

**A Classroom Quasi-experimental Study to  
Explore Processing Instruction**

**Hsin-Ying Chen**

**Doctor of Philosophy**

**University of York**

**Department of Educational Studies**

**September 2009**

## Abstract

Processing Instruction (VanPatten, 1996, 2002a, 2004) contains two types of input activity: Referential activities, which force learners to focus on a form and its meaning, and affective activities, which contain multiple exemplars of the target form but focus learners' attention on the meaning of the sentences in which the form is embedded. To date, these two types of PI activity have been treated as one pedagogical technique, and no study has been empirically conducted to investigate the instructional impact of them individually. Furthermore, whether or not PI activities can promote learners' implicit knowledge has not been addressed empirically.

120 12-year-old Taiwanese learners of L2 English were quasi-randomly assigned to four groups: Referential + Affective, Referential-only group, Affective-only and a Control. Pre, post and delayed post tests were administered to assess learning of the English 'ed' verb inflection. The measures included three tests aiming to elicit implicit knowledge: A timed grammaticality judgment test, an oral picture narration, and a short structured conversation. Following these tests, a self-report technique was employed to check whether or not learners drew on explicit knowledge. A gap-fill test without a time constraint and a written vocabulary test were also included to examine instructional impact.

Findings suggest that referential activities are responsible for the learning gains observed and that the gains are held for up to six weeks after completion of the intervention. However, the issues regarding the role of affective activities in vocabulary learning and PI's impact on implicit knowledge need further study. An implication of this study is that the claims of previous PI studies regarding the causative factors for its effectiveness require more refined exposition.

## Contents

Title.....	i
Abstract.....	ii
List of Contents.....	iii
List of Tables.....	x
List of Figures.....	xiii
Acknowledgements.....	xiv
Declaration.....	xv
Appendices.....	xvi

<b>Introduction.....</b>	<b>1</b>
--------------------------	----------

### Chapter 1 Context Review

Introduction.....	4
1.1 The English curriculums reforms in Taiwan.....	4
1.2 The belief and practice in grammar teaching and learning in Taiwan.....	6
1.2.1 Taiwanese teachers' viewpoints about grammar teaching.....	6
1.2.2 Taiwanese students' viewpoints about grammar learning.....	7
1.3 Justification for an investigation of PI in a Taiwanese primary school.....	8

### Chapter 2 The framework of Processing Instruction, its relevant theoretical backgrounds, and review of literature on Processing Instruction research

Introduction.....	11
2.1 What is Processing Instruction?.....	11
2.1.1 The framework of PI.....	12
2.1.1.1 The components of PI.....	12
2.1.1.2 The differences between these two structured input activities.....	18
2.1.2 The Nature and Uniqueness of PI.....	20
2.1.2.1 An input-based approach.....	20
2.1.2.2 A focus-on-form approach (FonF).....	21
2.1.2.3 Derivation from Input Processing (IP) theory to achieve better Form-Meaning Connection (FMC).....	22
2.1.3 The guidelines for the structured input activities.....	23
2.2 Relevant theories underpinning PI.....	26
2.2.1 Input Processing (IP) Theory.....	26

2.2.1.1	What is Input Processing Theory? .....	26
2.2.1.2	Key notions of IP theory .....	28
2.2.1.3	The IP principles and their empirical evidence .....	30
2.2.1.4	Challenges and unclear issues of IP .....	42
2.2.2	Form-MeaningConnections (FMCs) .....	46
2.2.2.1	Why are FMCs important? .....	46
2.2.2.2	The developmental processes of FMCs .....	47
2.2.3	Attention Theory .....	48
2.2.3.1	Characteristics of attention .....	48
2.2.3.2	Postulations of attention in SLA .....	52
2.2.4	Some evaluations and challenges of PI in terms of IP, FMCs, and attention ...	54
2.2.4.1	PI activities adherence to FMCs? .....	54
2.2.4.2	Redundancy obstructing FMCs? .....	55
2.2.4.3	The impact of different modalities on PI .....	55
2.2.4.4	The role of output .....	56
2.2.4.5	Practical issues .....	57
2.3	Literature Review of previous PI research: verified and unverified issues of PI, and the formation of the current study .....	58
2.3.1	Some verified and unverified issues of PI .....	58
2.3.1.1	What is the relative effectiveness of PI vs. other types of grammar instruction? .....	58
2.3.1.2	Can the positive effect of PI be generalised to other linguistic features in different languages? .....	62
2.3.1.3	What is the relative effectiveness of PI delivered by different modes? .....	63
2.3.1.4	Are the positive effects of PI studies attributable to the explicit information provided? .....	63
2.3.1.5	What is the long-term effect of PI? .....	65
2.3.1.6	What is the effect of PI on the less-controlled oral production task? ...	67
2.3.1.7	Can PI promote learners' implicit knowledge? .....	69
2.3.1.8	Do different PI activities have different instructional impact? .....	71
2.3.1.9	How trustworthy are the results of previous PI studies? .....	74
2.3.2	The formation of the current study .....	76
2.3.2.1	Motivations .....	76
2.3.2.2	The definition and operationalisation of implicit and explicit knowledge in the current study .....	78
2.3.2.3	The choice of measures to elicit implicit and explicit knowledge .....	79
2.3.2.4	The choice of linguistic feature for the current study .....	86
2.3.2.5	The research questions and hypotheses .....	87

### **Chapter 3 The methodological issues and design of the current study**

Introduction.....	90
3.1 A review of literature on carrying out an experiment in educational research .....	90
3.1.1 The methodological issues .....	90
3.1.1.1 Why a classroom-based quasi-experimental study?.....	90
3.1.1.2 Some challenges faced when conducting an experiment and how this study would handle them .....	91
3.1.2 Ethical considerations .....	95
3.1.2.1 Can the results of the education experiment inform practice?.....	95
3.1.2.2 Should the practising teachers be involved in this study?.....	95
3.1.2.3 Ethical considerations regarding the interpretation of the findings .....	97
3.1.2.4 The privacy of research participants .....	97
3.1.2.5 The right of the control group to be treated equally.....	98
3.2 The current study: a quasi-experimental design.....	99
3.2.1 The participants and the educational context .....	99
3.2.2 The interventional procedures.....	104
3.2.3 The instructional material packages.....	108
3.2.3.1 Design of the instructional materials.....	108
3.2.3.2 The administration of the interventions .....	115
3.3 The achievement assessments used in the current study.....	118
3.3.1 The Grammaticality Judgment Test (GJT).....	118
3.3.1.1 The design of timed GJT for this study.....	118
3.3.1.2 The obtainment of a given time for each individual sentence .....	120
3.3.1.3 The administration of the timed GJT for L2 learners. ....	122
3.3.2 The written production test: a gap-fill test .....	123
3.3.2.1 The design of the written production test.....	123
3.3.2.2 The administration of the written production test.....	124
3.3.3 The oral production tests .....	125
3.3.3.1 The design of the oral production tests .....	125
3.3.3.2 The administration of oral production tests .....	126
3.3.4 The vocabulary test .....	128
3.3.5 The retrospective self-report .....	128
3.3.5.1 The post-task written questionnaire .....	128
3.3.5.2 The structured interview of rule verbalisation .....	129
3.3.6 Two versions of each achievement assessment.....	130
3.3.6.1 Randomly assigning tests to groups as the pre-test and post-test .....	131
3.3.6.2 The comparability of the achievement assessments.....	131
3.3.7 The scoring procedures .....	132

3.3.7.1	The timed Grammaticality Judgment Test .....	133
3.3.7.2	The written gap-fill test.....	133
3.3.7.3	The oral tests .....	133
3.3.7.4	The vocabulary test .....	134
3.3.7.5	The post-task self-reports.....	134
3.4	The statistical analysis procedures of the achievement assessments .....	134
3.4.1	Parametric tests vs non-parametric tests .....	135
3.4.2	The tests to examine the fitness of parametric tests .....	136
3.4.3	Statistical significance test: setting the probability value .....	137
3.4.4	Statistical tests used in this thesis to examine the mean differences.....	138
3.4.4.1	The dependent and independent t-tests and their non-parametric counterparts .....	139
3.4.4.2	The ANOVA and its non-parametric equivalent .....	139
3.4.4.3	The planned contrast .....	141
3.4.4.4	The ANCOVA .....	142
3.4.5	The estimate of the magnitude of interventions: effect size.....	142
3.4.6	The correlation .....	143
3.4.7	The principal component analysis.....	145
3.4.7.1	Suitability/factorability of principal component analysis .....	146
3.4.7.2	The criterion for extracting components .....	147
3.4.7.3	To assist in the interpretation: factor rotation technique.....	148
3.4.8	Pearson's chi-square test for independence.....	149
3.5	The validity, reliability and comparability of the assessment tests.....	150
3.5.1	The validity of the achievement assessments.....	151
3.5.1.1	The validity results for the achievement assessments.....	152
3.5.2	The reliability of the achievement assessments .....	153
3.5.2.1	The test-retest method.....	154
3.5.2.2	Cronbach's alpha ( $\alpha$ ) .....	158
3.5.3	The comparability of the two versions of the achievement assessments .....	160
3.5.3.1	The comparability results of the two versions of the achievement assessments .....	162
3.6	The limitations of the achievement assessments.....	164
3.6.1	The elicitation tests for measuring implicit and explicit knowledge .....	165
3.6.2	Reservations concerning the post-task retrospective self-report.....	166
3.6.3	The validity of the achievement assessments.....	166
<b>Chapter 4 The results of the achievement assessments</b>		
Introduction.....		168



## **Chapter 5 The results of the questionnaires regarding the participants' bio-data and their attitudes towards the interventions, and from the ANCOVA**

Introduction.....	217
5.1 Analysis of the participants' bio-data and English learning backgrounds.....	217
5.1.1 Experience of travelling to English-speaking countries.....	218
5.1.2 Length of English learning experience.....	220
5.1.3 Extra English exposure outside school: attending extra English lessons.....	222
5.1.4 Contact with English native speakers outside school.....	225
5.1.5 Summary of this section.....	227
5.2 The results of the ANCOVA.....	227
5.2.1 Using participants' English learning length as a covariate.....	228
5.2.1.1 The results of ANCOVA on gap-fill tests at delayed post-test.....	228
5.2.1.2 The results of ANCOVA on vocabulary tests at post-test.....	229
5.2.1.3 The results of ANCOVA on vocabulary tests at delayed post-test.....	230
5.2.2 Using participants' extra exposure to English as a covariate.....	231
5.2.2.1 The results of ANCOVA on timed GJTs at delayed post-test.....	231
5.2.2.2 The results of ANCOVA on the gap-fill tests at post-test.....	232
5.2.2.3 The results of ANCOVA on the gap-fill test at delayed post-test.....	234
5.2.2.4 The results of ANCOVA on vocabulary tests at post-test.....	235
5.2.3 Summary of the ANCOVAs which take into account English learning length and Extra exposure to English as potentially confounding factors.....	236
5.2.3.1 The ANCOVA results using English learning length as a covariate.....	236
5.2.3.2 The ANCOVA results using extra English exposure as a covariate.....	236
5.3 Analysis of the attitudinal questionnaire.....	237
5.3.1 The operation of the computer.....	238
5.3.2 The motivation level of the intervention.....	239
5.3.3 The level of difficulty of the intervention.....	241
5.3.4 Willingness to carry out similar activities in the future.....	243
5.3.5 Summary of this section.....	244

## **Chapter 6 Discussion of the results and findings**

Introduction.....	246
6.1 Discussion of the findings of the timed GJT.....	246
6.1.1 The relative impact of the interventions on the timed GJT.....	246
6.1.2 Linkage of the results of the timed GJT to previous PI studies.....	249
6.2 Discussion of the findings of the gap-fill test.....	250
6.2.1 The relative impact of the intervention on the gap-fill test.....	250



6.2.2	Linkage of the results of the gap-fill test to previous PI studies.....	253
6.3	Discussion of the findings of the oral tests .....	254
6.3.1	The impact of interventions on the picture-based narration.....	254
6.3.2	Discussion of the impact of interventions on the structured conversation.....	257
6.3.3	Linkage of the results of the oral tests to previous PI studies .....	257
6.3.4	Why was the impact of the interventions not so promising in the oral tests? .....	258
6.4	Discussion of the findings from the vocabulary test.....	260
6.4.1	The relative impact of the interventions on the vocabulary test .....	260
6.4.2	Linkage of the results of the vocabulary test to previous PI studies .....	263
6.5	Discussion of the issues regarding implicit and explicit knowledge derived from the PI activities in this study .....	264
6.5.1	Discussion of the results of the elicitation tests .....	264
6.5.1.1	The results of the principal component analysis .....	264
6.5.1.2	The results of self-reports following the timed GJT and oral tests .....	265
6.5.2	Why the timed GJT in this study failed to elicit implicit knowledge .....	267
6.5.3	Did the oral tests tap into implicit knowledge?.....	269
6.5.4	What type of knowledge is derived from different interventions?.....	270
6.5.5	Linkage of the results to previous PI studies .....	271
6.6	Discussion of the relative effectiveness of interventions for the ‘-ed’ feature.....	272
6.6.1	From the perspective of FMCs.....	273
6.6.2	From the perspective of attention.....	275
6.6.3	Further explanations for the relative effectiveness of interventions .....	277
6.7	The purpose of the affective activities .....	279

## **Chapter 7 Summary and Conclusion**

7.1	Summary of the current study .....	281
7.2	Justification and originality of the current study.....	282
7.3	Findings and Discussion .....	283
7.3.1	RQ 1- 4 and H1 .....	283
7.3.2	RQ 5 and H2.....	286
7.3.3	RQ 6 and H3-5 .....	287
7.3.4	RQ 7 and H6 .....	290
7.4	The contribution of the current study.....	291
7.5	Limitations of the current study and implications for future research.....	294
7.5.1	Research design.....	294
7.5.2	The targeted linguistic feature.....	297
7.5.3	Achievement assessments .....	297

## List of Tables

2.1	A summary of PI studies including oral test and their results.....	68
3.1	The timescale of the intervention and assessment .....	105
3.2	The tally of the number of practice items in each instructional group.....	112
3.3	Total occurrences of vocabulary test items in the intervention materials.....	113
3.4	The number of glossed vocabulary items assessed in the vocabulary test.....	115
3.5	Summarisation of the counterparts of parametric and non-parametric tests used in this thesis to examine the mean differences.....	138
3.6	Summary of the investigation into validity, reliability, and comparability.....	151
3.7	Descriptive statistics in assessments to investigate validity .....	152
3.8	The results on the validity of the timed GJT and the gap-fill test.....	153
3.9	Descriptive statistics to investigate the test-retest reliability.....	156
3.10	The Wilcoxon signed-rank results for test-retest reliability.....	156
3.11	The Spearman's correlation for test-retest reliability.....	157
3.12	The internal reliability results on achievement tests.....	160
3.13	Descriptive statistics for the comparability of the tests .....	163
3.14	The results of the comparability of the tests .....	164
4.1	Descriptive statistics for the GJTs.....	171
4.2	The results of the Friedman test in the timed GJTs.....	172
4.3	The results of the post-hoc test for the Friedman test in the timed GJTs.....	173
4.4	Descriptive statistics for non-parametric tests in the timed GJTs.....	174
4.5	The magnitudes of instructional effect on the timed GJTs.....	175
4.6	The magnitudes of change from the pre-test to the post-test on the GJTs.....	175
4.7	The meta-analysis of previous PI studies on the GJT.....	176
4.8	Descriptive statistics for the gap-fill test.....	177
4.9	The results of the Friedman test on the gap-fill tests.....	179
4.10	The results of the post-hoc test for the Friedman test on the gap-fill tests.....	180
4.11	Descriptive statistics for non-parametric tests on the gap-fill tests.....	181
4.12	The magnitudes of instructional effect on the gap-fill test.....	181
4.13	The magnitudes of change from the pre- to post-tests on the gap-fill test.....	182
4.14	The meta-analysis of previous PI studies on the written production test.....	184
4.15	Descriptive statistics for the picture-based narration test.....	185
4.16	The results of the Friedman test on the picture-based narration test.....	187
4.17	Descriptive statistics for non-parametric tests on the picture-based test.....	187
4.18	The magnitudes of effect for the intervention on the picture narration test.....	188
4.19	The magnitudes of change from pre- to post- tests on picture narration test.....	189
4.20	The meta-analysis of previous PI studies on the oral test.....	189

4.21	Mean rate of suppliance in obligatory contexts for structured conversation....	190
4.22	Descriptive statistics for the vocabulary test.....	192
4.23	The results of the Friedman test on the vocabulary test.....	194
4.24	The results of the post-hoc test for the Friedman test on the vocabulary test...	195
4.25	Descriptive statistics for non-parametric tests on the vocabulary test.....	196
4.26	The magnitudes of change from the pre- to post- tests on vocabulary test.....	197
4.27	Descriptive statistics for the RA group at the post-test.....	199
4.28	Pearson correlation matrix for the tests of the RA group at the post-test.....	200
4.29	Principal component analysis of the RA group at the post-test.....	201
4.30	Loadings after the oblique rotation of the RA at the post-test.....	201
4.31	Descriptive statistics for the R group at the post-test.....	201
4.32	Pearson correlation matrix for the tests of the R group at the post-test.....	202
4.33	Principal component analysis of the R group at the post-test.....	203
4.34	Loadings after the oblique rotation of the R group at the post-test.....	203
4.35	Descriptive statistics for the A group in the post-test.....	203
4.36	Pearson correlation matrix for the tests of the A group in the post-test.....	204
4.37	Summaries of the principal component analysis results at the post-test.....	205
4.38	Descriptive statistics for the RA group at the delayed post-testl.....	205
4.39	Pearson correlation matrix for the RA group at the delayed post-test.....	206
4.40	The principal component analysis of the RA group at delayed post-test.....	206
4.41	Loadings after the oblique rotation of the RA group at delayed post-test.....	207
4.42	Descriptive statistics for the R group at the delayed post-test.....	207
4.43	Pearson correlation matrix for tests of the R group at delayed post-test.....	208
4.44	Descriptive statistics for the A group at the delayed post-test.....	208
4.45	Pearson correlation matrix for the A group at the delayed post-test.....	208
4.46	The principal component analysis of the A group at the delayed post-test.....	209
4.47	Loadings after the oblique rotation of the A group at the delayed post-test.....	209
4.48	Summaries of the principal component analysis results at delayed post-test....	210
4.49	Cross-tabulation of Group * post-task questionnaire at the post-test.....	212
4.50	Cross-tabulation of Group * post-task questionnaire at the delayed post-test...	212
4.51	The biserial correlation between post-task questionnaire and their test scores in the timed GJT.....	213
4.52	Cross-tabulation of Group * post-task interview at the post-test.....	215
4.53	Cross-tabulation of Group * post-task interview at the delayed post-test.....	215
4.54	The point-biserial correlation between post-task interviews and merged oral scores.....	216
5.1	The cross-tabulation of whether participants had travel experience in English- speaking countries * by group.....	218

5.2	The point-biserial correlations between experience of travel in an English-speaking country and scores in the achievements tests.....	219
5.3	Descriptive statistics of participants' English learning length.....	221
5.4	Spearman's correlation between English learning length and test scores at the post-tests.....	222
5.5	The cross-tabulation of extra English exposure * Group.....	222
5.6	Descriptive statistics of participants' extra exposure to English.....	223
5.7	Spearman's correlation between the participants' extra exposure to English after school and the test scores.....	225
5.8	Crosstabulation of contact with English native-speakers outside school * Group.....	225
5.9	Point-biserial correlations between whether or not contact was made with English native speakers and the tests scores .....	226
5.10	The results of the ANCOVA on the gap-fill test at the delayed post-test.....	229
5.11	The results of planned contrasts on the gap-fill test at the post-test.....	229
5.12	The results of the ANCOVA on the vocabulary test at the post-test.....	230
5.13	The results of the ANCOVA on the vocabulary test at the delayed post-test.....	230
5.14	The results of the ANCOVA on the timed GJT at the delayed post-test.....	232
5.15	The results of planned contrasts on the timed GJT at the delayed post-test.....	232
5.16	The results of the ANCOVA on the gap-fill test at the post-test.....	233
5.17	The results of planned contrasts on the gap-fill test at the post-test.....	233
5.18	The results of the ANCOVA in the gap-fill test at the delayed post-test.....	234
5.19	The results of planned contrasts on the gap-fill test at the delayed post-test.....	235
5.20	The results of the ANCOVA on the vocabulary test at the post-test.....	235
5.21	The cross-tabulation of operation on the computer * Group.....	238
5.22	The point-biserial correlation between whether or not participants felt that it was easy to operate the computer and the test scores.....	239
5.23	The cross-tabulation of attitudes towards the intervention* Group.....	240
5.24	The point-biserial correlation between whether or not participants felt an intervention interesting and the test scores.....	240
5.25	Attitudes of the participants towards the level of difficulty of the intervention* Group.....	241
5.26	Which activities were difficult.....	242
5.27	Point-biserial correlation between whether or not participants perceived an intervention difficult and the test scores.....	242
5.28	Willingness to carry out similar activities* Group.....	244
6.1	The number of occurrences of participants' reformulation in the oral tests.....	269

## List of Figures

2.1 The processing of second language acquisition.....	27
3.1 The experimental design.....	101
3.2 The interventioanl procedure.....	105
4.1 Scores on the timed GJTs over time .....	171
4.2 Scores on the gap-fill tests over time.....	178
4.3 Scores on the picture-based narration tests over time.....	186
4.4 Scores on the vocabulary tests over time.....	193

## **Acknowledgements**

There are many people that I would like to thank, as this thesis would not have been complete without them. Many thanks go to my supervisor at University of York, Dr. Marsden, for her constant support and insightful advice over the past four years.

I am also grateful to the staff, teachers, and students at Dong-Hai elementary school in Taitung, Taiwan, for their participation in this study.

I thank my family for their great love, substantial support, and endless encouragement over my many years as a student. I also thank my friends, my boyfriend Ben Dudson in particular, for their love and patience.

Finally, I thank myself for going through and surviving from this Permanent Head Damage (PHD) process.

## **Declaration**

I declare that this thesis is based on my own work. All published references are cited. At the time of submission, the results of this study have been submitted to the Journal of *Language Learning*.

## Appendices

Appendix 1	The official objectives of English curriculum and guidelines for constructing English teaching materials.....	300
Appendix 2	The ‘first noun principle’ and its corollaries in IP.....	304
Appendix 3	A tabular summary of studies related to PI .....	305
Appendix 4	Summary of PI-based studies’ findings on interpretation tasks .....	311
Appendix 5	Summary of PI-based studies’ findings on production tasks .....	314
Appendix 6	The consent of the headmaster of the participating school .....	317
Appendix 7	The raw scores of the 13 outliers .....	318
Appendix 8	The timetable of the current study.....	319
Appendix 9	The questionnaire regarding subjects’ English learning backgrounds ....	320
Appendix 10	The attitudinal questionnaire .....	322
Appendix 11	Handout given to participants during the instructional phases.....	323
Appendix 12	An example of referential activities for this study .....	324
Appendix 13	An example of affective activities for this study .....	325
Appendix 14	The test items of the timed GJT .....	326
Appendix 15	The answer sheet of the timed GJT .....	328
Appendix 16	The consent form for L1 participants .....	330
Appendix 17	Ten examples for the timed GJT .....	331
Appendix 18	The gap-fill test: Version A .....	332
Appendix 19	The gap-fill test: Version B .....	334
Appendix 20	Quick revision list for the gap-fill test: Version A .....	336
Appendix 21	Quick revision list for the gap-fill test: Version B .....	337
Appendix 22	The picture-based narration test: Version A .....	338
Appendix 23	The picture-based narration test: Version B .....	348
Appendix 24	Quick revision list for the picture-based narration test .....	358
Appendix 25	The vocabulary test: Version A .....	360
Appendix 26	The vocabulary test: Version B .....	361
Appendix 27	The post-task questionnaire.....	362
Appendix 28	The interview sheet for the post-task interview .....	363
Appendix 29	Results of the K-S test for achievement test versions .....	364
Appendix 30	The results of Levene’s test on both versions of achievement test to investigate the validity .....	365
Appendix 31	Results of the K-S test for the achievement tests .....	366
Appendix 32	The results of Levene’s test on achievement tests.....	368
Appendix 33	The results of parametric tests on the timed GJT .....	369
Appendix 34	The results of parametric tests on the gap-fill test.....	371



Appendix 35	The results of parametric tests on the picture-based narration test .....	373
Appendix 36	The results of parametric tests on the vocabulary test.....	374
Appendix 37	The results of parametric tests for the validity of tests.....	375
Appendix 38	The results of parametric tests for test-retest reliability of tests.....	376
Appendix 39	The results of parametric tests for the comparability of two versions of achievement tests.....	378
Appendix 40	The Histograms and Boxplots .....	379
Appendix 41	A sample transcription of a structured conversation .....	382
<b>Reference</b>	.....	<b>384</b>

## **Introduction**

This thesis sets out to investigate the roles of the two different types of structured input activities (SIA) in the framework of Processing Instruction (i.e. PI, a type of input-based grammar pedagogy package). The study set out to address some weaknesses of previous PI studies and to begin to address several new issues in the PI research agenda. A classroom-based quasi-experiment was carried out to investigate the effectiveness of different types of input activities within PI – referential and affective activities. 120 learners of English aged 12 from a Taiwanese primary school were allocated to four different groups: Referential + Affective group, Referential-only group, Affective-only group, and a Control group. Learning gains were examined by comparing the mean scores achieved by the participants at the pre-test and two post-tests, and by comparing the mean scores of the instructional groups and control group on the learning of an English verb inflection, the past tense ‘-ed’ feature. A range of elicitation tests were employed to measure the impact of the interventions, including a grammaticality judgment test (GJT) with a time constraint, a gap-fill written test, a picture-based narration oral test, a structured conversation and a short written receptive vocabulary test. This thesis is the first study which aims to separate the SIA. It is also the first study to date to explore the nature of the learning promoted by SIA by investigating whether the knowledge tended to be explicit or implicit knowledge. The study also investigated some potentially extraneous variables which have not been previously acknowledged in the published literature, including learners’ attitudes towards the intervention materials and their extra-curricular exposure to English.

The layout of this thesis is as follows:

Chapter 1 provides a brief contextual review of policy and research relating to English

grammar teaching and learning in Taiwan and the views of Taiwanese teachers and students about grammar instruction. The grounds for why PI fits in with the context are addressed. Chapter 2 mainly focuses on the literature review of PI research. It starts with the presentation of the framework of PI by introducing its components, articulating its nature and uniqueness, and describing how to construct PI activities. This chapter also addresses the theoretical frameworks which are claimed to underpin PI, including Input Processing (IP), Form-Meaning Connections (FMCs), and attention. Following that, some of the motivations for the current study are laid out by reviewing the relevant issues which emerge from prior PI-based studies. The motivation, the choice of the targeted linguistic feature and the rationale behind the outcome measurements are also described. This chapter closes by proposing the research questions and hypotheses for the current study. Chapter 3 starts with a review of literature about the implementation of a quasi-experiment in educational research, justifying some of the broad methodological decisions taken. A description of the quasi-experimental design for this current study is then provided, including the allocation of the participants, the intervention procedures, and the instructional material packages used. The design and administration of the achievement assessments and the statistical procedures applied to analyse the data are also described. In addition, the validity and reliability of the measures developed to assess learners' learning gains are reported in this chapter. This chapter closes with an acknowledgement of the limitations of some assessments used. Chapters 4 and 5 present the results obtained from the achievement assessments and the questionnaires. A critical analysis of the results of the achievement assessments at the pre-test, the post-test, and the delayed post-test are reported in Chapter 4. In order to identify any potential confounding variables exerting an influence on the impact of the interventions, Chapter 5 reports the results from the questionnaires concerning participants' bio-data and English learning backgrounds, and their attitudes towards the

interventions. Chapter 6 discusses the findings which emerged from the data. These findings are also linked to prior PI studies and relevant theoretical frameworks. The final chapter, Chapter 7, starts with a summary of the study and then reports concisely on and discusses the main findings of this study in the light of the research questions and hypotheses posed in the second chapter. A critique of this study in terms of its limitations as well as the implications for future research is also presented in this final chapter.

## Chapter 1 Context Review

### Introduction

This Chapter describes the context of English<sup>1</sup> grammar teaching and learning in Taiwan and then goes on to reason why the investigation of Processing Instruction (PI) fits into that context. The basic layout of this chapter is as follows. Section 1.1 briefly describes the English curriculum reforms in Taiwan by presenting some key relevant policies. Section 1.2 presents both teachers' and students' attitudes towards English grammar teaching and learning in Taiwan. The final section 1.3 then draws on these in order to give reasons why PI appears to deserve specific attention and how it is compatible with the current methods of English education in Taiwan.

### 1.1 The English curriculums reforms in Taiwan

The Taiwanese Ministry of Education (MOE henceforth) introduced English curriculum policies in the Nine-Year Integrated Curriculum, regulating that primary schools are required to provide students with English courses from the 5<sup>th</sup> grade (age 11) since 2001<sup>2</sup>. Subsequently, English education has been expanded on a larger scale by the announcement that compulsory English courses would be implemented from the 3<sup>rd</sup> grade, commencing from September 2005. In order to improve the feasibility of delivering the required levels of English education, the Taiwanese MOE developed and published guidelines for the implementation of English curriculums. The official objectives of the English curriculum and the guidelines to construct English teaching materials are provided in Appendix 1 (English translation provided). Due to the

---

<sup>1</sup> Note that PI can be applied to other second or foreign languages, not exclusively English. English was chosen here as a targeted language to examine this grammar pedagogy due to the fact that it is the most pervasive foreign language learnt in Taiwan - English education is compulsory and students are required to undertake it from the third grade in primary schools in Taiwan.

<sup>2</sup> Prior to the commencement of the Nine-Year Integrated Curriculum, English education formally started in schools from the 7<sup>th</sup> grade (age 13).

limitations of space, only the statements to construct English teaching materials which are relevant to the current study are translated and presented here. The English translation is as follows<sup>3</sup>:

When compiling teaching materials, every unit should be suitable to life's circumstances, integrate the main ideas and include sentence structures, useful words and phrases. Activities should be diverse, but must emphasise communication activities and cultivate a student's basic communicative ability. Every unit should include activities which are appropriate to the main topic, and should introduce vocabulary, phrases and sentence patterns. These should follow a step-by-step, easy-to-difficult model ... . A variety of lively and appropriate topics in the material should put the students in touch with a variety of language study experiences, which should promote their interest in the language and benefit their studies.

(Taiwanese MOE, 2006)

The statement reveals that teaching materials and activities should be constructed with the goal of cultivating students' ability in communication. This emphasis on the development of 'basic communicative ability' has led a number scholars and practitioners in Taiwan to believe that, among a variety of English teaching approaches, the communicative language teaching approach<sup>4</sup> (henceforth, CLT) should be the leading approach in the Taiwanese English reform scheme (Keng, 2008). Since CLT is

---

<sup>3</sup> The original text can be found by the following link: <http://teach.eje.edu.tw/9CC/3-2.php>

<sup>4</sup> According to Maley (1984, cited in Anderson, 1993, p. 471), the characteristics of the communicative approach are as follows: 1) focus on use and appropriateness rather than on language form; 2) fluency-focused rather than accuracy-focused activities; 3) communicative tasks are achieved via the language itself rather than simply exercises on the language; 4) emphasis on students' initiative and interaction rather than on teacher-centred direction; 5) sensitivity to learners' differences rather than a 'lockstep' approach; and 6) an awareness of variations in language use rather than simply attention to the language.

regarded as a basis for the English curriculum, English teachers are expected to offer lessons based on the CLT. The pupils in English classrooms are expected to be motivated by means of communicative activities which are student-centred, instead of by passively acquiring grammar knowledge from their teacher (see Appendix 1: basic concepts).

## **1.2 The belief and practice in grammar teaching and learning in Taiwan**

Given that the beliefs which teachers and students hold can influence what they do in the classroom (Hsu, 2007), any pedagogy introduced should take into account the needs and beliefs of teachers and students, and cultural differences. The following section presents the opinions and beliefs of Taiwanese teachers and students about grammar teaching and learning, as this is relevant to the justification for researching PI in Taiwanese schools.

