob	SE	df	asympt.LCL	asympt.UCL
1	.16	.07	Inf	.03	.29
2	.40	.06	Inf	.29	.51
3	.33	.07	Inf	.18	.48
4	.11	.05	Inf	.01	.20

time_point = t3, approach = novel, enjoyment_baseline = Unsure:

response	prob	SE	df	asympt.LCL	asympt.UCL
1	.19	.08	Inf	.04	.34
2	.42	.05	Inf	.32	.51
3	.30	.08	Inf	.15	.45
4	.09	.04	Inf	.01	.17

time_point = t2, approach = standard, enjoyment_baseline = Unsure:

response	prob	SE	df	asympt.LCL	asympt.UCL
1	.05	.03	Inf	-.00	.10
2	.20	.08	Inf	.05	.35
3	.43	.03	Inf	.37	.49
4	.32	.12	Inf	.09	.56

time_point = t3, approach = standard, enjoyment_baseline = Unsure:

response	prob	SE	df	asympt.LCL	asympt.UCL
1	.09	.05	Inf	.00	.18
2	.31	.09	Inf	.14	.48
3	.41	.05	Inf	.31	.51
4	.19	.08	Inf	.02	.36

Confidence level used: .95

Appendix Q.11

Pairwise Comparisons for Enjoyment Ratings of Pupils with a Positive Baseline by Approach and Time-point

time_point = t2, approach = novel, enjoyment_baseline = Yes:

response	prob	SE	df	asympt.LCL	asympt.UCL
1	.05	.02	Inf	.00	.09
2	.20	.06	Inf	.08	.33
3	.43	.03	Inf	.38	.48
4	.31	.10	Inf	.13	.50

time_point = t3, approach = novel, enjoyment_baseline = Yes:

response	prob	SE	df	asympt.LCL	asympt.UCL
1	.07	.03	Inf	.01	.13
2	.27	.07	Inf	.13	.41
3	.43	.03	Inf	.36	.49
4	.23	.08	Inf	.07	.38

time_point = t2, approach = standard, enjoyment_baseline = Yes:

response	prob	SE	df	asympt.LCL	asympt.UCL
1	.02	.01	Inf	.00	.04
2	.11	.04	Inf	.03	.19
3	.37	.06	Inf	.25	.48
4	.50	.11	Inf	.29	.71

time_point = t3, approach = standard, enjoyment_baseline = Yes:

response	prob	SE	df	asympt.LCL	asympt.UCL
1	.04	.02	Inf	.00	.07
2	.17	.06	Inf	.06	.28
3	.42	.04	Inf	.35	.49
4	.36	.10	Inf	.17	.56

Confidence level used: .95

Appendix R.1

Additional Pupil Quotations

Confidence:

Start of the study:

Anticipating difficulty in learning French:

(PQ1) Standard: “*i think i will not remember french that well in year 3*” [sic])

Expressing concerns about forgetting:

(PQ1) Novel: “*Because I'm a bit forgetful*”

Concern about not being good at languages / French:

(PQ1) Novel: “*I'm not very good a [sic] French*”;

Self-efficacy

Mid-point of the study:

(Group interview) Standard: “*now in April, erm, I'm going to go to Spain, and there's a French, erm, erm, friend there. I could speak French with that girl.*” (MLAA, SchB)

(Group interview) Novel: “*I definitely think, if I keep learning French at this pace, and at this knowledge, I think I might be able to read, like, some books in French ... Like maybe some short books, but then, when I get really confident, I may be able to read thicker and thicker and thicker books*”. (HLAA, SchA)

(Group interview) Novel: “*I might also be able to, erm, say paragraphs*”. (MLAA, SchB)

End-point of the study

(Group interview) Standard: “*I would be able to, like, write a couple of paragraphs, or just say ... so, erm, write a couple of paragraphs and I'll be able to say, say what I've writ ...*

[Read it aloud, you mean?]

Like, read it aloud ...” (HLAA, SchA)

(Group interview) Standard: “*I think when the new French teacher asks us something, that she doesn't know that we've already learned, then we will be able to say it ... She might actually be, like, 'Wow'”* (LLAA, SchC).

(Group interview) Novel: *“say other words to other people if I go to other countries”*.
(LLAA, SchA)

Enjoyment:

Start of the study

Familial connections:

(PQ1) Novel: *“my brother nose it”*.

(PQ1) Standard: *“Because my big brother is good at French”*.

Dislikes:

Mid-point

Classroom factors:

(Group interview) Standard: *some get, like – teachers – get really frustrated ... because the kids don’t listen. [laughs]*

[Does that happen in French?]

I know ... I think, like, yes and no, because, like, we sometimes, like ... [teacher] might get a bit, like ... when we’re all chatting really loud, he might get ... (voice trails off) (LLAA, SchE)

(Group interview) Standard

“[Do you think they like teaching you French?

[silence]

[No ?]

Cause they [pupils] keep talking over every single time when they talk, and they have ... erm, and they have to, erm, get told off and stuff”. (LLAA, SchF)

(Group interview) Novel:

“I don’t like when people, erm, like ... do something wrong, and then we have to stop. And then, like, for ages, and then we just go back on it again.

[When you say they do something wrong, do you mean like they do something wrong in French, so you have to go over it, or do you mean like ... they’re not behaving?]

They’re not behaving ... Because I forget everything”. (LLAA, SchA)

Appendix R.2

Examples from Audio Recordings Supporting Pupils' Enjoyment of Vocabulary Grid Activities

The following notes were made by the researcher when observing lessons taught by different teachers in the last four weeks of the Novel scheme of work.

Week	Teacher	Comment
26	T7	<i>Vocab grid – the pupils seemed quite excited at doing the vocab grid.</i>
30	T2	<i>Slide 11 – swapped to doing the vocab grid. Pupils seemed pleased to be doing this.</i>
28	T3	<i>Started the vocab grid activity at 37 mins. Pupils seemed to settle well to this.</i>
26	T9	<i>36 mins – gave out vocab grids. Pupils settled quietly to doing this.</i>

Appendix R.3

Full Quotation from T1 (Interview1)

I am a complete novice. Which I am equally really frustrated with ... heading up French in the school and also from my own personal point of view, I just can't understand how in this country we've not nailed it ... something isn't right ... we're doing something wrong. And yeah, I find it really frustrating ... There are certain words that we need to learn – I'm looking around now - like 'table', 'wall', 'computer' – but, to make it work alongside the grammar side, by doing the Novel approach it should, in theory, I think, it makes the topic, the words, the language clear. That's how I'm thinking with it. And that could be the difference between ... the Standard approach which is very topic-based, which is pretty much what we do here at school, which is just from memory – literally, 'You need to learn, you need to learn, you need to learn, why are you not learning?' And it's probably why I'm not learning, because it's the wrong way of learning. So, I'm hoping that the Novel approach will just kind of cushion us with a foundation. (T1)

Appendix S.1

Normality of Distribution for Pupil Perceptions of Teacher Enjoyment: Q1 “Do Your Teachers Enjoy Teaching You French?”

PQ2:

While kurtosis measures were acceptable for both Novel and Standard enjoyment ratings at PQ2, the results of the Shapiro-Wilk tests (Novel: $W = .71, p < .001$; Standard: $W = .72, p < .001$), and visual inspection of histograms indicated that data were not normally distributed in Novel (see [Appendix S.2](#)).

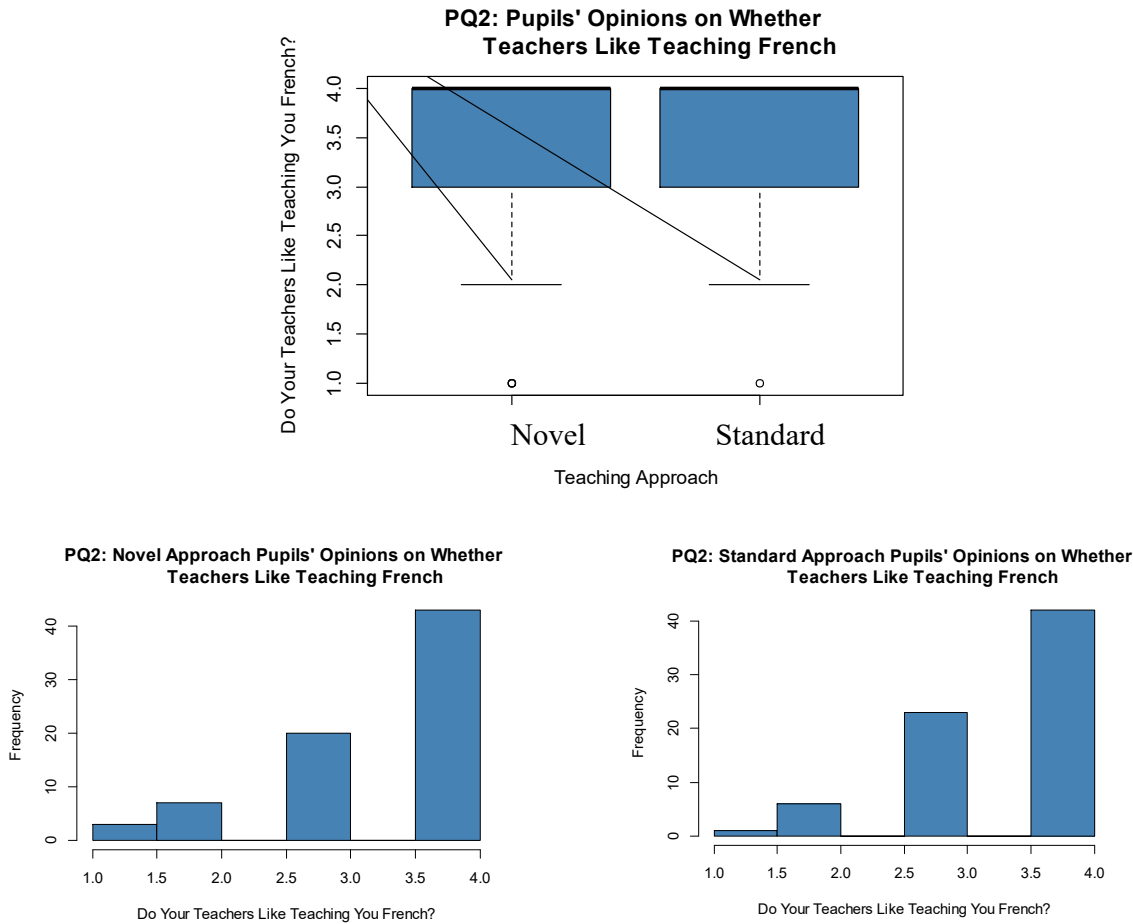
PQ3:

As with PQ2, visual inspection of the PQ3 enjoyment ratings (see [Appendix S.3](#)) indicated that the data were not normally distributed; this was supported by the results of the Shapiro-Wilk tests (Novel: $W = .78, p < .001$; Standard: $W = .73, p < .001$).

Return to [section 6.8.1](#).

Appendix S.2

Normality of Distribution for PQ2 Pupil Perceptions of Teacher Enjoyment: Q1 “Do Your Teachers Enjoy Teaching You French?”

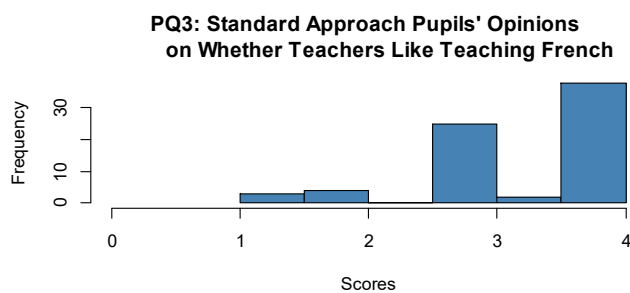
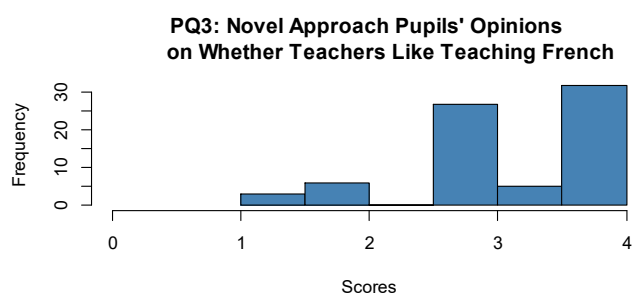
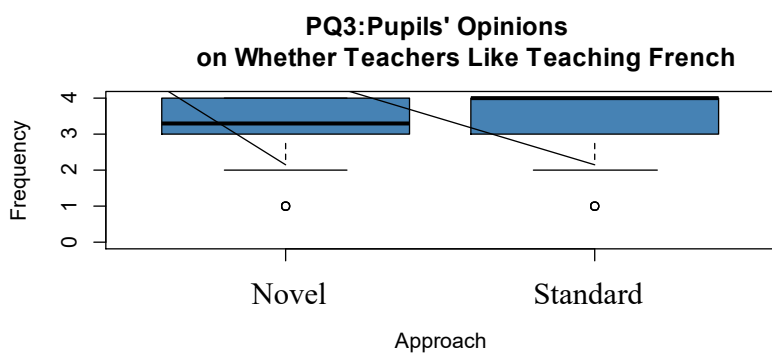


Approach	Pupil Perceptions of Teacher Enjoyment Q1 Ratings PQ2	
	Skew	Kurtosis
Novel	-2.31	.79
Standard	-2.08	.71

Return to [section 6.8.1](#)

Appendix S.3

Normality of Distribution for PQ3 Pupil Perceptions of Teacher Enjoyment: Q1 “Do Your Teachers Enjoy Teaching You French?”



Approach	Pupil Perceptions of Teacher	Enjoyment Q1 Ratings PQ3
	Skew	Kurtosis
Novel	-1.96	.85
Standard	-2.40	1.44

Return to [section 6.8.1](#)

Appendix S.4

Normality of Distribution for Pupil Perceptions of Teacher Enjoyment: Q2 “Do Your Teachers Try to Make You Excited to Learn French?”

PQ2:

Kurtosis measures were acceptable for both Novel and Standard enjoyment ratings at PQ2. However, the results of the Shapiro-Wilk tests (Novel: $W = .74, p < .001$; Standard: $W = .66, p < .001$), and visual inspection of histograms indicated that data were not normally distributed in Novel (see [Appendix S.5](#)).

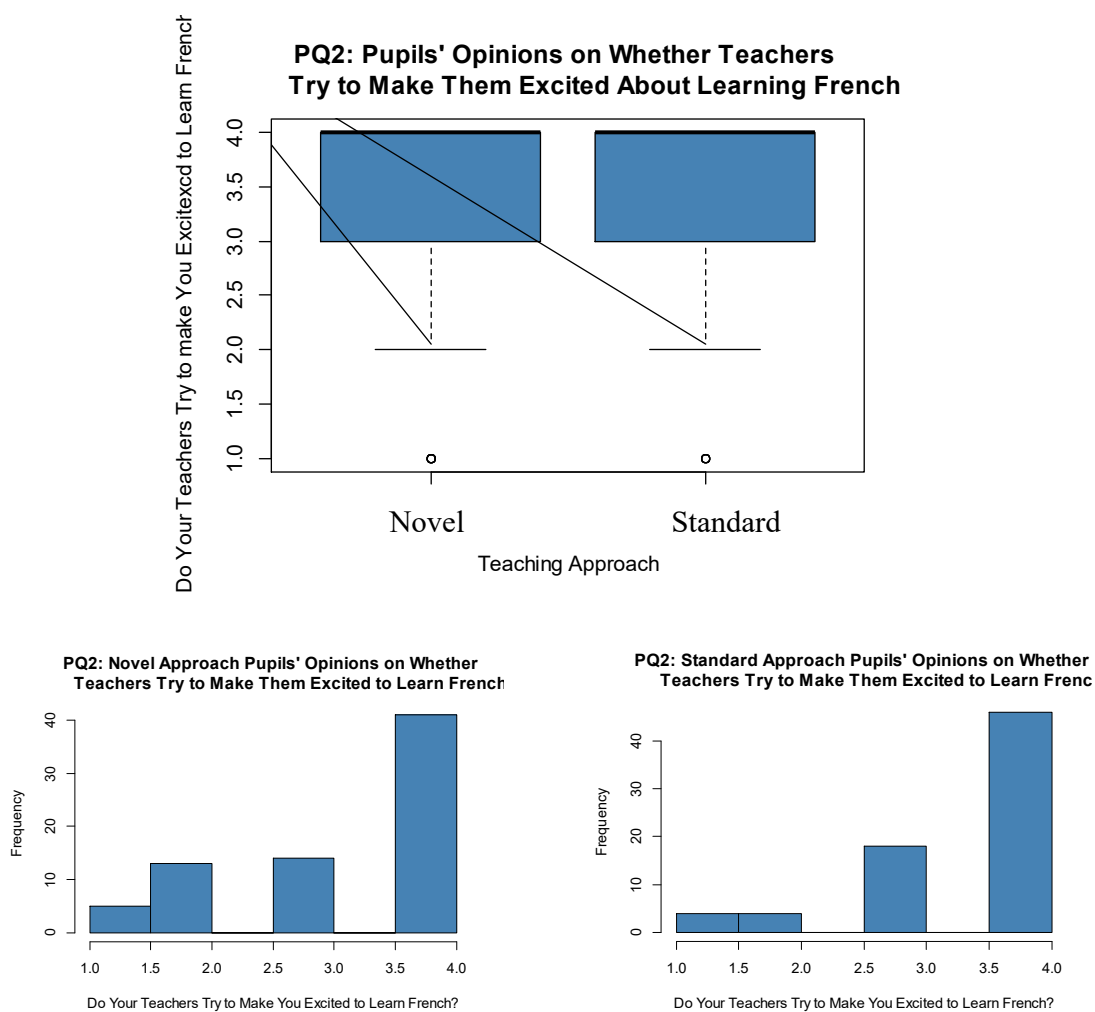
PQ3:

As with PQ2, visual inspection of the PQ3 enjoyment ratings (see [Appendix S.6](#)) indicated that the data were not normally distributed; this was supported by the results of the Shapiro-Wilk tests (Novel: $W = .7, p < .001$; Standard: $W = .70, p < .001$).

Return to [section 6.8.1](#).

Appendix S.5

Normality of Distribution for PQ2 Pupil Perceptions of Teacher Enjoyment: Q2 “Do Your Teachers Try to Make You Excited to Learn French?”

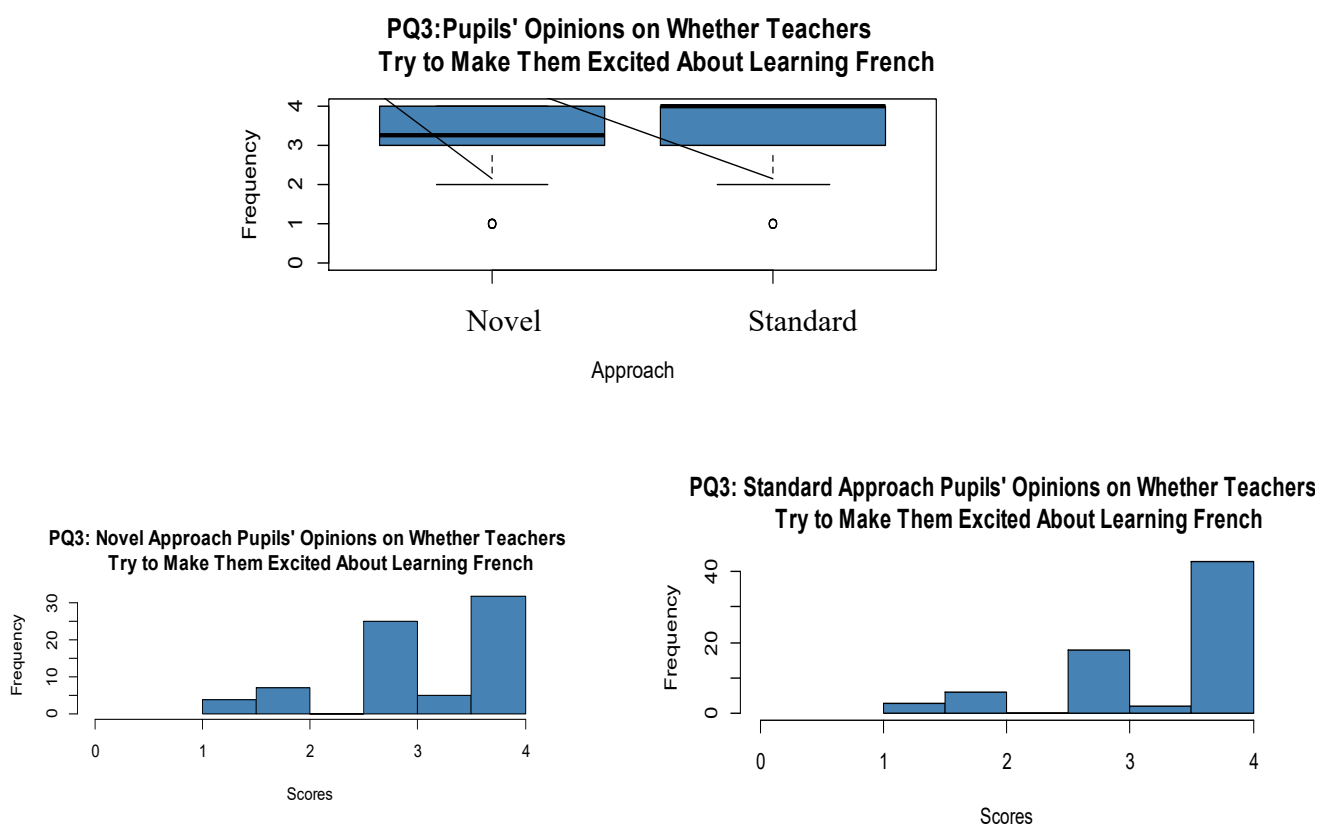


Approach	Pupil Perceptions of Teacher Enjoyment Q21 Ratings PQ2	
	Skew	Kurtosis
Novel	-1.66	-.43
Standard	-2.84	1.85

Return to [section 6.8.1](#).

Appendix S.6

Normality of Distribution for PQ3 Pupil Perceptions of Teacher Enjoyment: Q2 “Do Your Teachers Try to Make You Excited to Learn French?”



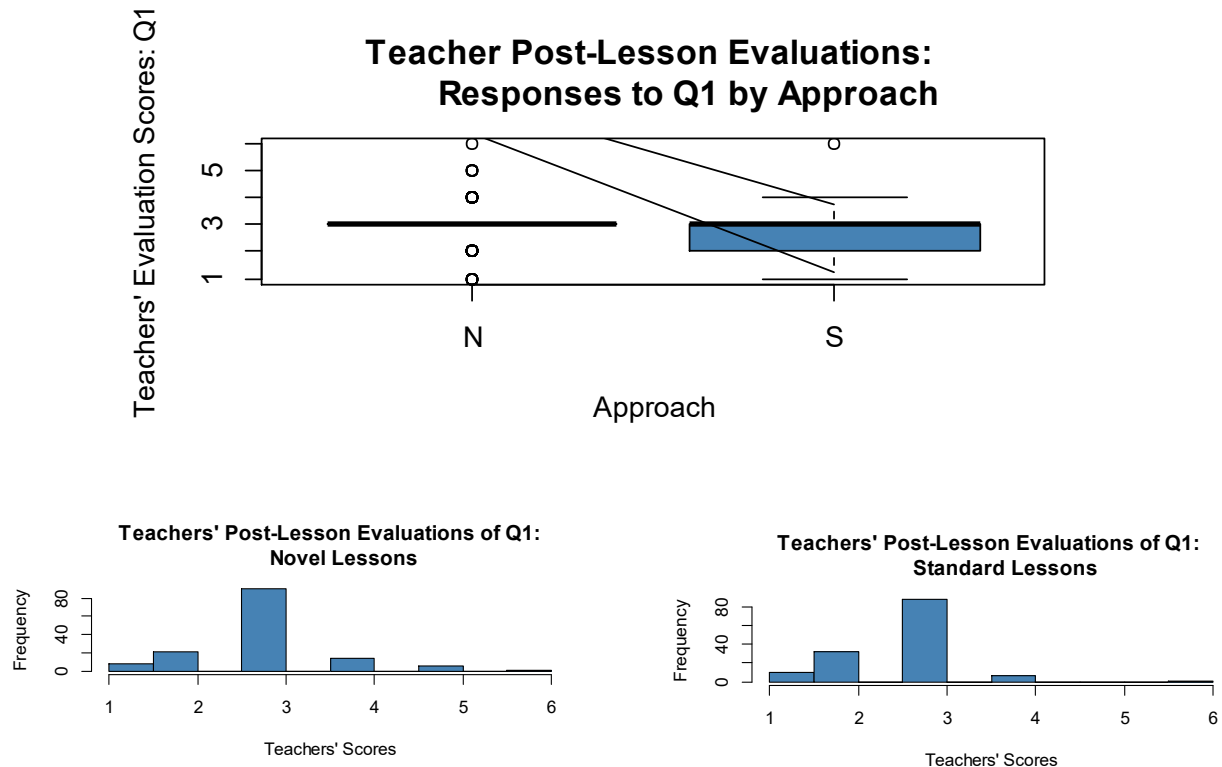
Approach	Pupil Perceptions of Teacher Enjoyment Q2 Ratings PQ3	
	Skew	Kurtosis
Novel	-1.90	.52
Standard	-2.53	1.18

Return to [section 6.8.1](#).

Appendix T.1

Normality of Distribution Post-lesson Evaluation: Item 1 “I enjoyed this lesson”

Normality of distribution was checked using visual inspections as recommended by Larson-Hall (2016), as well as through the Shapiro-Wilk test. The p -values in both approaches indicated that data were not normally distributed (Novel: $W = .80, p < .001$; Standard: $W = .75, p < .001$), supporting conclusions drawn from visual inspections of the data.

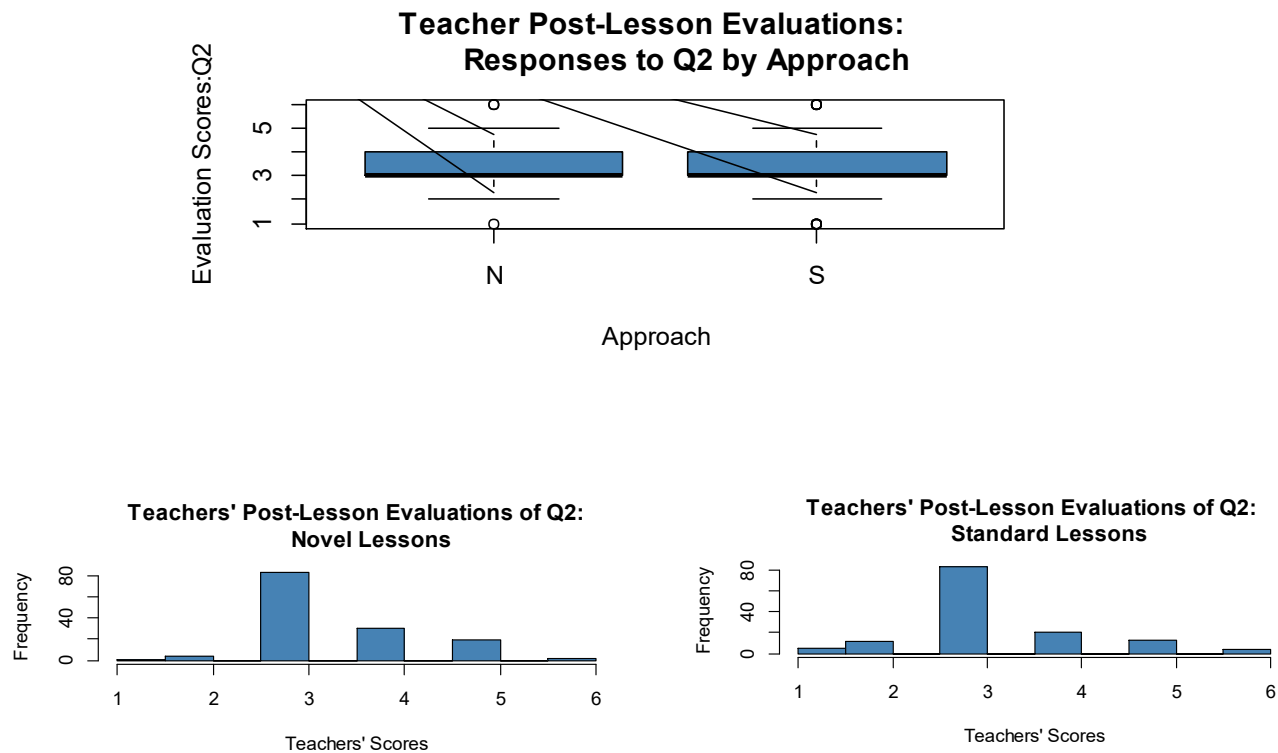


Return to [7.4.1](#).

Appendix T.2

Normality of Distribution Post-lesson Evaluation: Item 2 “I was tense during this lesson”

Normality of distribution was checked using visual inspections, as well as through the Shapiro-Wilk test. The histograms and the p -values in both approaches indicated that data were not normally distributed (Novel: $W = .79, p < .001$; Standard: $W = .83, p < .001$).

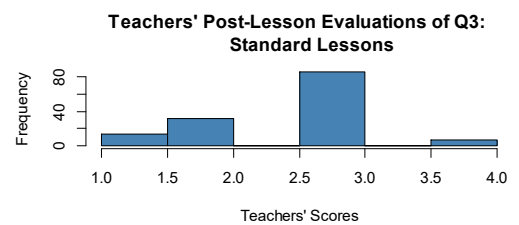
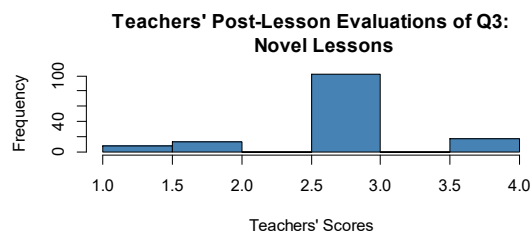
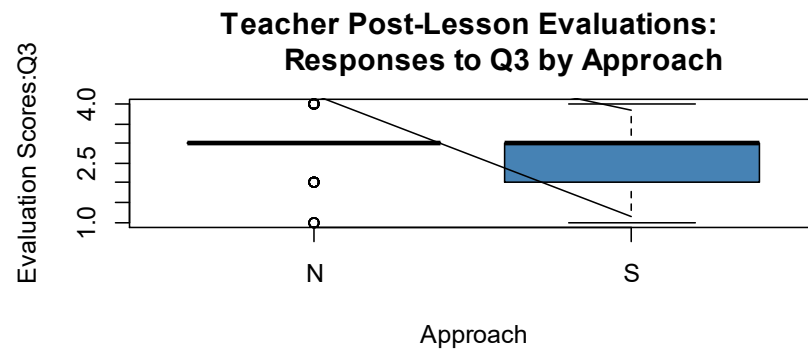


Return to [7.4.1](#).

Appendix T.3

Normality of Distribution Post-lesson Evaluation: Item 3 “Students understood the material during this lesson”

The Shapiro-Wilk tests indicated that data from both approaches were not normally distributed (Novel: $W = .69, p < .001$; Standard: $W = .76, p < .001$).

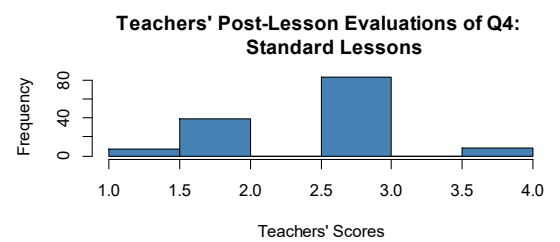
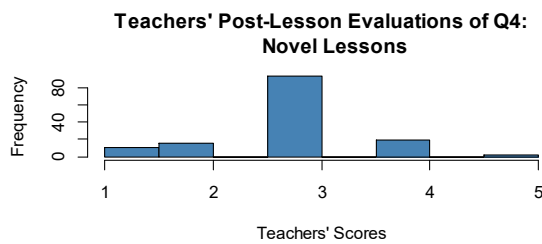
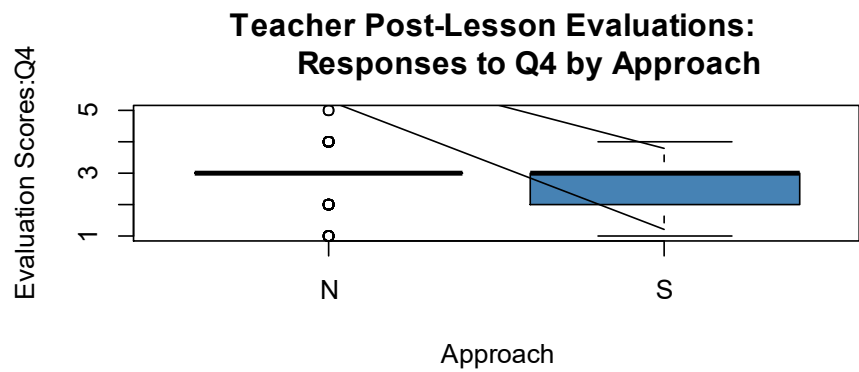


Return to [7.4.1](#).

Appendix T.4

Normality of Distribution Post-lesson Evaluation: Item 4 “Students were motivated in this lesson”

Visual inspections of data and the Shapiro-Wilk test indicated that data from both approaches were not normally distributed (Novel: $W = .77, p < .001$; Standard: $W = .78, p < .001$).

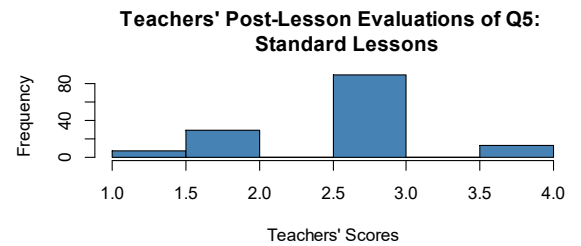
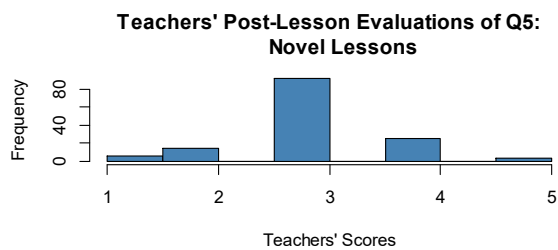
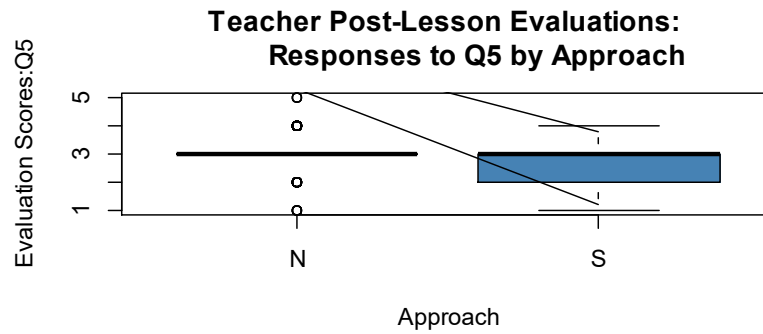


Return to [7.4.1](#).

Appendix T.5

Normality of Distribution Post-lesson Evaluation: Item 5 “Students were focused in this lesson”

Visual inspections of the data and the Shapiro-Wilk test suggested that data were not normally distributed (Novel: $W = .78, p < .001$; Standard: $W = .77, p < .001$).