### ***1.2.1 Taiwanese teachers' viewpoints about grammar teaching***

Lee (2005) investigated 159 Taiwanese primary school teachers' opinions about grammar instruction in classrooms using questionnaires. She concluded that the majority of teachers (82.4%) acknowledged the importance of grammar instruction to young learners, and that most of them put grammar lessons into practice in their classrooms. The English teachers also expressed the belief that primary school students at grade 5 and 6 are cognitively ready for grammar instruction, although grammar instruction is, generally speaking, comprised of complicated and abstract terminologies, structures and rules. However, Lee (2005) expressed her view about the counterproductive effect of instructing grammar in primary schools. She indicated that students might be demotivated by receiving grammar instruction at an early age.

Lai (2004) examined the grammar teaching beliefs of 199 Taiwanese high school English teachers through questionnaires and interviews. She found that most teachers held positive attitudes towards the inclusion of grammar instruction in the classroom on the condition that the main focus in the language classroom is on meaning and on exposure to the target language. Hsu (2007) investigated high school English teachers' beliefs about grammar teaching and their classroom practices by means of case study. He concluded that the teachers participating in his study tended to use the traditional grammar-teaching approach, in which the teacher dominated most of the discussion, and that the learners' first language (Chinese) was primarily used for the delivery and explanation of the instruction.

### ***1.2.2 Taiwanese students' viewpoints about grammar learning***

Lee (2005) investigated the viewpoints of 731 Taiwanese primary school students about learning grammar in English classrooms through questionnaires. She concluded that a high percentage of students (90.2%) favoured grammar instruction in the English classroom. Students considered that grammatical knowledge is beneficial for them to take examinations, and that it is also favourable for particular types of test such as translation and composition. Furthermore, due to the fact that learning grammar is emphasised in the English curriculum in junior high schools, some students were of the opinion that they should learn grammar at primary school in an attempt to prepare themselves for the language instruction at junior high school. On the other hand, the grounds given by those students who were against grammar instruction were as follows. First, learning grammar rules is discouraging, time-consuming, tedious and intimidating, which results in weakening the motivation to learn English. Second, communication could be impeded by grammar in the view of those who considered that fluency should take precedence over accuracy. Finally, some students reported that



grammar instruction produced no tangible benefits in helping them to use English in real-life settings. Lee also concluded that a discrepancy existed between the opinions of teachers and students about the ideal proportion of time distribution on grammar instruction. The students wanted more grammar learning (40% to 50% of class time) than the teachers (20% to 30%).

### **1.3 Justification for an investigation of PI in a Taiwanese primary school**

The investigation of PI (which is the main focus of the current study) in this context is justified on the following grounds. First, the introduction of PI corresponds to a great extent to the English curriculum policies of the Taiwanese MOE. For example, the policies state that when developing teaching materials, activities should be kept diverse by means of bringing in a variety of lively and appropriate topics, and should follow an easy-to-difficult route. The guidelines for the construction of PI activities advise that the design of PI activities must be aware of “keeping meaning in focus” by using diverse topics which are related to students’ real-life circumstances. Also, the guidelines for creating PI activities suggest that the activities should be moving from sentences to connected discourse, corresponding to the easy-to-difficult principle.

Second, the international research agenda relating to PI probably justifies investigating PI as one potential grammar teaching option for Taiwanese English teachers. According to Lai (2004), Taiwanese teachers believe that grammar instruction is beneficial for language learning if it is not at the expense of meaningfulness and communication. Lee (2005) reported that English teachers believe that teaching grammar could improve students’ accuracy in a language and lead them to regard grammar as a fundamental component in their classrooms. They do wonder, however, what kind of grammar instruction method is appropriate for teaching grammar to young learners. Although the

Taiwanese MOE has provided English curriculum guidelines, the guidelines appear not to be specific for elementary school instructors to follow, given that they are not clear on what grammar should be instructed, how much grammar should be introduced, and importantly how to construct a meaningful grammar activity (Lee, 2005, p. 75). Since PI is constructed out of integrating grammar into an input-based and meaning-based approach, it fits in with the basic constructs: the combination of form and meaning, and the integration of grammar into communicative tasks (VanPatten, 1993; Wong, 2004a).

Furthermore, to the author's best knowledge, only two studies have hitherto been carried out to investigate the effectiveness of PI in Taiwan (Wu, 2003; Xu, 2001). Wu (2003) concluded that PI is an effective technique for learning the English subjunctive mood compared to a traditional instruction (TI) (see Section 2.3.1.1), which is an output-based instruction. The learning gains were assessed by a reading comprehension task (the learners had to judge whether a statement was true or false according to whether the mood was subjunctive or indicative) and a written production task (a gap-fill test). Similar to Wu's (2003) study, Xu (2001) examined the instructional impact of PI compared to TI on learning the English Wh-question, specifically the two interrogative pronouns *who* and *what*. Learning gains were measured by a reading interpretation task (find a reasonable answer that matches a question) and a written production task (a scrambled task). Both studies found that PI could lead to a satisfactory improvement at the immediate post-test in the interpretation task and in the written production task. The desirable effect of these two PI studies appears to suggest that the introduction of PI is potentially applicable as it could provide an alternative for language teachers in Taiwan (Xu, 2001, p.75). However, only two studies have been conducted in the Taiwanese context, so there is a need for more studies investigating the effectiveness of PI on different structures, amongst different proficiencies and ages of

learners and comparing its effectiveness to other techniques.

Finally, the introduction of PI could cater for the needs of both Taiwanese teachers and students in English classrooms: teachers would like to carry out meaningful activities according to the curriculum guidelines provided by the MOE, while students want more grammar learning in order to pass examinations (Lee, 2005), because the higher education entrance examination in Taiwan is form-oriented (Huang, 2003). Although I was not able to find empirical evidence that Taiwanese learners are somehow grammatically deficient, nevertheless without mastery of the grammar, it may be difficult to fulfil Taiwanese teachers' and learners' preferences for teaching and learning English. In addition, the MOE curriculum stresses that communicative competence is the main aim of English teaching, and as grammatical competence is clearly part of this, it is relevant to investigate effective ways of teaching it. Following Green and Hecht (1992), grammatical competence is interpreted here as "the degree of accuracy achieved by learners when their attention is focused on form" (p. 169). Canale and Swain (1980) claimed that grammatical competence can serve as a catalyst for accuracy and fluency in second/foreign language learning and that it could be acquired in the context of meaningful communication. In sum, PI proponents provide guidelines for the design of PI activities which are in line with the guidelines produced by the MOE (i.e., to be meaning-based and communicative); at the same time, a grammatical focus is expected to be achieved during PI activities.

## **Chapter 2 The framework of Processing Instruction, its relevant theoretical backgrounds, and review of literature on Processing Instruction research**

### **Introduction**

Processing Instruction (PI henceforth) consists of pedagogical techniques which aim to help learners better process input in a second language (L2) or a foreign language (FL). It has been substantiated as an effective pedagogical package for learning grammar (see more detailed discussion in Section 2.3.1). Given that the current study aims to investigate the framework of PI (Processing Instruction), it is essential to review the related theories which underpin it. This chapter therefore primarily focuses on sketching out the basic framework of PI, the theoretical backgrounds germane to PI, and the literature review of PI-based studies. The first section will depict the framework of PI by presenting its nature and uniqueness. The subsequent section will make an attempt to address the theoretical backgrounds related to PI: Input Processing (IP) theory, Form-Meaning Connections (FMCs), and attention. The literature review of previous PI studies, setting out how the current study builds on some unverified issues in PI studies, is given in the third section. This chapter closes by posing research questions and hypotheses for the current study.

### **2.1 What is Processing Instruction?**

PI is an innovation based on Krashen's 'Input Hypothesis' (1985)<sup>5</sup>, and arises from the fact that most studies on Krashen's input hypothesis did not demonstrate the acquisition of accurate grammar (Sheen, 2005, cited in Sheen, 2007). Thus, PI is a pedagogical

---

<sup>5</sup> Krashen claimed that 'comprehensible input' alone is enough for language acquisition to take place and then the acquired knowledge can be used to produce that acquired language. According to Krashen, 'comprehensible input' is defined as input comprising  $i + 1$ , in which  $i$  is a learners' current language level and  $i + 1$  is a structure which advances slightly his/her current language level.

reaction to language comprehension situations, and it offers grammatical pedagogical techniques to help structure input of a second language or a foreign language. In order to provide a broad picture of what PI is, this section is divided into three sections. The first section dwells on the basic framework of PI by illustrating three components in PI. The second section aims to present the characteristics and uniqueness of PI and to elaborate on why PI is different from other grammatical instruction. The third section presents the guidelines for practitioners on how to create PI activities.

### ***2.1.1 The framework of PI***

#### *2.1.1.1 The components of PI*

According to VanPatten (1996, 2002a, 2004), the PI framework has three basic components: explicit grammar explanation, referential activities and affective activities. Referential and affective activities together in the framework of PI are often jointly termed ‘*structured input activities*’ (SIA) (referential *plus* affective activities) as these activities have been *structured* purposefully, with the aim of reducing learners’ ineffective input processing (VanPatten, 2004). It has been suggested that PI activities should begin with referential activities and then be followed by affective activities (VanPatten 1996; Wong, 2004a). The three components of PI are simply described as follows:

#### *a. Component one: explicit grammar explanation*

Like other explicit grammar instruction, the provision of explicit grammar explanation seeks to give language learners a brief outline of the properties of a specific grammatical feature. However, the explicit grammar explanation provided in PI also aims to inform learners of the specific faulty processing strategies that they may employ, based on the insights of Input Processing theory (see Section 2.2.1). The faulty

processing strategies could lead to a detrimental impact on the incipient development of learners' Form-Meaning connections (FMCs) during language comprehension. Thus, the presentation of explicit grammar explanation precedes both referential and affective activities. The explicit information with respect to what learners are thinking of during processing input and, why they make such errors, is given to the learners. The presentation of the explicit grammar explanation in PI is displayed in general terms in the following example, extracted from Marsden's (2006, p.560) study on the learning of the past tense in French. Added comments are underlined and indicated using [ ].

[To explain the grammar rule in order to make an initial FMC]:

“To talk about what somebody else did in the past, we usually add ‘a’ before the main verb. For example, *Il mange* (present tense) → *Il a mangé* (past tense).”

[To explain what errors learners tend to make]:

“Learners of French seem to find this hard – they miss out the ‘a’ and say things like *il mangé* or *il mange* - but these don't tell us they are talking about the past!”

[To explain why learners tend to make such errors]:

- Perhaps learners don't notice the ‘a’ because words like ‘*le weekend dernier*’ tell us that we are talking about the past.
- Learners may not notice the ‘a’ because the word ‘*il*’ or ‘*Paul*’ has already told us who we are talking about!
- Sometimes it can sound like there is “a” but the verb is in the present. For example, *il achète* (compared to *il a acheté* in the past).”

*b. Component two: the referential activities*

The second component of PI is the implementation of referential activities. In order to complete the referential activities successfully, learners are required to attend to the targeted grammatical form and to interpret its meaning. Learners are expected to rely on interpreting the meaning and function of the specific targeted feature so that the development of the FMC happens. In addition, referential activities are purposefully constructed in particular ways in that there are right and wrong options provided for learners to exercise. Subsequently, feedback is given to learners to check the correctness of their responses. The formulation of referential activity may appear to share characteristics with the ‘Garden Path’ (Tomasello & Herron, 1988, 1989) technique for language learning, although referential activity is input-based, whereas garden path activity is output-based. During instructional phrases, both activities attempt to place learners in a situation in which they are likely to make errors, leading to a failure-driven learning process (Carroll, 1999). Following that, feedback is provided such as ‘correct’ or ‘incorrect’ (Benati, 2005; Sanz & Morgan-Short, 2004), so that learners can check whether or not their responses are correct in a referential activity. PI’s essential idea of achieving FMCs also corresponds to the task-essentialness<sup>6</sup> requirements for constructing the grammar tasks (Loschky & Bley-Vroman, 1993), though VanPatten’s conception of FMCs is that they happen during input processing rather than output processing. Examples of referential activities in PI are displayed below, extracted from VanPatten & Cadierno’s (1993a, p.231) study of learning Spanish direct pronouns<sup>7</sup> (English translations provided) and Benati’s (2005) study of learning the English ‘-ed’ feature.

---

<sup>6</sup> According to Loschky & Bley-Vroman (1993, p. 132), task-essentialness is the necessity of designing grammar tasks in which learners are required to use the specific grammatical feature in order to perform the tasks successfully.

<sup>7</sup> Spanish has flexible word order such as SVO, SOV, OVS, OV. In a Spanish sentence such as “*Me llaman los padres*”, learners may have problems in interpreting which is the subject or object (i.e. My parents call me or I call my parents).

Referential activity 1:

Listen as your instructor reads a sentence. Select the best interpretation from the English renderings.

1. a. My parents call me.

b. I call my parents.

(Instructor reads aloud: *Me llaman los padres.*)

[and so on]

(VanPatten & Cadierno, 1993a)

Referential activity 2:

You will hear 10 sentences and you need to determine whether the action is taking place now (present) or has already taken place (past).

1) Student hears: I listen to music

a) Present      b) Past

2) Student hears: I walked to the park

a) Present    b) Past

(Benati, 2005)

*c. Component three: the affective activities*

The final component is the affective activities in which learners merely have to carry out the tasks in meaningfully-oriented contexts containing the targeted linguistic feature. Learners are required to respond to affective activities by expressing their own belief, opinions or feelings related to their own personal experience. The role of affective activities is, to date, less clearly articulated in the PI and IP literature. One role of the affective activity, claimed by the PI proponents, is to maintain PI in line with a



prominent tenet of communicative language teaching: a focus on the learner (VanPatten, 1993, p.439; Wong, 2004a, p.45). As Wong (2004a) stated, “by requiring learners to express an opinion or some other kind of personal response, we can keep instruction in line with an important tenet of communicative language teaching: *a focus on the learner*” (p.44-45).

Another claimed role of affective activities is to reinforce FMCs by offering learners more opportunities to hear and see the target feature appearing in a meaningful context (Wong, 2004a). Wong (2004a) stated that “the purpose of affective activities is to reinforce those connections by providing them with more opportunities to see or hear the form used in a meaningful context” (p. 44). Also, Marsden (2006, p.514-515) implied one potential purpose of affective activities by citing Schmidt’s (2001) viewpoints. Schmidt argued that once the initial mental representation has been achieved, the implicit internalisation of a specific form in a language-developing system is likely to be attained through subsequently being exposed to copious amounts of examples embedding the form. In this sense, it appears that one of the functions of affective activities could be to reinforce the FMCs established by the referential activities, and then to internalise the form into the language developing system (Wong, 2004a).

The other potential role of affective activities is to promote lexical learning (Marsden, 2004, 2006). Taking the English ‘-ed’ feature as an example, it is possible that a learner in referential activities can scan whether there is an ‘-ed’ attached to the end of a verb and then s/he can complete referential activities without trying to understand the meaning of a sentence. On the other hand, a learner in affective activities is required to show her/his own opinions (such as agree or disagree, interesting or boring) to

accomplish the task, so s/he needs to comprehend the lexical items. In affective activities, it does not matter whether or not s/he notices or knows what the ‘-ed’ means, as the verb stem itself is critical to the task rather than the ‘-ed’ feature.

Examples of affective activities in PI are given below, extracted from VanPatten & Cadierno’s (1993a, p.232) and Benati’s (2005) studies.

Affective activity 1:

Indicate whether or not each statement about your parents applies to you. Then share your responses with a classmate.

*Sí, se me aplica. No, no se me aplica.*

\_\_\_\_\_      \_\_\_\_\_      1. *Los llamo con frecuencia por teléfono.* (“I call them on the phone frequently.”)

\_\_\_\_\_      \_\_\_\_\_      2. *Los visito los fines de semana.* (“I visit them on the weekends.”)

[and so on]

(VanPatten & Cadierno, 1993a)

Affective activity 2:

Listen to the instructor making a series of statements and indicate whether you did the same thing at the weekend:

1) Student hears: I played sport

a) Me too      b) I did not

2) Student hears: I visited my friend

a) Me too      b) I did not

(Benati, 2005)

### *2.1.1.2 The differences between these two structured input activities*

Importantly, studies to date have used these two input activities together, labelling them ‘structured input activities’ (e.g. Benati, 2001, 2004a). However, there are some critical differences between them.

One of the prominent differences between referential and affective activities is that in referential activities the targeted form is often juxtaposed with other similar linguistic feature(s) (for further discussion, see Benati, 2004a, p.211; Marsden, 2006, p.519; VanPatten, 2002a, p. 767). For example, in a task on learning the English regular past tense, learners would be exposed to the past-tense verbs (the targeted feature) and present-tense verbs (a similar feature) in referential activities, whereas learners engaging in affective activities would simply be exposed to sentences containing past-tense verbs. Though not acknowledged elsewhere in prior PI studies or IP literature (discussed in section 2.2.1), it is likely that the juxtaposition of contrasting objects may be conducive to the development of an FMC because the properties of the target form may be ‘highlighted’ by comparison with other features.

The other critical divergence of these two structured input activities lies in the feedback provided to learners. In order to ensure that the learners are correctly making FMCs, feedback offered in referential activities explicitly indicates the correctness of their responses to the task, such as your answer is ‘correct’ or ‘wrong’. In addition, as referential activities place learners in the situation of making errors and experiencing a failure-driven process, so the context of referential activity provides learners with an

opportunity to be involved in both positive and negative evidence<sup>8</sup>. On the other hand, the completion of affective activities does not require a clear-cut answer, given that learners are required to express their opinion about the sentence in which the targeted feature is embedded (agree/disagree; interesting/not interesting; same for me/ not same for me). In this sense, affective activities are always carried out in a format such as input flood (i.e. learners are simply exposed to the input with the targeted feature), and only positive evidence is provided to the learners. Due to the different nature of these two types of structured input, presumably, the contrasting pairs and the feedback provided in referential activities may encourage ‘hypothesis testing’ of the targeted feature (Bley-Vroman, 1986; Tomasello & Herron, 1989) so that an FMC of the targeted feature is initiated.

Another difference between referential and affective activities is concerned with the meaningful-bearing basis in the context of a task. PI proponents have claimed that PI is in accordance with Focus-on-Form studies (VanPatten, 2000; Wong, 2004a)(see Section 2.1.2.2), in which learners’ attention focuses on a given linguistic form in the course of undertaking a communicative and meaning-bearing task. However, undertaking referential activities is likely to be less communicative and meaningful than carrying out an affective activity. Once a learner grasps the targeted linguistic form, he/she can successfully accomplish the activity by observing the occurrence of the targeted form without any understanding of the meaning of the whole utterance or context. Furthermore, learners are perhaps more likely to get bored in the course of a referential activity than in an affective activity, given that referential activities always involve deciding on either right or wrong options, and the feedback provided to their responses

---

<sup>8</sup> According to White (1991), negative evidence refers to “information about ungrammaticality” (p. 134). Positive evidence refers to the utterances in the input, similar to those provided for first language acquisition, like input flood, in which learners’ “incorrect hypotheses can be disconfirmed” (p.133-134).

is rather monotonous in that it always indicates correctness.

In contrast to referential activities, learners involved in affective activities can show their own opinions and the feedback provided could be varied. So it appears that the nature of affective activities is more communicative and meaning-oriented than that of referential activities. Nevertheless, Marsden (2007) indicated that “different activities are likely to have slightly different objectives, each entailing slightly different advantages and disadvantages” (p.575). Although it has not been empirically investigated by previous PI studies, it is possible that learners can complete affective activities without focussing their attention on the targeted feature. In this scenario, it is not obvious that affective activities further facilitate learning of a grammatical form. However, the nature of affective activities may be conducive to vocabulary learning (Marsden, 2006).

To sum up, the above discussion suggests that these two types of PI activity might have different pedagogical values. However, no study has yet isolated these two PI activities to test them. As VanPatten (2002a) has pointed out, although all the published guidelines for PI activities suggest using both of these structured input activities, the role of each in PI is worth investigating (p.784). Also, DeKeyser *et al.* (2002) provided a similar perspective in encouraging more studies to explore further the contribution and attribution of a variety of processing activities in terms of the learning of various kinds of features in various languages (p.820).

## ***2.1.2 The Nature and Uniqueness of PI***

### ***2.1.2.1 An input-based approach***

Proponents of PI believe that acquisition of an underlying grammar is input dependent

(VanPatten, Williams, & Rott, 2004). One of the characteristics of PI is that during the instructional phases learners are only engaged in input-based practices (i.e. learners only receive reading and listening practices) but at no time are learners engaged in output-based practices (i.e. no writing and speaking practices are given). Out of consideration for individual differences, the input practices given to learners are, according to PI guidelines (see section 2.1.3), required to be in both the written and aural modes. The input materials in PI, unlike the pedagogical implications of Krashen's 'Input Hypothesis', are delicately structured and manipulated. The concept of prioritising input rather than output in PI corresponds to Loschky & Bley-Vroman's (1993) perspective that comprehension should take precedence over production tasks. Loschky & Bley-Vroman maintained that "comprehension tasks are particularly well suited to hypothesis formation and to restructuring" (p. 143) of a new linguistic feature. Once a learner's language competence includes that linguistic feature, he/she will need production tasks to automatise it. The emphasis on input-only practice leading to gains in language *production*, has given rise to a number of empirical studies on the comparability of PI and other output-based instructions (see the discussion in 2.3.1.1) and this has, to date, been the main focus of PI-based research.

#### *2.1.2.2 A focus-on-form approach (FonF)*

According to Long & Robinson (1998, p.16), the focus-on-form approach entails a concomitant focus on meaning and a focus on form. In other words, learners' attention is directed to a specific linguistic feature in the input in the course of undertaking a communicative task. PI proponents have claimed that PI is in line with focus-on-form studies. As VanPatten (2000) has stated, "PI is a focus on form that serves as a supplement to existing communicative and acquisition-oriented approaches" (p.52). In addition, VanPatten (1996) and Wong (2004a) advised putting 'meaning' in the centre

when designing PI activities (see Section 2.1.3 regarding the guidelines for PI activities). In this sense, although PI is a grammar instruction, its activities, in principle, are supposed to be developed on the basis of meaningful-orientation.

However, the claim that PI is in line with focus-on-form studies may not be true in one of the PI activities (the referential activity). This issue is addressed in the current study (see Section 2.1.1.2), due to the possibility that the completion of referential activities does not necessarily call for learners' comprehension of an utterance.

#### *2.1.2.3 Derivation from Input Processing (IP) theory to achieve better Form-Meaning Connection (FMC)<sup>9</sup>*

VanPatten (2002a) argued that PI should not be regarded as the same as other types of input-based or focus-on-form instructions. What makes PI unique compared to other input-based or focus-on-form approaches is that PI is fundamentally informed by the theoretical model known as IP. The design of referential activities requires prior identification of the processing strategies as described in the IP model, which prevent learners from processing a specific feature or structure successfully. Once the ineffective input processing strategy is identified, a better Form-Meaning Connection is possible. According to VanPatten (2000, p.49), PI is a type of grammar instruction with the following three basic characteristics:

1. Learners are given information about a linguistic structure or form.
2. Learners are informed about a particular IP strategy that may negatively affect their picking up of the form/structure during activities.
3. Learners are pushed to process the form/structure during activities with structured input – input that is manipulated in particular ways so that

---

<sup>9</sup> The detailed presentation of IP and FMCs is addressed in sections 2.2.1 and 2.2.2 respectively.

learners become dependent on form and structure to get meaning and/or to privilege the form/structure in the input so that learners have a better chance of attending to it (i.e., learners are pulled away from their natural processing tendencies towards more optimal tendencies).

Since PI is derived from the insights of IP theory concerning L2 learners' psycholinguistic mechanisms, the design of PI instructional materials has to take learners' input processing into account. Other input-based or focus-on-form instructions that do not consider learners' input processing and make an effort to assist in their FMCs would not be considered to be PI.

### ***2.1.3 The guidelines for the structured input activities***

PI proponents have drawn up explicit guidelines on construction of the structured input activities (VanPatten, 1996; Wong, 2004a). VanPatten (1996, p. 67) noted that the application of these guidelines should be flexible. In addition, VanPatten has acknowledged that only guidelines b, e, and f are informed by the theory of IP, and the other remaining guidelines are formulated out of *practical and experiential* considerations. Note that these guidelines are presented here mainly because they informed the design of the PI activities in this study to maintain parity with other studies wherever possible. This study does *not* aim to test the validity of the guidelines.

#### ***a. Teach only one thing at a time***

VanPatten advocates breaking down the paradigms and rules into small parts. Although VanPatten has not discussed on what theoretical grounds he lay down this guideline, it sounds sensible if relating this guideline to the limited attentional resources, given that it can reduce learners' working memory load (Gathercole, 2008). Teaching more than



one thing may result in attentional resources being overloaded. However, although VanPatten has not mentioned it, it is sometimes unlikely to act on this principle when designing referential activities as its nature requires that the task incorporates contrasting pairs. The contrasting pairs could be some linguistic feature(s) either learners have learnt or not learnt when learning the targeted feature. Under this circumstance, learners may have to be instructed in more than one thing at a time.

*b. Keep meaning in focus*

Given that PI proponents have claimed that PI is in line with focus-on-form instruction (VanPatten, 2000; Wong, 2004a), the design of the structured input activities should be meaningful and communicative, not just only noticed solely the targeted form. As VanPatten (1996) states, “If meaning is absent or if learners do not have to pay attention to meaning to complete the activity, then there is no enhancement of input processing (p. 68).” In this sense, in the formulation of structured input activities, the notion of helping learners to connect a specific form to its meaning (i.e., the FMC) should be always kept in mind. VanPatten (2000) states when depicting this guideline that “all structured input activities include (1) the meaning of the form has to be processed or (2) the propositional meaning of the sentence and the form have to be processed” (p. 51). VanPatten argues that having learners circle verb stems in a passage is not regarded as a structured input activity, given that the meaning is not attended to. However, VanPatten has not discussed this guideline further. This guideline may not always be achievable when learners undertake referential activities, because the propositional meaning of the sentences may be neglected once the learners pick up targeted rule (see Section 2.1.1.2).

*c. Learners must do something with the input*

The structured input activities can not be counted as being effective if the learners are not paying attention to them. Thus, VanPatten suggests that the design of structured input activities should invite learners' active participation. This guideline could be fulfilled by creating various topics or contexts in which learners respond to the input. For example, requiring learners to show their agreement or disagreement, to fill in a survey, to choose from various options and so on could help learners to attend to the meaning of the input; whether learners are paying attention to the activities can be observed from their responses.

*d. Use both oral and written input*

Though VanPatten has not provided any empirical evidence to support this guideline about processing in different modalities, he suggests that a combination of aural and written input in structured input activities would produce a favourable impact. VanPatten indicates that this guideline is given under consideration of individual differences, since learners may prefer to 'see' or 'hear' the language to a different degree.

*e. Move from sentences to connected discourse*

This guideline is based on the competition between form and meaning during language input processing. VanPatten (1990) found that learners experience difficulty in attending to form and content simultaneously when processing input, so he hypothesised that attending to both form and content at the same time would result in a cognitive load. In addition, VanPatten proposed that the meaning of the content would take priority over the form during input processing. In order to release more attentional resources to attend to form, VanPatten recommends carrying out structured input activities firstly at sentence level to establish an initial FMC, and then move to discourse level. He points

out that starting an activity with the connected discourse level straight away may prevent learners from developing FMCs due to their limited attentional capacity.

*f. Keep the psycholinguistic processing strategies in mind*

This guideline is explicitly relevant to IP theory as it suggests that developing a structured input activity should always bear in mind what inefficient strategies learners may rely on (i.e. the IP principles). Therefore, the prerequisite for devising structured input activities in PI is to identify learners' processing strategies for a specific linguistic feature, and then design structured input activities to keep learners from using these inefficient processing strategies.

## **2.2 Relevant theories underpinning PI**

### ***2.2.1 Input Processing (IP) Theory***

This section sets out to elucidate briefly what IP theory is. Some key notions central to IP will be clarified. Then a set of principles of IP and the empirical evidence supporting them will be presented. Finally, some challenges and unclear issues of IP will be addressed.

#### *2.2.1.1 What is Input Processing Theory?*

VanPatten (1993, 1996, 2002a, 2002b and elsewhere) argued that the process of SLA could be briefly explained by Figure 2.1. VanPatten's Input Processing theory (IP henceforth) mainly stresses the process of how input<sup>10</sup> converts to intake<sup>11</sup> based on learners' psycholinguistic perspectives. Thus, IP theory works on how learners make *initial* FMCs when encountering a new linguistic form during input processing (i.e. the

---

<sup>10</sup> Input refers to the 'available target language' provided for language learners (Corder, 1967).

<sup>11</sup> According to VanPatten (2002a), intake is "that subset of filtered input that the learner actually processes and holds in working memory during on-line comprehension" (p. 761).

first arrow, Process 1 in Figure 2.1) in terms of acquiring *grammar* rather than other aspects of language. It is emphasised here that a detailed description of the second and third arrows is beyond the purpose of this study in that IP theory simply attempts to address the first arrow ‘input processes’ regarding how learners decode the input.

**Input → intake → Developing System → output**

**1            2            3**

1 = input processing

2 = accommodation, restructuring

3 = access

*Figure 2.1* The processing of second language acquisition

In addition, IP theory is developed on the basis of FMCs and a cognitive attention theory, which will be respectively discussed later in Sections 2.2.2 & 2.2.3. In brief, limited-capacity attention results in L2 learners’ being unable to process all incoming input; L2 learners have to ‘select’ what to process during input process. IP theory proposes that L2 learners tend to comprehend the ‘meaning’ of the input before they process the ‘form’. It is not until L2 learners process meaning without any cost to attention that they may have spare attentional resources to process the form.

To sum up, IP theory addresses questions about what factors affect the allocation of attentional resources if the input processing demands for attentional resources exceed supply, and why some linguistic features are favoured over others (i.e. why some FMCs are made before others). Thus, the set of IP principles is derived for making predictions about how learners utilise relevant strategies and mechanisms to make initial FMCs during input processing. It is noted that both VanPatten’s IP theory and Krashen’s Input

Hypothesis highlight the crucial role of input in acquiring a language. However, Krashen's Input Hypothesis does not explain why one linguistic form is acquired earlier than others. On the other hand, VanPatten's IP theory attempts to illustrate the actual processes involved in acquiring linguistic features and to make predictions about why some forms are processed earlier than others.

#### *2.2.1.2 Key notions of IP theory*

##### *a) Process a form = detect a form; process a form ≠ notice a form*

The term 'process' used in IP theory refers to the establishment of a connection between a given form and its meaning during the act of comprehension (VanPatten, 2002b, p.242). Although VanPatten and Schmidt (1990, 2001) appear to take a similar position, i.e. that attention is crucial for language learning, the cognitive process of 'processing' or 'attending' to a form in IP partially departs from Schmidt's conception of 'noticing'<sup>12</sup> (VanPatten, 2007, p.125 and elsewhere). Schmidt's 'noticing' refers to some kind of registration of a given form in working memory without necessarily connecting to its meaning or function. A form may get noticed (i.e. it is held somehow in the working memory), but not get processed (i.e. the connection between the form and its meaning fails). Accordingly, VanPatten emphasises that the term 'process' or 'attend' to a form used in IP corresponds more closely to the notion of 'detection'<sup>13</sup>, described by Tomlin & Villa (1994) rather than Schmidt's 'noticing'.

##### *b) Intake ≠ acquisition*

It is worth noting that "not all input becomes intake, not all intake matches the input, and not all intake is delivered to the developing system" (Lee & Benati, 2007, p.2).

---

<sup>12</sup> Schmidt's construct of 'noticing' is discussed in Section 2.2.3.2.

<sup>13</sup> Tomlin & Villa's construct of 'detection' is presented in Section 2.2.3.2.