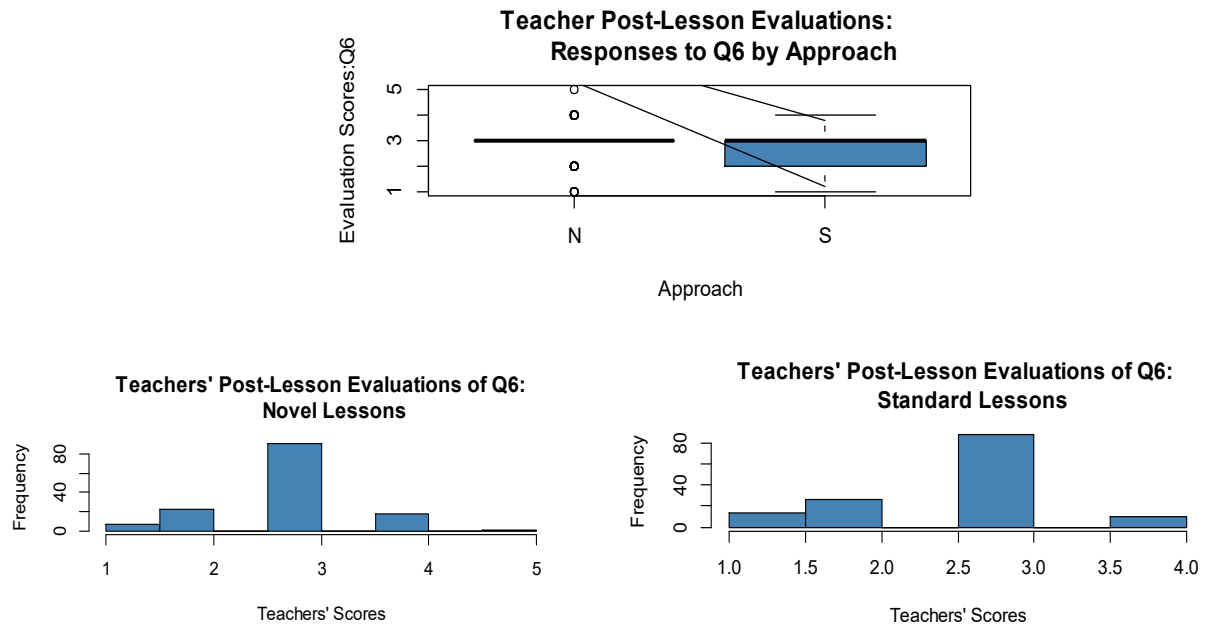


Return to [7.4.1](#).

Appendix T.6

Normality of Distribution Post-lesson Evaluation: Item 6 “I was satisfied with what the students knew by the end of the lesson”

Data appeared not to be normally distributed according to visuals and the Shapiro-Wilk test (Novel: $W = .79, p < .001$; Standard: $W = .73, p < .001$).

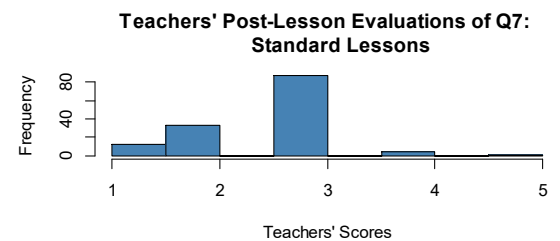
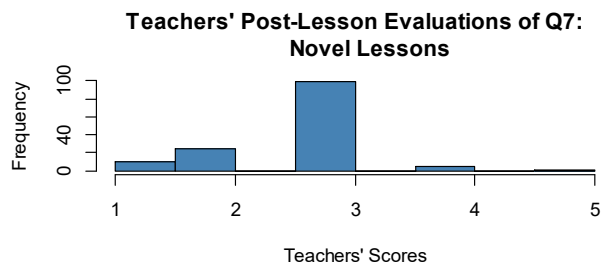
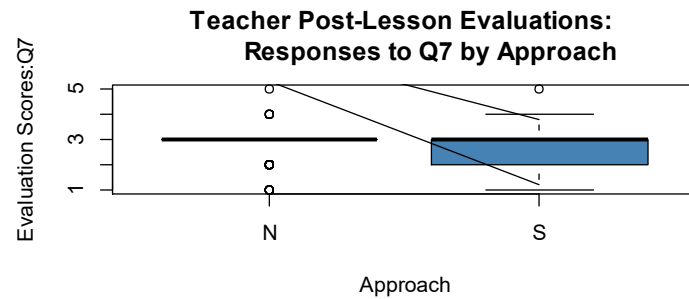


Return to [7.4.1](#).

Appendix T.7

Normality of Distribution Post-lesson Evaluation: Item 7 “I was satisfied that the French I taught was useful”

Visual inspections of the data and the Shapiro-Wilk test suggested that data were not normally distributed (Novel: $W = .70, p < .001$; Standard: $W = .76, p < .001$).

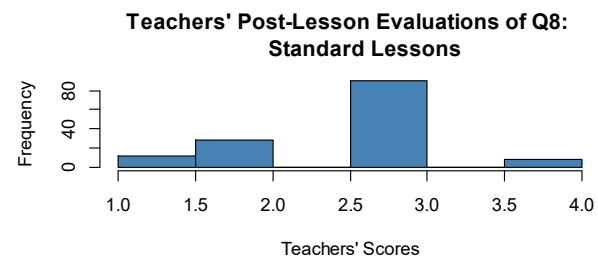
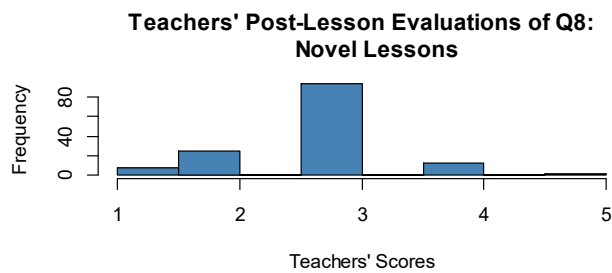
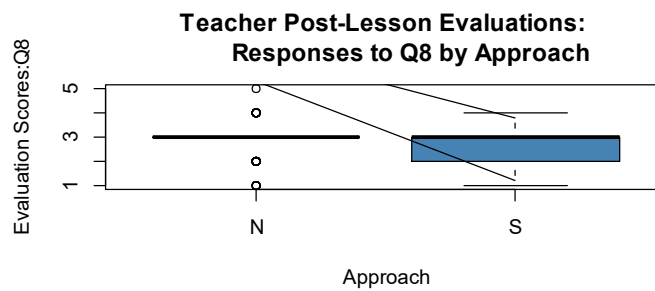


Return to [7.4.1](#).

Appendix T.8

Normality of Distribution Post-lesson Evaluation: Item 8 “The activities gave me sufficient evidence about students’ actual learning”

The visual inspections of data and the Shapiro-Wilk test indicated that data from both approaches were not normally distributed (Novel: $W = .76, p < .001$; Standard: $W = .75, p < .001$).

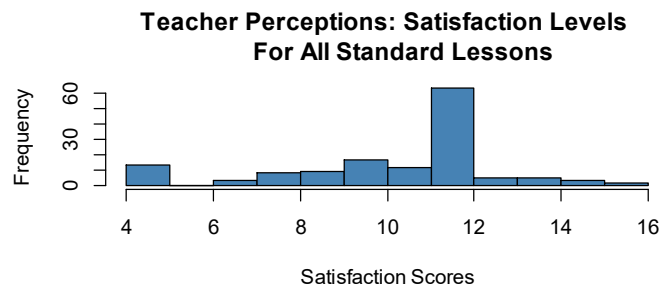
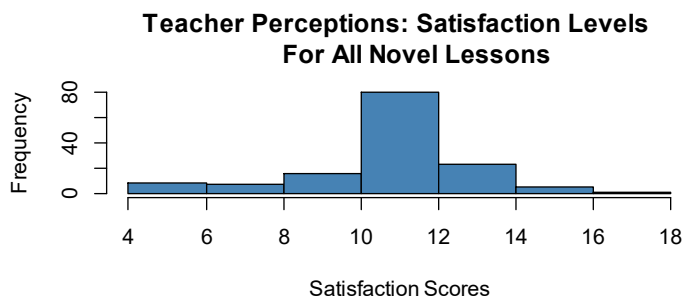
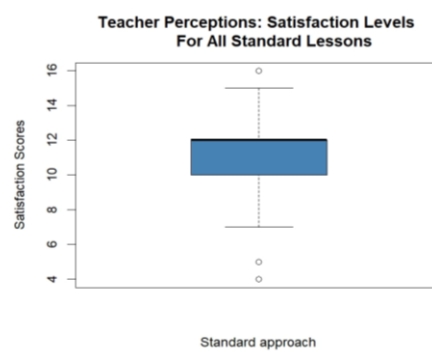


Return to [7.4.1](#).

Appendix T.9

Post-lesson Evaluation: Normality of Distribution for Teachers' Perceptions of Pupil Knowledge and Understanding, and of Usefulness of French Taught

Although means were similar, there were several outliers in the Novel approach at both ends of the scale. The visual inspections of the histogram for Standard and the Shapiro-Wilk test indicated that data from both approaches were not normally distributed (Novel: $W = .84, p < .001$; Standard: $W = .84, p < .001$).

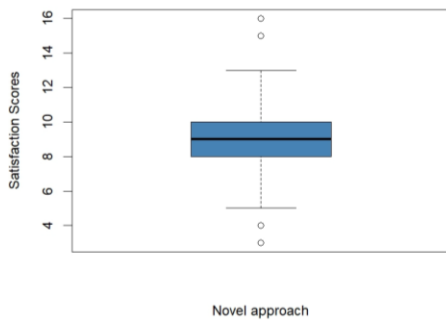


Appendix T.10

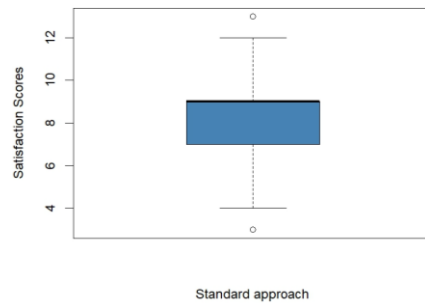
Post-lesson Evaluation: Normality of Distribution for Classroom Enjoyment

Although Novel data appeared normally distributed from visual inspection, this was not supported by the Shapiro-Wilk test (Novel: $W = .90, p < .001$). The Shapiro-Wilk test also indicated that data from Standard were not normally distributed (Standard: $W = .89, p < .001$).

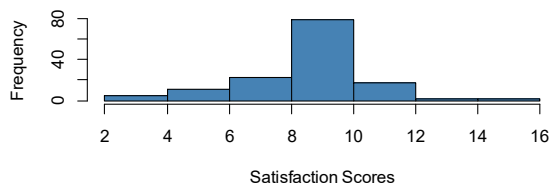
Motivation: Satisfaction Levels For All Novel Lessons



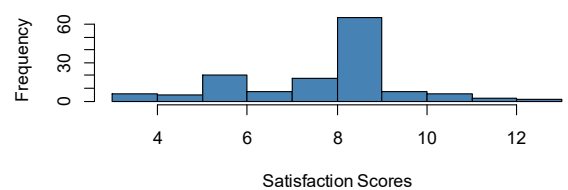
Motivation: Satisfaction Levels For All Standard Lesson



Motivation: Satisfaction Levels For All Novel Lessons



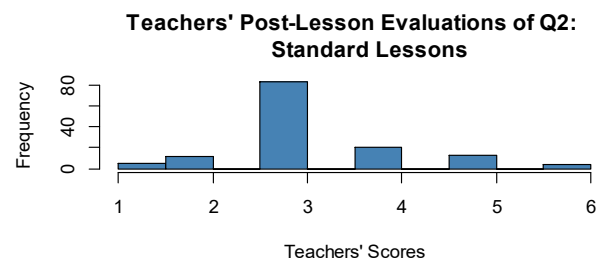
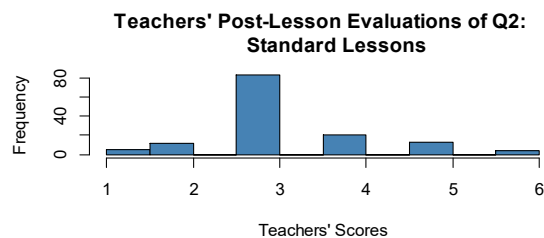
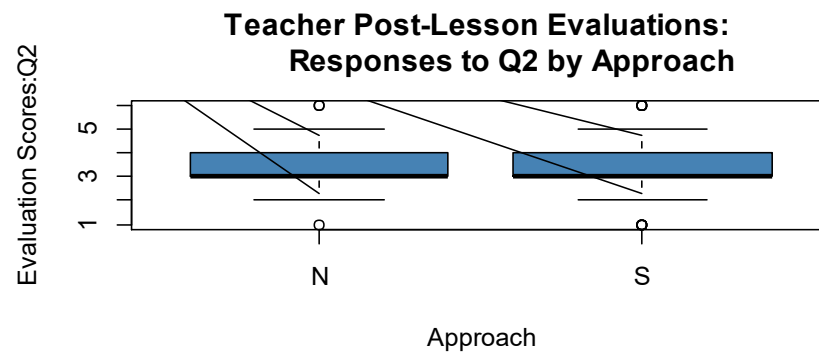
Motivation: Satisfaction Levels For All Standard Lessons



Appendix T.11

Normality of Distribution Post-lesson Evaluation: I2 “I was tense during this lesson”

Normality of distribution was checked using visual inspections, as well as through the Shapiro-Wilk test. The histograms and the p -values in both approaches indicated that data were not normally distributed (Novel: $W = .79, p < .001$; Standard: $W = .83, p = < .001$).



Appendix U.1

Individual Teacher Data for Teachers' Perceptions of Pupil Knowledge and Understanding, and Usefulness of the Language Taught

[Figure 86](#) shows individual teachers' post-lesson evaluation ratings for this factor, with descriptive and inferential statistics in [Table 62](#). Normality checks are provided in [Appendices V.1-V.8](#).

Higher perceptions of learner outcomes and linguistic usefulness were reported in Standard than Novel by T2, T3 (School A), T4, T7, T8 and T9, with this reaching statistical significance with a large effect size for T7 and T8 (although the d CIs dipped just below zero for T8). T1's and T3's (School C) ratings were slightly higher in Novel than Standard with medium and large effect sizes respectively, although these differences were not statistically significant. T3's ratings for the two approaches in School E were similar.

Figure 86.

Individual Teacher Ratings for Perceptions of Pupil Outcomes and Usefulness by Lesson (Post-lesson Evaluations)

All Teachers: Perceptions of Pupil Outcomes and Usefulness by Lesson

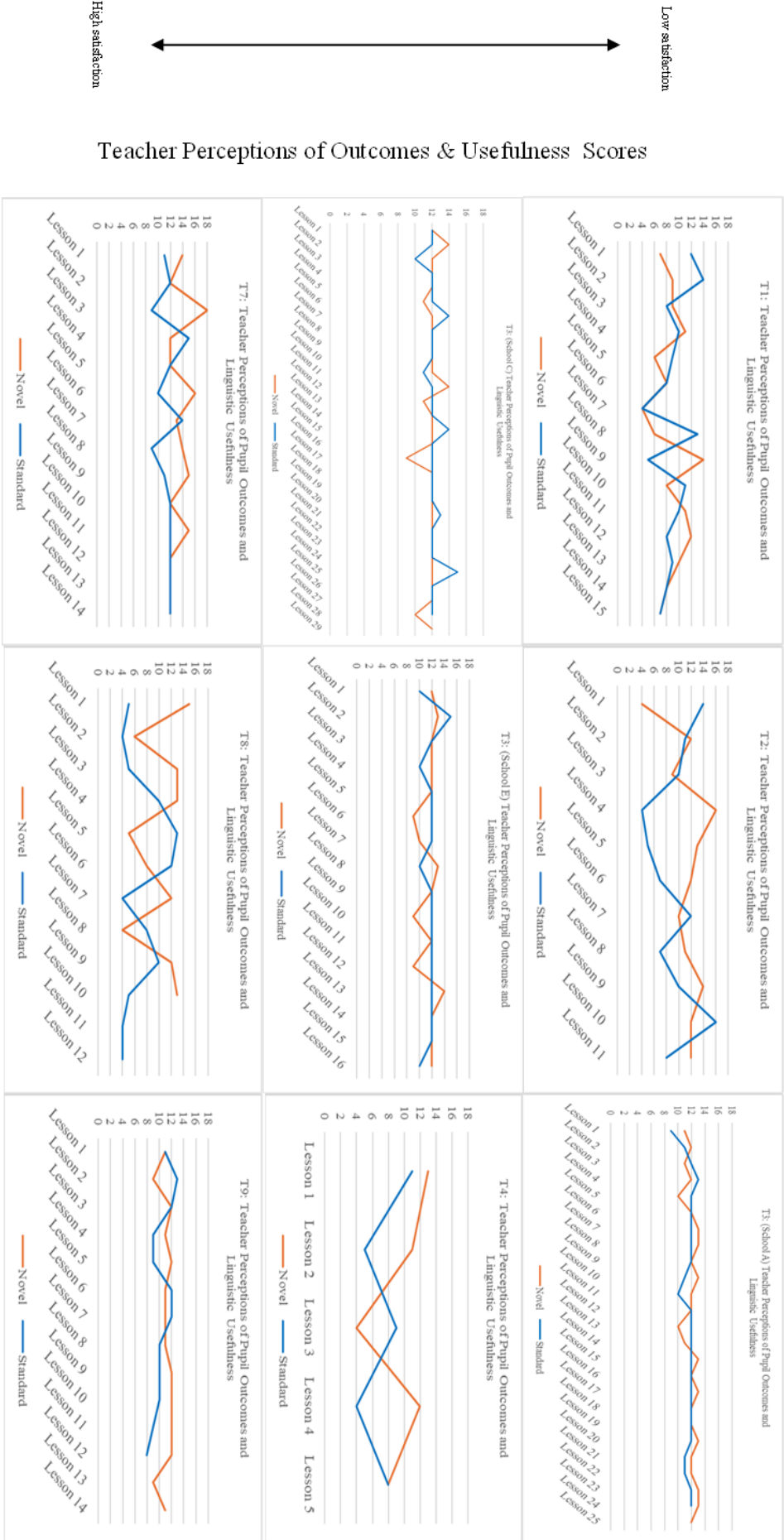


Table 62.