Although there should be some level of intake occurring for acquisition to take place, VanPatten (2000, p.48) stresses that ‘intake’ should not be equated to ‘acquisition’, and ‘no intake, no acquisition’. Put another way, input processing may cause some linguistic data to be held in the working memory, but it is not a corollary that those data will be internalised in a learner’s language system. When data is held in the working memory, it may fade within seconds. However, only the data registering in the working memory can possibly be further processed (i.e. accommodation and restructuring), and get into learners’ underlying language system, and then be accessed for language production.

*c) Communicative value*

VanPatten (1996, 2000, 2002a and elsewhere) posited the concept of ‘communicative value’ to address why some forms get processed earlier than others in his IP model. According to VanPatten (2002a), communicative value refers to “the meaning that a form contributes to overall sentence meaning” (p.759). Communicative value is based on two features: inherent semantic value and redundancy. Inherent semantic value refers to the referential meaning of a grammatical form; redundancy refers to grammatical form which encodes a meaning (or has a function) which is also coded in another feature (e.g. pastness is coded by the ‘-ed’ feature and temporal adverbials). VanPatten claimed that forms with semantic value always take precedence over redundancy. The higher the communicative value a given form has, the more likely the form is to be processed (i.e. form-meaning connection is to be attained). The degree of a grammatical form’s communicative value is based on: +/- inherent semantic value and +/- redundancy. As a result, communicative value could be classified into the following four categories from high to low: [+semantic value and - redundancy] (for example, the English progressive ‘-ing’), [+semantic value and + redundancy] (for example, the English past tense ‘-ed’), [-semantic value and - redundancy] (for example, adjective

concordance in Romance languages), and finally [-semantic value and + redundancy] (for example, complementisers such as ‘that’) (VanPatten, 2002a, p.759).

### *2.2.1.3 The IP principles and their empirical evidence*

VanPatten’s IP theory takes the position that the capacity of attention within the working memory is limited, so learners have to select what to attend to as they cannot intake all of the language input. Consequently, when attention is allocated to a specific stimulus, other stimuli might be overlooked or merely partially processed. Thus, the predictions of IP mainly concern why learners make some FMCs but not others during the comprehension of the input strings, and under what conditions learners may succeed in establishing FMCs where they have failed before.

VanPatten’s IP theory overall consists of two principles (Principle 1: ‘the Primacy of Meaning Principle’; Principle 2: ‘the First Noun Principle’). In this section, only the Primacy of Meaning Principle will be presented and discussed further because it is relevant to this study, whereas the First Noun Principle is not. For the sake of completeness, the First Noun Principle and its corollaries pertinent to syntactic parsing are attached in Appendix 2. This section will focus on presenting Principle 1 and its corollaries<sup>14</sup>, based on what VanPatten himself has claimed (1996, 2002a, 2004, 2007). The empirical evidence to support this principle is also given, based on what VanPatten himself has drawn on (1996, 2002a, 2004, 2007). Within this section I also provide some suggestions for additional empirical support for these principles which has not so far been put forward in published literature. A critique of these principles is presented later in this section.

---

<sup>14</sup> The IP principles listed here are based on VanPatten’s claims (2004). Some principles (such as Principle *b*, *c* and *d*) have been partially revised in his published book (2007), and these will be presented in the footnotes along with his original proposition.

**Principle 1: *The ‘primacy of meaning’ principle.***

Learners process input for meaning before they process it for form<sup>15</sup>.

The ‘primacy of meaning’ principle is largely based on the notion from cognitive psychology concerning the limited attentional resource. When the processing of one thing consumes a great large deal of attentional resource, the processing of others inevitably is deprived (Broadbent, 1958; Just & Carpenter, 1992). Thus, learners are forced to select what to attend to during input processing. Since the ultimate goal of learning a language is to communicate with others, the ‘primacy of meaning’ principle predicts that when learners are engaged in processing language input, their attention tends to be allocated to understanding the meaning first. As a result, grammatical linguistic forms will be overlooked during the processing of language input.

VanPatten claimed that his perspective on the competition for meaning and form during the processing of language is in line with that of other researchers in both the first (Peters, 1985) and second language (Klein, 1986; Sharwood Smith, 1986) (see more discussion in VanPatten, 1996, chapter 2). VanPatten argued that Peters’ proposal (1985) supported the notion of ‘meaning before form’ because it stated that L1 children tend to “pay attention to utterances that have a readily identifiable meaning and extract and remember sound sequences that have a clear connection to a clear context” (p.1034). In support of claims from the perspective of L2, VanPatten cited Sharwood Smith’s study (1986) and argued that there was a great deal of evidence suggesting that L2 learners were capable of understanding what they heard and read, but were not acquiring the

---

<sup>15</sup> VanPatten (1996) defined form as “surface features of language: verbal inflections, nominal inflections, particles, functors, and so forth” (p.18) to account for the IP principles.



language features which appeared in the input.

**Principle a. The ‘primacy of content words’ principle.**

Learners process content words in the input before anything else.

This principle makes a prediction that learners seek to grasp content words (that is, lexical forms) rather than non-content words during the input process. For example, if a learner is hearing or reading ‘The girl is crying’, he/she tends first of all to seek out content words (in this case ‘girl’ and ‘cry’) in the utterance, instead of non-content words (here the ‘-ing’). VanPatten cited both L1 (Peters, 1985; Radford, 1990) and L2 (Klein, 1986; Mangubhai, 1991) studies to support this principle. Peters’s data showed that children attended to isolated words and unanalysed chunks of language in the input, and then used them in their oral production. Radford noted that children’s language learning started from producing some elementary vocabulary without any grammatical properties. Based on the propositions of Peter and Radford, VanPatten (1996) reported that children’s L1 acquisition commences with “using single words or whole unanalysed chunks of language (which they treat as content words) in the early stage and then combine these to form utterances” (p.18).

The tendency of language learners to process content words rather than grammatical form has also been discussed in early L2 research. Klein (1986) concluded that adult learners of German tended to use content words as opposed to grammatical features in a sentence repetition task. Mangubhai (1991) found that the processing strategy which adult learners brought into full play in order to get meaning from the input was to resort to the lexical words. Also, VanPatten (1990) conducted an empirical study in order to determine the competition between form and meaning for learners’ limited attentional

resources when processing input. His empirical study indicated that beginning learners had a tendency towards processing meaning before form in the aural mode. VanPatten's claim was also substantiated by Wong's (2001) conceptually replicated study; however, the tendency towards 'meaning before form' was not found in the written mode. In addition, Dulay & Burt's (1978) claim, though not cited by VanPatten, lent support to this principle. Dulay & Burt indicated that "the late acquisition of grammatical morphemes compared to content words has become an established fact for second language learners" (p.75).

**Principle b. The 'lexical preference' principle.**

Learners will process lexical items for meaning before grammatical forms when both encode the same semantic information.

Note that this principle is central to the current study. It is common for the semantic notion to be expressed in *both* lexical words and the linguistic form. Principle *b* predicts that when processing language input, learners lean towards paying attention to the lexical words as opposed to the grammatical form in order to acquire meaning, given that lexical words are the easiest way to comprehend the meaning of incoming input. For example, the English regular past tense can be indicated by both the verbal inflection and the temporal adverbials. When learners encounter an utterance containing the meaning of pastness, they are prone to grasp it from the lexical words (such as 'yesterday' or 'last night') rather than from the grammatical form '-ed'. In this sense, the acquisition of the grammatical '-ed' inflection may be delayed due to interference from the temporal adverbials.

Furthermore, Principle *b* has been slightly revised by VanPatten (2007). In this revised

version of Principle *b*, VanPatten (2007) stated “If grammatical forms express a meaning that can also be encoded lexically (i.e., the grammatical marker is redundant), then learners will not initially process those grammatical forms until they have lexical forms to which they can match them” (p. 118). Put another way, a grammatical form will not be processed until the corresponding lexical form has been incorporated into a learner’s developing linguistic system. For example, the English plural marker ‘s’ will not be processed until learners have processed lexical words like ‘two’, ‘three’, ‘many’, and so on. To sum up, this principle predicts, in VanPatten’s words (2007), that “as long as comprehension remains effortful, learners will continue to focus on the processing of lexical items to the detriment of grammatical markers, given that lexical items maximize the extraction of meaning, at least from the learner’s point of view. Grammatical markers will be processed later, if at all”(p.119).

The empirical evidence cited by VanPatten to support this principle can be summarised as follows. Bardovi-Harlig (1992) reported that learners had a preference for marking the past tense by lexical items rather than grammatical features on the acquisition of tense. Evidence consistent with the lexical preference principle was also found by Pica (1985) on the acquisition of marking plurality (such as in ‘two dog’) and the third person singular (such as in ‘he sleep’). Additionally, some unpublished experimental studies (Cadierno *et al.*, 1991; Glass, 1994; Musumeci, 1989) were cited by VanPatten (1996, pp.22-23) to support this principle. Cadierno *et al.* (1991, cited in VanPatten, 1996) concluded that providing aural input at discourse level with temporal adverbials was more beneficial for L2 learners of Spanish to decide the temporal reference of an event than input without temporal adverbials present. Glass (1994, cited in VanPatten, 1996) asked learners of Spanish to reflect on how they decided the temporal reference of an event after listening to a passage. Glass reviewed participants’ introspective

responses and found that learners tend to assign tense by means of lexical information instead of verb inflections. Musumeci (1989, cited in VanPatten, 1996) reported that the presence or absence of a temporal adverbial is a key point in assisting learners to assign the correct tense. VanPatten claimed that the results of these studies suggest that L2 learners tend to rely on temporal adverbials rather than verb inflections to mark tense, which supports the Lexical Preference principle in IP. In addition, VanPatten (2007, p.126) mentioned the eye-tracking research to corroborate this principle. He concluded that the eye-pupil movements of native speakers and non-native speakers are greatly at odds. Native speakers are more likely to fix visually on the verb inflections, and non-native speakers fix more on temporal adverbials when comprehending the English regular past tense.

Additional supporting evidence for these principles can also be found in Lee *et al.* (1997) and Lee (1998), though they were not cited by VanPatten. Lee *et al.* (1997) reported the comparative effect of providing two types of discourse-level input (the presence or absence of temporal adverbials) in the aural mode for adult L2 learners of Spanish. Learning gains were examined on both free recall and tense-identification tasks. The results showed that learners who listened to the passage with adverbs present outperformed those who listened to the passage without adverbs. Lee *et al.* concluded that learners had a tendency to align attention on lexical cues to reconstruct the propositional content. Lee *et al.* claimed that although grammatical cues also received some attention, this was not sufficiently utilised to re-construct the propositional content. Likewise, Lee (1998) showed evidence to sustain the finding that lexical cues in the input string are more important than morphological cues for learners' comprehension during input processing.

Furthermore, the assertions of N. Ellis' (2006, 2007) support the Lexical Preference

principle to some extent, though he did not make any reference to this IP principle. N. Ellis indicated that the degrees to which L1 and L2 learners rely on lexical words such as temporal adverbials are different. The phenomenon of learners preferring a temporal adverbial to a grammatical form is much more commonly observed in second than first language acquisition. Ellis (2007) reasoned that, contrary to L1 learners, who acquire the meanings of temporal adverbials quite late in their language development (Dale & Fenson, 1996), L2 learners already know the functions of adverbials in utterances from their L1 learning experience, so they instinctively realise that they can acquire the temporal meaning readily from the adverbials when processing an L2. Also, these adverbials are “both salient and reliable in their communicative functions while tense markers are neither” (N. Ellis, 2007, p.83). However, prior knowledge of these adverbials could block L2 learners’ subsequent acquisition of other cues (including specifically the grammatical form) (N. Ellis, 2006, p.179, 2008<sup>16</sup>).

### **Principles c & d**

As the principles *c* and *d* are largely based on the notion of communicative value (see the discussion in Section 2.2.1.2), these principles are presented and discussed together. In brief, the crucial concept of communicative value is the nature of a linguistic feature, leading to the degree to which it gets processed by learners.

According to VanPatten, the higher communicative value a linguistic feature has, the more likelihood there is that it will be processed in the input. Note that principles *c* & *d* are not critical for this study. They are presented here for the completeness of principle 1 (i.e. the “primacy of meaning” principle).

### **Principle c. The ‘preference for non-redundancy’ principle<sup>17</sup>.**

---

<sup>16</sup> Ellis (2008) suggests this phenomenon in SLA is due to the attentional blocking of inflectional cues due to earlier entrenchment of reliance on lexical cues.

<sup>17</sup> According to VanPatten (2007), Principle *c* states that “Learners are more likely to process non-

Learners are more likely to process non-redundant meaningful grammatical forms before they process redundant meaningful forms.

**Principle d. The ‘meaning-before-nonmeaning’ principle<sup>18</sup>.**

Learners are more likely to process meaningful grammatical forms before nonmeaningful forms irrespective of redundancy.

Principles *c* and *d* account for how linguistic features get processed, particularly in relation to the notion of communicative value. It is noted that the term ‘meaningful’ used in these two principles simply refers to the semantic value that a linguistic feature conveys, not the overall meaning expressed in an utterance. Some linguistic features are meaningful (such as the English ‘-ed’ or ‘-ing’ forms) and some are not (such as the English article ‘the’).

Principle *c* suggests that when learners are exposed to language utterances, they are inclined to process those meaningful features which do not share semantic value with other elements in the utterances (in other words non-redundancy). On the other hand, meaningful features which do share semantic value with other expressions in the utterance tend to get processed later. VanPatten illustrates this principle with the ‘-ing’ and ‘-ed’ forms in English. The progressive expression of ‘-ing’ form is a meaningful and less redundant linguistic feature (i.e., it has higher communicative value) because, more often than not, no other information in an utterance co-occurs to indicate the progressive aspect. The regular past tense ‘-ed’ form is meaningful but is more redundant (i.e., it has lower communicative value), because it usually co-occurs with

---

redundant meaningful grammatical *markers* before they process redundant meaningful *markers*” (p.119).<sup>18</sup> The revised Principle *d* is “Learners are more likely to process meaningful grammatical markers before non-meaningful grammatical markers” (VanPatten, 2007, p.120).

temporal adverbs, or other contextual clues, to encode the pastness. When learners encounter these two linguistic features in utterances, Principle *c* predicts that the ‘-ing’ form without redundancy is supposed to be processed earlier than the ‘-ed’ form with redundancy.

Since learners are expected to be driven to acquire meaning firstly during the processing of input, Principle *d* predicts that the notion of the ‘semantic value’ of a linguistic feature always takes precedence over ‘redundancy’ during the processing of input. This principle may account for why the complementisers such as ‘that’ are acquired rather later by L2 learners of English, given that the complementiser ‘that’ does not encode any semantic meaning. Although the complementizer ‘that’ has a grammatical function – to join two sentences, it is categorised as a nonmeaningful grammatical form (VanPatten, 1996).

Bransdorfer’s studies (1989, 1991) were cited by VanPatten to support these principles. According to VanPatten (1996, p.26), Bransdorfer (1989) classified the Spanish possessive case ‘*de*’ as having higher communicative value than the definite article *la*. Neither of these linguistic features can be regarded as a content word, and they have similar syntactic positions in a sentence, usually preceding their nouns. Learners of Spanish would have problems in interpreting a sentence without the preposition *de* being present because *de* has an inherent semantic value of possession. On the other hand, learners would not have problems in comprehending a sentence with the absence of the definite article *la* due to its lower communicative value. In this sense, to process the form ‘*de*’ with its higher communicative value should be easier for learners than to process the form ‘*la*’ with its lower communicative value. Bransdorfer (1990) examined the ability of L2 Spanish learners to process meaning and form at the same time by

having them listen to a brief passage and note the occurrence of either ‘*de*’ or ‘*la*’. Bransdorfer’s results concluded that attending to ‘*la*’ whilst listening to the passage caused more trouble in comprehension than attending to ‘*de*’. VanPatten claimed that Bransdorfer’s results corroborated his theory on the impact of relative communicative value on learning an L2.

***Principle e. The ‘availability of resources’ principle.***

For learners to process either redundant meaningful grammatical forms or non-meaningful forms, the processing of overall sentential meaning must not drain available processing resource.

The principles already described suggest that the acquisition of a given form during input processing should depend on its relative communicative value. L2 learners are involuntarily driven to process those grammatical forms with higher communicative value than those with lower communicative value. However, L2 learners can still acquire those forms with low communicative value. Principle *e* sets out to explain this.

Principle *e* proposes that undertaking two tasks (such as comprehending the meaning and perceiving the grammatical features) simultaneously is possible on condition that the other task does not consume all of the attentional resources (i.e. comprehending the overall meaning does not use up all the attentional resources). Relevant to this principle is Just & Carpenter’s (1992) study. Just & Carpenter indicated that individuals differ in their attentional capacity in L1 processing, which affects on-line listening or reading during comprehension. Since the capacity is different for different individuals, the input processing patterns are supposed to be different between those with greater and lesser attentional capacity. In this scenario, learners with higher language proficiency may



release more attentional resources during input processing in that comprehending the sentential meaning occupies fewer attentional resources in comparison with lower-proficiency language learners. Although VanPatten has not explicitly accounted for the relationships between individual differences, proficiency, and input processing in this principle, it appears intuitively appealing that learners with higher proficiency are more capable of processing non-meaningful forms than those with lower proficiency during input processing.

VanPatten acknowledged that this principle is theoretically motivated rather than empirically substantiated, stating that there is “no solid experimental evidence that directly supports this principle” (p. 27). He cited Leow’s (1993) and Blau’s (1990) findings to illustrate this principle. Presumably, simplified input (such as making shorter sentences, or using familiar vocabulary) would induce less attentional load than unsimplified input. Leow (1993) concluded that participants engaged in reading simplified input outperformed those engaged in unsimplified input. Blau (1990) demonstrated the impact of three factors (speed, complexity, and pausing) on the comprehensibility of aural input in Puerto Rican and Polish learners of L2 English. Blau concluded that pauses did enhance the comprehensibility of aural input significantly better than slowing down the rate of speech or simplifying the syntax. As VanPatten (1996) noted, the implication from Blau’s study is that “less or non-meaningful grammatical features should be more easily detected when the input contains pauses that allow for processing time” (p.28) since pauses may help learners to detect non-meaningful features.

Additionally, N. Ellis (2006) cited the study of Matessa & Anderson (2000) to express a similar view to this principle, although he had no intention of validating this IP

principle. Ellis pointed out that language learners are apt to focus on only one cue at a time in the beginning stage. As they make progress and can track the use of the first cue, they “add a second cue to the mix and begin to use the two in combination” (p.169-170). By associating the perspective of Ellis with this IP principle, the first cue tracked down by the learner could be the lexical item, and the second cue could be the grammatical form (see the Lexical Preference Principle). At first, learners rely on the lexical item during initial input processing. Later on, learners start to process the grammatical form if they have no problem at all in comprehending the lexical items, and then they are able to use both cues together in their utterances.

***Principle f. The sentence location principle.***

Learners tend to process items in sentence-initial position before those in final position and those in medial position.

Apart from weighing communicative value to decide how a grammatical form gets processed, VanPatten has cited the studies of Barcroft & VanPatten (1997) and Klein (1986) to suggest that one potential factor affecting learners’ allocation of attentional resource is the relative location within the utterances. Barcroft & VanPatten found that items in the initial position in utterances are easier to process for beginner learners of L2 Spanish in an imitation task. Klein also stated that “With any speech sound sequence (which may represent an utterance) there are always some segments which are more readily available to analysis than others” (p.68). Klein gave the processing priority as follows: the opening segment(s) of an utterance, the concluding segment(s), and then the segment(s) immediately preceding and following any identifiable pauses.

***2.2.1.4 Challenges and unclear issues of IP***

The following section aims to address some challenges put forward by other researchers and to point out some unclear issues relating to IP. Note that it is beyond the purpose of this study to address these issues. They are mentioned here simply to indicate that it is acknowledged that IP theory is not fully validated and remains to be refined and/or changed.

*a) Adoption of an outmoded model of attention?*

The notion of attention being a limited capacity, adopted from cognitive psychology in IP theory, has undergone some critiques (DeKeyser *et al.*, 2002; Harrington, 2004; Long & Robinson, 1998). DeKeyser *et al.* (2002) specified that the limited-capacity attentional construct underpinning IP theory is outdated. DeKeyser *et al.* cited the opinions of Neumann (1996) and Robinson (2003) that attention is unlimited to argue against VanPatten's IP theory. Furthermore, VanPatten's argument concerning the competition between form and meaning during input processing has been called into question (DeKeyser, *et al.*, 2002; Long & Robinson, 1998). DeKeyser *et al.* suggested that "simultaneous attention to form and content is clearly possible" (p. 809) by arguing that attending to form and content during processing input is a single task and not a dual task. DeKeyser *et al.* rationalised the possibility of attending to form and meaning at the same time by citing the results of experimental studies (e.g. de Graaff, 1997; Robinson 2002) on incidental learning (such as that learners learn 'some' forms during processing for meaning). Long & Robinson (1998, p.39) argued that attentional resource is not a single and undifferentiated resource. Long & Robinson (1998) drew on Wickens's (1984, 1989) multiple resources proposal to suggest that form and meaning are not always necessarily competing for attention. In addition, Harrington (2004) pointed out that, although the limited-capacity idea is widely used in cognitive psychology, the IP theory, which is fundamentally based on this assumption, should show evidence to

account for the fact that L2 input processing was due to the “capacity limitation and not just to lack of L2 knowledge” (p. 89) in order to increase the explanatory power of the IP theory.

VanPatten (2002c, p.826), however, acknowledged that some ‘unlimited attentional’ models might exist, but further argued that these unlimited models are developed on the basis of L1 rather than L2 speakers. Additionally, the ‘language processing’ is not always the main focus of these studies seeking evidence for the unlimited attentional models. VanPatten cited Just & Carpenter’s (1992) study to defend his ground. Just & Carpenter proposed that human limited attentional resources during language processing is comprehension-oriented. Their proposal that language comprehension would rob learners’ attentional resources and have an influence on their reading and listening skills was demonstrated by L1 learners. VanPatten argued that if L1 learners’ attentional resources are robbed when comprehending their native language, how much more would a new language deplete L2 learners’ attentional resources during input processing? Further, IP has *never* claimed that simultaneously attending to form and content/meaning is impossible. What IP posits is that L2 learners ‘*tend*’, ‘*prefer*’, or are ‘*more likely*’ to process content words before grammatical form. In fact, the IP principle *e*, ‘the availability of resources principle’, implies the possibility of attending to form and meaning at the same time during input processing.

*b) Generalisability of IP to all types of input processing?*

According to Wong’s (2001) study, the IP Principle on ‘the primacy of content words’ does not seem to be supported if input is in a written mode. Wong’s results (2001, p.358) suggested that the learners’ limited attentional capacity is not constrained in the same way during input processing in the aural and the written modes. Wong found that

learners could attend to form and meaning in the written mode but not in the aural mode. However, VanPatten did not illustrate whether or not the IP principles are applicable to different input modalities, although he did suggest using two modes (the listening and reading modes) in PI activities.

Furthermore, VanPatten developed his IP theory with beginning and intermediate learners in mind. However, the issue of whether PI is applicable to high-proficiency language learners has not been clearly delineated. Although the principle of the availability of resources might be extended to elucidate this issue to some extent, in that it implies that the higher a learner's proficiency, the more likely s/he is to process a linguistic feature during the input process, VanPatten has not satisfactorily addressed this issue. What is more, if processing form and meaning simultaneously is possible for those learners with high proficiency, are the IP principles applicable to them when they encounter a *new* linguistic feature?

*c) Too much simplicity?*

The notion underpinning how a linguistic grammatical form gets processed (i.e. the establishment of form-meaning connections) in IP theory largely rests on the communicative value, which is the nature of the linguistic grammatical form. If the concept of FMCs is at the heart of IP theory, then the potential factors affecting the initial FMCs should be taken into consideration. Apart from Principle *f*, 'the sentence location principle' (an intrinsic characteristic of a given form), predicting how a grammatical form gets processed from its relative location in the utterance, IP theory so far has not sought to expound how other intrinsic characteristics (for example, phonological and perceptual saliency (Goldschneider & DeKeyser, 2001), complexity (DeKeyser, 2008, p.8), and so on) and extrinsic characteristics (such as frequency

(Goldschneider & DeKeyser, 2001)) of the grammatical form affect learners' input processing. Some other factors may also exert an influence on the establishment of FMCs, such as learners' L1 (Cadierno & Lund, 2004), and their language proficiency (Gass, 2004; Shirai, 2004). Harrington (2004) criticises IP theory for being too narrow to explain the initial processes of grammar: Harrington (2004) states that IP theory provides "no account of how the input processor might interact with other language processing mechanisms and existing knowledge" (p.89).

VanPatten (1996, 2007) acknowledged that these potential factors may interact with IP principles and affect the input processing for language acquisition to take place. Nevertheless, VanPatten (1996, p.31) somewhat defended that the factors, such as the frequency or complexity of a given form, are more relevant to the strength of long-term storage as opposed to short-term storage. However, although VanPatten repeatedly stressed that IP centres on form-meaning connections take place during input processing, these other factors remain to be investigated.

In addition, some researchers have pointed out that IP is not an exhaustive theory of L2 grammar development, given that issues regarding the grammatical development of L2 fluency or accuracy do not seem to be addressed (DeKeyser *et al.*, 2002; Harrington, 2004). Although IP intends to explain why some FMCs develop earlier than others, how these FMCs ultimately turn into part of the learner's developing or developed language system is not explicitly specified.

### **2.2.2 Form-Meaning<sup>19</sup> Connections (FMCs)**

---

<sup>19</sup> According to VanPatten *et al.* (2004, p.2-3), *form* refers to "a surface feature of language or a surface manifestation of an underlying representation", such as lexicon, inflections, particles and the like. *Meaning* refers to a concrete referential meaning (for example, 'fish' in English means a creature which

One of the characteristics that make PI distinct from other grammar approaches is that PI proponents stress the construct of FMCs, which urges learners to understand the properties and meanings of the targeted grammatical forms. It is noted that the construct of FMCs can be employed in lexicon or grammar learning, and it is claimed that the development of FMCs for lexical items is easier than for grammatical forms (Gass, 2004; VanPatten 1990). The concept of FMCs discussed in this thesis is essentially concerned with the linkage between a specific *grammatical* L2 form and its referential meaning, unlike FMCs in lexical learning. This section will begin by briefly delineating why the concept of FMCs is important in grammar learning, and then move on to the developmental processes of FMCs.

#### *2.2.2.1 Why are FMCs important?*

Krashen's (1982, 1985) proposal of the 'Input Hypothesis' suggested that language acquisition takes place when learners are engaged in comprehensible input. Krashen (1982) argued that for acquisition to occur, learners are required to understand the input, in which "understand means that the acquirer is focussed on the meaning and not the form of the message" (p.21). However, Krashen's proposal has been challenged due to the results of a range of immersion and naturalistic acquisition studies, suggesting that learners did not develop target-like accuracy by being entirely exposed to comprehensible input (Harley, 1992; Harley & Swan, 1984; Spada & Lightbown, 1989) and meaning-oriented input (for example, the enriched input in Marsden, 2006; input flood in Trahey & White, 1993).

One of the potential problems in Krashen's proposition may lie in his dissociation of 'form' from 'meaning'. As N. Ellis (2004) stated, "SLA is the learning of constructions

---

lives in water and has a tail and fins) or an abstract referential meaning (for example, un- at the beginning of a term means 'no' or 'not') such as number, temporal, agency, lexical reference and so on. Thus, FMC is viewed as "a situation in which a form encodes some kind of its referential meaning".

relating form and meaning” (p.50). Ellis pointed out that the provision of opportunities for language learners to associate form and its meaning is important in enabling them to develop target-like associations. If one considers that successful language acquisition should not merely achieve fluency but also achieve accuracy, the establishment of FMCs will be essential in language acquisition. Note that this particular FMC construct is being discussed here because of the purpose of this study, namely to explore Processing Instruction, which necessarily involves the notion of FMCs. Further investigation into potential factors that affect the establishment of FMCs (e.g. L1, frequency, the complexity of a form, and so on) is beyond the purpose of the current study.

#### *2.2.2.2 The developmental processes of FMCs*

VanPatten *et al.* (2004) outlined three stages that learners go through to achieve it: 1) making the initial connection; 2) subsequent processing and strengthening; and 3) accessing the connection for use. The first step in establishing an FMC is to initiate the initial connection between the form and its meaning. However, the incipient FMC may not be established immediately because it may be weak or incomplete. The incipient FMC may eventually diminish from memory if it lacks the subsequent input to reinforce it. In order to consolidate a weak or incomplete FMC, the subsequent process to make it more solid may be vital. Once the FMC has been solidly established, further access to comprehend or to produce the targeted form will be attainable. Note that the developmental processes of FMCs are not unidirectional (VanPatten *et al.*, 2004). Access to a form can strengthen the association between the form and its meaning, given that the process involves a form being detected and providing opportunities for FMCs to be reinforced.



### **2.2.3 Attention Theory**

This section presents some characteristics of attention, namely that attention is a limited capacity, effortful and selective. Following this, two attentional postulations (i.e., Schmidt's 'Noticing Hypothesis' (1990) and Tomlin & Villa's fine-grained attention (1994)), which VanPatten noted, are described. These attention issues are discussed here because VanPatten's IP theory relates to them, but they are not the focus of this study.

#### *2.2.3.1 Characteristics of attention*

##### *a) Attention has a limited capacity<sup>20</sup>*

The proposal that attention is a finite resource has cropped up for decades in cognitive psychology (Broadbent, 1958; Kahneman, 1973; Treisman, 1964) (see more discussion on Robinson (1995a)). The limited resource metaphor for attention has been extensively used to elucidate why humans are incapable of processing the entire stimuli in a given time, and why, consequently, only parts of stimuli register in the working memory for further processing. There has also been a general acceptance of the notion of attention as a limited capacity within L2 research (Lee, Cadierno, Glass, & VanPatten, 1997; Schmidt, 2001; Tomlin & Villa, 1994; VanPatten, 1990, 1994).

VanPatten (1996, p. 46) cited Dienes, Broadbent & Berry's (1991) study to strengthen this construct. Dienes *et al.* carried out an experiment in which subjects were exposed to an artificial grammar-learning task and were required to complete a concurrent task (i.e. to indicate the appearing letter string's grammaticality). Their results showed that subjects' ability to judge the grammaticality of the strings in a post-test was noticeably

---

<sup>20</sup> In this thesis, attentional capacity is defined as the maximum amount of activation available in the working memory. Attention is defined as the registration of input strings in the working memory, so it is a subcomponent of the working memory. The working memory is involved in two functions (i.e., the processing and the storage), but attention is only involved in processing (Just & Carpenter, 1992).

reduced, suggesting that the processing of letter strings was vitiated during the learning phase. Although the applicability of an artificial language to SLA is debatable due to the ‘inherent complexity of natural language’ (e.g. the lack of morphological forms in an artificial language, Hulstijn, 1997, p.139), VanPatten (1996) argued that “if subjects reveal a limited capacity to process rather simple strings of letters, logically it follows that to process new forms and perform all the mental operations required to map form onto meaning in second language acquisition must be at least as taxing if not more” (p.46). Note that although some studies are in favour of unlimited attentional capacity (see DeKeyser *et al.*, 2002 commenting on this issue), VanPatten (2002a) argued that the notion with respect to the attentional resource being limited has broadly been taken on board in SLA research. After all most of the time learners cannot manage to attend to all the stimuli during the on-line attempt to comprehend incoming input.

*b) Attention is selective*

It is proposed that attention is selective in cognitive psychology (Broadbent, 1958; Norman, 1968; Treisman, 1964) and in SLA (Lee, *et al.*, 1997; Robinson, 2003; Schmidt, 2001; VanPatten, 1989, 1990). Given that attention is of limited capacity, when an individual deals with incoming stimuli, the information process may be involved in the competition for limited resources. In this respect, selecting the incoming stimuli is a corollary. In cognitive psychology, Broadbent’s (1958) ‘filter theory’ argued that information processing has to undergo an attentionally selective mechanism whereby a decision is made on which incoming sensory information is further processed (i.e. detected, and then encoded in the short-term memory). Norman (1968) proposed that whether or not sensory information got processed depended on its importance. The more important information is judged to be, the more likely it is to get processed.

The idea of attention being limited and selective has stimulated some SLA researchers to come up with a question: what do learners tend to attend to while processing incoming input (Lee *et al.*, 1997; Robinson, 2003; VanPatten, 1989, 1990). VanPatten (1996) cited Broadbent's 'filter theory' from cognitive psychology to support IP theory and carried out classroom-based experiments on L2 processing to explore this issue. VanPatten (1989, 1990) argued that in most L2 processing learners have to 'select' what to process during input processing due to limited attentional capacity. During a Spanish listening task, VanPatten (1990) allocated 202 learners into four groups. They were instructed to process information under four different conditions: whilst listening to the passage for meaning only (the control group), listening to the passage and noting any lexical items, listening to the passage and noting any definite articles, or listening to the passage and noting any morphological markers. The participants' comprehension of the passage was then assessed. His results showed a decrease in comprehension when learners were required to pay attention to the grammatical forms (i.e., the Spanish article or morphological markers). VanPatten (1990, p. 296) concluded that it is difficult for L2 learners, particularly for beginners, to pay attention to form and meaning simultaneously in terms of aural mode. With the aim of understanding the meaning of the contexts, L2 learners are involuntarily primarily directed to attend to those elements that carry the meaning of the message (mainly, lexical items). Only later when they have spare attentional resources is it likely that they will process the grammatical form, given that it makes less contribution to their understanding of the meaning of the grammar marker (Lee, *et al.*, 1997; VanPatten, 1990, 1996, 2002a and elsewhere). An important caveat is that VanPatten's (1990) study was only borne out in the aural mode. Whether his results can be generalised to the visual mode is not clear.

*c) The processes of attention*

In respect of attentional processes, it has been argued that the attentional system is involved in two types of processes, namely automatic and controlled processes (Robinson, 1995a, 2003; Schmidt, 2001; Tomlin & Villa, 1994). According to Tomlin & Villa (1994),

“automatic processes require little or no attention and thus do not interfere with other activities. Controlled process requires attention, which is of limited quantity, and therefore these processes will interfere with other processes that also require attention” (p.189).

The automatic process can occur if the tasks are similar and therefore interfere less with each other, so that little extra attention is required. On the other hand, if these tasks are rather different then this involves more controlled attention, and it is difficult for people to undertake them concurrently. In cognitive psychology, Wickens (1984, 1989) pointed out that performing two tasks at the same time is possible either when two tasks simultaneously draw on different pools of attentional resources, entailing no competition for resources between tasks, or when one of the tasks is automatized. Wickens further suggested that attentional resource allocation would be affected by individual difference and task demand. Though VanPatten has not directly addressed the processes of attention in his IP theory, the concept of IP Principle *e* (i.e. the ‘availability of resources’ principle) is related to it. VanPatten (1996, 2004) suggested that for L2 learners to attend to a linguistic feature, comprehending the meaning of the overall content should not detract from their attentional resources.

### *2.2.3.2 Postulations of attention in SLA*

#### *a) Schmidt’s ‘Noticing Hypothesis’*

Schmidt (1990) came up with the construct of the ‘noticing hypothesis’<sup>21</sup>, suggesting that *conscious* noticing is both necessary and critical for learning to occur. Schmidt (2001) further indicated that the role of attention is required for learning nearly every aspect of a second or foreign language. The construct of this ‘noticing hypothesis’ was initially developed from the results of Schmidt & Frota’s (1986) diary study based on Schmidt’s own personal attempts to learn Portuguese. However, Schmidt & Frota’s diary study has been criticised due to the likelihood of mismatch occurring between the processing of L2 input while coding diary entries and the processing of incoming input in natural interaction (Leow, 2001, p.118; Tomline & Villa, 1994). The validity of the self-report technique has also been questioned (Leow, 2001; Robinson, 1995a).

In addition, Schmidt (1990, 1995) proposed two levels of awareness: awareness at the lower level of noticing, and awareness at the higher level of understanding. Schmidt (1995, p.29) referred to ‘noticing’ as “conscious registration of the occurrence of some event” which is a surface-level phenomenon (i.e., it is briefly registered in the short-term memory). On the other hand, Schmidt (1990, p.132) expressed the view that ‘understanding’ is related to “recognition of a general principle, rule or pattern” such as hypothesis and rule formulations, which is a deeper-level phenomenon, takes place in the long-term memory and pertains to the ability to analyse and compare the linguistic input. Note that Schmidt’s noticing hypothesis is in contrast to Krashen’s (1985) claim that conscious awareness is *not* necessary for the linguistic input to be incorporated into a developing linguistic system.

*b) Tomlin & Villa’s fine-grained analysis of attention*

---

<sup>21</sup> Schmidt (1995) regarded ‘noticing’ as attention accompanied by some low level of awareness and he stressed that ‘noticed’ in his term is nearly isomorphic with ‘attention’ (p.1).

Tomlin & Villa (1994) proposed that the attentional system consists of three separate but interconnected components, namely, alertness, orientation and detection. Alertness refers to the learners' readiness to tackle the incoming stimuli or information (p.190). In the field of SLA, alertness in attentional system acts as L2 learners' motive for interest in learning a piece of L2 knowledge before receiving any instruction. Orientation is "the specific aligning of attention on a stimulus" (p.191), and "a heightened sensitivity to a specific feature of some incoming stimulus" (p.197). Attentional orientation can be purposefully directed to some types of stimuli by means of abandoning others. As for its role in SLA, the issue regarding whether orientation works well or not is related to the techniques that instructors use to draw learners' attention to a certain type of stimulus during activities. The final component detection involves the processes of selection, or engagement in specific stimuli or information, which determines whether incoming stimuli can register in the working memory or not. Once the incoming stimuli or information is detected (i.e. it is registered in the working memory), further processing is possible, such as hypothesis formation or testing. Tomlin & Villa (1994, p.197) argued that detection is the key to acquiring an L2. They further explained that the relationship between these three components is relevant but not causal. Detection may be strengthened when a learner is more alert or is oriented towards a specific type of stimulus. Even so, neither alertness nor orientation is necessary to induce detection.

Note that Schmidt's noticing refers to merely 'noticing', or being aware of, the presence of a linguistic feature, but does not require the learners to understand its meaning. On the other hand, Tomlin & Villa's detection refers not only to attending to a form but also to linking it to its meaning, which is the so-called establishment of form-meaning connection (p.198). As noted earlier, VanPatten discarded Schmidt's 'noticing' but embraced Tomlin & Villa's 'detection' to amplify IP theory. Due to the emphasis on the

importance of developing FMCs in IP and PI, VanPatten (2002b, 2007) claimed that ‘processing a form’ could be equal to ‘detecting a form’, but not to ‘noticing a form’. The manipulation of the PI structured input activities aims to orientate learners’ attention to the targeted feature, and then enable ‘detection of the targeted feature’ to happen more easily. However, VanPatten has not clearly addressed whether this is the case for both types of PI activity.

#### ***2.2.4 Some evaluations and challenges of PI in terms of IP, FMCs, and attention***

This section attempts to evaluate PI on the basis of the relevant theories and concepts that PI has claimed underpin it. It is noted that the viewpoints put forward here are outside the author’s area of interest, and it is beyond the scope of the current study to explore all the issues, although some of them will be discussed later.

##### ***2.2.4.1 PI activities adherence to FMCs?***

The construct of ‘pushing learners to make an FMC’ is at the heart of PI. However, whether or not the two types of PI activities succeed in achieving FMCs is questionable, given that the completion of *affective* activities does not consequently entail learners interpreting the meaning of a specific form (Marsden, 2004, 2006) (see Section 2.1.1.2). One of the purposes of affective activities is to reinforce an FMC established by the referential activities, and then to internalise the form into the language developing system (Wong, 2004a). However, it is possible that learners can complete the affective activities successfully without noticing or detecting the targeted feature embedded in the tasks. If the role of affective activities is to strengthen an FMC initiated by the referential activities (Marsden, 2004, 2006; Wong, 2004a), can affective activities achieve that without forcing learners to make the FMC?

##### ***2.2.4.2 Redundancy obstructing FMCs?***

When designing PI activities, it is advisable to remove the redundant element in the context to assist the learners in making better FMCs. However, some researchers are not convinced about the role of redundancy that PI proponents addressed (Batstone, 2002; Harrington, 2004). Harrington (2004) pointed out that the redundancy can “facilitate communication by lowering the processing load through minimising the amount of new information the system has to deal with” (p.88). From a discourse-oriented perspective, Batstone (2002) argued that a lexical item could be used as an ‘anchor’ to assist learners in learning a new language, at least in the initial stage of learning, given that the referential meaning of the targeted feature might be reinforced by the appearance of the lexical item in an utterance. However, VanPatten (2002b) responded to Batstone’s critique and argued that the feedback provided to PI learners could act as a supplement to facilitate learners’ understanding of the discourse. However, as noted earlier in 2.1.1.2, the feedback provided through different types of PI activities is not the same in essence. It is not clear whether or not the feedback provided via a referential activity and an affective activity can exert its influence as VanPatten claimed.

#### *2.2.4.3 The impact of different modalities on PI*

PI provides learners with both visual and aural input; however, it is suggested that the processes of visual and aural input are rather different (de Jong, 2005). Wong (2001) argued that processing aural input is more difficult than processing visual input because the former places more restriction on learners’ attentional capacity than the latter. From a cognitive psychology perspective, Baddeley (1986) proposed a model of the working memory, which comprises three components, namely a supervisory system (the central executive), and two subsidiary slave systems (i.e., visuo-spatial sketchpad and phonological loop). This could suggest that the modality could be a potential variable affecting how a learner processes input, since processing the aural and visual inputs is



not the same. Although VanPatten suggested employing the two modes to cater for individual differences when creating PI activities, he has not further addressed how the two modes are related to the effectiveness of PI.

#### 2.2.4.4 *The role of output*

In spite of SLA research generally agreeing that input is necessary, whether input is ‘sufficient’ for language acquisition to happen remains a contentious issue (VanPatten & Williams, 2007, p.10). PI regards output practice as a means of getting access to what has been developed during input processing. Consequently, it appears that PI gives a different role for output practice in the early stage of learning an L2. VanPatten (1996) argued that having learners produce meaningful language in the early language learning phase is like “putting the cart before the horse”. A number of studies, however, have argued the facilitative role of output in the development of FMCs (Salaberry, 1997; Toth, 2006) and language acquisition (Izumi, 2002; Swain, 1998). Salaberry (1997, p.440) suggested that “the distinction between input and output processing is not consequential for language development, because both processes are involved in the development of form-meaning connections” (p.440). Toth (2006) claimed that PI necessitates learners reacting through *indicating* the meaning of a given form; on the other hand, output-based activities necessitate learners reacting through *producing* the meaning of the form. The establishment of FMCs can be achieved by both input- and output-based activities; it is just accomplished by different modes.

In addition, PI proponents have always claimed that at no point do PI learners become involved in producing the targeted feature. However, it is highly possible that learners spontaneously experience subvocal rehearsal when they are engaged in processing

input. Although subvocal rehearsal does not necessarily involve producing the actual sounds, it could still be regarded as a type of output.

#### *2.2.4.5 Practical issues*

Some practical issues emerge when constructing PI activities. One is the design of tasks for learning a targeted linguistic form without any inherent meaning. Since the form has no ‘meaning’, how and to what extent can the PI activities manage to facilitate the formation of FMCs? VanPatten has not specifically addressed this practical issue or offered guidelines on how to devise PI activities relating to the non-meaningful form.

In addition, in order to push learners to make better FMCs, VanPatten suggests that the input sentences should be manipulated so that the target feature is located as near to the start as possible according to the ‘sentence location’ principle of IP. In this sense, it appears that ‘fragmentation’ of input may be desirable. However, practical difficulties may arise when putting this principle into effect designing PI activities. For example, the bound inflection clearly cannot be placed in an initial position (e.g. the English ‘-ed’ or third person singular ‘-s’ features). Furthermore, Collentine (2002) criticised Farley’s (2001) manipulation of PI activities by placing Spanish subjunctive forms in utterance-initial positions. Collentine (2002) stated “in authentic language, if learners do hear the subjunctive in authentic input in an utterance-initial position, it more than likely connotes coercion” (p.883-884). Although PI proponents have provided guidelines to develop PI activities (see Section 2.1.3), some practical issues as discussed above related to how to construct PI activities according to its theoretical backgrounds require further exposition.

## **2.3 Literature Review of previous PI research: verified and unverified issues of PI, and the formation of the current study**

This section reviews previous PI studies. The first section focuses on setting out some well and less researched issues within the PI research agenda. The subsequent section will concentrate on explaining how this current study is built up, and will present the motivations behind this study and discuss the gaps in previous PI studies. Also the reasons for choosing the targeted linguistic feature and the measure to elicit it for the current study will be given; after this, the hypotheses and research questions are put forward.

### ***2.3.1 Some verified and unverified issues of PI***

This section is a review of PI studies, and aims to discuss some verified and unverified issues that have emerged from previous PI studies. To date, a great deal of research has empirically investigated the effectiveness of PI (Benati, 2001, 2004a, 2004b, 2005; Cadierno, 1995; Cheng, 2004; Farley, 2004a, 2004b; Marsden, 2006; VanPatten & Cadierno, 1993a, 1993b; Wong, 2004b). Because of the limited space of this thesis, Appendix 3 provides a tabular summary of detailed information about studies related to PI. The findings of these PI-related studies on the interpretation tests and the production tests are provided in Appendix 4 and Appendix 5 respectively.

#### *2.3.1.1 What is the relative effectiveness of PI vs. other types of grammar instruction?*

Studies have, to date, been conducted to compare PI with output-based grammar instruction such as traditional instruction (TI henceforth) (Benati, 2001; Cadierno, 1995; Cheng, 2004; VanPatten & Cadierno, 1993a, 1993b; VanPatten & Wong, 2004; Wu, 2003; Xu, 2001), meaning-output instruction (MOI henceforth) (e.g., Benati 2005; Farley, 2001, 2004a; Morgan-Short & Bowden, 2006), and communicative output

instruction (COI) (Toth, 2006). So far, only one study has been carried out to compare PI with an input-based instruction, namely enriched input-based instruction (EnI) (Marsden, 2006). The following will demonstrate the relative effectiveness of PI in comparison with other grammar instruction.

*a) PI vs. Output-Based Instruction*

Note that that the comparison between PI and output-based instruction has been the main focus of previous PI-based research, but it is *not* central to this study. However, it is described here as a background context to the present study and to show how PI relates to other instruction.

TI<sup>22</sup> has been regarded by PI researchers as a common grammar teaching approach adopted in USA language classrooms (VanPatten & Wong, 2004, p.100). VanPatten and Cadierno's (1993a) study was the first one which set out to investigate the relative effectiveness between PI and other type of instruction, namely TI. They concluded that PI was superior to TI, given that the PI group made significant improvement in both the comprehension and production tests, whereas the TI group only made significant improvement in a production test. Following VanPatten and Cadierno's (1993a) study, some studies (e.g., Benati, 2001; Cadierno, 1995; Cheng, 2002; VanPatten & Wong, 2004) were conducted to replicate their results. VanPatten (2002a) summarised the results with the claim that "In general, it seems that the conclusions of VanPatten and Cadierno hold overall, namely, that PI is superior to TI" (P. 790).

---

<sup>22</sup> TI is composed of explicit grammar explanation plus output-based practices in which it moves through a sequence of being mechanical, meaningful and then communicative, as suggested by Paulston (1972).

However, some researchers have queried the fact that PI yielded better learning gains than TI was due to the absence of mechanical practices (Farley, 2004a). So subsequent studies have set out to compare PI with meaning-based output instruction (MOI), in which the mechanical practices were removed. The results of these comparative studies were not as clear as those of PI vs. TI when it came to PI vs. MOI. It was found that PI was superior to MOI in the interpretation task, but performed as well as the MOI in the production task (Benati, 2005; Farley, 2001; Lee & Benati, 2007a). Farley (2004a) reported that PI performed equally as well as MOI on the acquisition of the Spanish subjunctive in both interpretation and production tests. Farley (2004a, p.163) argued that learners in the MOI group also received IP-like input language (i.e., incidental input) through interaction between the instructor and learners, or through interaction between peers during the instructional phase. Farley argued that the ‘incidental input’ occurring in the course of MOI might have strengthened learners’ performance in the tests. Based on the results comparing PI with TI and MOI, VanPatten (2004) claimed that “although it is not clear that all output-based approaches always make a difference, PI always does. Our claim is that the consistently positive effects of PI are due to the effect(s) it has on learner processing of input”(p. 96).

As for the relative effect of PI vs. communicative output instruction (COI henceforth, Toth, 2006), although MOI and COI have some common characteristics (they are both meaning-oriented, and both involve the removal of mechanical and non-meaningful interactions), Toth (2006, pp.330, 341) stressed that communicative output instruction should not be considered as a replication of Farley’s (2001a, 2004a) MOI research due to the disparate operationalisation of the output practice. The implementation of communicative output practices follows Swain’s (1998) ‘pushed output’ rather than Lee & VanPatten’s (2003) ‘structured output practices’, which is what Farley’s studies were

based on. Statistically speaking, the results of Toth's study showed that both PI and COI performed equally on a timed grammaticality judgement<sup>23</sup>, but that the COI group outperformed the PI group in the guided written production test, though PI outperformed the control group in both tests.

However, a few studies suggested that PI was not superior to output-oriented instruction (see Allen, 2000; Collentine, 1998; DeKeyser & Sokalski, 1996; Salaberry, 1997), PI proponents argued over the validity of these studies on operationalising PI (see Farley, 2002; Sanz & VanPatten, 1998; VanPatten, 2002a; VanPatten & Wong, 2004). VanPatten (2002a) further stated that the implication from the results of comparisons of PI with output-based instruction is that "as long as classes and materials are meaning-oriented and avoid mechanical and display language, acquisition is fostered, and PI is no better than any other meaning-based instruction with a form focus" (p.798).

#### *b) PI vs. Input-Based Instruction*

In terms of empirical studies comparing PI with input-based instruction, only one study has been carried out so far to compare their relative effectiveness (Marsden, 2004, 2006). Marsden compares PI with Enriched Input Instruction (EnI henceforth)<sup>24</sup>. To some extent, the basic framework of EnI resembles PI in that it is composed of a brief grammar explanation and input-based activities. However, EnI merely requires learners to be exposed to a number of examples of the targeted feature, which is similar to input flood activities and affective activities. EnI did not force learners to make FMCs, just like affective activities. Marsden's (2004, 2006) results suggested that learners who

---

<sup>23</sup> In Toth's (2006) study, learners were told to finish the entire test within 25 minutes on the timed grammaticality judgement test. Note that each test item was not separately timed.

<sup>24</sup> EnI was considered to share characteristics of listening and reading activities frequently seen in UK modern foreign language (MFL) textbooks and done in UK classrooms. EnI in Marsden's study also included explicit information, which could have raised learners' awareness of the target feature (also in line with some kinds of metalinguistic instruction observed in MFL classrooms).

received PI treatment improved more than learners of EnI on the acquisition of French verb inflections, suggesting that the pushing learners to establishment of an FMC in practice activities is conducive to learning grammar. Furthermore, Marsden (2006) argued her study “went some way to exploring the two types of structured input activity in PI: referential and affective activities” (p.548-549), given that EnI was fundamentally similar to PI affective activities: forcing learners to interpret the meaning of the targeted feature was not essential. She argued that these two types of PI activities may have different instructional impacts, and that referential activities may be the cause of gains in learning a grammatical feature but that affective activities may have a role in promoting lexical knowledge. However, this issue has never arisen in previous PI studies.

### *2.3.1.2 Can the positive effect of PI be generalised to other linguistic features in different languages?*

VanPatten (2002a) claimed that PI-based studies have offered evidence that “the results of VanPatten and Cadierno (1993) are generalizable to other structures and in different languages and at least that the effects of PI alone are generalizable to other structures” (p. 775). To date, it has been demonstrated that PI can work well on Spanish direct object clitics (VanPatten & Cadierno, 1993a, 1993b), the Spanish simple past tense (Cadierno, 1995), the Spanish subjunctive (Farley, 2004a), the Spanish *ser* and *estar* (Cheng, 2004), the Spanish anti-causative *se* in the passive, middle voice, and impersonal constructions (Toth, 2006), French verb inflections in the perfect and present tenses (Marsden, 2006), the French causative (VanPatten & Wong, 2004), English past tense (Benati, 2005), the English simple present versus progressive (Buck, 2000, cited in VanPatten, 2002b), the English Wh-questions (Xu, 2001), the English subjunctive mood (Wu, 2003), and the Italian future tense (Benati, 2001). As far as learning

different aspects of grammar is concerned, the targeted linguistic features used in previous PI studies have substantiated its positive effect in morphosyntactic features (such as Spanish direct object clitics), in morphological features (such as the English ‘-ed’ form and the Italian future tense), and in syntactic features (Farley, 2004a; Wu, 2003; Xu, 2001).

#### *2.3.1.3 What is the relative effectiveness of PI delivered by different modes?*

PI materials can be delivered by computers or in a teacher-fronted classroom. Feedback in PI activities can be delivered by the computer, the instructor, or peers. For example, Sanz & Morgan-Short (2004) used computers to deliver PI materials, though the mode of delivery was not the focus of their research design, it simply enabled reliable delivery of the feedback types. Lee & Benati (2007a) reported two studies (Lee *et al.*, 2007; Lee & Benati, 2007b) which explored the relative effectiveness of PI delivered by different modes (the classroom setting vs. computers). Lee *et al.* (2007, cited in Lee & Benati, 2007a) found that two PI groups, in which the teaching materials were delivered either by computers or by the paper-and-pencil format, all made improvement in the recognition task on learning Spanish, and that no difference was observed between these two groups. Also, Lee & Benati (2007b, cited in Lee & Benati, 2007a) concluded that there was no significantly instructional difference in PI delivered either in the regular classroom or by computers on the acquisition of the French and Italian subjunctives. Their results led them to conclude that PI is an effective instruction for learning both the French and Italian subjunctives, and that it does not matter how it is delivered, by computers or in normal classroom settings.

#### *2.3.1.4 Are the positive effects of PI studies attributable to the explicit information provided?*



As presented in 2.3.1.1, PI has been empirically demonstrated to be an effective pedagogical package in learning grammar. However, some may doubt that the positive effect of PI may be generated by the more explicit information<sup>25</sup> offered to its learners in comparison with other types of instruction. In order to investigate this issue, VanPatten & Oikarinen (1996) isolated the explicit information and structured input activities (SIA) in PI. Their results indicated that it was the SIA, as opposed to explicit information, that was responsible for learners' improved performance. The generalisability of VanPatten & Oikarinen's findings can be observed in some conceptually replicated studies (Benati, 2004a, 2004b; Farley, 2004b; Wong, 2004b). It appears that the role of explicit information in PI is not as beneficial as the SIA for language improvement. Furthermore, Benati (2004a) and Wong (2004b) found that the impact of the SIA alone on both interpretation and production tests is equivalent to that of a full PI, suggesting that the SIA *alone* would be necessary and sufficient to generate learners' improved language performance.

Furthermore, Sanz & Morgan-Short (2004) investigated the effectiveness of explicit information provided *before* (i.e. explicit grammar explanation) and *during* (i.e. explicit negative feedback<sup>26</sup>) the exposure to input by means of PI-based tasks. According to their findings, all of the groups made significant language improvement after the treatments. Importantly, the most 'implicit group', in which the learners received no explicit information about the targeted form *before* and *during* the course of PI but were

---

<sup>25</sup> Based on the related literature review of previous PI studies, explicit information is concerned with explicit grammar explanation given *prior* to the PI activities, and the feedback serving as the reminder of learners' defaulting to less effective processing strategies *during* the activities. Thus, explicit information refers here to information about how the targeted linguistic feature works (i.e. grammar rules and the processing strategies), provided *before* and *during* exposure to the input.

<sup>26</sup> According to Sanz & Morgan-Short (2004, p. 55-56), explicit negative feedback indicates learners' ineffective processing strategies and gives an explanation for the error. The implicit group, without provision of explicit negative feedback, is still offered feedback, but the feedback merely indicates whether their answer is correct or incorrect. No further explicit information is provided, such as metalinguistic grammar explanation.

just told whether their response was correct, showed equal learning gains in the interpretation and production tests compared with the other three groups, which had received either explicit grammar explanation or explicit negative feedback. One crucial implication from Sanz & Morgan-Short's study is that PI could give rise to learners' language improvement by the provision of the SIA *plus* the implicit feedback, as opposed to explicit feedback.

It is worth noting here that a common misinterpretation of this strand of studies (i.e. comparing explicit information vs. the SIA in Benati, 2004a; VanPatten & Oikkenon, 1996; Wong, 2004b and so on) is that explicit information is not necessary and not beneficial in an instruction. An appropriate interpretation should be that explicit information is favourable, perhaps even necessary, for some instruction but not for all (Benati, 2004a, p.217). It is also noted that although the structured-input-activities-only group (i.e. in Sanz & Morgan-Short's most 'implicit group' and previous PI studies) did not receive any explicit information before and during the treatment, the feedback provided to learners was rather explicit, at the very least in the referential activities, given that it indicated whether learners' response was correct or incorrect. As Terrell stated (1991, p.53), "the use of instructional strategies to draw the students' attention to, or focus on, form and/or structure" could be considered as explicit grammar instruction. Also, DeKeyser *et al.* (2002) pointed out that the feedback provided in PI was explicit, although explicit information was not given.

#### *2.3.1.5 What is the long-term effect of PI?*

Most previous PI studies have examined the immediate or short-term learning gains, usually within an interval of between three and four weeks. However, long-term delayed post-tests are recommended to investigate the impact of a specific instruction (Norris &

Ortega, 2000; Truscott, 1998). Truscott (1998) argued that whether or not long-term follow-up testing is introduced in a study is a valid criterion by which to evaluate the effectiveness of the intervention, because the immediate benefits, undoubtedly, can fade within a few months. To the best of my knowledge, there have been, so far, three PI studies (Marsden, 2006; VanPatten & Fernández, 2004; Xu, 2001) which have included a *long-term* delayed post-test, ranging over a few months.

Xu (2001) found that the language retention of PI on learning English wh-questions could be observed in a written production test, but not in an interpretation test in a five-month delayed post-test. Xu attributed the regression to some affective factors, such as low motivation and impatience to do the delayed post-test (p. 65-66).

VanPatten & Fernández (2004) used the same instructional package as was used in VanPatten & Cadierno (1993a). Their participants received a pre-test, an immediate post-test, and an eight-month delayed post-test to examine their learning gains with respect to Spanish subject pronouns and direct object pronouns. The results of VanPatten & Fernández's study indicated that learners' post-instructional learning gains were sustained over an eight-month period compared with the pre-test, in spite of a decline from the immediate post-test to the delayed post-test. The long-term effect of PI has also been corroborated in the study by Marsden (2006), in which L2 learners of French demonstrated their language retention in a delayed post-test taken 14-16 weeks after completion of the intervention. These studies have empirically confirmed that PI, aiming to alter learners' processing mechanisms, could have a long-term instructional impact, lasting for at least 4 months after the intervention. However, Collentine (2004) has commented that one of the questions remaining unclear regarding PI's effect is

whether “the learner’s developing system is responding differently to authentic input” (p.179).

#### *2.3.1.6 What is the effect of PI on the less-controlled oral production task?*

As PI proponents have claimed, one promising effect of PI is that PI can alter learners’ underlying developing language system so that PI learners can perform both interpretation and production tasks (VanPatten & Cadierno, 1993a). As VanPatten (2002a) commented, “altering the way learners process input can alter their developing systems” because the “processing group showed evidence of this on both interpretation and production tests” (p.771). Based on this review of previous PI literature, however, it has been found that the production tasks used in previous PI studies were rather controlled, and most of them were written production tests which required learners to write a single sentence. As Benati (2004a) suggested, “Further studies should be conducted, including different forms of assessment (e.g. timed tasks) that would reduce the ability to monitor” (p. 217).

To the best of my knowledge, there have been only six studies which have applied oral test(s) to investigate the effect of PI (Benati, 2001, 2004b; Erlam, 2003; Marsden, 2006; Salaberry, 1997; VanPatten & Sanz, 1995). The types of oral tests used in these six studies and the results are summarised in Table 2.1. Examination of Table 2.1 shows that the impact of PI on performing oral tasks was marginal (Erlam, 2003; Salaberry, 1997; VanPatten & Sanz, 1995) in that PI groups did not outperform the control group in the post-instructional oral task. Although Marsden (2006) concluded that PI had an impact on learners’ oral performance in one of her experimental schools, based on gains of the amalgamated oral scores (i.e. combining the scores in two types of oral task), between the pre-test and the post-test, and between the pre-test and the delayed post-

test, her results were “approaching statistical significance at the 90% confidence level,  $p = .105$  ( $p.537$ )” rather than the more rigorous threshold of the 95% confidence level, which is commonly accepted for the significance test (see Section 3.4.3 about setting the probability value). Furthermore, the difference between the pre-test and the delayed post-test in School 1 was of borderline statistical significance,  $p=.072$ . Accordingly, the finding that PI contributes to learners’ oral improvement in Marsden’s study is not considered to be completely valid. On the other hand, Benati (2001, 2004b) has found empirical evidence that PI is conducive to learners’ oral performance. On the whole, the answer to whether or not PI leads to learners’ improvement in oral performance is not certain and this issue needs further investigation.

Table 2.1

*A summary of PI studies including oral tests and their results*

Studies	Types of oral test	Results on oral test
VanPatten & Sanz (1995)	a. video retelling b. structured interview	a: $PI > C$ b: $PI = C$
Salaberry (1997)	video retelling narration	$TI = PI = C$ (pt and one month dpt)
Erlam (2003)	picture-based narration	pt: $MOI > C$ ; $PI = C$ dpt: $MOI = PI = C$
Benati (2001)	picture-based narration	pt: $PI = OI > C$ dpt: $PI = OI > C$
Marsden (2006)	a. picture-narration b. guided conversation	School 1: <u>PI group</u> (the results amalgamated a and b): pt > pre- test dpt > pre- test School 2 at the post-test <sup>27</sup> : $EnI = PI = C$
Benati (2004b)	an oral commentary-based production	$PI = SIA$ $PI > EI$ $SIA > EI$

Note: pt = post-test; dpt = delayed post-test; C=control group; TI=traditional grammar instruction; PI=processing instruction; OI=output-based instruction; MOI=meaningful output instruction;

<sup>27</sup>The delayed post-test was not conducted in School 2.

EI= explicit information only group; SIA=structured input activities only group

### *2.3.1.7 Can PI promote learners' implicit knowledge?*

R. Ellis (2005) argued that it is important to distinguish between learners' implicit and explicit knowledge of an L2, but there have been few empirical studies to examine this issue because researchers have failed to take these constructs into account (p.168).

There is, however, little comment in existing PI literature regarding whether PI leads to implicit knowledge after its treatments, and the two types of knowledge have not been clearly distinguished in previous PI studies.

PI proponents have claimed that PI affects learners' underlying language developing system, given that even though PI learners were only engaged in input-based activities, they could show improvement on the production tasks. In the words of VanPatten & Cadierno (1993b), "Theoretically, altering input processing should have a significant impact on changing the internalized knowledge" (p. 46-47). VanPatten (1994) stated that PI could assist in "building up an implicit knowledge of the language via intake facilitation" (p.34). According to the results on long-lasting effects of PI, VanPatten & Fernández (2004) argued that "other approaches may cause temporary performance improvement and the subsequent decline in performance may be due to the fact that the instruction did not affect the mechanisms used for processing and acquisition (e.g., DeKeyser & Sokalski, 1996). PI, however, deliberately attempts to affect the processing mechanisms"(p. 277). Their claim suggested associating the knowledge gained from PI with implicit knowledge, given that implicit knowledge is often considered to be less sensitive to corruption over time than explicit knowledge.