Individual Teachers' Perceptions of Pupil Outcomes and Linguistic Usefulness Ratings in Post-lesson Evaluations

Teacher	Approach	(<i>k</i> = 4)	Descriptive		statistics				Inferential	statistics
		# data points	<i>Mean</i> /24	<i>SD</i>	# data points	Cohen’s <i>d</i>	CI	Wilcoxon W	<i>p</i> -value	
T1	Novel	60	8.7	2.6	Novel - Standard	120	-.15	-.90-.60	100.50	.63
	Standard	60	9.1	2.7						
T2	Novel	44	11.4	3.1	Novel - Standard	88	.56	-.35-1.47	82.00	.16
	Standard	44	9.5	3.7						
T3	Novel	100	12.0	.9	Novel - Standard	196	.44	-.14-1.02	377.00	.08
(Sch A)	Standard	96	11.7	.8						
T3	Novel	116	11.9	.9	Novel - Standard	228	-.31	-.85-.22	354.50	.25
(Sch C)	Standard	112	12.2	.9						
T3	Novel	64	11.6	1.5	Novel - Standard	128	-.09	-.81-.63	132.50	.86
(Sch E)	Standard	64	11.7	1.3						
T4	Novel	20	9.6	3.7	Novel – Standard	40	.67	-.83-2.17	17.50	.34
	Standard	20	7.4	2.9						
T7	Novel	64	13.5	1.9	Novel - Standard	128	1.04	.21-1.87	151.50	.009
	Standard	64	11.6	1.7						
T8	Novel	40	10.10	4.0	Novel - Standard	88	.85	-.09-1.78	89.50	.05
	Standard	48	7.00	3.4						
T9	Novel	64	11.14	1.0	Novel - Standard	112	.56	-.27-1.39	107.0	.23
	Standard	48	10.42	1.6						

Note. Bold text indicates the approach with the lower mean rating (and therefore more positive perceptions).

Appendix U.2

Individual Teacher Data for Enjoyment

For the construct of enjoyment, [Figure 87](#) shows individual teacher post-lesson evaluation ratings and [Table 63](#) presents descriptive and inferential statistics. For normality of distribution checks, see [Appendices V.9-V.17](#).

The ratings of T1 and T3 (in School C) were higher for the construct of classroom enjoyment in Novel than Standard, both with large effect sizes and the latter reaching statistical significance; however, these CIs crossed zero ([Table 63](#)). The classroom enjoyment ratings of the remaining teachers (T2, T3 Schools A/E, T4, T7, T8, T9) were higher in Standard than Novel, with statistically significant differences for T3 (School A), T7 and T8 and large effect sizes. They were not significant for T2, T3 (School E), T4 or T9, all of which had large effect sizes, although the CIs crossed zero.

Figure 87.

Individual Teacher Ratings for Enjoyment Construct by Lesson (Post-lesson Evaluations)

Individual Teachers : Enjoyment by Lesson



Lessons from Start to End of Project

Table 63.

Individual Teachers' Perceptions of Classroom Enjoyment Ratings in Post-lesson Evaluations

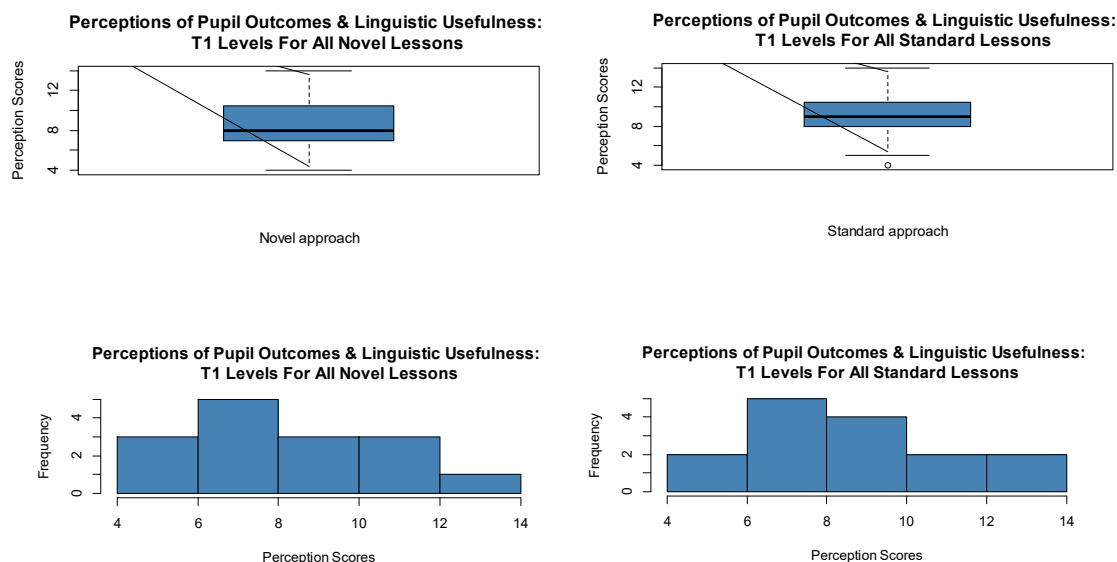
Teacher	Approach	(k = 3) # data points	Descriptive statistics			# data points	Cohen's d	CI	Inferential statistics	
			Mean /18	SD					Wilcoxon W	p-value
T1	Novel	45	6.0	1.9	Novel - Standard	90	-.37	- 1.13- .38	94.00	.45
	Standard	45	6.7	2.1						
T2	Novel	33	9.2	2.9	Novel - Standard	66	.62	-.29- 1.53	84.50	.12
	Standard	33	7.5	2.6						
T3 (Sch A)	Novel	75	9.2	1.1	Novel - Standard	147	.84	.24- 1.44	431.00	.003
	Standard	72	8.3	1.2						
T3 (Sch C)	Novel	87	9.0	.7	Novel - Standard	171	-.23	-.76- .30	.49	<.001
	Standard	84	9.1	.9						
T3 (Sch E)	Novel	48	9.4	1.5	Novel - Standard	96	.43	-.30- 1.16	161.00	.21
	Standard	48	8.7	1.7						
T4	Novel	15	6.6	2.2	Novel - Standard	30	.33	- 1.14- 1.79	16.00	.50
	Standard	15	6.0	1.4						
T7	Novel	42	11.0	2.4	Novel - Standard	84	1.09	.26- 1.93	151.00	.009
	Standard	42	9.0	.7						
T8	Novel	30	9.3	2.3	Novel - Standard	66	1.07	.11- 2.02	93.00	.03
	Standard	36	6.8	2.5						
T9	Novel	42	8.9	.4	Novel - Standard	78	.37	-.45- 1.19	96.00	.45
	Standard	36	8.5	1.4						

Note. Bold text indicates the approach with the lower mean rating (and therefore higher perceptions).

Appendix V.1

Normality of Distribution Checks for Teacher 1's Ratings for Perceived Pupil Outcomes and Linguistic Usefulness

Data appeared to be normally distributed in both approaches, as evidenced by visual inspection of the histograms and the results of the Shapiro-Wilk tests: Novel: $W = .98, p = .97$; Standard: $W = .97, p = .88$. A parametric t -test was therefore used.

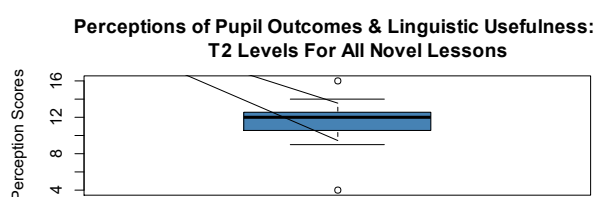


Approach	T1 Perceived Pupil Outcomes and Linguistic Usefulness	
	Construct	Descriptive Statistics
	Skew	Kurtosis
Novel	.22	-.32
Standard	.01	-.34

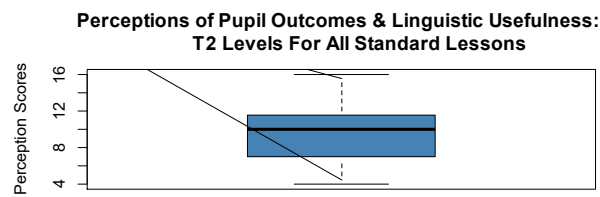
Appendix V.2

Normality of Distribution Checks for Teacher 2's Ratings for Perceived Pupil Outcomes and Linguistic Usefulness

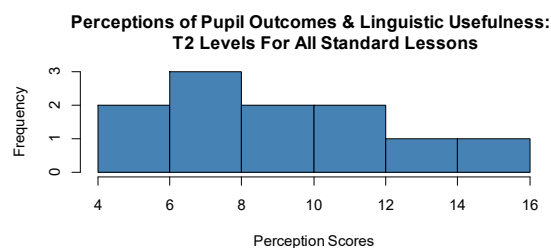
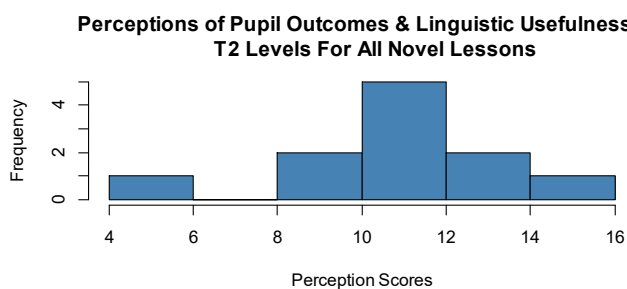
A parametric *t*-test was used as, despite the slightly platykurtic distribution of the data in Standard approach, the *p*-values from the Shapiro-Wilk tests (Novel: $W = .90, p = .16$; Standard: $W = .98, p = .93$) indicated that data did not deviate from normal distribution.



Novel approach



Standard approach

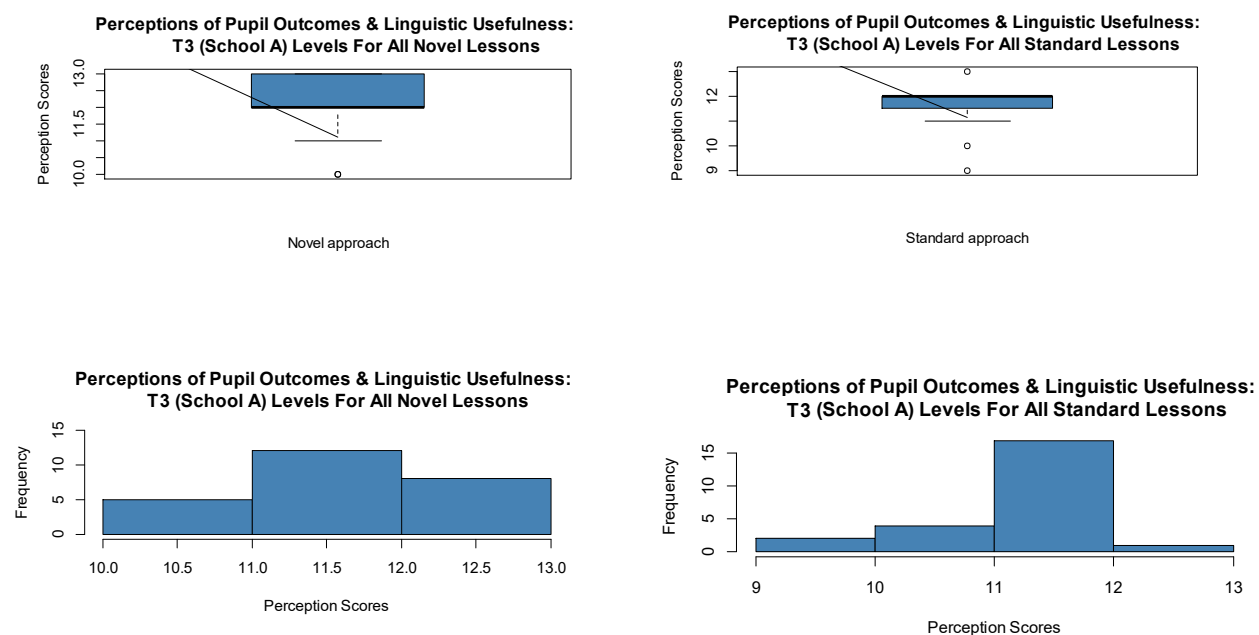


Approach	T2 Perceived Pupil Outcomes and Linguistic Usefulness Construct Descriptive Statistics	
	Skew	Kurtosis
Novel	-.69	.22
Standard	.15	-.48

Appendix V.3

Normality of Distribution Checks for Teacher 3's (SchA) Ratings for Perceived Pupil Outcomes and Linguistic Usefulness

The p -values from the Shapiro-Wilk tests (Novel: $W = .82, p < .001$; Standard: $W = .68, p < .001$) indicated that data were not normally distributed and therefore a non-parametric Wilcoxon rank sum test was conducted.

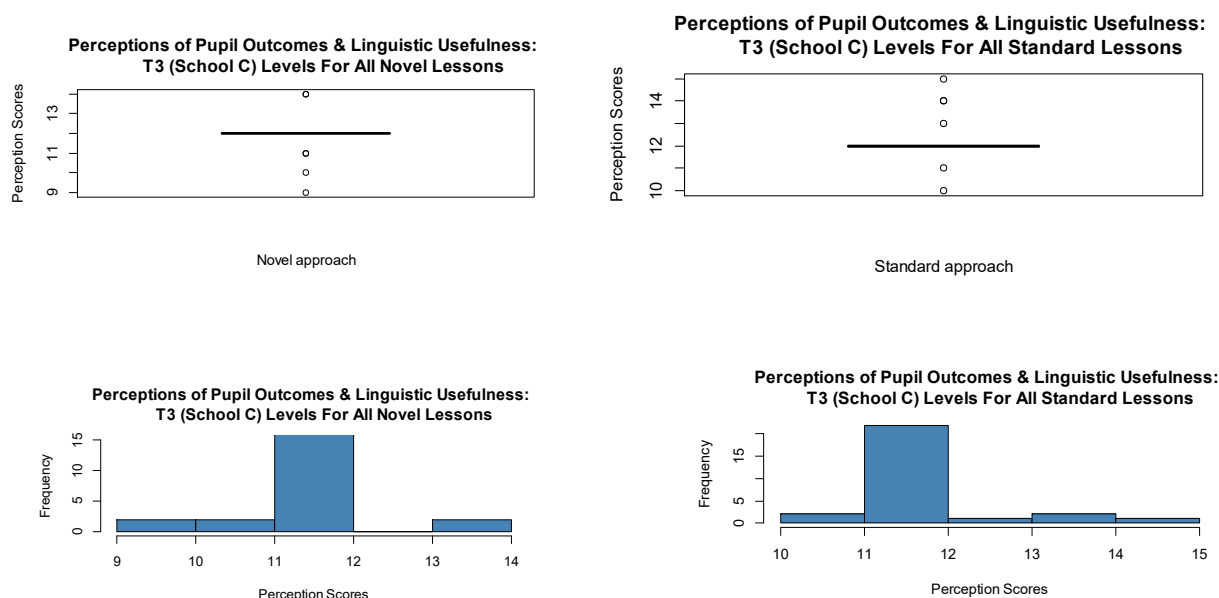


Approach	T3 (SchA) Perceived Pupil Outcomes and Linguistic Usefulness Construct Descriptive Statistics	
	Skew	Kurtosis
Novel	-.82	-.07
Standard	-1.76	1.55

Appendix V.4

Normality of Distribution Checks for Teacher 3's (SchC) Ratings for Perceived Pupil Outcomes and Linguistic Usefulness

As the p -values from the Shapiro-Wilk tests were $< .05$ (Novel: $W = .63, p < .001$; Standard: $W = .64, p < .001$) and the histograms showed leptokurtic distribution (also seen in the outliers in the boxplots), the data appeared to deviate from normal distribution and therefore a non-parametric Wilcoxon rank sum test was conducted.

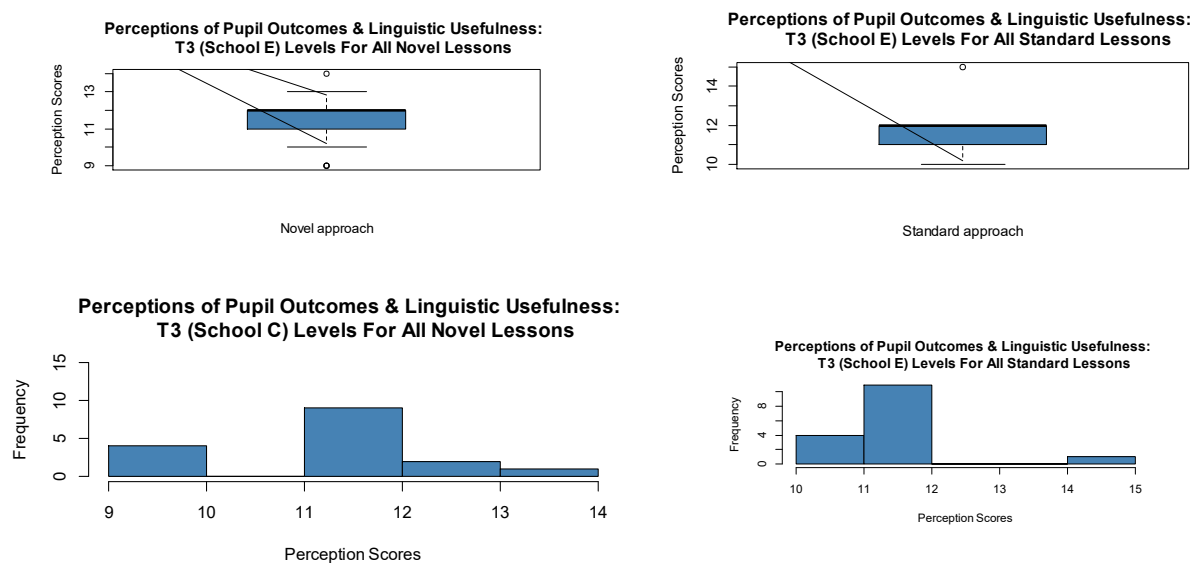


Approach	T3 (SchC) Perceived Pupil Outcomes and Linguistic Usefulness Construct
	Descriptive Statistics
	Skew Kurtosis
Novel	-.76 2.07
Standard	1.26 1.68

Appendix V.5

Normality of Distribution Checks for Teacher 3's (SchE) Ratings for Perceived Pupil Outcomes and Linguistic Usefulness

It was evident from the Shapiro-Wilk tests and visual inspections that data were not normally distributed: Novel: $W = .81, p < .003$; Standard: $W = .71, p < .001$. Therefore, a non-parametric Wilcoxon rank sum test was conducted.