Some research is relevant to whether implicit learning can occur. For example, Jiménez & Méndez (1999) argued that implicit learning can occur if the target form is attended to sufficiently. Williams' empirical study (2005) concluded that FMCs can be made implicitly if the learners pay attention to both form and meaning but are not aware of their relevance to each other, if the target feature is part of the L1 system. In this scenario, it is possible to form FMCs implicitly. However, it is not clear whether it can occur in PI or not. In order to make the above issues more transparent, it is valuable to identify precisely what knowledge learners derive from PI activities. It is possible that the affective activities in PI have a role in promoting implicit knowledge. For example, Wong (2004a, p.44) claimed that the purpose of affective activities is to reinforce the representation of a form which has been established during referential activities. DeKeyser *et al.* (2002) also suggested that “the many examples in the structured input” (p.813) may interact with the learning which has occurred during referential activities. Furthermore, Marsden (2006, p.514-515) discussed the possibility that affective activities may promote implicit strengthening of a form, drawing on Schmidt's (1994, 2001) notion that once an initial mental representation is established, implicit reinforcement of this form can occur without the learner's consciously noticing.

PI studies have not, however, clearly and empirically demonstrated that PI could promote learners' implicit knowledge. As DeKeyser *et al.* (2002, p.819) have pointed out, “very little, if any, research on PI can even claim to address acquisition and not just the learning of monitored knowledge” (p.189). VanPatten & Oikkenon (1996) called for further research to examine whether learners develop some sort of conscious knowledge due to the interaction between structured input and feedback. Benati (2004b, p.217) also called for further study to include different forms of assessment (e.g., timed tasks) to reduce learners' ability to monitor the targeted language. In addition, de Jong (2005)

had reservations about PI's claim with regard to it being able to alter underlying implicit knowledge. De Jong (2005) pointed out that PI proponents have not demonstrated that PI learners' performance was "based on implicit knowledge, and it is also not likely that implicit knowledge was acquired. ... In all of these studies, there is a feasible alternative explanation that attributes the results to practice with explicit knowledge ... Therefore, it cannot be claimed that the same implicit knowledge was used in comprehension as well as production" (p.211).

Furthermore, DeKeyser *et al.* (2002, p.813) argued that learners must have engaged in explicit learning in the structured-input-only group (e.g., in VanPatten & Oikkenon's (1996) study), even though the explicit information was not given to them. DeKeyser *et al.* postulated the possibility for learners to figure out the grammatical rule from the feedback or the practice which they received during the PI treatments. DeKeyser *et al.* posited that learners may induce their explicit knowledge from PI and use this explicit knowledge to monitor their production of that knowledge during testing phases. However, DeKeyser *et al.* have not provided any empirical evidence to attest this. In short, the possibility that PI activities promote implicit knowledge has not been researched to date, and it is not clear whether affective activities can promote it, either alone or following referential activities.

#### *2.3.1.8 Do different PI activities have different instructional impacts?*

As noted in Section 2.1.1.2, the nature of these two types of PI activity is inherently different so that they might lead to different instructional impacts. Marsden postulated that affective activities are more beneficial for learning vocabulary than referential activities, as the focus of the two activities is quite different. Referential activities appear to channel learners' attention to the targeted feature and the affective activities



direct learners attention to the semantic meaning. Note that PI was created to assist in promoting learners' grammatical knowledge instead of lexical knowledge. The fully detailed description of vocabulary learning theory is beyond the scope of this thesis. In addition, it is argued that knowing a word requires a variety of types of knowledge to master it such as orthographical and phonological form, grammatical behaviour, collocation (Nation, 1990). However, this thesis is purely concerned with the acquisition of a word's referential meaning and its form, so other categories of lexical knowledge have not been investigated. The following short overview of lexical learning is thus limited in this scope.

In general, it is argued that there are two ways that vocabulary learning can take place in L2 classrooms, namely direct learning from explicit teaching (Nation & Waring, 1997; Schmitt & McCarthy, 1997), and indirect learning such as incidental learning (Hulstijn, Hollander, Greidanus, 1996) and making inferences/guessing from context (Sökmen, 1997). Nation and Waring (1997) argued that direct learning of a word could take place in a non-contextual way such as through the use of word cards. At a more general level, they argued that "the research evidence supporting the use of such an approach as one part of a vocabulary learning programme is strong" (p. 12), though it may not be in line with a communicative language learning approach. Indirect learning of a word can occur incidentally during input activities while learners' attention is on understanding the meaning of what they hear or read (Hulstijn et al. 1996, p.327) - where learning vocabulary is not the main focus (Nation & Waring, 1997) or where learners have to infer or guess the word meaning from context (Sökmen, 1997). However, Sökmen concluded that "guessing from context does not necessarily result in long-term retention" (p. 238).

In terms of vocabulary learning in the current study, it is noted that unfamiliar words learners encountered during the instructional period were given by glosses with their Chinese equivalent meaning and syntactic category (see Section 3.2.3.1). Therefore, learning words from contextual inference and guessing was unlikely to occur. In theory, learners were engaged in direct learning to some extent, given that the unfamiliar word was provided with the gloss, though learners did not learn vocabulary directly in a non-contextual fashion and they learnt words through activities. Learners could readily refer to the glosses if they wanted to. Additionally, it was possible for incidental vocabulary learning to occur inasmuch as learning vocabulary was not the main focus of these PI activities. Learners were never explicitly instructed to pay attention to vocabulary, though glosses of some words were provided. Furthermore, the degree or frequency with which learners looked up the glosses is unknown and it varied between individuals, in that learners carried out these activities at their own pace; no further investigation was conducted. Some might have used glosses all the time, while others might have resorted to them occasionally. In this sense, thus, incidental learning of a word was to a certain extent in this study.

There are two possibilities that may account for referential activities being less effective in promoting lexical knowledge than affective activities. Firstly, learners may not notice the presence of unfamiliar words and glosses, as they can merely rely on the inflection of a verb stem to complete the activities, without comprehending the meaning of the whole sentence. Secondly, learners may notice the unfamiliar words and their glosses, but they do not pay attention to them or they decide to ignore them, because mastery of unfamiliar words is not critical to undertaking the tasks once they have grasped the ‘-ed’ rule. On the other hand, learners in affective activities may benefit from the provision of glosses. In principle, the completion of an affective activity requires that learners

understand the whole sentence meaning in order to accomplish the tasks. Because of this, learners in affective activities are more likely to attend to unfamiliar words and to use the glosses compared with those in referential activities.

#### *2.3.1.9 How trustworthy are the results of previous PI studies?*

Although the results from previous studies appear to suggest that PI is effective in helping learners to learn a grammar, there are a number of issues which could potentially affect the interpretation of its effectiveness. The limitations of previous PI studies are discussed as follows:

1. The targeted language: most previous studies have involved Romance languages (e.g. French, Spanish, and Italian) (see Section 2.3.1.2). Apart from Romance languages, so far only two studies (Wu, 2003; Wu, 2001) have investigated PI's effectiveness on English. PI's effectiveness for non-Romance languages is therefore less well established.
2. The sample: samples have tended to involve learners in higher education except for Allen's (2000) high school students, Erlam's (2003) 14-year-old learners, Benati's (2005) 12-13 year olds, Marsden's (2006) 13-14 year olds, VanPatten & Oikkenon's (1996) secondary school learners, Wu's (2003) 16-year-olds, and Xu's (2001) 13-year-olds students.
3. The duration of the intervention: There are few examples in the literature of long interventions. For example, Collentine's (1998) intervention lasted a total of approximately 100 minutes, Erlam's (2003) about 135 minutes; Farley's (2001 & 2004a) about 90 minutes and 100 minutes respectively, VanPatten & Wong's (2004) about 45 minutes, and Xu's (2001) about 120 minutes.

4. A control group: there has often been no control group (For example, Benati, 2004a & 2005; Farley, 2001, 2004a & 2004b; Keating & Farley, 2008; Sanz & Morgan-Short, 2004; VanPatten & Fernández, 2004; VanPatten & Oikkenon, 1996; Wu, 2003).
5. The sample size: several studies have had small sample sizes, with fewer than 15 participants in each instructional group (e.g., Benati, 2001, 2004a, 2004b & 2005; Marsden, 2006; Morgan-Short & Bowden, 2006; Salaberry, 1997).
6. A delayed post-test: several studies have not included a delayed post-test to examine language retention beyond the intervention (e.g., Benati, 2004b, 2005; Collentine, 1998; Sanz & Morgan-Short, 2004; VanPatten & Oikkenon, 1996; VanPatten & Sanz, 1995; VanPatten & Wong, 2004; Wong, 2004b).
7. The type of elicitation tests: in terms of the measures used to assess learning gains, aural interpretation and written production tests have been commonly used in previous PI studies. However, most of them have been controlled sentence-level tests (e.g., listen to a sentence and then make a judgement, or fill a gap). Such measures have been criticised by SLA researchers on the ground that they tend to elicit explicit knowledge (see Marsden, 2004). Only a few studies have included a less controlled oral test to examine instructional impact (see Section 2.3.1.6). Moreover, even where an oral test has been used, investigating whether or not PI promotes learners' implicit knowledge has never been the focus of study.
8. The validity and reliability of elicitation tests: there has been little justification of the validity and reliability of measures used, except for Erlam's (2003) study reporting on the validity and reliability of her measures, and Marsden's (2006) study reporting on the comparability of two versions of measures.

9. The statistical tests for data analysis: most studies have employed parametric tests without justifying whether it is valid to do so, except for Marsden's (2006) study. If there is no attempt to justify the selection of a parametric or non-parametric test, the validity of results becomes questionable.
10. Effect size: reporting on effect size has been overlooked in previous PI studies, except for Marsden's study, though reporting it was strongly recommended by Norris and Ortega (2000).

The limitations of previous PI studies listed above might to some extent cast doubt on the trustworthiness of the results reported, but it is not the intention of the current study to argue that their results are totally untrustworthy. After all, there are inherent problems in carrying out a classroom-based experimental study which are difficult to overcome. For example, both the recruitment of participants and obtaining permission to carry out interventions are problematic, inasmuch as the interventions can interfere with participants' regular classes. Nevertheless, it is important to try and overcome limitations 1 – 10 (above) as far as possible in future PI studies, including the present one.

### ***2.3.2 The formation of the current study***

#### *2.3.2.1 Motivations*

##### *a) Innovation in grammar teaching in Taiwan is needed*

As described in Chapter 1, the introduction of PI in the Taiwanese context appears to satisfy the current needs of English education in Taiwan. The reasons why PI deserves specific attention in the Taiwanese context were given in Section 1.3. In brief, although PI is a grammatical instructional package which aims to promote learners' grammatical knowledge, PI design guidelines require that activities be meaning-bearing and communicative in accordance with the principles of CLT, which has been advocated by

the Taiwanese Ministry of Education. The desirable effects observed in PI studies (see Section 2.3) suggest that the introduction of PI is applicable as it could provide an alternative for language teachers in Taiwan. Moreover, PI not only meets the expectation of English teachers and students but also follows the policies of the concerned authorities.

*b) Motivation arising from unverified issues in previous PI studies*

According to the review of previous PI-based studies, what has been verified so far is that PI is an effective grammar instructional tool for a range of target features in different languages as discussed in 2.3.1.1 and 2.3.1.2. In addition, PI proponents have claimed that the main causative factor for learners' improved performance is the SIA (i.e. components two and three) rather than explicit information provided (See Section 2.3.1.4). Nevertheless, PI studies so far have regarded these two components (referential activities and affective activities) as one entity and have not separated them to examine their individual impacts. After all, these two types of structured input activity are rather different in nature (see Sections 2.1.1.1 and 2.1.1.2). According to her results comparing PI and EnI, Marsden (2006) suggests that referential activities may be more beneficial in helping learners with grammatical linguistic features than affective activities, and that affective activities might be more favourable to learning vocabulary than referential activities. A separate issue concerns to what extent PI could affect learners' developing language system, and promote their implicit knowledge (see Section 2.3.1.7).

These issues discussed above have not been researched yet. The purpose of this study is to identify the roles of these two types of PI activity, and to investigate the impact of PI activity in order to improve the understanding of this framework.

### 2.3.2.2 *The definition and operationalisation of implicit and explicit knowledge in the current study*

Although the design of measures to elicit implicit and explicit knowledge is difficult (R. Ellis, 2005), the two types of knowledge are distinguishable constructs (N. Ellis, 2005). Han & Ellis (1998) proposed two criteria to discriminate between implicit knowledge and explicit knowledge, namely accessibility and awareness:

“Implicit knowledge is easily accessed in tasks that call for fluent language performance. In contrast, explicit knowledge can be accessed only with controlled effort and, thus, is typically used in tasks that allow for careful planning and monitoring. Whereas implicit knowledge is unanalysed, and constantly held without awareness, explicit knowledge is analysed and model-based, and thus represents consciously held insights about language” (p.6).

This study, however, holds a slightly different perspective of the criterion of awareness, given that implicit knowledge may occur in a scenario in which language users are aware of the properties of a specific linguistic feature, and they can use it spontaneously *without* controlled effort in either input-based or output-based practice. This perspective corresponds to the strong interface position<sup>28</sup> (see DeKeyser, 2003). Based on the above, the definition of implicit knowledge in the current study is language behaviour which occurs *without* planning and monitoring that language, and though learners might have knowledge of a language, they do not consciously use it. On the other hand, the definition of explicit knowledge is users’ language behaviour which occurs *with* planning and monitoring of that language.

---

<sup>28</sup> According to DeKeyser (2003), the strong interface position states that “explicitly learned knowledge can become implicit in the sense that learners can lose awareness of its structure over time, and learners can become aware of the structure of implicit knowledge when attempting access” (p.315).

As for how these definitions are operationalised in the outcome measures in this study, the manifestation of explicit knowledge is operationalised as the significant performance of learners on a measure without a time constraint or learners' ability to verbalise the targeted grammatical rule immediately after taking a measure with a time constraint, according to R. Ellis' (2004) conceptualisation of L2 explicit knowledge being "generally accessible through controlled processing" (p.237) and "potentially verbalizable"(p.239). On the other hand, the manifestation of implicit knowledge is operationalised as learners' significant performance on a measure with a time constraint and subsequent self-reports of not using the targeted grammatical rule during the measure. The choice of outcome measures to elicit these two types of knowledge in this study will be discussed in the following section.

### *2.3.2.3 The choice of measures to elicit implicit and explicit knowledge*

Since the current study set out to explore the impact of the interventions on the development of two types of language learner knowledge (implicit and explicit knowledge), the criteria for choosing measures to elicit explicit and implicit knowledge are given here. Previous research has indicated that task demands and the time allowed can affect the use of these two types of knowledge (Bialystok, 1982; Butler, 2002; R. Ellis, 2004, 2005; DeKeyser, 2003; Purpura, 2004; Skehan, 1998). A test is assumed to measure explicit grammatical knowledge if planning time is granted (Purpura, 2004). Skehan (1998) expressed the view that "planning will predispose learners to try out 'cutting edge' language..., monitoring is more likely to be associated with greater accuracy (p.74)." Bialystok (1982) suggested that a written task would elicit learners' explicit knowledge as learners have time to spot errors and then to correct them. Thus,



for the current study a gap-fill test without time constraint was created, with the intention of measuring learners' explicit knowledge.

On the other hand, Bialystok (1982) suggested that an unplanned oral communication might tap into learners' implicit knowledge as it is less controlled (learners might not have sufficient time to ponder thoroughly), so that learners are less likely to employ their explicit knowledge. R. Ellis (2005) pointed out that when it comes to implicit knowledge, "spontaneous production tasks are probably the best means of elicitation, but, again we cannot be sure that learners do not access at least some explicit knowledge, especially when the task involves writing (p. 147)." R. Ellis (2005) also found empirical evidence that a timed grammaticality judgement test is valid for eliciting implicit knowledge. Roehr (2008) stated that "time pressure in combination with certain task types, for example tasks that focus learners' attention on meaning and require oral production, are likely to encourage the use of implicit knowledge" (p.191). Purpura (2004) stated that an extended-production task<sup>29</sup> (such as the structured conversation in this study) is hypothesised to measure implicit grammatical knowledge. Doughty (2004) pointed out that the effects of PI have largely been assessed using controlled tasks which required either comprehension or production of the targeted feature without a time constraint. However, VanPatten and Fernández (2004) noted, agreeing with Doughty (2004), "a task which pushes participants to produce language at the discourse level in a confined time period is the best indication that learners are tapping their underlying system and not a conscious knowledge source" (p.285). Based on the preceding suggestions, the oral tests were predominantly used to measure learners' 'productive' implicit knowledge, and a timed grammaticality judgement test

---

<sup>29</sup> According to Purpura (2004), "extended-production tasks present input in the form of a prompt instead of an item. The input can involve language and/or non-language information" (p.139).

(GJT) was adopted as a measure to assess participants' 'receptive' implicit knowledge of language for the current study.

The oral tests in this study were also used to maintain some parity with a communicative task in language testing, whereby a task should to some degree involve attaining a communicative goal in a real-life situation (Rea-Dickins, 1991; Robinson & Ross, 1996). However, the real-life context was not incorporated in the timed GJT and the gap-fill test, and these two assessments were less communicative compared with the oral tests.

In sum, the performance elicited from the measures used in this study encompassed three types of expected responses, following Purpura (2004, p.123), namely the selected-response task (i.e. the timed GJT), the limited-production task (i.e. the gap-fill test), and the extended-production task<sup>30</sup> (i.e. the structured conversation). It was hoped that the application of multi-faceted measures in the current study (e.g. different modalities (reading/ speaking/ writing), allotment of time, and expected responses) would help to collect more representative data and produce more informative clues to what learners can do with what they have learned as well as inform the researchers of the extent of the interventions' impact, as suggested by VanPatten & Sanz (1995).

Up to now, PI studies have not attempted to explore these two types of knowledge derived from PI activities, so the timed GJT and the retrospective self-report used in this study were an unprecedented step in investigating the impact of PI. Thus, a brief review

---

<sup>30</sup> According to Purpura (p.123-124, p.127), the selected-response task measures a test-taker's ability to recognise or recall the grammatical feature; the limited-production task requires the test-taker to speak or write from a word to a sentence; the extended-production task requires the test-taker to speak or write more than two sentences.

of the GJT and the self-report in the SLA are given in the following.

*a) A brief review of the GJT used in SLA*

The application of the GJT has provoked debates in SLA research with respect to measuring learners' language competence (Davies & Kaplan, 1998; Goss *et al.*, 1994; Gass, 1994; Mandell, 1999; Munnich *et al.*, 1994). Some studies have challenged the use of GJT in measuring learners' language competence (Davies & Kaplan, 1998; Johnson *et al.*, 1996; Munnich *et al.*, 1994). Davies & Kaplan (1998) reported that English adult L2 learners of French used different strategies (e.g. guessing or translation) when taking the GJT. In addition, Davies & Kaplan (1998, p.199) postulated that the strategies which learners employed to judge L2 sentences would get closer to those they used for L1 judgement with the increasing development of L2 proficiency. Munnich *et al.* (1994) pointed out that the GJT might not be a sensitive measure of a learners' developing linguistic ability, and they argued that GJT "does not elicit linguistic behavior per se but rather a response indicating the learner's belief about the L2 grammar" (p. 229). However, Munnich *et al.* (1994, p.239) suggested that an aural GJT might be a more sensitive tool than a written GJT for examining learners' developing language abilities. Johnson *et al.* (1996) found that adult learners of English performed inconsistently in the aural GJT. Goss *et al.* (1994) indicated that the problem of the GJT in measuring an individual's language competence is due to the problems in control of extra linguistic factors, such as L1 equivalents, or translation.

Ellis (1991, p.163), on the other hand, pointed out that the GJT allows researchers to investigate learners' linguistic 'competence' as some phenomena emerge either rarely or not at all which are not accessible to be observed in production data. Gass (1994) examined the reliability of GJT by means of a test-retest format delivered to the same

group with a one-week interval, and the results from the correlational analysis of the data showed that overall the participants' performances at two different testing times were consistent. Consequentially, Gass (1994, p.320) suggested that the GJT is a reliable measure in SLA research if used properly and appropriately. Mandell (1999) claimed that his results indirectly lent support to Gass' finding, suggesting that GJT is a reliable measure of learners' L2 competence through the correlational analysis of the test scores of a timed GJT and a "dehydrated" sentence test. Leow (1996, cited in Mandell, 1999, p.76) also found a significant relationship between learners' performance on the GJT and written and oral production tests. The brief review of literature regarding the use of the GJT above mainly focused on whether GJT is valid (Davis & Kaplan, 1998) and reliable (Gass, 1994; Johnson *et al.*, 1996; Mandell, 1999) in SLA research. Although the GJT has been criticised for having problems in the control of extra linguistic factors as described above, it is commonly used to measure the impact of specific language instruction in L2 research (e.g. de Jong, 2005; Doughty, 1991; Lightbown & Spada, 2000).

In most studies, the GJT used in L2 research is the 'standard' GJT, which is without time constraint (Davies & Kaplan, 1998). Bialystok (1979) suggested that the length of time allowed to respond in the GJT was one of the important factors related to learners' use of implicit or explicit knowledge because it dictated a response out of intuition or conscious analysis. Han & Ellis (1998) and R. Ellis (2005, 2007) suggested that a timed GJT could potentially tap into participants' implicit knowledge because it is likely to encourage the use of feeling or intuition, and to suppress the access to explicit knowledge. Nevertheless, Isemonger (2007) questioned the validity of using the timed GJT to measure implicit knowledge. Isemonger argued that the essence of GJT entails participants focusing on 'form' rather than on 'meaning', as the GJT invites participants

to judge the correctness of a sentence, which spontaneously dictates the use of explicit knowledge. Ellis & Loewen (2007), however, argued that “speakers of a language are perfectly able to decide whether a particular usage is grammatical without any explicit knowledge of the rule or feature involved. Indeed, the plethora of studies that have utilised GJTs have been based on precisely this assumption” (p. 124).

*b) A brief review of the retrospective self-report*

DeKeyser *et al.* (2002) suggested further research to “document learners’ subjective experience of the treatment received” (p.814). In the field of SLA, different formats of self-report have been adopted to explore if the learners draw on explicit knowledge. Green & Hecht (1992) and Hu (2002) employed a written rule verbalisation task, which required their participants to explain the grammar rule in a GJT. Robinson<sup>31</sup>(1995b) used a post-task questionnaire to explore learners’ subjective experiences after an intervention. Butler (2002) asked the learners the reasons for their choices through a structured interview. DeKeyser (1995) conducted an experiment to explore the implicit-inductive and explicit-deductive learning, and a retrospective self-report interview was conducted to explore learners’ subjective experience.

Given this previous use of retrospective self-report technique, in the current study a retrospective questionnaire was given immediately following the timed GJT at the post-tests. Also, a brief structured interview was carried out immediately following the oral tests at the post-tests to investigate the extent to which the participants were aware of having used explicit knowledge. In order to examine whether or not the implicit tests drew on explicit knowledge, the questions in the questionnaire and the interview asked

---

<sup>31</sup> Though Robinson did not set out to explore the issues of implicit and explicit knowledge/learning, his post-task questionnaire required learners to state the rule after being exposed to a task in order to investigate the issue of level of awareness.

participants to state whether or not they were using a grammatical rule to undertake these implicit tests.

There were some reservations about the validity of retrospective self-reports. Bialystock (1979) pointed out decades ago that learners' verbalisation of rules yields a conservative result regarding what they know explicitly. Learners may not be able to verbalise what they have experienced due to their lack of skills to verbalise it (R. Ellis, 2004; Hu, 2002) or due to forgetfulness. However, Butler (2002) found Japanese learners of English could explain their use of articles in the fill-in-the-article test. Green & Hecht (1992) asked the German learners of English to judge the ungrammatical sentences, to correct them, and then to provide the rules for the correction. Butler and Green & Hecht's studies indicated that learners were capable of verbalising the rules. Also, they indicated that learners' ability to correct the ungrammatical sentences was better than their ability to verbalise the rules. This may suggest that self-reports underestimate explicit knowledge.

An additional criticism of the retrospective self-report is that it is not sensitive enough to reflect subjects' on-line cognitive processes. However, the purpose of the post-task self-report in the current study was to explore learners' 'learning outcomes' (i.e. explicit or implicit knowledge) as opposed to 'learning processes' (i.e. explicit or implicit learning). Also, a concurrent self-report (e.g. a think-aloud task) would have increased access to explicit knowledge (R. Ellis, 2005) and cannot in any case be done during a timed assessment. The self-report conducted after the timed GJT and oral tests in this study was a supplementary measure to inform us about the nature of the knowledge drawn on in these tests.

#### *2.3.2.4 The choice of linguistic feature for the current study*

One important criterion affecting the choice of a linguistic form for instruction is whether this form is problematic for learners. Ellis (1993) argued that a choice of linguistic forms to teach can be derived from observing whether or not learners have difficulties in producing them (for example, in writing or speaking). From my past English teaching experience in Taiwan, I have found that Chinese L2 learners of English have problems in the use of the English regular past tense, namely the ‘-ed’ ending. One possible explanation for this is that Chinese does not have inflections, and always relies on temporal adverbs, word order, or context to indicate time relations (Chang, 2001, p.315). Benati (2005) also pointed out that Chinese L2 learners of English might have difficulty in attaching pastness as “they may borrow the concept of past tense in their L1 as the starting point” (p.76). The other possible explanation is that the ‘-ed’ feature is lacking in ‘perceptual saliency’, because it is always attached to the end of a verb stem. As a result, Chinese L2 learners of English may have serious problems in dealing with English tenses and aspects.

In addition, IP theory predicts that this targeted feature will be problematic for L2 learners as follows. The ‘Lexical Preference Principle’ in IP theory predicates that learners tend to process input for lexical items before they process for the targeted form if both encode the same meaning (VanPatten, 1996, 2004). It is predictable that learners may have problems in processing the targeted ‘-ed’ form. In a sentence such as ‘I walked to school yesterday’, although both the lexical item ‘yesterday’ and the ‘-ed’ verb ending communicate the past tense, the Lexical Preference Principle of IP predicts that the learners may prefer to rely initially on the adverb ‘yesterday’ as the indication of pastness in the sentence. In this scenario, the learners may not notice the ‘-ed’ attached to the end of the verb, or if they do notice it, they might not give it any

significance. Accordingly, the L2 learners of English would not process the targeted ‘-ed’ form if they were left alone to process the input, and if the input they encountered was not structured. (see Section 2.2.1.3, principle b).

Another reason why the English past tense ‘-ed’ feature was chosen as the targeted feature for the current study was that it allowed the control of participants’ outside exposure to be taken into account. As Mackey & Gass (2005) stressed, “one aspect that all researchers grapple with in second language research is how to control for outside exposure to the language” (p.148). The specific targeted feature (the English regular past tense ‘-ed’ form) selected for this study was not integrated into the scheduled English syllabus at the participating school and had not been on previous year’s curricula. This study introduced the targeted feature about one year earlier than scheduled in the normal English syllabus plan (they were supposed to learn the targeted feature in grade 7 in the second semester). Although many students are sent to private cramming schools after school, these schools normally reinforce the school curriculum. Therefore, as the target feature was not on the school curriculum, this reduced the likelihood of it being taught in the cram school.

#### *2.3.2.5 The research questions and hypotheses*

The purpose of the current study was to investigate the roles of two types of structured input activity within PI by isolating them, partly building on Marsden’s studies (2004, 2006), in order to investigate the impact of referential and affective activities on learning a grammatical form and on vocabulary learning. These issues have not been empirically investigated to date. The current study also set out to investigate the impact of PI on less-controlled tasks, given that the results of PI’s impact on learners’ oral performance to date have been mixed (see Section 2.3.1.6). The issue regarding to what



extent PI activities could affect learners' underlying language system was not clear (see Section 2.3.1.7). PI proponents tended to suggest that PI can affect learners' underlying language system i.e. promoting implicit knowledge (VanPatten & Cadierno, 1993b; VanPatten, 1994; VanPatten & Fernández, 2004). However, this claim has been challenged by some researchers (DeKeyser *et al.*, 2002; de Jong, 2005; Ellis *et al.* 2009, p.340 for similar interpretations). Furthermore, the current study was also concerned with the retention of knowledge following PI in a delayed post-test. Consequently, research questions (RQ) and the hypotheses (H) which guided the design of the current study are as follows:

RQ 1: Are referential activities more beneficial for learning the English past tense '-ed' feature than affective activities in a timed Grammaticality Judgement Test (GJT)?

RQ 2: Are referential activities more beneficial for learning the English past tense '-ed' feature than affective activities in a gap-fill test?

RQ 3: Are referential activities more beneficial for learning the English past tense '-ed' feature than affective activities in a picture-based narration test?

RQ 4: Are referential activities more beneficial for learning the English past tense '-ed' feature than affective activities in a structured conversation?

H1: Referential activities are beneficial for twelve-year-old L1 Chinese learners' interpretation and production of the English regular past tense.

RQ 5: Are affective activities more conducive to learning vocabulary than referential activities?

H2: Affective activities lead to more vocabulary learning than referential activities in twelve-year-old L1 Chinese learners.

RQ 6: What kind of knowledge do the four tests (i.e. the timed GJT, the gap-fill test, the picture-based narration test, and the structured conversation) tap into and what is the relationship between this knowledge and the intervention type that the learners received?

H3: The gap-fill test without a time constraint will elicit explicit knowledge and the other three tests with a time constraint will elicit implicit knowledge.

H4: Referential activities will promote learners' explicit knowledge of the English past tense '-ed' feature.

H5: Affective activities, either alone or following referential activities, will promote learners' implicit knowledge of the English past tense '-ed' feature.

RQ 7: Are PI learners' improved performances retained in a delayed post-test six weeks after the instruction?

H6: The effect of PI on twelve-year-old L1 Chinese learners' ability to interpret and produce the English regular past tense will be retained beyond the time of instruction.

## **Chapter 3 The methodological issues and design of the current study**

### **Introduction**

This chapter is a review of the literature on the methodological issues relating to the current study. This chapter is divided into six sections. The first section will focus on presenting the research approach utilised in this study, justifying why this approach has been chosen and clarifying what challenges may be encountered when using this research approach. The second section will briefly discuss the current study, including the participants, the design and implementation of the intervention and instructional materials. The third section will be chiefly concerned with the achievement assessments used for this study, including the design and implementation of the tests, and the scoring procedure. The fourth section will describe a variety of statistical procedures applied for analysing the achievement assessments. The statistical procedures involve parametric and non-parametric tests to compare the mean scores across experimental groups, the computation of the effect size to measure the magnitude of interventions, the correlation, and the principal component analysis. The fifth section describes the validity and reliability of the achievement assessments. As the delivery of the achievement assessments was a split-block design and each achievement assessment had two versions, the comparability of the two versions of the achievement assessments will also be reported in this section. This chapter will close by acknowledging some limitations of the achievement tests and design in order to avoid undue interpretations.

### **3.1 A review of literature on carrying out an experiment in educational research**

#### ***3.1.1 The methodological issues***

##### ***3.1.1.1 Why a classroom-based quasi-experimental study?***

The choice of a specific approach among the various research approaches greatly

depends on a researcher's field and his/her research questions. No matter what research approach is decided upon, "fitness for purpose is the key" (Gorard, 2002a, p. 354). The underlying characteristic of an experimental design is that researchers "deliberately control and manipulate the conditions which determine the events in which they are interested (Cohen *et al.*, 2000, p. 211)". Because this current study is interested in identifying the impact of two types of teaching activity (i.e. the referential and affective activity in the same pedagogical package, namely PI), the researcher believes that undertaking an experiment is the appropriate way for this study to answer the research questions.

Apart from the suitability consideration, one advantage of undertaking an educational experiment for this study is that the findings can be used to test a theory (i.e. IP) as well as being useful to the investigation of the pedagogical package (i.e. PI). As Marsden (2007) argues, "to test a specific learning theory, the multi-faceted nature of pedagogical packages can be problematic (p. 572)". As argued earlier, only the referential activity seems to adhere to the FMC theory. The affective activity has seemed not to underpin the theory in that it possibly fails to channel learners' attention to the targeted form. However, previous PI studies have never made an attempt to isolate them. An experiment is an appropriate method to isolate and investigate these two types of structured input activity.

*3.1.1.2 Some challenges faced when conducting an experiment and how this study would handle them*

*a) Educational experiments are artificial and fallible*

Conducting experiments in the field of social science is different from those in science (Cohen *et al.*, 2000). In the field of social science, it is inherently unlikely that

replicating an experiment will generate exactly the same results, since it is impossible to completely control all of the variables relevant to subjects. However, this is not to suggest that doing educational experiments can offer nothing so that we should discard the value of the experimental design; there are methods that researchers can adopt to help counter these problems. For example, researchers can make an effort to identify and control the extraneous variables in order to reduce their effect. Researchers should give detailed information in their report for those who may replicate their studies (Cohen *et al.*, 2000; Norris & Ortega, 2000; Torgerson & Torgerson, 2003a). Also, researchers should strive to avoid excessive inferences from, and false interpretation of, their research results. In addition, a range of methods can be employed to yield deeper insight, such as carrying out a battery of tests, delivering questionnaires, or conducting interviews.

*b) Can we do an experiment when the theory underpinning it is too naïve?*

Another criticism of experiments is that the theory underpinning the intervention may be too naïve (Cobb *et al.*, 2003; Moore, 2002; Trochim, 1998). Moore (2002) suggested that experiments should be based on developed theory and that the researchers should be fairly sure of what the findings will be in order to avoid wasting resources. However, some researchers have argued that it is unnecessary to fully understand a theory and be able to predict the results before undertaking an experiment, given that the main purpose of doing an experiment is to figure out something that we do not understand (Scriven, 1998; Tymms & Fitz-Gibbon, 2002). Tymms & Fitz-Gibbon (2002) pointed out that sometimes what drives a researcher into his/her study could just be a hunch. Tymms & Fitz-Gibbon also argued that researchers may face some ethical difficulties (e.g. how to allocate their participants) if they have some predictions before undertaking their studies. In this case, the operationalisation and results of the studies may not be

objective if the researchers are sure of what their results will be.

As for this study, the principle (i.e. the lexical preference) of the theory underpinning the experimental design has been successfully manipulated and tested in previous PI studies (e.g. Benati, 2005; Marsden, 2006). Although this study was based on a developed IP theory, the predictions of the research questions were not so clear-cut, since they have not yet been researched.

### *c) Extraneous variables in experiments*

It is difficult to control or eliminate confounding variables in experiments with human participants (Moore, 2002; Torgerson & Torgerson, 2001, 2003a & 2003b). In theory, objective and informative experimental results depend on the rigorous control of the experiment's validity (Hammersley, 2001; Mackey & Gass, 2005; Torgerson & Torgerson, 2003a;). A range of types of validity are usually seen to be the most critical elements of a good experiment. Internal validity refers to "the extent to which the results of a study are a function of the factor that the researcher intends" (Mackey & Gass, 2005, p.109). In other words, internal validity is how truly the results are attributable to the interventions and not to other potential variables. External validity refers to "the implications that go beyond the confines of the research setting and participants" (Mackey & Gass, 2005, p.119). Also, it refers to whether the experiment is reasonably realistic. If an experiment is so controlled and artificial, its findings will have no real meaning for practitioners (i.e. low external validity). It has been argued that high internal validity is a prerequisite for external validity (Mackey & Gass, 2005; Torgerson & Torgerson, 2003b; Campbell & Stanley, 1963; Pilliner, 1973, cited in Cohen *et al.*, 2000, p.128; Trochim, 2006). Based on the above perspective, ensuring internal validity is essential in an experimental study.

One of the requirements for strong validity in an experiment is the “standardized and uniform delivery of interventions to all participants” (Gorard *et al.*, 2004, p.584; Moore, 2002; Moore *et al.*, 2003). Sanz (2000) states, “delivering treatments and testing components of experimental studies via computer allows for tighter control of individual and environmental variables as well as finer measures of the effects of treatment” (p. 27). In addition, some researchers have called for the careful operationalisation of the variables, such as the instructors and feedback provided (Marsden, 2006; Morgan-Short & Bowden, 2006; Sanz & Morgan-Short, 2004). Morgan-Short & Bowden claimed that having different instructors taking part in an experiment could be a confounding variable due to the idiosyncrasies each instructor has. The adverse effect can be reduced by swapping over the different instructors who take part in the interventions (Marsden, 2006). Additionally, when collecting and interpreting the data, researchers normally assume that all participants make their best efforts in the interventions (Mackey & Gass, 2005). However, it is possible that some participants’ reactions or interactions with the interventions might be affected by omission or inattention due to the feedback provided. These ignored or inattentive participants would not reap benefits from the feedback. Thus, the validity of the study (i.e. assuming the all participants make their best efforts) could be adversely affected by participants’ inattention.

The computer was introduced in the current study to deliver different types of intervention and feedback, with the purpose of increasing internal validity, by standardising the intervention and feedback and thereby reducing bias. Learners simply interacted with the computer, which could prevent them from producing the targeted feature accidentally in the course of interacting with peers or an instructor. It needs to be clarified that although the use of computer-based delivery of instruction is not

ubiquitously seen in formal language teaching and learning settings, it is believed to be a feasible approach to control some confounding variables, to establish the internal validity of this study, and is best suited to answer the research questions.

### ***3.1.2 Ethical considerations***

#### *3.1.2.1 Can the results of an education experiment inform practice?*

One issue raised regarding the main contribution of educational research is whether the results of research can be used to inform practice (for a review of this debate see Marsden, 2007; Gorard, 2002b). One crucial problem is that practitioners do not understand the research findings, because not every practitioner is trained to interpret the results of elaborate statistical analysis. Thus, Torgerson & Torgerson (2003a, p.75) encourage researchers to incorporate the concept of the “Numbers Needed to Teach” (NNT) into their research report. The NNT is a concept in which the effect of the intervention is expressed in a way that is easier for teachers to understand. For example, a researcher concluded that a given English teaching approach has improved the learning scores of 5% of the participants. As a result, the NNT is 20, which means that giving 20 students this specific teaching approach would result in one more student attaining a better grade. Torgerson & Torgerson argue that incorporating NNT into research reports helps practitioners to decide whether a given approach is worthy of practice.

#### *3.1.2.2 Should the practising teachers be involved in this study?*

Due to the ‘research informing practice’ consideration, it has been suggested that teachers should be involved in the research (for discussion see Marsden, 2007; Hiebert *et al.*, 2002). Given that teachers are working in the “frontline” and are in the position of deciding how to teach, they are therefore those who most understand learners’



difficulties in learning, and those who need to receive and be inspired by a variety of teaching approaches. The benefit of practising teachers' involvement is that a specific teaching approach can be introduced to teachers where it may increase the opportunities for teachers to accept it and then use it. Also, the practising teacher might give different viewpoints of the interventions due to their teaching experience. For example, Marsden (2004) reported that the practising teacher in her study suggested the affective activities may be more beneficial for learners learning vocabulary than learning grammar. In this respect, it seems useful to get some valuable viewpoints from teachers with respect to the feasibility and usefulness of research.

However, some researchers have cautioned against the involvement of the practitioner or political control in research in that they may have an unexpected impact on the quality of the research (Gorard, 2002b; Marsden, 2007). It is possible that the involvement of teachers may put the study's validity under threat because teachers may 'contaminate' the process of delivering the intervention and collecting the outcome data by not sticking to the study protocol owing to personal belief, lack of knowledge or regulation about a given intervention, and so forth. Consequently, a researcher has to think about whether or not to invite the practitioners to take part in his/her research before embarking on it, as they could influence the internal validity of the study.

After taking account of afore-mentioned points, I decided not to invite participants' regular English teacher to get involved in the current study. She was not informed of the targeted feature through the interventional and testing periods, given that she might have 'contaminated' this study if she had known what the intervention was intended to assess. For example, it is possible that the teacher may remind learners of what they should pay attention to in the intervention. In this sense, the participants' awareness of the targeted feature is raised and the chance of participants' exposure to the target

feature outside the intervention is increased. As a result, the teacher did not take part in this study. However, a succinct report about the current study and its findings was promised to the teacher and the participating school after the completion of this study. This is so that the teacher and school can adopt PI in the future if beneficial effects are found.

#### *3.1.2.3 Ethical considerations regarding the interpretation of the findings*

Clearly, different tasks or interventions have different educational objectives. In this sense, it is unethical to entirely dispute the potential effectiveness of a comparison treatment if its results are not as effective, on one particular measure, as those produced by the experimental group. For example, the results of Marsden's study (2006) concluded that the PI group outperformed the Enriched group in the acquisition of French verb inflections in the perfect and present tense. However, she argued that the value of enriched input-based intervention should not be refuted as it may be beneficial for learners in acquiring vocabulary. In this sense, a researcher should carefully interpret the findings of a piece of experimental research and avoid making an all-or-nothing judgment.

#### *3.1.2.4 The privacy of research participants*

A vital ethical consideration with regard to participants in an experimental study is not to place them in a situation where there is a risk of harm (Trochim, 2006). In order to prevent any potential harm to the participants, Trochim argued that the privacy of participants should be taken seriously. Trochim suggested two principles in the effort to achieve this: a) anonymity b) confidentiality. The former principle is stricter than the latter. Trochim suggested that participants' anonymity should be guaranteed from the beginning to the end of the research, even to the researchers. However, Trochim pointed

out that the principle of anonymity is sometimes difficult to achieve, particularly where participants have to be assessed at different time points. The latter principle assures participants that any information related to their identity will not be revealed to anybody who does not directly partake in this research.

However, it is noted that the principle of anonymity was not adhered to in the current study due to the following practical reasons. The researcher was the instructor during the instructional and testing phases, and administered the post-tests. In addition, the achievement assessments and the exit questionnaire were not filled in anonymously due to the need to match up participants' responses in questionnaires and their performances on achievement assessments. On the other hand, confidentiality was guaranteed. The participants were informed that only the researcher (myself) and my supervisor could directly examine their performances on the achievement tests. Any information regarding their participation in this study would be kept confidential. In any case if it were necessary to reveal their information or performance to a third party, their identifying information would be anonymous.

#### *3.1.2.5 The right of the control group to be treated equally*

This issue concerns a person's right to service (Trochim, 2001). The inclusion of a control group as a comparison to the experimental groups has been encouraged (Norris & Ortega, 2000). The degree of a control group's involvement in a study depends on the research design. Some control groups are engaged in a treatment but in a different way from the experimental groups; some are not engaged at all. When a piece of research requires the use of a non-instructional control group, the instruction(s) to other experimental groups may have beneficial effects. In this circumstance, the non-instructional control group may be denied the right of equal access to the instruction.

This ethical consideration has inevitably arisen in the current study, and so several measures to address this issue were taken, including obtaining the consent of the headmaster of the participating school (Appendix 6). Additionally, delayed access to the intervention was allowed for all the participants. The training materials were offered to the school after the end of this study so that the participants could get access to them if they wished to.

### **3.2 The current study: a quasi-experimental design**

#### ***3.2.1 The participants and the educational context***

##### *a) Background of the participants and the participating school*

Four classes (grade 6) of a primary school in Taitung, Taiwan, were recruited for this study from the end of February to the middle of May 2007. It is noted that the participants in the current study were younger (12-year-old primary school students) than those in any previous PI studies. Most of the participants in other PI studies were adult learners, recruited from universities, except for the following seven studies: Erlam's (2003) school-aged learners were about 14 years old, Benati's (2005) 12-13 years old, Marsden's (2006) 13-14 years old, VanPatten & Oikkenon's (1996) secondary school students, Wu's (2003) 16 years old, Xu's (2001) 13 years old, and Allen's (2000) high school students. The participants in this study had been learning English for at least three years prior to the intervention, having started to take formal English lessons at school once (40 minutes a lesson) a week in grade 3 and grade 4. The frequency of English lessons increased to twice a week in grade 5 and 6. Prior to the intervention, the participants had received about 115 hours of English instruction in a formal school setting. They were classified as English beginning learners<sup>32</sup> in a foreign-language

---

<sup>32</sup> Norris & Ortega (2000, p. 454) called for reporting on the initial proficiency level of learners such as scores on TOEFL or IELTS. However, the participants in this study had not taken any English proficiency test. The classification as 'English beginner' is based on the English syllabus at school (participants had

instructional setting in this study. All participants in the four classes had been instructed by the same English teacher since grade 3.

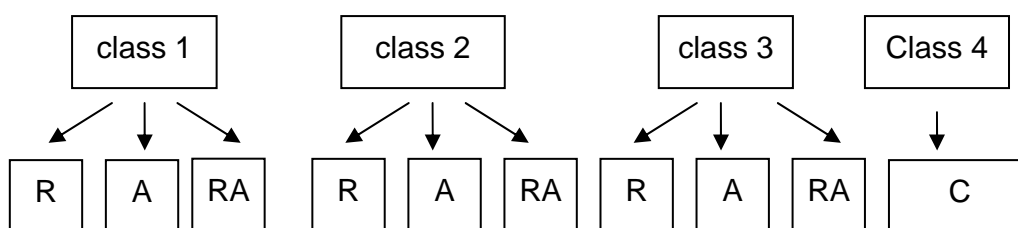
In theory, the four classes contained mixed-ability students since the students were randomly assigned by the school to the four classes, when they were enrolled at the beginning of grade 1. There were two steps in the participating primary school's procedure for assigning the newcomers. The first step was to distribute students into different sub-groups according to their parents' occupations. The second step was to randomly allocate individuals within each sub-group evenly into four classes using the computer. The purpose of the first step was to avoid accidentally distributing many students with similar family background (i.e., socio-economic status) into the same class. The students were supposed to stay in the same class from grade 1 to grade 6, and no further re-grouping would be carried out as a result of their academic achievement at school.

*b) The allocation of participants to interventions*

Because of the school's policy and regular teaching program, it was not possible to randomise all of the participants into the four groups and thus the students were required to remain intact in their regular classes during the intervention time slots. However, as this study was designed to deliver the instructional materials via computers, split-class design was achievable. In the current study students remained in their intact class, allocated into different experimental groups, and received different instructional interventions at the same time in the same class. Figure 3.1 shows the experimental design of the study during the interventional period.

---

been learning interrogative sentences in the present tense at school during the instructional phases) and their amount of exposure to English at school.



\* RA=group of referential + affective activities; R=group of referential activities only;  
 A=group of affective activities only; C= control group

Figure 3.1. The experimental design

The control group was randomly chosen out of the four classes. Thus, the control group in this study is a ‘non-equivalent control group’ (Cohen *et al.*, 2000, p. 215)<sup>33</sup>. The students in each of the remaining three intact classes were placed in rank order and then assigned to the three experimental groups based on their combined pre-test scores in the grammaticality judgment test and gap-fill test. The allocation of the participants into three experimental groups was as follows: the highest scorer was allocated to the RA group, the second highest scorer to the R group, and the third highest scorer to the A group, then the fourth highest scorer allocated to the A group, the fifth highest scorer to the R group, the sixth highest scorer to the RA group and so on. The process continued going back and forth: RA–R–A–A–R–RA. The second class was allocated according to the pattern of R–A–RA–RA–A–R, and the final class A–RA–R–R–RA–A. The participant allocation procedure for this study was adopted in order to create well-matched experimental groups, to ensure that each group had a similar prior knowledge of the target feature, and to reduce the possibility that any differences shown between the groups in the post-tests could be attributed to the initial imparity of the groups.

<sup>33</sup> Although the participants in the control group were not allocated in the same way as the instructional groups, a pre-test was administered to ensure the equivalent language knowledge about the targeted feature (see the pre-test results presented in Chapter 4).

It is noted that the control group in this study was not a ‘true’ control group in that the control group was an intact class (i.e. the participants were not individually randomly allocated to the control group). As a result, it is acknowledged that the current study was a ‘quasi-experimental design’ instead of an ‘experimental design’, given that the allocation of participants was not fully randomised. In brief, the most significant difference between these two types of study is the lack of random selection in the quasi-experimental study (Campbell & Stanley, 1963; Cohen *et al.*, p.212; Mackey & Gass, 2005, p.146). Although true experiments are more desirable than quasi-experiments, it requires a large number of samples for the randomisation to have an effect. Otherwise, chance can make the groups unequal. However, most educational research studies in educational settings are quasi-experiments rather than true experiments, given that randomisation of the participants in educational settings, particularly in classroom experiments, is not easily achievable.

*c) The sample sizes*

One issue worthy of noting is attrition. Only those who took part in all the phases of the intervention (i.e. instructional sessions and assessments) were included in the final data pool. Initially, a total of 136 participants (average age 12) were recruited for this study. However, sixteen participants were excluded from the final data pool: two being absent at either the post-test or the delayed post-test; one with a learning disability; thirteen participants being identified as the outliers<sup>34</sup>, given that they scored relatively high at the pre-test compared to other participants. The raw scores of the excluded 13 participants are given in Appendix 7. Consequently, one hundred and twenty participants were included in the final data pool: 31 in the RA group, 29 in the R group,

---

<sup>34</sup> The approach adopted by this study for checking the outliers was to inspect the Boxplot produced via SPSS (see Pallant, 2007, p.62-63). Note that only those who were identified as outliers in *both* the timed GJT and the gap-fill test at the pre-test were removed.

30 in the A group, and 30 in the control group. The results of the K-S test, the histograms, and the boxplots including the 13 outliers are provided in Appendix 40, so that readers can check comparability of the inclusion and exclusion of the outliers. Due to space limitations, only those of the timed GJT are presented.

*d) The pilot study*

The teaching materials and achievement assessments were piloted six weeks before the formal commencement of this study (i.e. the administration of the pre-test). A total of 13 participants, the same grade (i.e. grade 6) as those participating in the main study, were invited to partake in the pilot study. Note that the 13 participants did not take part in the main study, nor were they involved in the examination of the validity and reliability of achievement assessments (see Section 3.5). Initially, the piloting school agreed to spare 5 sessions (40 minutes a session) to assist in the pilot study in two consecutive weeks, but only 3 sessions within a week were allowed in the end. Due to the time limitation, the pilot study could only examine the feasibility of instructional materials and achievement assessments. Therefore, the participants in the pilot study did not go through all of the 4-session instructional materials, and only half of the teaching material for each intervention was piloted. The adjustments of teaching materials and achievement assessments as a result of the pilot were as follows:

- i) Some subjects' interactions with computers were interrupted by the unknown vocabulary appearing in the instructional materials. These participants kept raising their hands and asking its meaning, which might have affected others' concentration on the tasks. Therefore, a decision was made to provide the Chinese equivalent meaning of unfamiliar vocabulary on the same slide along with the instructional items.



- ii) Some introductions of the topic of the activities seemed to be lengthy. One subject reflected that he lost the patience to go over them. Therefore, the introductions of the activities were revised and made as simple and short as possible. The modified versions were checked by a primary school teacher to ensure the phrases and sentences used in the introductions were appropriate for the expected participants.
- iii) In the pilot study, the starting time of the timed GJT was not synchronised because participants could decide their start time simply by clicking a mouse. However, it was found that two subjects went back to check previous test items as they knew how to operate the Microsoft PowerPoint software. Consequently, a broadcast system was used to deliver the timed GJT in sync, to ensure the equality of the timed GJT for participants (i.e. everyone started and finished this test at the same time, and no one had extra exposure to the test items).
- iv) After taking the gap-fill tests, three participants complained that the gap-fill test was overburdened due to too many test items. As a result, the original 20 test items were cut to 15 in the final, revised gap-fill test.
- v) The teaching material was delivered by means of computers in a computer laboratory, where access to the internet was usually possible. A school teacher suggested disconnecting the internet during instructional phases in case the accessibility of the internet distracted participants' attention.

### ***3.2.2. The interventional procedures***

The experiment was conducted in the participants' regular computer laboratory during regular class hours, which were scheduled for computer lessons rather than English lessons. The same instructor – the researcher – carried out all of the interventions during the instructional period, and she was not the participants' regular classroom instructor.

The regular classroom instructor was also present during the interventions in accordance with the participating school's regulations. The instructor did not actually get involved in the intervention – he was not informed of the targeted feature at all. Although he was present during the intervention, he mainly turned up to maintain classroom discipline. In addition, a teaching assistant was present, who acted as a facilitator during the interventions, helping to prepare the facilities and solve any problems participants had when using the computers.

The timetable of the current study is attached in Appendix 8. Figure 3.2 below was the interventional procedure for this study. The timescale of the intervention and assessments is given in Table 3.1.

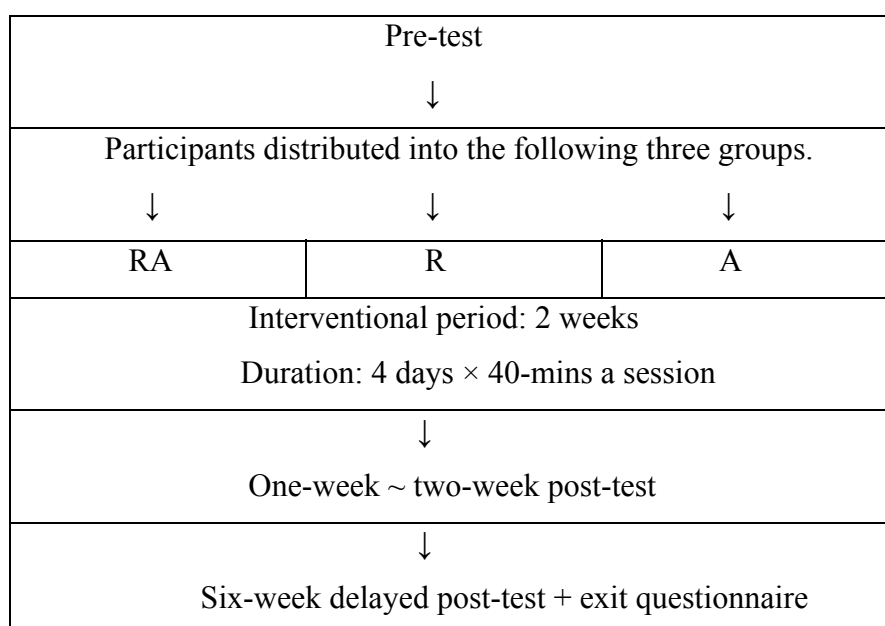


Figure 3.2. The interventional procedure

Table 3.1

*The timescale of the intervention and assessments*

	<b>Pre-test</b>	<b>Intervention</b>	<b>Post-test</b>	<b>Normal instruction</b>	<b>Delayed post-test</b>
<b>Duration</b>	2 weeks	2 weeks	2 weeks	4 weeks	2 weeks
<b>Cumulative</b>					

<b>length</b>	2 weeks	4 weeks	6 weeks	10 weeks	12 weeks
---------------	---------	---------	---------	----------	----------

A pre-test was carried out approximately two weeks before the interventions. The duration of the intervention itself was about 2.5 hours (40 minutes for each of the 4 classes) in two consecutive weeks. It is acknowledged that the duration of the treatments is rather short according to Norris & Ortega's (2000) meta-analysis of L2 instruction. Norris & Ortega concluded that the length of instruction, whether it is short (1-2 hrs), medium (3-6 hrs), or long (over 7 hrs), did not influence the observed instructional effectiveness. A post-test was administered a week after the interventions to examine the impact of the interventions. In addition, a delayed post-test was conducted 6 weeks after the interventions, because a delayed post-test has been recommended in order to measure the retention of an instructional impact (Mackey & Gass, 2005; VanPatten & Sanz, 1995). Mackey & Gass (2005, p.148) suggested that another way to prevent the risk of additional exposure was to avoid long intervals between the testing sessions. Although the long-term follow-up post-test was recommended for the examination of an intervention's impact, as discussed above in Section 2.3.1.5, the six-week delayed post-test is considered a good compromise.

An exit questionnaire was also delivered along with the delayed post-test. The exit questionnaire was intended to identify potential extraneous variables, and it consisted of two sections: a) participants' language learning background, b) participants' attitudes towards the intervention. Section one focused on the collection of data germane to participants' English learning backgrounds, such as experience of staying in any English-speaking countries before the intervention, and their extra English exposure outside the classroom during the interventional period. Section two aimed to collect participants' attitudinal data to probe whether there was any attitudinal difference

observed between the interventions, given that participants' attitudes towards the intervention (e.g. the level of motivation) might be a potential variable affecting their learning gains (Marsden, 2004). The attitudinal questionnaire adopted a two-option rating scale (yes or no). No neutral option was provided as the intention was to compel them to manifest their preference. Note that the exit questionnaire was not filled in anonymously due to the wish to match up participants' performance on achievement tests. The exit questionnaire regarding participants' English learning background is attached in Appendix 9. The attitudinal questionnaire is given in Appendix 10. The analysis of the exit questionnaire will be discussed in Chapter 5.

During the post-test and the delayed post-test, the oral assessment was carried out earlier than other achievement assessments. The order of the tests was: oral tests, the timed Grammaticality Judgment Test (GJT), the written gap-fill test, and the vocabulary test. However, at pre-test the oral test was administered a week after other tests (though the other tests were carried out in the same order as at the post-test and the delayed post-test). This arrangement was because the allocation of participants into different interventions was based on participants' combined test scores in the timed GJT and the gap-fill test.

The order of the tests was as above so that the implicit measures preceded the explicit measures (except at the pre-test). In addition, the timed GJT, gap-fill test, and vocabulary test were given to the whole class and were completed in a single 40-minute session. The oral tests were administered through one-to-one meetings (i.e., one test administrator and one test taker) during participants' 40-minute lunch break, not in their regular instructional session. Each test taker spent approximately 5 minutes completing the oral tests, so it took around a week to collect all of the oral data at each of the testing

periods (the pre-test, the post-test, and the delayed post-test).

I was responsible for the administration of the achievement assessments, including the timed GJT, the written gap-fill production test, and the vocabulary test. The oral tests were carried out by me and two research assistants, who were third-year undergraduate students recruited from the local university. Both of them met with me prior to the pre-test and two post-tests, at which times they were given the explicit written guidelines about the procedures for conducting the oral test, such as how to use the audio-recording software and how to conduct the interview immediately after the post-test and the delayed post-test. At the end of each meeting, they were asked to demonstrate the procedures of the oral test to confirm that they were familiar with the procedure for the task.

### ***3.2.3 The instructional material packages***

#### ***3.2.3.1 Design of the instructional materials***

All of the three instructional materials (R, A, and RA) in this study were developed by the researcher to suit the participants' English proficiency. So far only one prior PI study (Benati, 2005) has used the English regular past tense as the target feature, which is the target feature used in the current study. The researcher contacted Benati to acquire his teaching and test material as a reference, but they were unavailable due to a computer crash. Therefore, the instructional material was designed specifically by the researcher.

Due to the fact that the nature of the RA instruction requires its learners to go through both referential and affective activities sequentially, the RA instructional materials simply combined the referential activities of the R group with the affective activities of the A group. So the actual design of the RA instructional material will not be further addressed here. The following description primarily elaborates the design of the two

instructional materials (the R group and the A group).

The instructional materials were developed on the basis of explicit guidelines for PI instructional material design, suggested by VanPatten (1996) and Wong (2004a). Note that the current study does not aim to test these guidelines. The reasons for following these guidelines are the consideration of parity with other studies, they are intuitively appealing to teachers, and they are likely to be accepted by teachers and pupils. The guidelines are as follows (see Section 2.1.3 in Chapter 2):

1. Teach only one thing at a time.
2. Keep meaning in focus.
3. Learners must do something with the input.
4. Use both oral and written input.
5. Move from sentences to connected discourse.
6. Keep the psycholinguistic processing strategies in mind.

The design of instructional materials for each experimental group adhered to the guidelines 2,3,4,5, and 6. All of the training items in each piece of instructional material for this study were at sentence level, and never moved to discourse level in that the participants in this study were regarded as beginning learners of English. The topics of activities included a diary written by somebody, an interview given by a famous pop star, or simply an email sent by a friend, and so on, all of which involved something that happened in the past. The purpose of creating different topics, which are akin to real life, was to excite the interest of the participants in carrying out the training tasks and to make those tasks more 'authentic', as though the participants were facing a real situation.

However, adherence to guideline 1 (i.e. teach only one thing at a time) was not achieved

in the design of the referential activities. By nature the referential activities require the training items to be juxtaposed with the contrasting pairs (for further discussion, see Benati, 2004a, p.211; Marsden, 2006, p.519; VanPatten, 2002, p. 767). Therefore, the targeted training items of referential activities were mixed with other features such as English present or future tenses. In this sense, learners could not employ a strategy that judged all sentences to be in the past tense, once they had realised the grammatical focus that they were learning. In addition, participants in the R group were, in principle, required to recognise the targeted feature ('-ed') and to complete the task by interpreting its meaning. The participants' attention was expected to be directed to the verb ending as the indicator of tense. An example of the instructional material for the R group was as follows:

**Example of referential activities:**

Some of Delia's diary entries have got smudged. Decide whether Delia has written about an event that happened in her *previous summer holidays* or if she is referring to something she *usually does in the summer holidays*.

1. I learn Spanish.
  - a. last summer
  - b. usually does
2. My family visited Paris.
  - a. last summer
  - b. usually does
3. I play tennis with my friends.
  - a. last summer
  - b. usually does

(and so on)

Note that the introduction of activities in each instructional package was written in Chinese rather than English out of consideration for the participants' developmental

stage as beginning learners of English. However, the options provided for the participants were in English, e.g. ‘last summer’ and ‘usually does’.

On the other hand, the affective activities for the A group required participants to show their own opinions or feeling towards events that happened in the past. The participants were merely exposed to training items with the targeted feature, so no contrasting pairs were present. In contrast to the referential activities, no right or wrong answer was required from the participants. An example of the instructional material for the A group is as follows:

**Example of affective activities:**

Delia has written a diary entry about her family’s *last summer holidays*. What do you think about her activities?

1. My family visited Paris.
  - a. interesting    b. boring
2. I learned Japanese.
  - a. interesting    b. boring
3. My family painted the wall.
  - a. interesting    b. boring

(and so on)

Note that both referential and affective activities provided extensively for binary choices (e.g. yesterday/tomorrow in the referential activity, or true/false in the affective activity). It is noted that although the ‘availability of resources’ principle in IP suggests that the input sentences should be manipulated so that the targeted feature is located as near to the start as possible, clearly the targeted ‘-ed’ feature cannot be placed in the



initial position.

Each of the three instructional packages (the R, A, and RA packages) consisted of 10 activities. In both the R and the A instructional packages, each activity was composed of one reading and one listening practice. An attempt was made to balance the content of activities and to ensure the same amount of exposure to the target feature in both the R and the A instructional materials. In each training session, the topics of the activities designed for both sets of instructional materials were similar. Also, an attempt was made to balance the vocabulary and sentence structures used in the instructional materials in most of the tasks. As in the examples of the referential and affective activities presented above, both activities have the same training item ‘my family visited Paris’. As for the RA group, the nature of the RA required the participants to go through both referential and affective activities, so an RA activity was composed of a referential reading practice, a referential listening practice, plus an affective reading practice. As a result, the participants in an RA activity would experience two reading practices and one listening practice. The RA group would be exposed to more of the targeted feature than those in the R and A groups, and the duration of the tasks would be greater than that in both the R and A groups. The RA group did not receive all the affective activities (i.e. reading plus listening activities) in the A group because of the need to minimise the difference in the duration of instruction. The tally of practice items in each instructional group is displayed in Table 3.2.

Table 3.2

*The tally of the number of practice items in each instructional group*

	<b>RA</b>	<b>R</b>	<b>A</b>
The number of practice items in <i>reading</i>	214	135	79
The number of practice items in <i>listening</i>	121	121	74
The total practice items	335	256	153

The total targeted practice items (only containing target feature)	232	153	153
--	-----	-----	-----

Note that the time required to complete the activities of R and A groups was slightly different. In both instructional packages, the numbers of instances of exposure to the target feature were identical, a total of 153 targeted practice items. As discussed earlier, the nature of the R required the participants to encounter the targeted practice items and contrasting ones, which did not contain the targeted feature, to complete the tasks. On the other hand, the participants in the A group did encounter the practice items with the targeted feature, but they did not have to notice the targeted forms. The nature of affective activities only required the participants to notice key vocabulary in order to complete the task. Given that the R training materials were made up of more practice items (a total of 256) than those of the A practice materials (a total of 153), the duration of the R tasks would be longer than that of the A tasks. As a result, the R group was exposed to more verb stems and their instruction lasted longer than the A group.

Furthermore, due to the differences in the number of practice items described above, it was inevitable that the items in the vocabulary test occurred in different amounts in different conditions. Table 3.3 provides the total occurrences of vocabulary test items in the intervention materials. It is noted that although the RA material had the most occurrences of vocabulary test items, and the A material had the least, the differences in quantity did not predict the results.