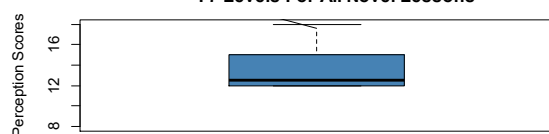
Approach	T3 (SchE) Perceived Pupil Usefulness Construct	Outcomes and Linguistic Descriptive Statistics
	Skew	Kurtosis
Novel	-.54	-.37
Standard	.50	.42

Appendix V.6

Normality of Distribution Checks for Teacher 7's Ratings for Perceived Pupil Outcomes and Linguistic Usefulness

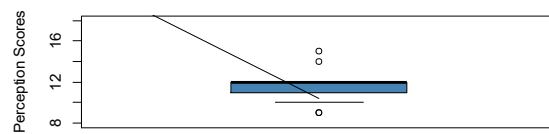
The Shapiro-Wilk tests (Novel: $W = .80, p < .005$; Standard: $W = .89, p = .009$) indicated that data were not normally distributed, and this was supported by visual inspection of boxplots and histograms, which showed positive skew in Novel. A non-parametric t -test was therefore conducted.

**Perceptions of Pupil Outcomes & Linguistic Usefulness:
T7 Levels For All Novel Lessons**



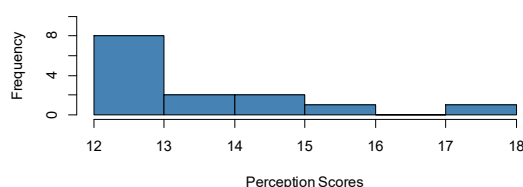
Novel approach

**Perceptions of Pupil Outcomes & Linguistic Usefulness:
T7 Levels For All Standard Lessons**

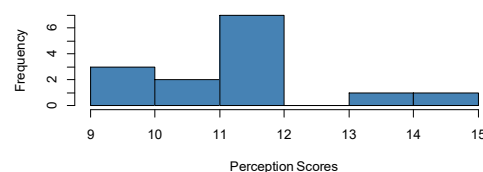


Standard approach

**Perceptions of Pupil Outcomes & Linguistic Usefulness:
T7 Levels For All Novel Lessons**



**Perceptions of Pupil Outcomes & Linguistic Usefulness:
T7 Levels For All Standard Lessons**

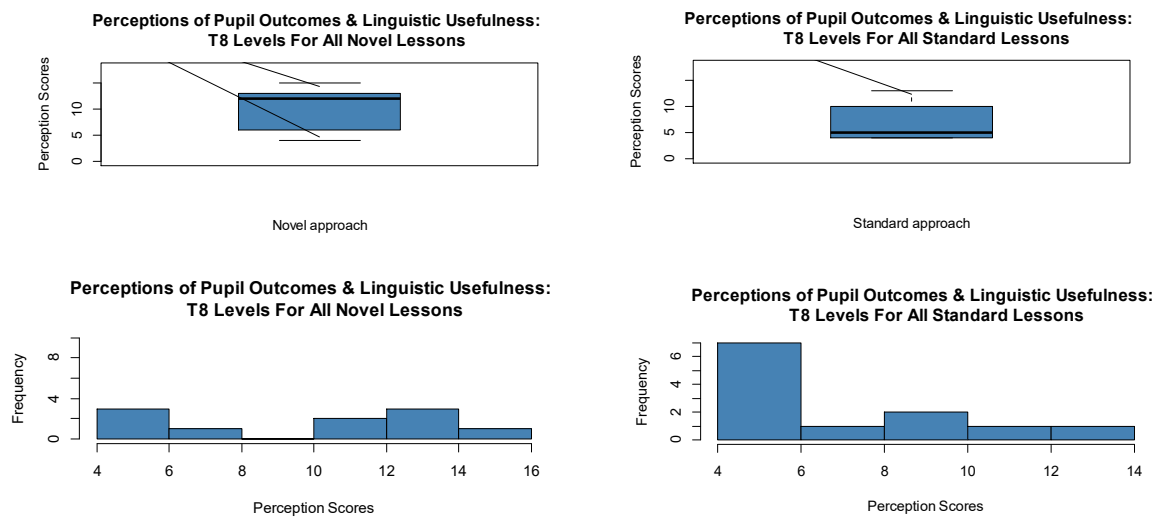


Approach	T7 Perceived Pupil Outcomes and Linguistic Usefulness Construct Descriptive Statistics	
	Skew	Kurtosis
Novel	.77	-.15
Standard	.13	-.19

Appendix V.7

Normality of Distribution Checks for Teacher 8's Ratings for Perceived Pupil Outcomes and Linguistic Usefulness

The Shapiro-Wilk tests (Novel: $W = .87, p = .09$; Standard: $W = .82, p = .01$) indicated that data were not normally distributed in Standard, and this was supported by visual inspection of boxplots and histograms. A non-parametric t -test was therefore conducted.

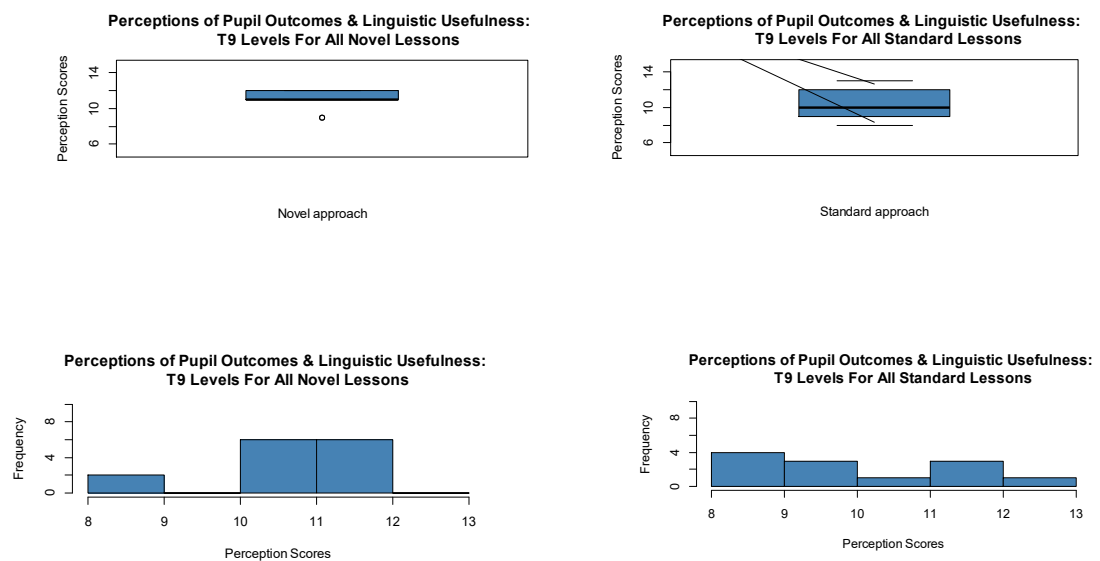


Approach	T8 Perceived Pupil Outcomes and Linguistic Usefulness	
	Construct Descriptive	Statistics
	Skew	Kurtosis
Novel	-.27	-.64
Standard	.43	-.60

Appendix V.8

Normality of Distribution Checks for Teacher 9's Ratings for Perceived Pupil Outcomes and Linguistic Usefulness

The Shapiro-Wilk tests (Novel: $W = .74, p = .001$; Standard: $W = .93, p = .40$) indicated that data were not normally distributed in Novel approach, and this was supported by visual inspection of boxplots and histograms. A non-parametric t -test was therefore conducted.

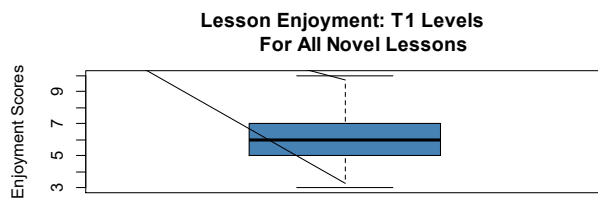


Approach	T9 Perceived Pupil Outcomes and Linguistic Usefulness Construct Descriptive Statistics	
	Skew	Kurtosis
Novel	-.88	-.04
Standard	.11	-.60

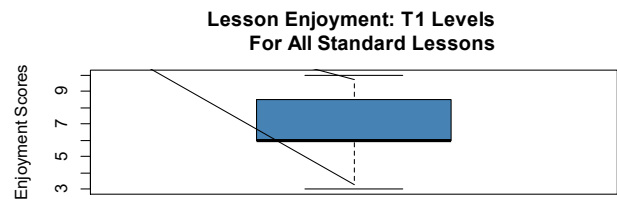
Appendix V.9

Normality of Distribution Checks for Teacher 1's Ratings for Enjoyment Construct

The p -values in the Shapiro-Wilk test in both approaches indicated that data were normally distributed (Novel: $W = .95$, $p = .47$; Standard: $W = .93$, $p = .23$). Although the histograms were not symmetrical around the centre, skew and kurtosis values were acceptable and a parametric t -test was used.



Novel approach



Standard approach

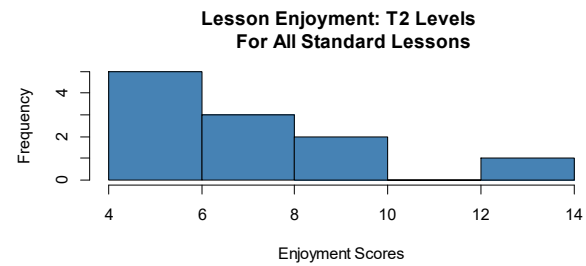
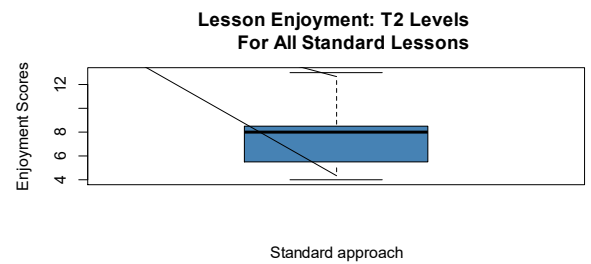
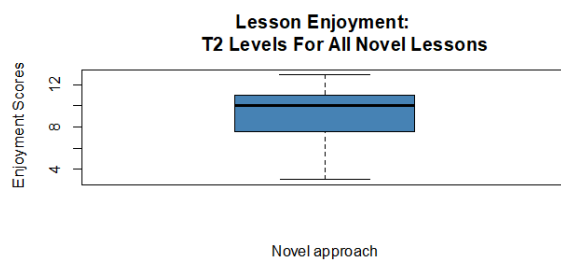


Approach	T1 Enjoyment Construct	Descriptive Statistics
	Skew	Kurtosis
Novel	.11	-.18
Standard	.10	-.48

Appendix V.10

Normality of Distribution Checks for Teacher 2's Ratings for Enjoyment Construct

Although p -values in the Shapiro-Wilk test in both approaches indicated that data were normally distributed (Novel: $W = .93, p = .41$; Standard: $W = .94, p = .49$), the histogram in the Standard approach appeared slightly positively skewed, with most data points clustered on the left side. For this reason, a non-parametric test was used.

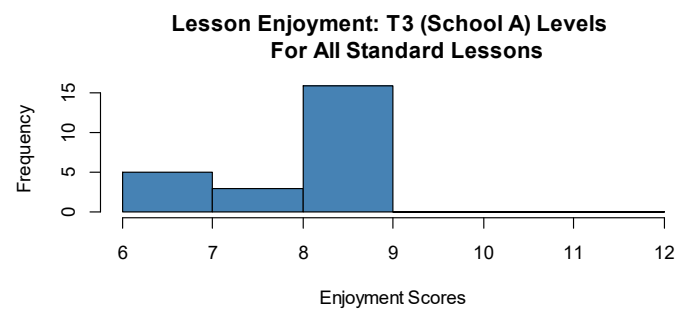
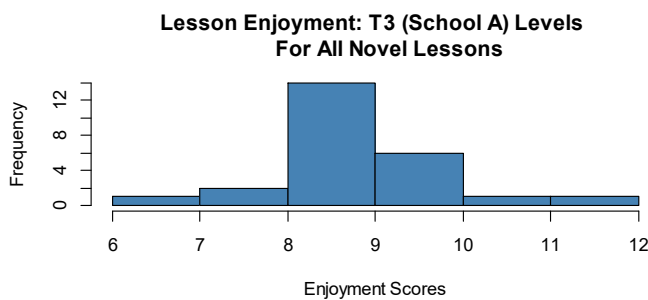
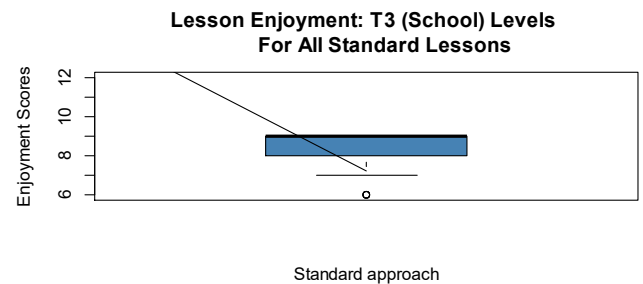
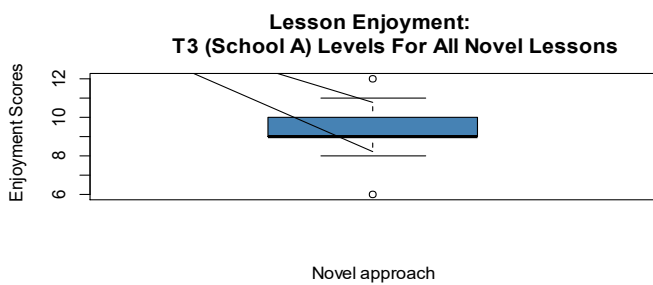


Approach	T2 Enjoyment Construct	Descriptive Statistics
	Skew	Kurtosis
Novel	-.52	-.26
Standard	-.44	-.26

Appendix V.11

Normality of Distribution Checks for Teacher 3's Ratings for Enjoyment Construct (School A)

In the Shapiro-Wilk test, p -values in both approaches indicated that data deviated from normal distribution (Novel: $W = .33, p < .001$; Standard: $W = .63, p < .001$), and this was supported particularly by the histogram in the Standard approach which appeared negatively skewed. Therefore, a non-parametric test was used.

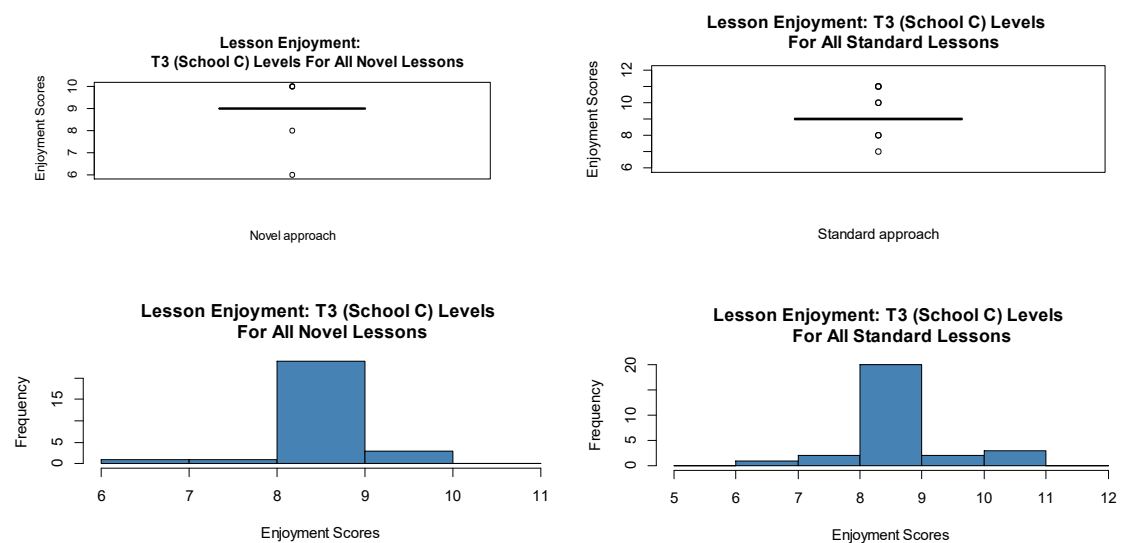


Approach	T3 (SchA) Enjoyment Construct Descriptive Statistics	
	Skew	Kurtosis
Novel	-.30	1.23
Standard	-1.26	-1.67

Appendix V.12

Normality of Distribution Checks for Teacher 3's Ratings for Enjoyment Construct (School C)

In the Shapiro-Wilk test, p -values in both approaches suggested non-normal distribution (Novel: $W = .50, p < .001$; Standard: $W = .73, p < .001$), and this was supported particularly by the box plots and the leptokurtic distribution evident in the histogram for Standard approach. Therefore, a non-parametric test was used.

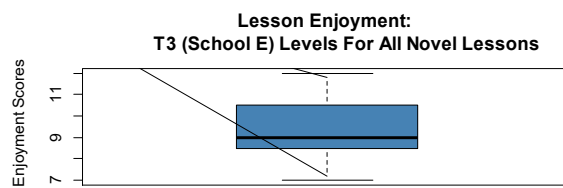


Approach	T3 (SchC) Enjoyment Construct Descriptive Statistics	
	Skew	Kurtosis
Novel	-2.99	5.99
Standard	.54	.71

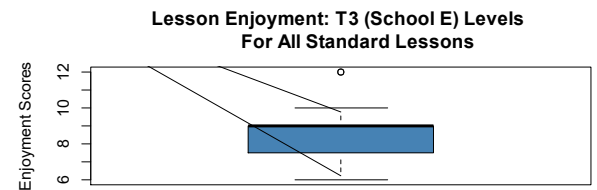
Appendix V.13

Normality of Distribution Checks for Teacher 3's Ratings for Enjoyment Construct (School E)

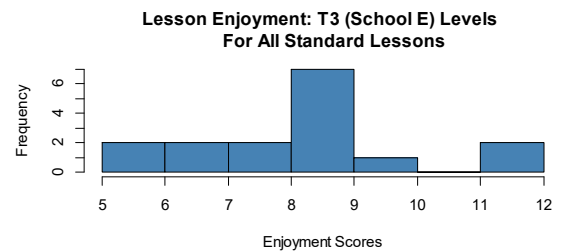
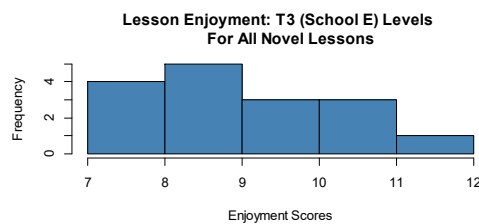
In the Shapiro-Wilk test, p -values in both approaches suggested that data were normally distributed (Novel: $W = .95$, $p = .46$; Standard: $W = .90$, $p = .08$). Although the histogram for Novel appeared to be slightly negatively skewed, skew and kurtosis scores were acceptable, and the same was true for Standard. Therefore, a parametric t -test was used.



Novel approach



Standard approach

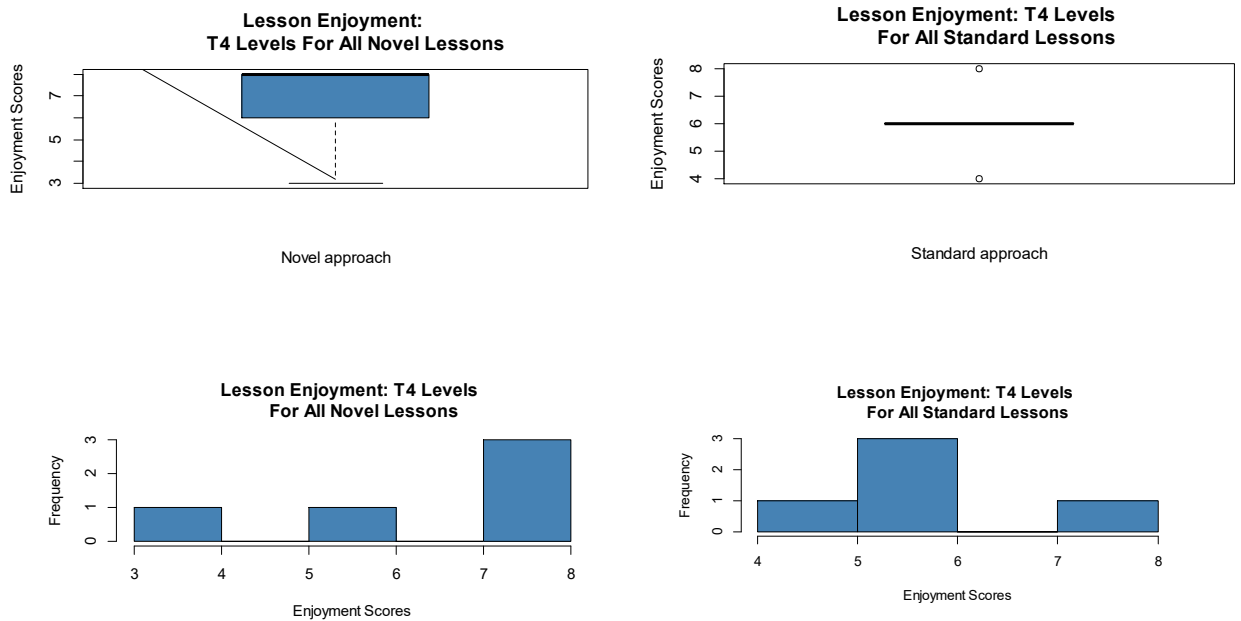


Approach	T3 (SchE) Enjoyment Construct Descriptive Statistics	
	Skew	Kurtosis
Novel	-.01	-.48
Standard	.27	-.24

Appendix V.14

Normality of Distribution Checks for Teacher 4's Ratings for Enjoyment Construct

In the Shapiro-Wilk test, although data appeared to be normally distributed in Standard approach ($W = .88, p = .33$), this was not the case in Novel ($W = .75, p = .03$). This was supported by visual inspection of the histogram where data appeared to be negatively skewed. Thus, a non-parametric test was used.

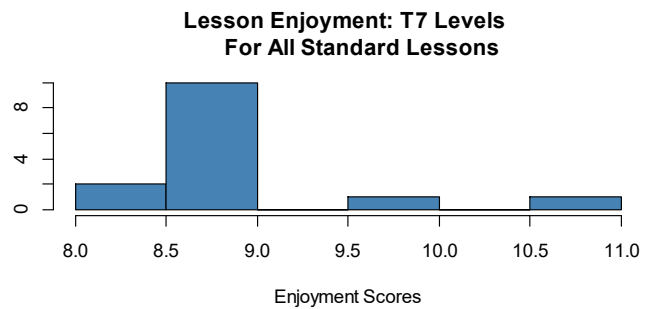
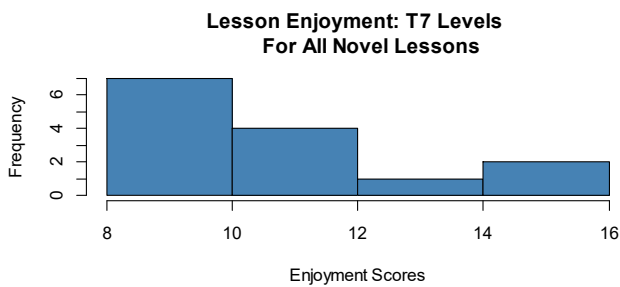
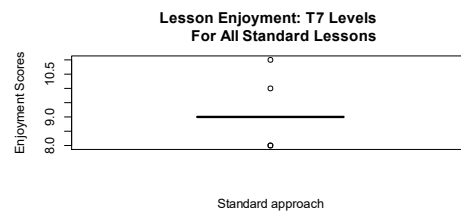
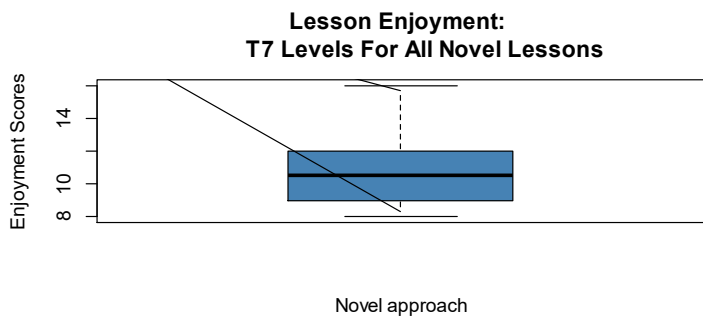


Approach	T4 Enjoyment Construct	Descriptive Statistics
	Skew	Kurtosis
Novel	-.40	-.36
Standard	.00	-.35

Appendix V.15

Normality of Distribution Checks for Teacher 7's Ratings for Enjoyment Construct

In the Shapiro-Wilk test, p -values suggested that data were normally distributed in Novel but not in Standard (Novel: $W = .91$, $p = .13$; Standard: $W = .72$, $p < .001$). This decision was supported by visual inspection of the histograms where data appeared to be positively skewed in both approaches. A non-parametric test was therefore used.

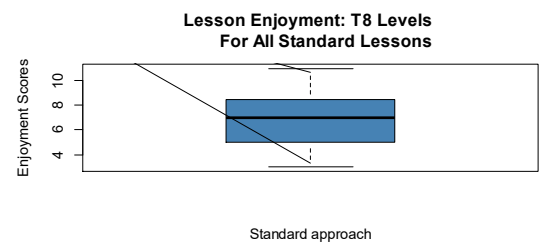
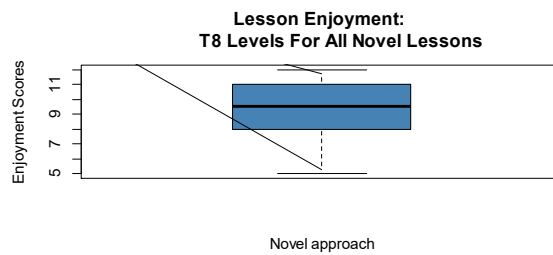


Approach	T7 Enjoyment Construct	Descriptive Statistics
	Skew	Kurtosis
Novel	.61	-.30
Standard	.85	.56

Appendix V.16

Normality of Distribution Checks for Teacher 8's Ratings for Enjoyment Construct

In the Shapiro-Wilk test, p -values in both approaches indicated normal distribution (Novel: $W = .94, p = .52$; Standard: $W = .96, p = .78$), and skew and kurtosis values were acceptable. A parametric test was therefore used.

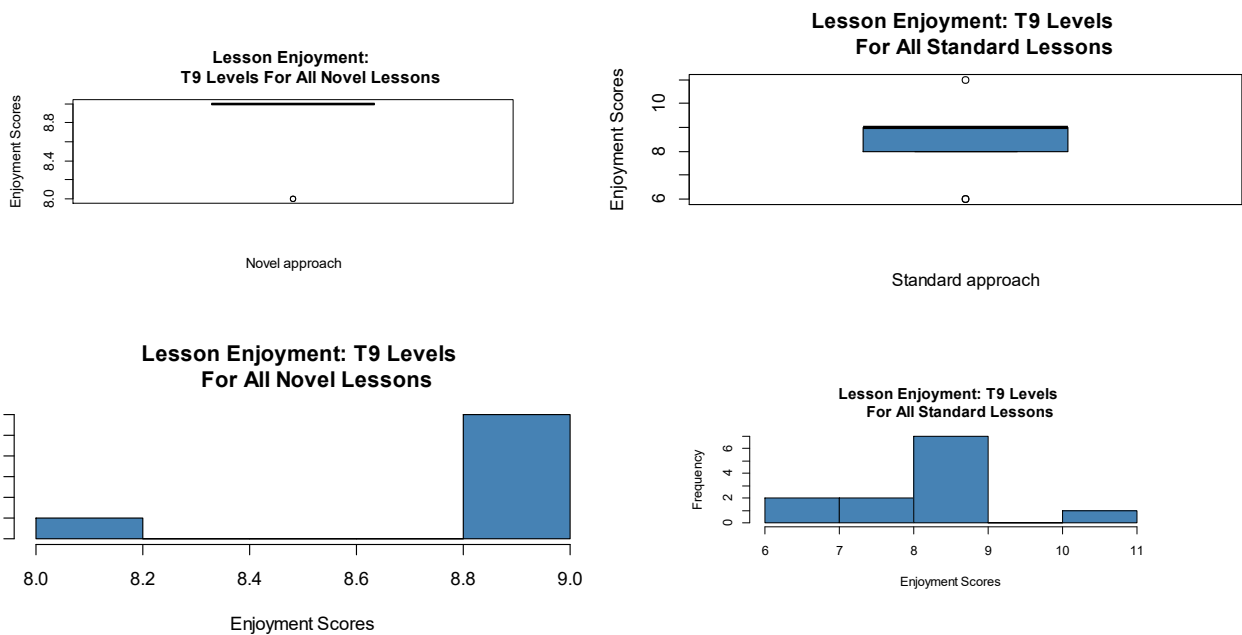


Approach	T8 Enjoyment Construct		Descriptive Statistics	
	Skew		Kurtosis	
Novel	-.27		-.47	
Standard	-.06		-.47	

Appendix V.17

Normality of Distribution Checks for Teacher 9's Ratings for Enjoyment Construct

It was evident from visual inspection of the histograms that data deviated from normal distribution. The Shapiro-Wilk p -values in both approaches confirmed this: Novel: $W = .43$, $p < .001$; Standard: $W = .81$, $p = .01$). A non-parametric test was therefore used.

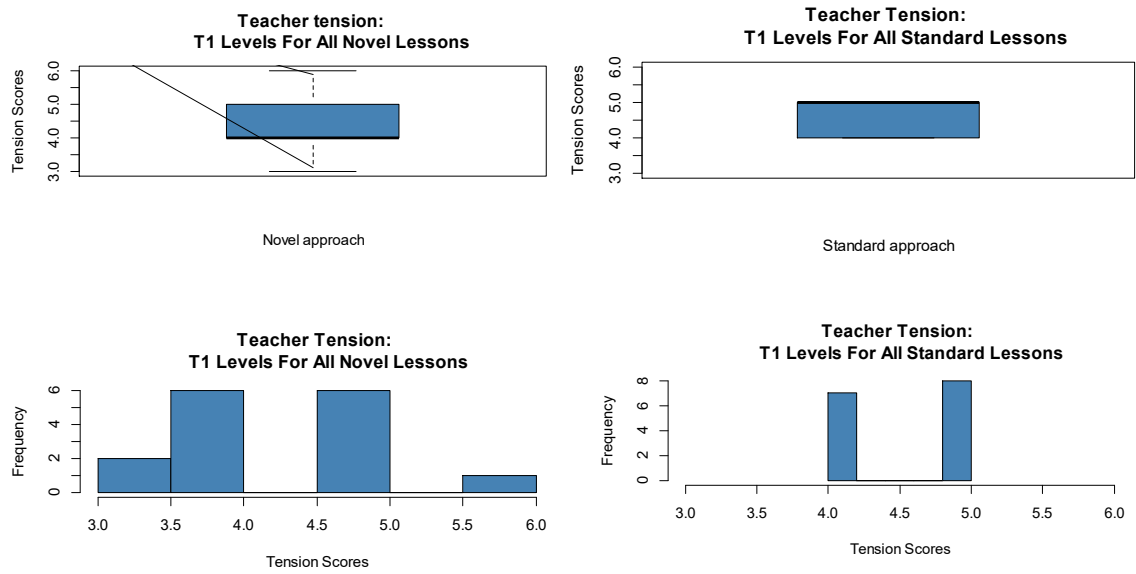


Approach	T9 Enjoyment Construct	Descriptive Statistics
	Skew	Kurtosis
Novel	-1.53	.63
Standard	-.37	-.12

Appendix V.18

Normality of Distribution Checks for Teacher 1's Ratings for Tension

Data appeared not to be normally distributed in either approach, as evidenced by visual inspection of the histograms and the results of the Shapiro-Wilk tests: Novel: $W = .88$, $p = .05$; Standard: $W = .64$, $p = < .001$. A non-parametric t -test was therefore used.

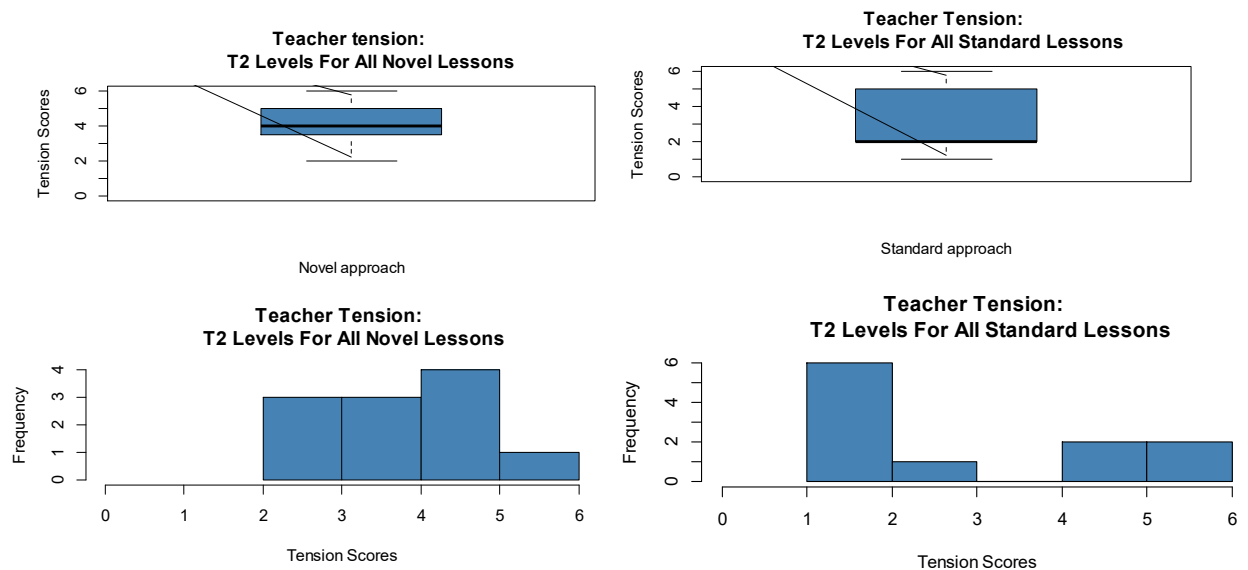


Approach	T1 Tension Construct	Descriptive Statistics
	Skew	Kurtosis
Novel	-.05	-.38
Standard	-.10	-.94

Appendix V.19

Normality of Distribution Checks for Teacher 2's Ratings for Tension

Although the p -value in the Shapiro-Wilk test indicated normal distribution in the Novel approach, in Standard it was $< .05$: Novel: $W = .93, p = .45$; Standard: $W = .84, p = .04$. This was supported by the histogram visual, indicating a positive skew to the data. A non-parametric t -test was therefore used.