Table 3.3

*Total occurrences of vocabulary test items in the intervention materials*

	<b>RA</b>	<b>R</b>	<b>A</b>
Vocabulary version 1	64	42	30
Vocabulary version 2	67	49	32

The ‘availability of resources’ principle in IP theory predicts that learners could process the redundant meaningful or non-meaningful forms if their processing resources are not used up when processing the overall sentential meaning. Based on the afore-mentioned perspective, using fewer unknown verb stems or fewer novel lexical items in the input strings is more likely to release attentional resources to process the targeted grammatical form. Thus, an attempt was made to use the verb stems which the participants had learnt prior to the intervention, as far as possible. Apart from referring to the school textbooks to ensure that the vocabulary was familiar to the participants, their regular English teacher was also invited to check the familiarity of the vocabulary used in the instructional materials. However, the participants of the current study were beginners in English, so the numbers of English verb stems which they had been taught prior to the intervention were not sufficient for the design of the instructional materials. Therefore, it was inevitable that some verb stems or lexical items would be used that the participants had not been taught prior to the intervention. During the instructional session, each unknown word (i.e. the nouns, verb stems and so on) was glossed with its Chinese equivalent meaning and syntactic category in the same slide as the practice items appearing on the computer screen.

Inevitably, the number of glosses provided varied across instructional groups. These differences were because the different input types required different numbers of practice items as described above. By the same token, the vocabulary test items were glossed in different numbers across instructional groups. The tally of the number of glossed vocabulary items, which were tested, is given in Table 3.4. In sum, the A materials contained the lowest number of glossed items that were then tested, and the RA materials contained the most. One main reason for reporting the above discrepancies in the number of glossed vocabulary items was to exclude the possible speculation that the

learning gains observed in the later vocabulary test were attributable to these differences in quantity; indeed, the differences in quantity did not predict the results (see Sections 4.1.5.3 & 4.1.5.4).

Table 3.4

*The number of glossed vocabulary items assessed in the vocabulary test*

	<b>RA</b>	<b>R</b>	<b>A</b>
Vocabulary version A	39	30	20
Vocabulary version B	41	30	19

### ***3.2.3.2 The administration of the interventions***

The method of allocation to three experimental groups was unknown to participants, in that they were merely informed that they were split into three similar groups of equivalent levels, which were called ‘Doraemon’, ‘Snoopy’, and ‘Hello Kitty’. The R group was referred to as the ‘Doraemon’, the RA group as the ‘Snoopy’, and finally the A group as the ‘Hello Kitty’. The three famous cartoon characters (i.e. Doraemon, Snoopy, and Hello Kitty) were well-known by school kids in Taiwan. The intention behind naming the groups after cartoon characters was to motivate the participants. In addition, by using the names of popular cartoon characters, I/the researcher expected that participants could readily pick up which group they were assigned to after the first allocation in the first instructional session. The training materials for each intervention were specifically saved in the allotted computers. During the instructional sessions, every participant was required to operate the same computer from the first interventional session to the last one in order to avoid the risk of being exposed to different interventions by accident.

The instructional materials were delivered by computer by means of the Microsoft

PowerPoint Software to achieve an element of interactivity in the training materials. In principle, participants were familiar with the operation of Microsoft PowerPoint software in that they had computer lessons as a compulsory course, and the application of PowerPoint software had been instructed in grade 5.

The researcher was the instructor for all four instructional sessions from the very beginning to the end of the interventional period. During the instructional phases, participants were required only to interact with the computers and they were not allowed to discuss the activities with their classmates. They were welcome to ask for assistance from the researcher and the research assistant if they encountered any problems in the operation of the computers and software. Each computer was equipped with a headset to facilitate the listening practice. In addition, a handout displaying various English temporal adverbials (i.e. the future tense (e.g. tomorrow, next year), the present tense (e.g. every year, every day), and the past tense (e.g. yesterday, last year) was distributed to all participants during the instructional phases (see Appendix 11), given that participants were required to recognise the temporal adverbials to undertake the achievement assessments after the interventions. In particular, the nature of referential activities required the participants to recognise the temporal adverbials to respond to the practice items. In the beginning of each instructional session, the researcher gave explicit instruction, lasting approximately 2-3 minutes only, introducing the meaning of the temporal adverbials, without explaining the grammar rule. The participants were only told that they were going to learn English expressions about the tenses, and they were allowed to look at the temporal-adverb handout during the instructional phases. The temporal adverb handouts were collected at the end of each instructional session.

During the instructional phases, every activity started with a slide displaying the introduction of it (e.g. an interview of a TV host). The participants clicked the 'next' button shown in the right corner, and then moved to undertake the practice items. While performing the referential activities (see Appendix 12), the participants responded to practice items by clicking one of the binary options (e.g. yesterday or every day). Then a feedback would show up on the screen merely indicating whether or not their response was correct or incorrect (e.g. 'well done' or 'sorry! Wrong answer'), no explicit grammar explanation being provided according to their responses. The participants undertaking the affective activities also responded to the practice items by clicking one of the binary options, and then a cartoon character would show up on the screen, acting as an interlocutor to provide feedback according to their responses (see Appendix 13). The feedback was written in Chinese in a conversational bubble, showing agreement or disagreement with the participants' responses, such as 'I think it's interesting as well' or 'I do not agree with you'.

Note that the instructions described above were given to all participants except the control group. The participants in the control group merely took part in the achievement assessment at different testing times, and they did their scheduled regular school activities during the instructional phases. Apart from the control group, all instructional groups received feedback when undertaking the instructional activities, but the feedback provided was different to some extent in different groups. Also, the participants in each group were never involved in producing (speaking and writing) the English regular past tense '-ed' form during the instructional period, they were only engaged in reading and listening practices. Furthermore, given that previous PI studies have shown evidence that explicit grammar explanation is not the causative factor for the effectiveness of PI (Benati, 2004a, 2004b; Farley, 2004b; and Wong, 2004b; VanPatten & Oikkenon, 1996),

at no point did participants in instructional groups receive an explanation of the targeted linguistic feature during the instructional phases.

### **3.3 The achievement assessments used in the current study**

This section will present the design and administration of the achievement assessments adopted in the current study in the following order:

- 1) the timed GJT;
- 2) the written production test (a gap-fill test);
- 3) the two types of oral production tests (picture-based narration task and the structured conversation)
- 4) the vocabulary test
- 5) the self-report of participants (a post-task written questionnaire and a structured interview) which followed the implicit measures (the timed GJT and oral tests).

#### ***3.3.1 The Grammaticality Judgment Test (GJT)***

##### *3.3.1.1 The design of timed GJT for this study.*

The timed GJT was composed of 40 sentences. The test items of the timed GJT are listed in Appendix 14. Twenty out of the 40 test items were the targeted sentences with the targeted feature, and the remaining 20 non-targeted sentences were distractors.

Among the 20 targeted test items, 10 were grammatical sentences and the other 10 were ungrammatical ones. Only the results of the 20 targeted test items were computed for further statistical analysis.

In order to maintain the linguistic complexity of the GJT, the length of each sentence was controlled and ranged from five to six words on average. The noun phrase used in the GJT was kept simple (i.e. determiner + noun). Furthermore, an attempt to devise sentences that were semantically plausible was made. The grammaticality of the 40 test

items was checked by two English native speakers to ensure the grammatical and semantic accuracy. Two English native speakers were invited to read the sentences and to reflect on whether or not these sentences corresponded to normal English usage. If they detected something incorrect or inappropriate, they were encouraged to correct or to express their opinions of sentences. Furthermore, the application of the third person singular (e.g. she/he/it/John, and so on), serving as the subjects of the ungrammatical sentences in the targeted test items, was avoided in order to reduce extraneous factors which might affect the judgment of participants. Given that the participants had been learning the present tense prior to the intervention, it was possible that their recently taught explicit knowledge about the usage of the third person singular might affect their judgments. For example, in a sentence like ‘John walk to school yesterday’, the participants would possibly judge the sentence ungrammatically and think ‘walks’ would be the right usage instead of ‘walked’. As a result, the use of the third person singular as the subjects in the ungrammatical targeted test items was avoided.

As Sorace (1996) argued that L2 learners’ interlanguage was pervaded with indeterminacy, the more constrained the informant’s responses are, the more unlikely it is that the researcher will acquire informative or valid responses. Sorace, therefore, suggested that the adoption of a scale of more than three points would be more statistically reliable and possibly produce a more accurate solution (p. 398). In order to allow for the indeterminacy, each sentence offered five options for participants to respond. The participants were required to decide whether the sentence they had read was ‘correct’ or ‘incorrect’ by circling one of five numbers (+2, +1, 0, -1, -2): circling +2 meant that they were 100 % sure the sentence was correct; circling + 1 meant less sure that it was correct. By the same token, circling – 2 meant that the participant was 100% sure the sentence was incorrect;- 1 meant that they were less sure it was incorrect.



Circling 0 meant that participants really could not tell whether it was correct or not. The answer sheet for participants to show their responses is attached in Appendix 15.

### *3.3.1.2 The obtainment of a given time for each individual sentence*

Although a timed GJT is likely to tap into learners' implicit knowledge, a question emerges regarding the allotment of time (e.g. how much time should elapse between each individual sentence). This issue has not been seriously considered in the literature (Murphy, 1997). In general, studies using the timed GJT allocated the same length of time to each individual sentence (e.g. Bialystok, 1979; Mandell, 1999). However, R. Ellis (2004) argued the time for judging each sentence in a timed GJT may vary on the basis of the grammatical complexity or the length of the sentence. Thus, an attempt was made to get the specific time constraint for each individual sentence in the GJT of this study.

The procedure for obtaining the specific time constraint for each individual test item was in accordance with R. Ellis' study (2005) regarding how to measure implicit and explicit knowledge. R. Ellis (2005) demonstrated that the timed GJT is a feasible test for measuring implicit knowledge. He administered test items to a group of L1 native speakers, calculated the length of time which subjects spent on each individual sentence via computers, and then obtained the mean of the time for each individual test item. Taking L2 learners' slower processing speed into consideration, Ellis added on 20% of the average response time taken by the group of L1 native speakers for each sentence in order to provide a greater length of time for L2 learners to elicit their implicit knowledge. It is acknowledged that the 20% of added time for L2 learners was arbitrary and arguable (Isemonger, 2007, p. 109). Also, "the perception of speed is individualistic (Purpura, 2004, p. 116)". However, to the best of my knowledge, no study so far has made an attempt to provide a clear criterion for deciding the time length for each

individual test item in a timed GJT for L2 learners, except for Ellis' study. Accordingly, Ellis' criterion of 20% additional time given to L2 learners was adopted to design the timed GJT in the current study.

Ten English native-speaker students, 10 years old on average, were recruited to take part in the one-to-one meeting in order to acquire a specific time constraint for the Chinese L2 learners of English. The purpose and the procedure for conducting this test were explained to the participants, and the need for them to respond to the test items as quickly as they could was also stressed before they started to take the test. Ten exemplary test items unrelated to the targeted feature were given to the L1 participants to practice before the commencement of the timed GJT. In consideration of the ethical issue involved, a consent form for participation in this study was distributed to the participants to substantiate their genuine willingness to take part in this study. The consent form is attached in Appendix 16.

The timed GJT was carried out on computers, so the L1 group was required to judge the 40 sentences which appeared one by one on the computer screen, and then show their responses by circling on the 5-point scale answer sheets. The test administrator sat next to the L1 participant, and as soon as the subject finished circling an option and then looked up at the screen, the test administrator would press 'enter' to move on to the next test item. Once the test administrator had pressed 'enter', the computer would automatically record the time the student had spent on a given test item. Following R. Ellis' (2005) approach to adding 20% of the mean L1 learners response time, this led to a mean time of 7.2 seconds (SD=1.14) for each test item for L2 learners, with a range

from 5 to 10<sup>35</sup>. Included in the time given for each item was 3 seconds to circle a response during which the sentence was *not* on the screen.

### *3.3.1.3 The administration of the timed GJT for L2 learners.*

The timed GJT was administered by using the broadcast system already installed in the computer laboratory at the participating school. It was given to the whole class at the same time, but each participant looked at his/her own screen rather than sharing with others. The use of the broadcast system was to ensure that the display of each test item to participants was synchronised in order to avoid participants going back to check prior test items. Full instructions were given before the test was formally carried out, explaining how to respond to the five-point scale on the answer sheet, how to change their answers if they circled the wrong option and so on. Also, the participants were advised that these sentences were de-contextualised (i.e. each sentence was irrelevant to the others), and they were not told specifically to pay attention to the grammatical accuracy in order to possibly preclude the elicitation of explicit knowledge. They were merely asked to judge the sentences either appropriate or inappropriate, such as ‘looks like a good English sentence’, or ‘something seems to be wrong in the sentence.’ After the explicit explanation of the procedures, participants were required to demonstrate the test by working through ten examples (see Appendix 17), which bore no relation to the targeted features.

Forty test items were displayed one by one on a computer screen and each sentence was shown for a given number of seconds on a slide. Immediately after a test item had been shown for a given time, a slide marking the test number which required the participants to show their responses would pop up. Three seconds were allowed for the participants

---

<sup>35</sup> The times were rounded to the nearest second as PowerPoint does not permit fractions of seconds. The time constraints in Ellis (2005) study were shorter (ranging between 1.8 and 6.24 seconds), probably because participants were adults and responses were given via pressing the button.

to circle their response on the five-point answer sheet. The 3-second response time was piloted by five Chinese learners at the same grade as the participants in the main study. In addition, there was a possibility that the participants would fail to answer or miss a subsequent test item if they spent too much time pondering the test item. To remind participants of the presence of the next test item, a ‘beep’ sound was emitted as soon as the subsequent test item was displayed.

### ***3.3.2 The written production test: a gap-fill test***

#### *3.3.2.1 The design of the written production test*

The design of the gap-fill test was intended to investigate the impact of PI input-based activities on a written production test. In addition, the gap-fill test is suggested being able to tap into explicit knowledge (Macrory & Stone, 2000) and is usually used to measure learners’ knowledge of grammatical forms and grammatical meanings (Purpura, 2004, p. 135). The written gap-fill production test was assessed in a paper-and-pencil format and it comprised fifteen test items, eight targeted items and seven distractors. Although the targeted feature in this study was the English regular past tense, the gaps did not have to be filled in with verbs alone. Apart from verb stems of the present tense, the distractors included nouns and adjectives in case the participants employed the test-taker strategy (e.g. attaching ‘-ed’ to the end of every verb stem.) or were aware that it was necessary to do something different to the verbs. The two versions of the gap-fill test are given in Appendix 18 & 19.

A quick revision list, which displayed all the lexical items to be filled in (i.e. the verb stems, adjective, and nouns) and their prompt pictures, was created to ensure that the participants produced the expected target-like words (see Appendix 20 & 21). For example, when filling in a sentence such as ‘Mother \_\_\_\_\_ dinner for us last night’, the

participants could write down ‘made’ (the non-target production) rather than ‘cooked’ (the target production), though both of them are correct in English usage. In addition, a prompt picture<sup>36</sup> with its equivalent Chinese meaning, corresponding to what the participants needed to fill in a gap, was provided at the end of each individual sentence since the participants were regarded as beginning learners of English. The provision of prompt pictures was intended to promote greater recognition of the words which the participants had reviewed in the quick revision list a minute before.

Nevertheless, it is acknowledged that there might be a pitfall in the provision of prompt pictures and their equivalent Chinese meaning, given that the participants could rely on the prompt meaning and not bother to go over the sentences. However, one possible way to examine if this was the case was by scrutinising their responses to distractors. If the participants knew the grammatical focus and they filled in the gap merely by depending on the prompt provided, without reading the sentences, they would fill in all verb stems with ‘-ed’ attached. In addition, the gaps, which need to be filled in, were vocabulary that participants had learnt prior to the intervention. This was aimed at precluding the possibility that participants’ failure to produce the targeted words was due to the spelling difficulty or unfamiliarity with the words instead of not knowing the correct usage of the target feature.

### *3.3.2.2 The administration of the written production test*

The quick revision list was distributed to each participant to review for about one minute prior to the formal administration of the test. After the reviewing time was up,

---

<sup>36</sup> The pictures and drawings in the quick revision list or serving as a prompt were freely downloaded from the Center for Technology Enhanced Language Learning, Department of Foreign Language and Literatures, Purdue University. <http://tell.fl.purdue.edu/JapanProj//FLClipart/>

all of the quick revision lists distributed to the participants were collected by the researcher and research assistant. At no time did the participants have any opportunities to refer to the quick revision list after the gap-fill test had been distributed to the participants. During the testing phase, participants had to read the incomplete sentences with one word missing in each sentence and they were required to write the word in the space provided. The implementation of this test was without a time constraint.

### ***3.3.3 The oral production tests***

#### *3.3.3.1 The design of the oral production tests*

The prior PI studies had showed mixed results regarding PI's impact on learners' oral performance (see Section 2.3.1.6), and PI has been criticised for its lack of less controlled conditions (DeKeyser *et al.*, 2002; Marsden, 2004). In addition, R. Ellis (2005) suggested that a viable way to elicit learners' implicit knowledge is through spontaneous oral tests, because learners *may* not have time to access their explicit knowledge and monitor their accuracy. Benati (2004b) suggested that measuring learners' communicative behaviours may avoid learners' access to a monitoring system because spontaneous tasks leave less opportunity for monitoring. Given that the oral test was regarded as a viable approach for suppressing the occurrence of language monitoring, the participants' oral performance was assessed via two types of oral tests: a) a more controlled picture-based narration task b) a less controlled structured conversation.

#### *a) Picture-based narration task*

The picture-based narration task has been used in some studies to measure the impact of PI (e.g., Benati, 2001, 2004b). As with the gap-fill test, a quick revision list was developed and the reasons for creating it were as in the afore-mentioned discussion of

the gap-fill test (i.e. to produce target-like verb stems and to consider learners' productivity of the target inflection). A total of 8 pictures were used as prompts, with the aim of producing 8 target verbs. The 8 target verbs, which participants were expected to pronounce, were chosen on the basis of what the participants had been taught before the intervention. In order to make the picture-based task more like a story-telling task, the pictures were in sequence by virtue of each picture being accompanied by a drawing of a clock, indicating the time that the character performed the activity. The participant was told to imagine that the character in these pictures was his/her friend, and they had to tell the test administrator what his/her friend did either last Sunday or yesterday. In addition, all of the pictures were presented in a small booklet and the test administrator was responsible for flipping the pages to ensure that the participant did not miss any of the pictures. The two versions of the picture-based narration task are given in Appendix 22 & 23, and the two versions of the quick revision list are provided in Appendix 24.

#### *b) Structured conversation*

The use of a structured conversation test to measure the effectiveness of PI was found in the prior PI study (VanPatten & Sanz, 1995). VanPatten & Sanz found that their participants performed poorly in this test. As the structured conversation is less controlled than the picture-based narration task, it was expected that this test would be more likely to tap into participants' implicit knowledge than the picture-based narration task. A sample transcription of a structured conversation is given in Appendix 41.

#### *3.3.3.2 The administration of oral production tests*

Thirty-seven participants were selected to take part in the oral tests (RA(n=10), R(n=9), A(n=9), and control (n=9)). The selection of participants was carried out with the assistance of the students' regular English teacher. It is acknowledged here that the

sample size for the oral tests was fairly small (i.e., fewer than 15 in each group) due to practical difficulties in recruiting more participants. The sample size did not meet limitation 5 discussed in Section 2.3.1.9; therefore, the oral test results should be interpreted with care. It was agreed that the participants in each group were chosen by the English teacher on the basis of her observation of their normal English performance in regular English classes. The English teacher selected those who were more active in English classes, because being shy and inactive would be an obstacle to performance of the oral tests. The participants did not interact with their regular English teacher in the test, but with some unfamiliar test administrators (i.e. the researcher, and two research assistants). Note that the participants in the oral tests at the pre-test, post-test, and delayed post-test interacted with the same test administrator.

The oral tests were audio recorded via laptops, and proceeded on the basis of one-to-one meetings. Before the participants formally began the picture-based narration task, a quick revision list was delivered to them to look through within a minute and familiarise themselves with the vocabulary before starting the oral test. The participants were encouraged to ask the test administrator about the pronunciation of the verb stems. No further information related to the target form was showed to the oral test taker. After the quick review of the words, a participant formally commenced the task by reading out the time adverbial on the first page (*yesterday* or *last Sunday*), intended to remind her/him that what she/he was going to describe had happened in the past. As the oral test was designed to elicit participants' use of implicit knowledge, it was sensible to prohibit the participants from having much time to give their utterance. Once a participant paused and produced nothing after seeing a picture for about 3 seconds, the test administrator would move on to the next page. At no point was the test administrator allowed to give participants any prompts related to the target verbs, but some



interactions with participants were granted (such as giving the participants some verbal awards or comforting them if they could not carry on the task, and so on). Immediately after the picture-based task, participants were required to talk about what he/she did either yesterday or last Sunday and to express themselves promptly to the test administrator.

### ***3.3.4. The vocabulary test***

As Marsden (2006) suggested that affective activities might be more beneficial to vocabulary acquisition than referential activity, a vocabulary test was designed to look into this issue. The words that served as the test items in the vocabulary test were those that were displayed in all three types of instructional material. The participants, in principle, had not learnt them at school prior to the intervention, as was verified by checking the textbooks that participants had used, and enlisting the assistance of their regular English teacher. This was a paper-and-pen and L2- to- L1 test with a total of 10 test items, and it was administered after the gap-fill test. In this test, the participants would read the English words and were required to write down their equivalent Chinese meanings. The two versions of the vocabulary test are given in Appendix 25 & 26.

### ***3.3.5 The retrospective self-report***

#### ***3.3.5.1 The post-task written questionnaire***

As the timed GJT was designed to elicit participants' implicit knowledge, a post-task questionnaire was attached in the final page of the GJT answer sheets at the post-test and the delayed post-test to explore participants' subjective experience during the GJT, namely whether or not explicit knowledge was tapped<sup>37</sup>. The questionnaire was written

---

<sup>37</sup> The construct of implicit knowledge in this study was operationalised as learners might be aware or unaware of the rule of grammatical focus, but they did not resort to it while performing the assessments. On the other hand, the construct of explicit knowledge was operationalised as learners were aware of the

in Chinese and it required the participants to think about how they were judging the grammaticality of the sentences during the GJT testing phase, by circling two options: a) by feeling, b) by some rules. If the participants circled b), they would go on to the sub-question, which required them to write down the grammar rules or to provide examples of how they judged the sentences during the testing phase. No time constraint was set for the participants to complete the post-task written questionnaire. The participants responded to the questions by using their L1. The post-task questionnaire is attached in Appendix 27.

#### *3.3.5.2 The structured interview of rule verbalisation*

A brief structured interview was held following the picture-based narration and structured conversation at the oral post-test and delayed post-test. The whole course of the interview was audio-recorded with participants' oral permission. The interview was conducted in participants' L1 and notes were taken by the interviewer, who was also the oral test administrator. The interview sheet for the interviewer to note down participants' responses in the post-task interview is given in Appendix 28. During the interview, the participants were involved in recalling whether or not they were thinking of some specific rules whilst performing the oral tests. If a participant self-reported that 'no, s/he was never thinking of any rules during the oral testing phase', the interviewer would terminate the interview. On the other hand, if the participant self-reported that s/he was thinking of the rule(s) during the testing phase, the interviewer would request her to provide the grammar rule(s) or examples. Straight away after the provision of the rule or examples, the participants were asked to confirm again whether or not they were thinking of using the rule while they were doing the oral test, not when performing the interview.

---

rule, and they relied on it to take the assessments.

Note that the interview was conducted immediately after the completion of the picture-based narration and the structured conversation, given that an attempt to require participants to verbalise the rule after the picture narration task might entail eliciting their use of explicit knowledge in the structured conversation. However, a methodological weakness is acknowledged here. The interviewer did not ask participants to identify which oral tests (either the picture-narration task or the structured conversation) they were recalling. Thus, the interpretation of the interview results was bound to be very cautious and the analysis of the interview would be suggestive rather than conclusive.

### ***3.3.6 Two versions of each achievement assessment***

Two versions (A and B) of each assessment were designed. The intention behind creating two versions of the same assessment was to avoid participants' memorising, or being more familiar with, some test items (i.e., test effects), which might lead to their improvement in the post-test or the delayed post-test. Because each test item in the GJT was timed, it is acknowledged that the only difference between the A and B versions in the timed GJT of this study lies in the order of the test items. Note that although the two versions for the gap-fill test and the picture narration test had different test items, the tests were identical in overall length with the same numbers of targeted test items and distractors.

However, creating different versions of tests might lead to a potential problem with respect to the comparability in difficulty across the tests (Mackey & Gass, 2005).

Different levels of difficulty across the test might result in an artificially greater or smaller improvement in the post-tests. Thus, Mackey & Gass (2005, p. 149) suggested

two techniques to avoid this bias. One is to design different versions of tests and then randomly assign them to groups as the pre-test and post-test. The other is to include other groups to test the comparability of the tests. The two techniques, suggested by Mackey & Gass, were applied in this study and are discussed as follows.

#### *3.3.6.1 Randomly assigning tests to groups as the pre-test and post-test*

For the control group, half of the class received version A, and half received version B; those who received version A at the pre-test would receive version B at the post-test and vice versa as a split block design. For the three instructional classes, two classes received version A and one class received version B. The class which had received either version A or B at the pre-test received the other version as the post-test. For the delayed post-test, the participants received the same version as in the pre-test. Note that all of the participants remained in their regular class to have the intervention, but the participants in each class were allocated to different groups to receive a specific intervention. Thus, the RA, R, and A groups all received the test as a split block design.

#### *3.3.6.2 The comparability of the achievement assessments*

Two groups of participants at different developmental stages were recruited to examine the comparability of the two-version assessments and the validity and reliability of the achievement assessments (see more discussion and results in section 3.5). Neither of the two groups took part in the main study or the pilot study. The results of comparability in the two versions of the tests (i.e. the gap-fill test, the picture narration test, and the vocabulary test) showed no evidence of a difference in these tests across different versions. Therefore, any difference of test scores observed in the post-test and delayed post-test when compared to those of the pre-test should not be attributed to the use of two-version tests in terms of the gap-fill test, picture narration test and vocabulary test.

The results of the two versions of tests on comparability will be presented in the section 3.5.

### ***3.3.7 The scoring procedures***

In this section, the scoring criteria used to judge the correctness of responses of the achievement assessments will be addressed. The scoring procedures were presented in the following order: the timed GJT, the written gap-fill test, the oral tests, and the vocabulary test. The timed GJT, gap-fill test, and vocabulary test were all marked by the researcher. The oral tests were marked by an English native speaker. Another native speaker was also invited to score the oral data, and the Kappa value for the inter-rater agreement was .752, which is a good agreement. According to Peat (2001, p.228, cited in Pallant, 2007, p.220), “a value of .5 for Kappa represents moderate agreement, above .7 represents good agreement, and above .8 represents very good agreement”.

In addition, the right-wrong scoring method, as opposed to partial-credit scoring method, was adopted to score participants’ answers in the gap-fill test, oral tests, and vocabulary test. According to Purpura (2004, p. 117), the criterion for the right-wrong scoring method is definite and the test item will be marked as either right or wrong. On the other hand, partial-credit scoring gives some credit for partially correct responses. For example, one might want to assess the grammatical knowledge of ‘form’ and ‘meaning’ of the English past tense. A response such as ‘goed’ would be assigned partial credit (zero credit for ‘form’ and full credit for ‘meaning’). However, the current study merely measured the knowledge of grammatical ‘meaning’ rather than ‘form’ of the targeted feature. In this scenario, the partial-credit scoring method was not adopted. The reason to disregard the scoring of orthographic form was based on the fact that the participants had not been instructed in the English irregular past tense prior to or during

the interventional phase. A full credit should be awarded if a learner gave a response such as 'goed' or 'writed'. Consequently, the key criterion for the right-wrong scoring method adopted in this study was whether the targeted '-ed' feature was attached in a response, demonstrating some language development, regardless of the target-likeness of the past tense formation in spelling or pronunciation.

#### *3.3.7.1 The timed Grammaticality Judgment Test*

The GJT consisted of 40 test items, 20 target items, and 20 distractors. Only the scores of the 20 target items were computed for further data analysis. As the GJT used a 5-point scale (2, 1, 0, -1, and -2), if a participant circled '2' in a grammatical sentence, 2 points would be awarded; if he/she circled '1', 1 point would be awarded; no point would be given if circling '0', '-1', or '-2'. Similarly, 2 points would be awarded if a participant chose '-2' in an ungrammatical sentence; 1 point for choosing '-1'; and 0 point for choosing '0', '1', or '2'. Consequently, the maximum score in the GJT was 40.

#### *3.3.7.2 The written gap-fill test*

The gap-fill task consisted of 15 test items in total: 8 target items, and 7 distractors. Only the scores of the 8 target items were calculated for the later statistical analysis. One point would be given for a target item if a correct answer was provided. However, if a participant misspelled a vocabulary item but correctly attached the target feature (-ed ending), such as 'studyed' or 'shoped', a full score (one point) would still be awarded. As a result, the maximum score was 8 in this test.

#### *3.3.7.3 The oral tests*

##### *a) The picture-based narration test*

Eight target verbs were expected from each individual participant in this test as each

picture provided an obligatory context for a verb in the past tense. One point would be awarded if a participant attached the ‘-ed’ form at the end of a verb stem. Therefore, the maximum score for a participant in the picture-based narration test was eight.

*b) The structured conversation*

The participant would receive one point if the ‘-ed’ form attached at the end of a verb stem was clearly identified. For each student, the percentage of verbs which had the ‘-ed’ ending was then calculated. The mean rate of appliance in obligatory contexts was obtained by dividing the sum of these percentages by the number of participants.

*3.3.7.4 The vocabulary test*

The test was administered in a paper-and-pencil format and was composed of 10 test items in total. One point would be given if its Chinese equivalent meaning was written down. The maximum score was 10.

*3.3.7.5 The post-task self-reports*

The self-report was designed to explore whether or not the participants were tapping their explicit knowledge of the language during the testing phases. Thus, only those participants who stated that they had used the targeted grammatical rule and specified it (i.e. mentioned the ‘-ed’ feature) or gave the correct example of the targeted feature, were counted as ‘rule users’. If a participant reported that s/he used the rule, but s/he did not provide any targeted rule or example with the targeted feature, s/he was not regarded as the rule user.

### **3.4 The statistical analysis procedures of the achievement assessments**

This section presents a range of statistical analysis tests applied in chapters 4 & 5 to

examine the impact of the interventions by checking the test scores of achievement assessments among the instructional groups and the control group at different testing times. Note that it is beyond the scope of this thesis to spell out exhaustively the mathematical equations behind the statistical procedures.

### ***3.4.1 Parametric tests vs. non-parametric tests***

The parametric tests, including analysis of variance (ANOVA), *t*-test, and Pearson product-moment correlation, have been commonly used in previous PI studies and second language research. In general, parametric tests are regarded as being more powerful than non-parametric ones in detecting the differences existing among the groups, but this statement is tenable only if four parametric assumptions are met (Field, 2005, p. 533): the normal distribution, homogeneity of variance, interval data, and independence (Field, 2005, p.64). As the parametric tests assume where the populations being examined are normally distributed and variances among the populations being compared are similar, the first two assumptions (i.e. the normality and homogeneity) could be checked by looking at the distribution of the sample data via the Kolmogorov-Smirnov test (K-S test) and Levene's test respectively, which are addressed in the following section. The last two assumptions (i.e. the interval data and independence) could be examined by common sense. It is therefore essential to examine the assumptions before determining which statistical test (i.e. parametric or non-parametric test) is applied, given that using a parametric test when the assumptions are not satisfied will produce inaccurate results (Field, 2005).