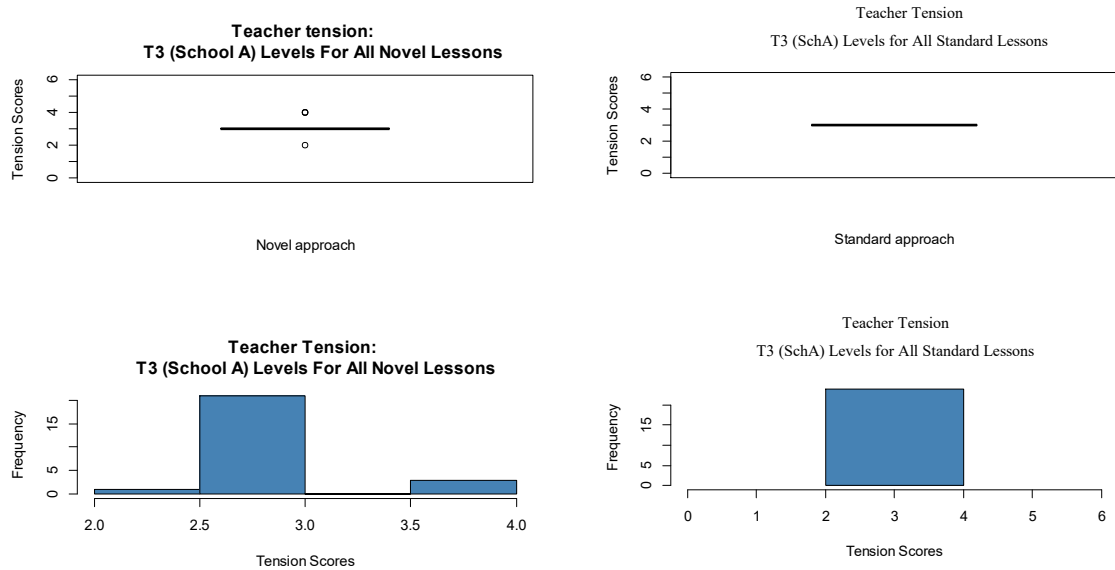


Approach	T2 Tension Construct	Descriptive Statistics
	Skew	Kurtosis
Novel	-.24	-.42
Standard	.28	-.67

Appendix V.20

Normality of Distribution Checks for Teacher 3's (SchA) Ratings for Tension

The p -value in the Shapiro-Wilk test indicated non-normal distribution in the Novel approach: Novel: $W = .55$, $p < .001$. This was supported by the histogram visual which indicated leptokurtic distribution. A non-parametric t -test was therefore used. Non-normal distribution was also found in the Shapiro-Wilk test for Standard: $W = .84$, $p = .04$.

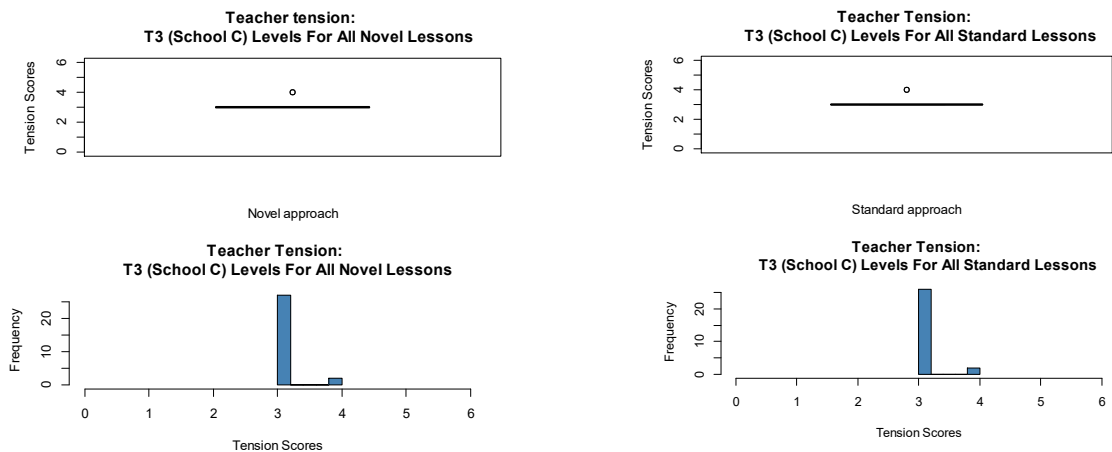


Approach	T3 (SchA) Tension Construct	Descriptive Statistics
	Skew	Kurtosis
Novel	.72	1.38
Standard	Unable to calculate	Unable to calculate

Appendix V.21

Normality of Distribution Checks for Teacher 3's (SchC) Ratings for Tension

The p -values in the Shapiro-Wilk tests indicated that data were not normally distributed in either approach (Novel: $W = .28$, $p < .001$; Standard: $W = .29$, $p < .001$) and this was confirmed by visual inspection of boxplots and histograms. A non-parametric t -test was therefore used.

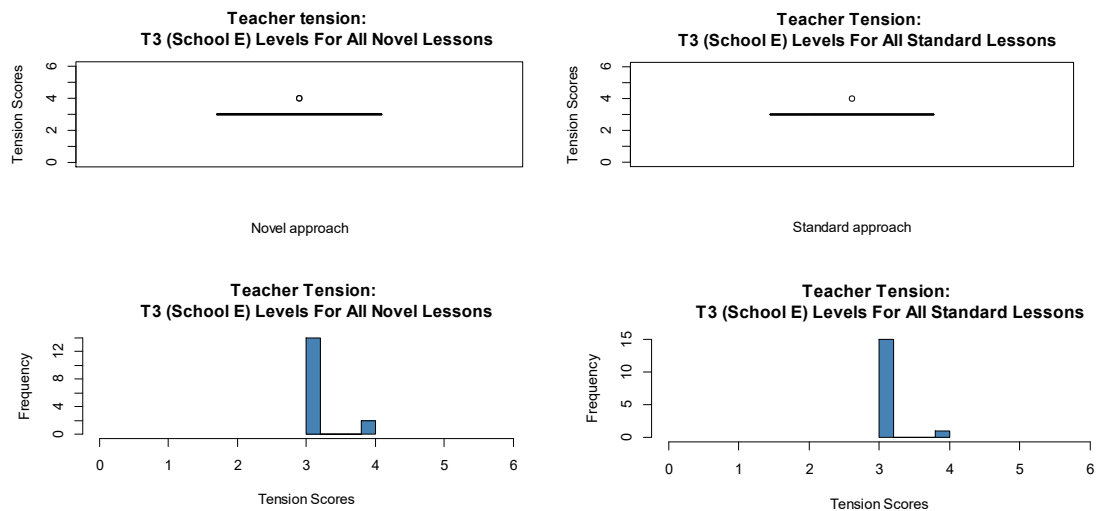


Approach	T3 (SchC) Tension Construct	Descriptive Statistics
	Skew	Kurtosis
Novel	3.72	5.16
Standard	3.58	4.79

Appendix V.22

Normality of Distribution Checks for Teacher 3's (SchE) Ratings for Tension

The p -values in the Shapiro-Wilk tests indicated that data were not normally distributed in either approach (Novel: $W = .40, p < .001$; Standard: $W = .27, p < .001$), confirmed by visual inspection of boxplots and histograms. A non-parametric t -test was therefore used.

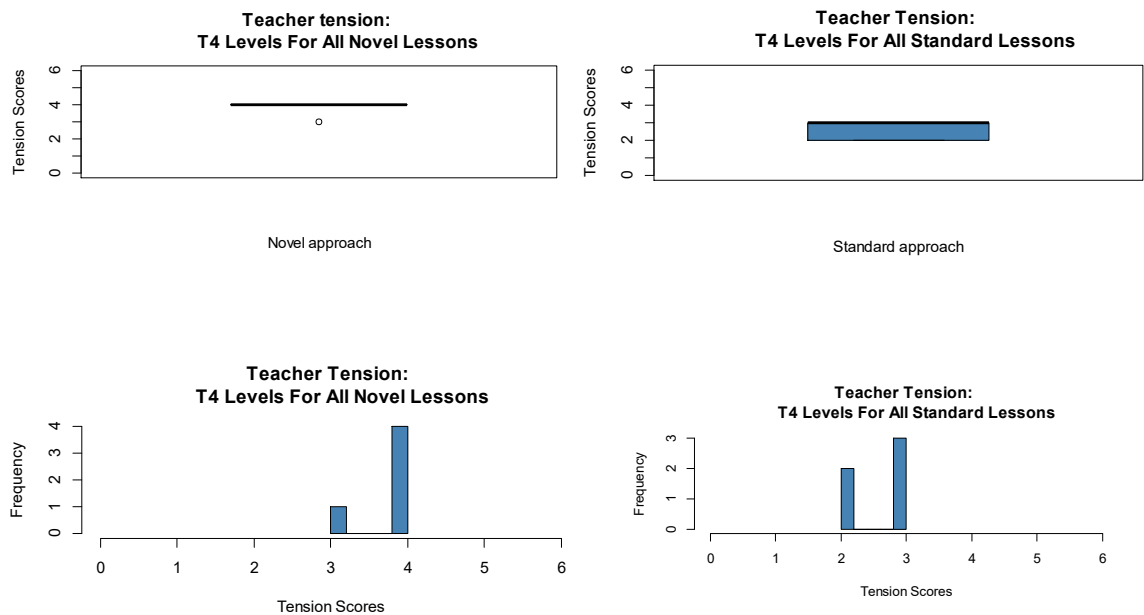


Approach	T3 (SchE) Tension Construct	Descriptive Statistics
	Skew	Kurtosis
Novel	1.82	1.10
Standard	2.91	4.29

Appendix V.23

Normality of Distribution Checks for Teacher 4's Ratings for Tension

The Shapiro-Wilk tests (Novel: $W = .55, p < .001$; Standard: $W = .68, p = .006$) indicated that data were not normally distributed in either approach, and this was supported by visual inspection of boxplots and histograms, which were slightly left-skewed in both approaches. A non-parametric t -test was therefore conducted.

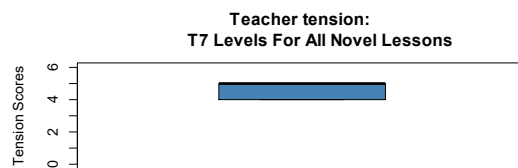


Approach	T4 Tension Construct	Descriptive Statistics
	Skew	Kurtosis
Novel	-.59	-.23
Standard	-.16	-.56

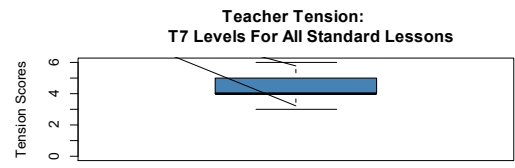
Appendix V.24

Normality of Distribution Checks for Teacher 7's Ratings for Tension

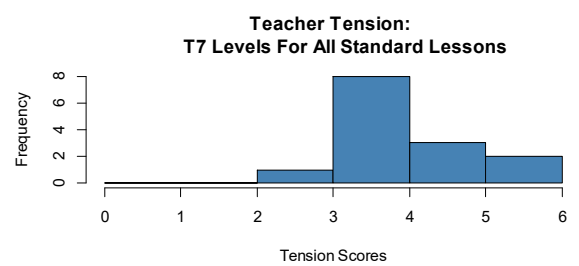
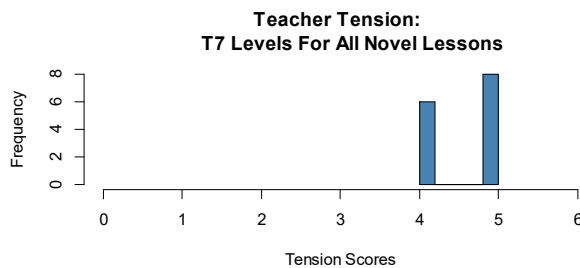
The Shapiro-Wilk tests (Novel: $W = .64, p < .001$; Standard: $W = .83, p = .01$) indicated that data were not normally distributed in either approach, and this was supported by visual inspection of boxplots and histograms, which showed a slight right-skew in Standard. As a result, a non-parametric t -test was conducted.



Novel approach



Standard approach

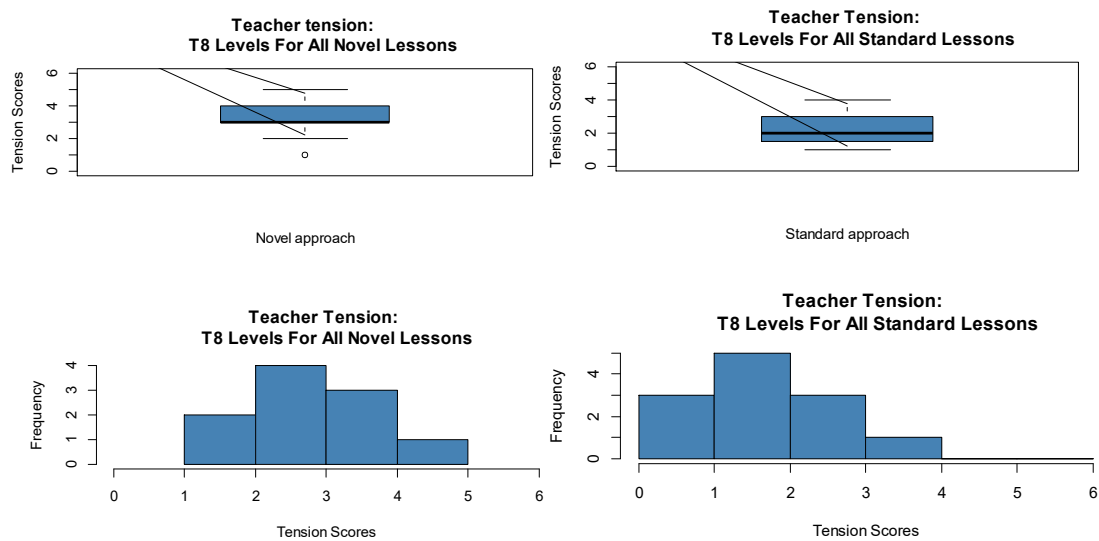


Approach	T7 Tension Construct	Descriptive Statistics
	Skew	Kurtosis
Novel	-.22	-.51
Standard	.46	-.30

Appendix V.25

Normality of Distribution Checks for Teacher 8's Ratings for Tension

As the histograms and the p -values in the Shapiro-Wilk tests indicated that data were normally distributed in either approach (Novel: $W = .93, p = .48$; Standard: $W = .89, p = .12$) a parametric t -test was used.

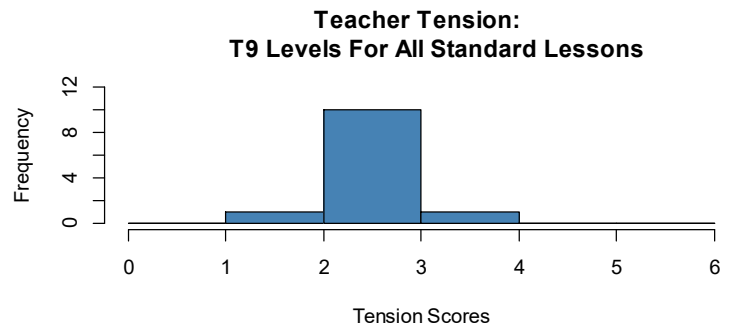
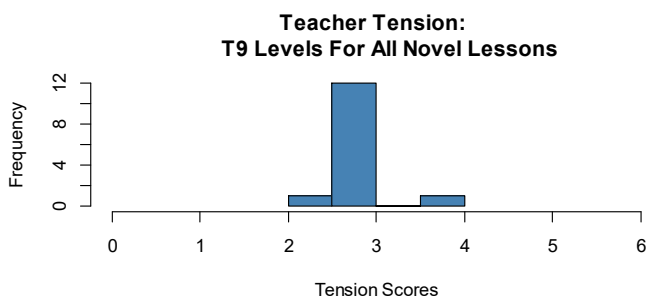
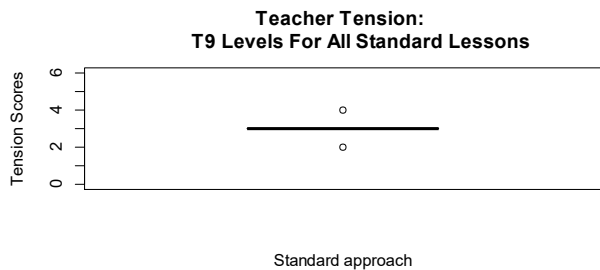
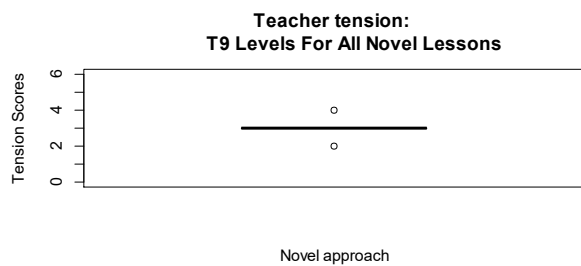


Approach	T8 Tension Construct	Descriptive Statistics
	Skew	Kurtosis
Novel	-.25	-.28
Standard	-.25	-.42

Appendix V.26

Normality of Distribution Checks for Teacher 9's Ratings for Tension

The histogram for Standard approach indicated some leptokurtic distribution. The p -values in the Shapiro-Wilk tests indicated that data deviated from normal distribution in both approaches (Novel: $W = .55$, $p < .001$; Standard: $W = .60$, $p < .001$). Therefore, a non-parametric t -test was used.



Approach	T9 Tension Construct	Descriptive Statistics
	Skew	Kurtosis
Novel	.00	1.32
Standard	.00	.83

Appendix W.1

Interview2 Teacher Comments on Pupils' Progress

Comments in blue are positive comments on pupil progress, comments in red are negative. Some comments do not refer to either approach specifically and have therefore been placed in the middle column.

Novel	Both	Standard
	T1: <i>"I can see the progress that the children have made in both"</i>. T1: <i>"they seem to have generally made a lot of progress in each"</i>.	
T2: <i>"when I've got the ... Novel lesson, I feel ... it's building upon their language skills a little bit better, whereas Standard I feel sometimes it can be very ... a bit more random ... whereas I can kind of ... see with the Novel approach ... the kind of, erm, way of thinking behind it – 'OK, well, this is structured in this way' - and it does make a bit more sense to me"</i>. T2: <i>"I can recall ... the children can recall that information from before, they can recall all of these sorts of things where ... 'OK, well, we learn about this in week 1, we're learning that information from week 1, we're learning this from last lesson, so actually we can build on that."</i> T2: <i>"I do find sometimes with the Novel approach ... especially if they've missed lessons ... it is hard for them to build upon that, whereas ... the children who've constantly been in week in, week out, they've definitely got that better knowledge in the Novel, because they can build week upon week upon week"</i>. T2: <i>"even the ... kids who struggle with French, that we taught previously, ... I think they probably found it a bit easier just because, like I said, it's that building of ... it seems a bit more of a slow build-up of knowledge rather than ... it doesn't seem as fast-paced sometimes. It seems like I can explore it a bit more"</i>. T2: <i>"the Novel approach is so different to the way that we normally teach here at school, I think some of the kids do struggle with that, because ... they're used to that routine, but some of the kids really are flying with this different one that we're doing now"</i>.	T2: <i>"it's [progress] down to individual kids, really. I think in general, I would say, I would say [they are making more progress in] the Novel approach ..."</i>	T2: <i>"Standard I feel sometimes it can be very ... a bit more random ..."</i> T2: <i>"sometimes I feel like, with the, erm, Standard approach, it is a bit kind of, like, 'Well, this, this area in France, this language for this, let's try and put them together'. It doesn't seem as kind of structured, as it were."</i>
T3: <i>"I thought I would prefer the more topic-based but ... I'm beginning to think that ... there's more grammar content in the Novel and that's helping the children. They seem to be able to do more than I thought they would, like, say whole sentences, for example"</i>. T3: <i>"the other thing that I like about the Novel approach is that ... you revisit – I know both schemes revisit, but you revisit and they give you some words – say you're practising a</i>	T3: children have made more progress than expected in both approaches because of the time allocation. <i>"Yes, I would say they have learnt more. ... I'm surprised at the level. I think it's quite a good high standard I would think, for year 3 ..."</i> <i>I would be, I would be very interested to see what the results are"</i>.	T3: <i>"I mean they do do verbs obviously in the Standard but it's more, there seems, it seems more topic-y vocabulary ..."</i> <i>they are getting muddled up between 'avoir' and 'être', I think that that ... they'll need to ... I'm sure it will be built in ... all the practice with that. Erm, but they are getting there"</i>.

particular sound, and then you get new words - and to begin with, I thought this is too much for the children. ... They've got these six words, or eight words, now they've got another four, but it's amazing how they do ... it's dripped in sort of thing. They do pick it up quite quickly". T3: "I think it's the Novel scheme where you've got, erm, Yves and Pierre and all those different people and then ... they've got to do sometimes 'être' and 'avoir' and I'm thinking "Gosh, they're going to find that difficult". Erm, but ... they did to begin with, but they have got quicker at it, better at it, so whether ... their knowledge is better or we've got better at it generally ... It could be both really". T3: "And they're getting better at that too [creating simple sentences] definitely. I think possibly the Novel – I think they have more tools at their fingertips, particularly with the verbs". T3: "I think ... they didn't like at first the grids where they had to write down ... sometimes it's the French, sometimes it's the English, and they have to get 15 points ... but they're beginning to like those now, because they realise that they can get 15 points, quite a lot of them. <i>Some don't like it because they don't feel they can remember anything. And obviously some children don't remember much ...</i> But on the whole, I think ... they quite like that". T3: "I think it seems to be quite a systematic revisiting, adding a bit more, revisit, add it, add a bit more ... Building in sentences more than the other scheme. That's my perception".		
T4: "it's really difficult with my class because I would say I think the, the children doing the N are doing better. But I also think it's a better group, if you know what I mean". T4: "it could be the group of children have been easier. I don't know. I feel like it's broken down more ... and I also feel like the children – again, it could be the group of children – but I feel like the children are getting more from it. They seem to be progressing quicker". T4: <i>"But there's a few in the N group as well ... that really struggle. So, maybe I'm being a bit unfair. It just seems to me that the N group ... they seem to be progressing and picking things up much quicker ... and they seem to be more focused. Erm, [long pause] maybe it is the French? Maybe.</i>	T4 (without reference to either approach): "also, where they have to write in French ... I found that really interesting because they've actually done really well with it. So, I've been surprised at what they've been able to do. I think that's the bit that I've enjoyed". T4: "and the pronunciation, ... they're really, they're not worried about ... speaking in French. And some of them have been, like, saying whole sentences. Which is – after six weeks ... or however long it's been ... the ten weeks, or whatever, you know, it's amazing ... For one ... lesson a week for that long. I think it is the same in both approaches actually".	T4: "... there are children in the S group that are obviously doing really well. <i>But there are also children in the S group that will struggle ... there are children in the S group that also have, erm, a great deal of focus; they just don't seem to be picking things up as quickly</i>".
T7: "some of the slides will have like questions in the top and, erm, the children will try to ... decipher what it ... might say. That's really impressed me with them actually, how much they can pick up from that. I mean sometimes they say it and I think "Oh, you know, I didn't even think of that ... word, being like another word, or, you know, how they've come to that." Erm, you know, it's quite impressive ... maybe slightly more [apparent] in the Novel".	T7: "Erm, I would say it's similar, er, the amount of progress, but in different ways. I ... would have thought the Novel would have picked up more on the kind of sentence structure of the language from what they've learnt, whereas the topic probably more so ... what we're more familiar with, with normal teaching of just having lines of ... some French, of being able to put out some nouns. The progress is there. But I would say they're probably similar, just in different ways".	

T8: “I’d say that the Novel group have ... come on more [in phonics] - generally there are my higher spellers ... that just happen to be in that group. I’ve got ... some of the very low spellers in that group ... but today one of ... the children – she’s assessed as one of my low abilities – was flying through the lesson ... so, in that sense, it, I am seeing a slight difference in the Novel. Standard are still working to a good level, but ... yeah, definitely Novel are slightly higher”. T8 (refers to feedback provided by the activities): “I find that those [matching] games work quite well – a bit like the ones today with the tick if it’s ‘un’/‘une’ (in Novel) ... That’s helping them a lot, cos I can then clearly see who’s got it, who hasn’t got it ... And they find it quite nice cos ... you can mark yourself, and ... if you’ve got it correct ... it’s a bit of a confidence boost for them too”. T8: “the instinct is to kind of go that Novel I’d say are making a bit more progress. Just, they seem more confident in themselves with the French; Standard – you can see it’s a bit of a 50:50 ... you’ve got some that are still quite – they’ll let everyone else join in, they’ll sit back and see – and you’ve got the really, really confident ones. But yeah, with Novel I’d say the whole group are progressing together”.	T8 (without reference to either approach): “And the children, I’ve noticed, ... it’s been helping their phonics sounding out as well. Because they’re making the links to the English and the French sounding ... [It is impacting on their English] cos they’re able to understand - some of the words that come up in the [Year] 3/4 spelling list have French origins and they’re able to now spell them better because they’ve had the French lessons ... especially with the [sh]/[ch] sound”.	T8: “cos last week we were doing, erm – I can’t remember what it was last week in standard – there was definitely sentences towards the end. And course, they’re doing it all the time now – they’re saying ‘I have’, ‘you have’. Erm, so, yeah, they’re much more confident building sentences. Some of them ... they surprise themselves, I think, by the end they go “oh, I was able to say ‘I have two bottles’” or something”. T8: “Standard – you can see it’s a bit of a 50:50 ... you’ve got some that are still quite – they’ll let everyone else join in, they’ll sit back and see – and you’ve got the really, really confident ones”.
	T9: “I think we’ve got an even split of ones which are doing really well across both those groups. Again, and also that thing of ‘Oh, I wouldn’t have expected you to do well’, so I think they’re both working”.	T9: (refers to feedback provided by the activities) “It feels like it’s easier to monitor the progress in Standard. ... it feels like the range hasn’t been as broad. I don’t know if that’s correct or not, but it feels like ... you’re following fairly similar lines with the Standard, ... you’re not jumping around so much. It doesn’t necessarily make it better. It’s just if you’re monitoring progress, I think ... the games again, it’s ... going back to the games again ... it’s quite a nice way of informally monitoring their progress”.

Appendix W.2

Interview3 Teacher Comments on Pupils' Progress

Comments in blue are positive comments on pupil progress, comments in red are negative. Some comments do not refer to either approach specifically and have therefore been placed in the middle column.

Progress in Novel	Progress in both	Progress in Standard
Impact of focus on sentence structure T1: "... with the frequency of words, you can apply it to the language, whereas just learning words is OK, but you're not learning the sentence structure or ... masculine or feminine to that level, that I think that you probably are to Novel ... as we got towards the end, where they [pupils] were writing the simple sentences, they were really good at learning the sentence structure".	Importance of regular lessons: T3: "I think the children have got a lot out of it – whether that's to do with the schemes or more to do with the frequency with which ... they've had French ..."	Less good recall: T2: "I think with the Standard, that I didn't really feel that as much. I didn't feel that ... they could recall the information as well as the Novel approach".
Overall better progress T1: "I'm trying really hard not to just think of individual children who are the high ability children. I'm trying to think of the group more generally - otherwise, it's an unfair answer I would give. I think, no ... it's difficult to tell because of the mixed abilities. I don't ... mmm, I think maybe Novel [pupils have made more progress]".	Struggling with pronunciation: T3: "I think ... some children found the pronunciation very difficult ... and even at the end there's definitely some children that really struggled. I did try to model it as much as I could within the, the time frame. I think that was an issue. I: Did you notice that being more of an issue in one approach or the other? T3: I don't think it ... no, I can't see that it was one or the other".	Possibly better performance in vocab recall: T3: "Perhaps the Standard approach, they might be better on the vocabulary though. Maybe. I don't know".
Better foundation; better preparation to write at sentence level. T1: "I think possibly ... with the Novel approach they have learned a good foundation. I think it's a very good foundation with Standard as well, but I think if I were to ask them to write and ... - obviously with scaffolding and support - but if we were to structure sentences and do more writing, I think the Novel approach would ... help more there".	Grasping grammar: T7: "I was quite impressed how much they remembered, the masculine and feminine, actually - I thought they were really good with that".	Impact on confidence: T8: "Standard children as well, they, they've made progress ... but they weren't always as confident sharing their knowledge ... They were less confident to share their knowledge".
Better recall in Novel: T2: "I definitely feel, even though I think I preferred Standard, I think the knowledge ... I think the bit that surprised me most, I think it must have been through the Novel approach, where the things that some of the kids were coming out with, I think there was one of the lessons where they recognised – 'Oh, we did this one ...' and I was really surprised ... I think it was through the Novel approach where it was that grid at the end, some of the words that we'd learned weeks and weeks ago ... they still knew"	High standards compared with experience of teaching Year 3 before: T7: "I was thinking that they're Year 3 and what I would normally do, that's, you know, quite impressive, I think even further up the school. I mean they, they manage it, they do it, but ... which is great, and it's a shame that that won't carry on for them. You know, they'll probably dip, but ... because they've amazed me how much they do actually".	Better retention of vocab: T8: "I think ... they possibly had a better understanding of, well, I'd say their vocabulary was almost ... they had a different vocabulary knowledge to ... the Novel. They picked up, they ... kept words that the Novel children hadn't ... they really, really loved the days of the week and numbers and things, whereas the Novel children, if one of, if one of the Standard children went up to them and said: "Oh, I can tell you the seven days of the week", the Novel, it would, it would take them a minute and they would hesitate to join in sometimes, because they would have forgotten one or two days".
Overall progression: T3: "I could definitely see the children's progression".	Surprised that children can work at sentence level: T7: "I would be teaching just really the key words, rather than worrying about putting them into a sentence. And I think I would be hitting more on that as well,	Lower achievers struggle less in Standard: T9: "I think the good ones have made equally as good progress in both groups ... but I think the less able ones find the Standard approach easier than the less able ones in the Novel".

	<i>actually ... cause they've shown me this year they can do that".</i> <i>"Rather than just teaching, 'Right, these are animals, and we're going to learn the names for them'. And that's the end of the lesson. You know, let's try and put it into a sentence, or let's ... say a sentence".</i>	
Able to say whole sentences: T3: <i>"I did get pleasure out of seeing ... the pleasure on the children's faces in the Novel approach when they realised that they'd said a whole sentence, for example ... that was amazing".</i>	High standards compared with experience of teaching Year 3 before / importance of regular teaching: T7: <i>"I've gen-, genuinely I was really impressed with the progress all of them made actually with both ... you know, when I was talking to a teacher here ... and I was saying what we were covering, they couldn't believe, like, the things that we were doing. You know, that they could do in Year 3, so I feel they've really been kind of pushed but were able to do it. Yeah ... they really impressed me actually with it, with how much they can do, you know, when given that kind of regular teaching".</i>	
Able to construct whole sentences: T3: <i>"I think the scheme, particularly the Novel scheme, enabled you to do that [build sentences]. I definitely think the children got a lot of ... satisfaction from that".</i>		
Impact on focus: T3: <i>"I think some children in the Novel approach probably gave up a bit particularly if they struggle anyway".</i>	Overall progress: T7: <i>"generally speaking, they've all made ... really, really good ... progress".</i>	
Improved overall performance in Novel: T3: <i>"I suspect that the Novel approach will have more progress. That's what ... I think may happen. I think it will be definitely more noticeable in the grammar and the phonics".</i>	Progress did not differ according to approach: <i>"Do you think they've made progress in different areas according to the, the approach that they were allocated to?"</i> T7: <i>No, I don't ... think so. I don't think so".</i>	
Increased ability to speak sentences: T8: <i>"in the Novel I quite liked how ... they did a lot of practice conversations. So, they were looking at what was going on the board. Sometimes it was the reading, sometimes they were listening, but they were constantly exposed to conversations in French. And I do find that those, that children in the Novel were more confident talking to me ... in French in sentences".</i>	Progress did not differ according to approach: <i>"Do you think they've made progress in different areas in each approach?"</i> T2: <i>... No, not necessarily ... I don't really think so, no".</i>	
Quicker to pick up vocab: T8: <i>"Novel - they picked up a lot quicker (the vocabulary) and the Standard, it took them a little bit longer, I think. I wouldn't be able to tell you why, I just ... I think both groups just approached it a little bit differently. But the Novel group definitely picked it up quicker".</i>		
Increased confidence and understanding / High standards compared with experience of teaching Year 3 before: T8: <i>"I would say Novel [made more progress]. I'd seen a lot of progress in the children ... just ... their confidence with French, their understanding surpassed anything I would have expected for a year 3 French lesson ... By the end, they were really, really confident in lessons. Standard children as well, they, they've made progress, erm, but they weren't always as confident sharing their knowledge".</i>		

Less good retention of (thematic) vocabulary: T8: "... the Novel children, if one of ... the Standard children went up to them and said: 'Oh, I can tell you the seven days of the week', the Novel, ... it would take them a minute and they would hesitate to join in sometimes, because they would have forgotten one or two days".		
Better foundation: T9: "Going forward as an academic subject, the Novel, obviously, I think gives you a better grounding ... It'd be quite interesting going forward, if the ones who've done the Novel with the more grammar focus, but ... the heavier grammar focus, actually go to be better at French with that, with the start they've had or not ...".		
Lower achievers struggle more in Novel: T9: "I think the good ones have made equally as good progress in both groups ... but I think the less able ones find the Standard approach easier than the less able ones in the Novel".		

Appendix W.3

Additional Quotations Supporting Growth in Teacher Confidence (Interview2)

"I feel much more confident now ... definitely my confidence is ... growing the more I teach" (T1);

"I definitely feel like, from this ... I'm becoming – slowly, but I am becoming – more confident" (T2);

"I feel a lot more confident teaching French than I have been in the past" (T8).

Additional Quotations Supporting Growth in Teacher Confidence (Interview3)

"I do think I would feel more confident teaching it next year ... where you're doing it ... as its own subject, which I'm not audio recording ... I think I probably would feel more confident now. I mean, even teaching the scheme that we have here, I think I would feel like my subject knowledge was a bit better, so I would feel more confident doing it" (T7);

"I think it's just it will give me that ... more confidence and ... the subject knowledge will help as well ... I think before, I've always felt like I've really got to work to try and stay ahead of them [the learners]. I feel like I would be in a better place at the start of them ... I think I would be more confident teaching it" (T7);

"I think overall I feel a lot more comfortable now and since February ... I'm into the routine. I knew what I was doing, and the lessons just became a bit easier and I found myself - even just the language side of it – I was more confident correcting and speaking French with them" (T8);

"I feel much more confident in delivering French lessons than I did at the start of the year, undoubtedly ...and I think that's due to both of them, to be honest" (T9).