Contrary to parametric tests, non-parametric tests are known as assumption-free or distribution-free tests. Researchers have argued strongly for the use of a non-parametric test if the assumptions of parametric tests are violated (Field, 2005; Pallant, 2007; Qin,



2005). Note that the non-parametric tests are based on the principle of *ranking* the data instead of the actual data themselves. The approach to ranking the data is to arrange the scores in ascending order, find the lowest score and label it as 1, then to find the next highest score and label it as 2, and so on. In this scenario, the high scores represent large ranks, and the low scores represent small ranks. Since the non-parametric test is based on ranks rather than actual scores, the mean rank and the median are reported when non-parametric tests are used in this study.

Although non-parametric tests are generally considered to be less powerful than their parametric counterparts due to the possibility of increasing the chance of type II error<sup>38</sup>, Field (2005, p. 533) emphasised that the claim that non-parametric tests are less powerful is cogent only when it is used in data where the assumptions of the parametric test are upheld. Due to the fact that the assumptions of parametric tests were violated in terms of the normality and/or homogeneity of variance in most instances within the current study, non-parametric tests were predominantly carried out to analyse the test scores of achievement assessments. However, the presentation of results of parametric tests has been ubiquitously observed in published journal papers or unpublished works in the field of social science, SLA, and prior PI studies without justifying the fitness to perform parametric tests.

#### ***3.4.2 The Kolmogorov-Smirnov test and Levene's test to examine the fitness of parametric tests***

The Kolmogorov-Smirnov (K-S test, from now on) is generally used to check the normality of the data. Levene's test is used to examine whether the variances in the

---

<sup>38</sup> Type II error is to wrongly accept that no differences between the groups observed, when the differences do exist among the groups in reality.

groups are equal. Both the K-S and Levene's tests assess the null hypothesis, assuming that the distribution of the population is normal, and variances between groups are equal. If the result of the K-S test is non-significant ( $p > .05$ ) (the probability value,  $p$  value, is discussed in the following section 3.4.3), it suggests that the distribution of the data does not significantly deviate from normality; whereas if the result is significant ( $p < .05$ ), the distribution of data deviates from normality. A deviation from normality suggests that the application of parametric tests is not tenable. In a similar way to the indications produced by a K-S test, if the results of Levene's test are significant ( $p < .05$ ), the variances are not significantly equal among groups; therefore, the assumption of homogeneity of variances has not been satisfied, and vice versa. Before the analysis of each achievement assessment, the results of the K-S test and Levene's test for the given achievement assessment are presented to justify the use of parametric or non-parametric tests.

### ***3.4.3 Statistical significance test: setting the probability value***

It is notable that the probability value (i.e.  $p$  value) of the statistical significance test set up for this current study is at the 5% level<sup>39</sup>. If the probability value of a result is small (i.e. less than the 5% level), the result is believed to be true and typically be reported as a statistically significant finding, because there is less than 5% probability that the result is being produced by chance.

Although the statistical significance test is broadly applied, there are some problems in the interpretation and the application of it. It is important to note that the result of the

---

<sup>39</sup> The  $p$  value at the 5% level was principally drawn by Fisher (1925, see more discussion in Field, 2005, p. 24). Although setting the probability value at the 5% level to confirm or falsify the hypotheses in an experimental study is arbitrary, it has been widely taken as a criterion when inferential statistics are carried out. According to Field (2005, p. 25), 'we're just prepared to believe that it is!'

statistical significance test cannot tell us ‘the importance of an effect’ (Field, 2005, p. 27; Norris & Ortega, 2000, p. 493). Put another way, if the result of a treatment is statistically significant, it merely tells us that the result is not a chance finding, but it does not necessarily suggest that the effect is important. The importance of an effect could be observed by the computation of effect size (see discussion in Section 3.4.5) instead of the statistical significance test. The other interpretive problem arises from the non-significant results being observed. A likely misinterpretation is to consider the non-significant effect as no effect. As Field (2005) states, “all that a non-significant result tells us is that the effect is not big enough to be anything other than a chance finding – it doesn’t tell us that the effect is 0 (p. 28).” In addition, Norris & Ortega (2000, p. 493) stressed that though the results of statistical significance tests are frequently reported (i.e. whether a result is significant or not), some types of information are omitted such as the results of descriptive statistics, and the inferential data (e.g. exact p value, degree of freedom, and the inferential statistics tables, etc). The omitted information might cause the loss of practical information and create a difficulty in conducting future meta-analysis. Based on Norris & Ortega’s suggestion, the results of descriptive statistics and inferential data are provided along with those of statistical significance tests in this thesis.

#### ***3.4.4 Statistical tests used in this thesis to examine the mean differences***

As this thesis incorporated both parametric and non-parametric tests to investigate the dependent variables, to prevent the reader’s being confused by a battery of tests, the counterparts of parametric and non-parametric tests are summarised in Table 3.5

Table 3.5

*Summarisation of the counterparts of parametric and non-parametric tests used in this*

*thesis to examine the mean differences.*

<b>parametric tests</b>	<b>non-parametric tests</b>
The independent <i>t</i> -test	The Wilcoxon rank-sum test
The dependent <i>t</i> -test	The Wilcoxon signed-rank test
One-way independent ANOVA	The Kruskal-Wallis test
One-way repeated-measure ANOVA	Friedman's test
Two-way mixed design ANOVA	?

*3.4.4.1 The dependent and independent t-tests and their non-parametric counterparts.*

The *t*-test is commonly used when the researcher intends to compare two means of two conditions or two groups of people. There are two types of *t*-tests, namely the dependent and independent samples *t*-tests. The choice of either of the *t*-tests depends on how the data are collected. The dependent *t*-test is used when the two means to be compared are obtained from the same population (i.e. repeated measure), whereas the independent *t*-test is used when the two means are based on different populations.

The *t*-test is a parametric test, so the parametric assumptions, in theory, should be upheld before performing it. In this thesis, if the parametric assumption is not supported, the Wilcoxon rank-sum test, the non-parametric equivalent of the independent *t*-test, is carried out to examine two groups of means, in which these two groups have no relationship. The Wilcoxon signed-rank test, the non-parametric equivalent of the dependent *t*-test, is used to compare two groups of means in repeated measure. A detailed discussion of the theory behind the Wilcoxon rank-sum and Wilcoxon signed-rank can be found in Field (2005, p.522-541)

*3.4.4.2 The ANOVA and its non-parametric equivalent*

The analysis of variance (ANOVA henceforth) is used to look at differences between

more than two conditions or groups of people. To put it another way, an ANOVA can be used “to analyse situations in which there are several independent variables and it tells us how these independent variables interact with each other and what effects these interactions have on the dependent variable” (Field, p. 309). Like the *t*-test, there are ‘dependent ANOVA’ and ‘independent ANOVA’ according to whether the several means with which the researcher intends to make comparisons come from the same group (i.e. repeated measure) or different groups.

In addition, ANOVA can be used to compare more than one independent variable. The total numbers of independent variables in a design are usually referred to as ‘the numbers – way.’ For instance, one-way ANOVA means that only one independent variable is involved; two-way ANOVA means two independent variables, and so on. Furthermore, it is possible that when conducting a two-way ANOVA, one independent variable measures the same participants whereas the other measures different participants. In this case, the word ‘mixed’ is used to indicate the condition when both the ‘dependent’ and ‘independent’ measures are used. The design is called two-way mixed ANOVA. As far as this thesis is concerned, a one-way independent ANOVA is applied to examine the pre-test scores of different instructional types to ensure the parity between groups prior to the interventions. A two-way mixed design ANOVA is performed to explore the impact of the interventions over time, the two independent variables being the instructional types as a between-subjects variable (independent measure) and the timing of tests as a within-subjects variable (repeated measure). Given that ANOVAs are parametric tests, the results are robust when the parametric assumptions are sustained. Once the parametric assumptions are not upheld, the non-parametric counterparts of ANOVAs would be carried out to evaluate the data in this thesis.

The non-parametric counterpart of the one-way independent ANOVA used in this thesis is the Kruskal-Wallis test. As for the non-parametric equivalent of the two-way mixed design ANOVA, the researcher/I did not find one in the literature on statistics. As a result, the Friedman's test, a non-parametric equivalent of the one-way repeated-measures ANOVA, was carried out. In this thesis, Friedman's test was used to examine each individual group's performance at the pre-test, the post-test, and the delayed post-test, when the parametric assumptions were violated in order to investigate the impact of the interventions. Note that since Friedman's test functions as a one-way repeated-measure ANOVA, it is not viable to be used to compare the impact across different types of interventions (see more discussion of non-parametric tests in Field, 2005, Chapter 13).

#### 3.4.4.3 *The planned contrast*

An ANOVA merely reveals whether or not statistically significant differences between groups exist, but it does not indicate where the differences between groups lie (Field, 2005, p. 325). After a significant difference is found, it is, therefore, necessary to carry out further analysis to examine which groups differ. Two possible approaches are suggested for performing further analysis of an ANOVA: *post hoc* tests and planned contrasts. The difference between these two tests is similar to that of two-tailed and one-tailed<sup>40</sup> tests. In *post hoc* tests, the hypotheses are non-directional; on the other hand, the hypotheses are directional in planned contrasts. As the current study derived from specific hypotheses based on the literature review before the data was collected (e.g. the

---

<sup>40</sup> The use of one- or two-tailed tests depends on a hypothesis of the study, developed before the collection of data. If the study makes a prediction with respect to what will happen with direction, a one-tailed test should be applied. A two-tailed test is used according to the hypothesis without predicting the direction of results.

affective activities might not be as helpful as referential activities in learning the English past tense marker), the planned contrasts approach, instead of the *post hoc* test, would be applied when the ANOVA detected differences between groups.

#### *3.4.4.4 The ANCOVA*

Analysis of covariance (ANCOVA henceforth) is an extension of ANOVA. ANCOVA allows us to explore differences between groups whilst statistically controlling for additional variables (aka covariates), which are not part of the main experimental manipulation but we suspect that they may have an influence on the dependent variables (see Field, 2005, Chapter 9; Pallant, 2007, Chapter 22). Note that the variable chosen as a covariate should be a continuous variable, and should correlate significantly with the dependent variable. An additional assumption, homogeneity of regression slopes, is required to check the suitability of the application of an ANCOVA. This assumption concerns the overall relationship between the dependent variable and the covariate(s). If the homogeneity of regression slopes is not achieved, the results of ANCOVA are misleading, and therefore the ANCOVA should not be conducted (Stevens, 1996, p.323, 331; Tabachnick & Fidell, 2007, p.202, cited in Pallant, 2007, p.293). Given that two confounding variables were observed in this study (see Chapter 5), an ANCOVA was conducted to examine the differences between instructional groups while controlling the covariate.

#### *3.4.5 The estimate of the magnitude of interventions: effect size*

Reporting the effect size in empirical studies involving quantitative and statistical procedure is strongly advocated (R. Ellis, 2000; Field, 2005, p.32-33; Norris & Ortega, 2000, p.442-443). As mentioned earlier, the effect size offers an objective estimation of the importance of an effect and it can be used to compare the magnitude of the observed

effects across different studies in which they examine different independent variables, or adopt a different scale of dependent variables (Field, 2005). Additionally, unlike the statistical significance test, the interpretation of effect size is not affected by the sample size (Cohen, 1990; Ellis, 2000; Norris & Ortega, 2000, p.425). The computation of the effect sizes in this thesis is based on Cohen's  $d$ , which only requires the fundamental descriptive statistics, namely the group sample sizes, means of dependent variables, and standard deviations of the two contrasted groups. As the SPSS does not have the options for calculating Cohen's  $d$ , the computation of Cohen's  $d$  is displayed in equations (1) and (1.1), which were suggested and applied in the study of Norris & Ortega (2000, p. 442-443).

$$d = \frac{(\text{mean}_e - \text{mean}_c)}{S_w} \quad (1)$$

$$S_w = \frac{[(N_e - 1)S_e + (N_c - 1)S_c]}{(N_e - 1) + (N_c - 1)} \quad (1.1)$$

In the above equations,  $\text{mean}_e$  and  $\text{mean}_c$  represent the means of experimental and control group respectively.  $N$  stands for the sample size, so  $N_e$  and  $N_c$  are the sample sizes for experimental and control groups.  $S$  stands for the standard deviation, and  $S_e$  and  $S_c$  are the standard deviations of experimental and control groups. Regarding guidelines for the strength of effect size, Cohen (1988, cited in Norris & Ortega, 2000, p.465) suggested that a specific intervention would be considered as having a small effect ( $.2 < d < .5$ ), a medium effect ( $.5 < d < .8$ ), or a large effect ( $d > .8$ ).

### **3.4.6 The correlation**

A correlation is typically used to examine the relationship between variables. The correlational coefficient ranges from -1 to 1. A zero value of correlational coefficient



means that the variables do not relate at all. A positive coefficient means a positive relationship between the variables and vice versa. The more the correlational coefficient approximates to the extremes (i.e. +1 or -1), the stronger the variables' association. The application of correlation in this study aims to investigate the relationship between the self-report of participants (i.e. whether using the rule to undertake the assessments) and their performances on the implicit measures (i.e. the timed GJT and oral tests) to find out if the implicit measures actually have any relationship with their self-reported use of explicit knowledge. Additionally, the correlation is conducive to exploring the interrelationships between the different achievement assessments used in this study. Presumably, the implicit measures would positively associate with other implicit measures, but not necessarily with the explicit measure.

As for the selection of different correlation methods used in this thesis, only the bivariate correlation is applied instead of partial correlation<sup>41</sup>. The bivariate correlation used in this thesis includes Pearson's product-moment correlation coefficient ( $r$ ), Spearman's rho ( $r_s$ ), point-biserial correlation ( $r_{pb}$ ), and biserial correlation coefficient ( $r_b$ ).

Pearson's correlation and Spearman's rho are distinguished by whether the parametric assumptions are violated, in that Pearson's correlation is a parametric statistic and Spearman's rho is a non-parametric statistic. Both of the point-biserial and biserial correlations are applied when one of the two variables is categorical and dichotomous. The difference between the two correlations lies in whether the dichotomous variable is

---

<sup>41</sup> A bivariate correlation merely copes with two variables, whereas a partial correlation looks at two variables when additional variable(s) is controlled (see Field, 2005, Chapter 3).

‘discrete’ or ‘continuous’ (see discussion in Field, 2005, p.131-132 for more details). The point-biserial correlation is used when the relationship between the dichotomous variables is clearly distinguishable, such as gender, and it was widely used in the current study when analysing the questionnaire regarding participants’ English learning background. The computation of point-biserial correlation is the same as that of Pearson’s correlation, so SPSS can perform this.

On the other hand, the biserial correlation coefficient is used when the relationship between the dichotomies is ‘continuous’<sup>42</sup>. As far as this study was concerned, the participants’ self-reports were dichotomously categorised as ‘using the rule’ or ‘not using the rule’ during the testing phases, and the dichotomous categories were not so discrete e.g. the participants in the ‘rule-user’ group might differ to some degree in how often they used the rule. In this regard, the biserial correlation was used to explore the relationship between participants’ self-report and their test scores. Due to the fact that SPSS cannot carry out biserial correlation, the equation for calculating the biserial correlation coefficient is shown in equation (2), as suggested by Field (2005, p.133).

$$r_b = \frac{r_{pb} \sqrt{(P_1 P_2)}}{Y} \quad (2)$$

The  $r_{pb}$  stands for point-biserial correlation coefficient.  $P_1$  and  $P_2$  are the proportions of cases that fall into the given dichotomies, respectively.  $Y$  is the ordinate of the normal distribution, obtained by checking the values of  $P_1\%$  and  $P_2\%$  in the table provided by Field (2005, p.751-754).

### ***3.4.7 The principal component analysis***

---

<sup>42</sup> The fact of being continuous means that even though the data are categorised into two different groups, the relationships between data in each individual specific group may differ to some extent.

In order to examine what types of knowledge each achievement assessment test elicited, a principal component analysis was performed, adhering to R. Ellis' (2005) procedure for measuring implicit and explicit knowledge. In brief, a principal component analysis is a 'data reduction' approach. It looks at the intercorrelations of a set of variables, and then reduces the variables into smaller numbers of factors/components, based on the relationships between variables (Kinnear & Gray, 2000). As far as the current study is concerned, the principal component analysis was used to examine the intercorrelations of dependent variables (i.e. the achievement assessments), and then to extract the underlying dimensions of the variables (i.e. tapping explicit or implicit knowledge).

#### *3.4.7.1 Suitability/factorability of principal component analysis*

Before a principal component analysis is performed, Pallant (2007, p.180-181) suggests checking the suitability of the data for principal component analysis by taking two issues into consideration, namely the sample size and the strength of the intercorrelations between the variables. The Kaiser-Meyer-Olkin (KMO) measure is used to check the sampling adequacy and it ranges from 0 to 1. The greater the KMO value, the more adequate the sampling. As for the criteria for deciding the sampling adequacy, a KMO value with .5 is suggested as the barely acceptable value for a good principal component analysis (Kaiser, 1974, cited in Field, 2005, p.640).

In terms of the strength of the intercorrelations between the variables, since the purpose of conducting the principal component analysis is to reduce the variables into smaller number components, there should be some relationships existing between the variables for principal component analysis to work out. Therefore, it would make no sense if the variables correlate too highly with all others (imaginably, only one component would be extracted in the end) or variables do not correlate with any other variables (the number

of variable equalises those of extracted components). Two statistical measures, the determinant of the R-matrix and Bartlett's test of sphericity, are used to inspect the intercorrelations between the variables.

The determinant of the R-matrix is used to detect whether the extreme multicollinearity (i.e. variables are highly correlated) and singularity (i.e. variables are perfectly correlated) exist. The determinant of the R-matrix is recommended to be greater than .00001(Field, 2005, p.641). The Bartlett's test of sphericity is used to inspect whether the population correlation matrix approximates to an identity matrix (i.e. all correlation coefficients are close to zero). As Bartlett's test of sphericity tests the null hypothesis, its result should be significant ( $p < .05$ ) for the principal component analysis to be considered suitable. The three statistical measures (i.e. the KMO, the determinant of the R-matrix, and the Bartlett's test of sphericity), used to inspect the suitability of principal component analysis, could be generated by SPSS and they are reported to examine the factorability in Chapter 4.

#### *3.4.7.2 The criterion for extracting components*

There are a number of approaches that could be used as a criterion to decide the number of extracted components (see discussion in Pallant, 2007, p. 182-183). One of the most common approaches is to look at the eigenvalue. The eigenvalue is used to indicate the substantive importance of a given component, and it tells us the amount of the total variance explained by that component. Kaiser's criterion of eigenvalue-greater-than-one is commonly applied to decide whether a specific component is retained (Field, 2005, p. 633; Pallant, 2007, p. 182). Jolliffe (1972, 1986, cited in Field, p.633) recommended retaining factors with eigenvalues over .7. Kaiser's suggestion of using the eigenvalue over 1 as the criterion was adopted in the current study, given that "Kaiser's criterion is

accurate when the number of variables is less than 30” (Field, 2005, p. 633). Because a total of four variables were involved in this study (i.e. a total of 4 achievement assessments designed to measure implicit and explicit knowledge), the eigenvalue-greater-than-one criterion was applied to decide the number of components to be extracted.

#### *3.4.7.3 To assist in the interpretation: factor rotation technique*

Basically, an unrotated principal component analysis is used as a first step to determine the number of extracted components. However, most variables, in general, have high loadings on the most important component, and small loadings on the rest of the components. Under this circumstance, the interpretation of results becomes difficult. Thus, the technique of ‘factor rotation’ is used to better discriminate between components by maximising the loading of the variables onto one component and minimising onto the remaining component(s) in order to improve the interpretation (Field, 2005; Pallant, 2007, p.183).

There are two types of factor rotation, namely orthogonal rotation and oblique rotation. The difference between these two factor rotations merely depends on the relationship between the components (see Field, 2005, p.634-636). Orthogonal rotation is used when all components are independent (i.e. they are uncorrelated); oblique rotation, on the other hand, should be used when the components are related (i.e. they are allowed to correlate). Therefore, the choice of rotation relies on whether there are good theoretical grounds to reason that the components are unrelated or related.

In the current study, oblique rotation, as opposed to orthogonal rotation, is considered to be more appropriate to perform the factor rotation. Note that the SPSS offers two

oblique rotation methods (i.e., the direct oblimin and Promax). The direct oblimin method was selected for further analysis, because the data sets collected in this study were considered small (see Field, 2005, p. 637). The justification for the use of oblique rotation is as follows. Field (2005, p.637) commented that orthogonal rotation should not be used for any data incorporating humans. Also, Cattell & Scheuberger (1978, cited in Kline, 1994, and also cited in Isemonger, 2007, p.106) argued that orthogonal rotation should not be used when components involved psychological phenomena. In addition, R. Ellis (2005, p.153) and Roehr (2008, p.191) pointed out the unlikelihood of constructing tests that could be used purely to measure explicit and implicit knowledge (i.e. they are likely to be related to each other to some extent as these two constructs are not entirely separable). Isemonger (2007, p.106), in the exchange of ideas on R. Ellis' study (2005), recommended that the two factors (i.e. implicit and explicit knowledge) should be correlated.

#### ***3.4.8 Pearson's chi-square test for independence***

Pearson's chi-square test for independence was broadly used in this thesis to analyse the questionnaires germane to participants' English learning background and their attitude towards the interventions. This test is used to examine the relationship between two categorical/nominal variables by comparing the observed frequencies or proportions in each of the variables. It is noted that the chi-square test is a non-parametric test, but two other fundamental assumptions require to be upheld for the chi-square test to be accurate. The assumptions are that a chi-square test should not be used on a repeated-measured design, and either the expected count should be greater than 5 or at least 80% of cells should have expected count greater than 5 (Field, 2005, p. 686; Pallant, 2007, p. 216). The violation of the assumption might cause a loss of statistical power (i.e., the test may fail to detect a genuine effect). In the current study, the first assumption was

upheld because the participants merely received the questionnaires once. However, the second assumption was not satisfied in some cases due to the inadequate sample size, when performing the chi-square test in Chapter 4 & 5. To improve this, further study is needed to increase sample size, but for these results of this study the expected frequencies are reported along with the results of the chi-square test.

### **3.5 The validity, reliability and comparability of the assessment tests**

The validity and reliability of the testing instruments appear to have been overlooked by SLA researchers (Douglas, 2001). Norris & Ortega (2000) stressed the necessity of considering “the validity of dependent variables in terms of the kinds of interpretation to be based on them: estimate and report the consistency or reliability of the use of outcome measures” (p. 498). Thus, this section focuses on addressing how the concepts of validity and reliability in this study are conceptualised and then shows the evidence of validity and reliability of the achievement assessments used in the current study. Subsequently, the evidence to prove the equivalence of two versions of the achievement assessment is displayed. It is acknowledged that not all validity and reliability of the tests used in this study were demonstrated. Due to some practical reasons, the validity of the oral tests and the vocabulary test was not executed. The failure to report the validity of these tests mainly lay in the difficulty in recruiting the participants and time limitation.

Table 3.6 below encapsulates whether or not a specific assessment used in this study was examined in terms of validity, reliability and comparability. The tests marked o were assessed in terms of their validity, reliability, and comparability, whilst those marked x were not. Two classes recruited from two different schools, neither of which ever took part in the intervention of the main study, were invited to examine the validity,

test-retest reliability and comparability of the assessments. Class 1 consisted of 27 grade 6 participants who, in principle, had not learnt the target feature before taking the tests. Class 2 consisted of 25 grade 8 participants who had learnt the target feature prior to the tests.

Table 3.6

*Summary of the investigation into the validity, reliability, and comparability of the achievement assessments*

	<b>GJT</b>	<b>Gap-fill<sup>43</sup></b>	<b>Oral 1</b>	<b>Oral 2</b>	<b>Vocabulary</b>
<b>a) validity</b>	o	o	x	x	x
<b>b) reliability</b>					
<b>i) test-retest</b>	o	o	x	x	x
<b>ii) Cronbach's <math>\alpha</math></b>	o	o	o	x	o
<b>c) comparability of two versions</b>	x	o	o	x	o

*Note: Oral 1: the picture-based narration test; oral 2: the structured conversation  
(o=assessed, x=not assessed)*

### **3.5.1 The validity of the achievement assessments**

According to Carmines & Zeller (1979), the validity of an assessment refers to “the extent to which any measuring instrument measures what it is intended to measure” (p.17). Because the testing instruments used in this study were aimed at investigating the effectiveness of the interventions (e.g. the improvement after receiving the interventions), the validity of the tests in this study was operationalised by comparing the performances of two populations (one had learnt the targeted linguistic feature, the other had not) on the testing instruments as demonstrated in Erlam’s study (2003). An

<sup>43</sup> It is noted that the scoring procedure for the gap-fill test in Section 3.5 was slightly different from those used for the gap-fill test in Chapter 4 & 5. In Section 3.5, two points would be given for a target item if a correct answer was provided. If a participant misspelled a vocabulary but correctly attached the target feature (-ed ending), such as ‘studyed’ or ‘shoped’, one point would be rewarded. As a result, the maximum score was 16 instead of 8. However, the difference in the scoring procedure would not affect the validity, reliability, and comparability of the assessments used in this study as they should be valid and reliable regardless of the differences.



achievement assessment would be regarded as being valid if the two populations performed significantly differently and vice versa.

### 3.5.1.1 The validity results for the achievement assessments

#### i) Can a parametric test be used?

Due to the fact that none of the tests (i.e. the timed GJT and the gap-fill test) in both Class 1 (grade 6 learners) and Class 2 (grade 8 learners) met the parametric assumptions of normality (see Appendix 29) and homogeneity of variance (see Appendix 30), the Wilcoxon rank-sum test, a non-parametric equivalence of independent *t*-test, was carried out to examine the performances of the two populations on these tests. In order to maintain the parity with other studies, the results of parametric tests for assessing the validity are given in Appendix 37. The pattern of statistically significant results did not differ between the non-parametric and parametric tests.

#### ii) The results of the Wilcoxon rank-sum test for assessing the validity

The mean rank and the median for the assessments to examine the validity are given in Table 3.7.

Table 3.7

*Descriptive statistics for the non-parametric test in assessments to investigate validity*

<b>Assessment</b>	<b>Population</b>	<b>N</b>	<b>Mean Rank (MR)</b>	<b>Median (Md)</b>
<b>Timed GJT</b>	Class 1	27	19.04	11.00
	Class 2	25	34.56	20.00
<b>Gap-fill test A</b>	Class 1	27	18.26	.00
	Class 2	25	35.40	12.00
<b>Gap-fill test B</b>	Class 1	27	17.83	.00
	Class 2	25	35.86	8.00

Class 1= grade 6; Class 2=grade 8

Table 3.8 summarises the results of the Wilcoxon rank-sum test. The results of the Wilcoxon rank-sum test indicated that both classes performed significantly differently on the timed GJT ( $W_s = 514.00$ ,  $p = .00 < .01$ ), version A of the gap-fill test ( $W_s = 493.00$ ,  $p = .00 < .01$ ), and version B of the gap-fill test ( $W_s = 481.50$ ,  $p = .00 < .01$ ). On examination of the mean rank and median in Table 3.7, it was observed that the significant differences lay in that Class 2 outperformed Class 1 on the timed GJT and two versions of the gap-fill test. To sum up, the results were considered as convincing evidence in terms of the validity of the timed GJT and the gap-fill test. The use of both tests to investigate the effectiveness of the interventions on the acquisition of the English regular past tense ‘-ed’ form was justified.

Table 3.8

*The Wilcoxon rank-sum results on the validity of the timed GJT and the gap-fill test*

<b>Assessment</b>	<b>Paired of population</b>	$W_s$	<b>Sig. (2-tailed)</b>
<b>Timed GJT</b>	C1 vs. C2	514.00	.000
<b>Gap-fill test A</b>	C1 vs. C2	493.00	.000
<b>Gap-fill test B</b>	C1 vs. C2	481.50	.000

*Note: C1 = Class 1 (grade 6, n=27); C2 = Class 2 (grade 8, n=25)*

### ***3.5.2 The reliability of the achievement assessments***

The reliability of an assessment concerns whether or not a participant’s performance is consistent in a test. According to Carmines and Zeller (1979), “Reliability concerns the extent to which an experiment, test, or any measuring procedure yields the same results on repeated trials”(p. 11). Field (2005, p. 666) pointed out that ‘reliability just means that a scale should consistently reflect the construct it is measuring’. The reliability of the assessments in this study was examined by two methods, namely the test-retest method (Field, 2005, p.666; Carmines & Zeller, 1979; Kline, 2000, p. 8) and the

internal consistency method, Cronbach's alpha (Field, 2005, p.668; Carmines & Zeller, 1979, p.43). The following briefly presents these two methods for estimating the reliability of achievement assessments, and their criterion for judging whether a test is reliable. The reliability results of the achievement assessments used in the current study are then presented.

#### *3.5.2.1 The test-retest method*

Test-retest reliability is based on the concept that a measure delivered to subjects should produce approximately associated scores if it is carried out at two different timings. The computation of test-retest reliability was performed via two types of statistical tests, namely comparing the two sets of means of the measure, and computing the correlation of the two sets of scores conducted at two different timings. If a measure is considered to be reliable, a mean difference acquired at two different timings of the measure should not be observed, and the correlation coefficient between the two sets of scores should be positive. The use of correlation to measure the reliability of tests was observed in some studies (Gass, 1994; Mandell, 1999). A minimum correlation of .8 is suggested by Kline (2000, p.11) as the criterion for test-retest reliability.

In this study, the interval between the two testing phases was one week. The advantage of the short testing interval is the possibility of reducing the extra exposure to the targeted feature. On the other hand, the disadvantages of the short testing interval are the memory (e.g. subjects may remember the test items) and the test effect (e.g. familiarity with the tests or realisation of what the test is for). However, the handicap of a short interval between the two testing phases might be lessened, given that subjects involved in the examination of test reliability in the first testing phase were not told that they would take the tests again one week later. Both Class 1 and Class 2, involved in

examining the validity of the achievement assessments, were recruited to conduct the test-retest reliability. Both Class 1 and Class 2 received a version of the timed GJT, and both versions of the gap-fill test together at one session. One week later, the same achievement assessments were given to both classes.

*a) The test-retest reliability results*

*i) Can a parametric test be used?*

Since the results of the K-S test showed that these tests violated the assumption for performing the parametric tests (see Appendix 29), the non-parametric Wilcoxon signed-rank test and the Spearman's correlation coefficient  $\rho, r_s$ , were used to examine the test-retest reliability for these assessments. The results of parametric tests are provided in Appendix 38. There was no different pattern of statistically significant results observed between the non-parametric and parametric tests.

*ii) The results of the non-parametric tests for assessing test-retest reliability*

The mean scores and the median for the assessments examining the test-retest reliability are given in Table 3.9. Table 3.10 summarises the results of the Wilcoxon two-tailed signed-rank test in assessing the test-retest reliability of the timed GJT and two versions of the gap-fill test. In terms of the timed GJT, the results indicated that no statistical differences between the two sets of test scores were observed in Class 1 ( $z=-1.095$ ,  $p=.273>.05$ ) and Class 2 ( $z=-1.620$ ,  $p=.105>.05$ ) at two different testing times. Similarly, no significant differences were found in both A and B versions of the gap-fill test, which were retaken with a one-week interval, in both Class 1 (A version:  $z=-1.000$ ,  $p=.317>.05$ ; B version:  $z=-1.000$ ,  $p=.317>.05$ ) and Class 2 (A version:  $z=-.1434$ ,  $p=.151>.05$ ; B version:  $z=-.852$ ,  $p=.394>.05$ ). As a whole, the results of the Wilcoxon signed-rank test indicated that there were no significant differences in the timed GJT

and A & B versions of the gap fill test delivered to the same participants twice at a one-week interval.

Table 3.9

*Descriptive statistics in assessments to investigate the test-retest reliability*

Assessment	Population	N	Mean	Median (Md)
<b>Timed GJT1</b>	Class 1	27	13.78	11.00
<b>Timed GJT2</b>			14.52	13.00
<b>Timed GJT1</b>	Class 2	25	25.04	20.00
<b>Timed GJT2</b>			26.36	29.00
<b>Gap-fill test A1</b>	Class 1	27	.33	.00
<b>Gap-fill test A2</b>			.37	.00
<b>Gap-fill test B1</b>	Class 1	27	.30	.00
<b>Gap-fill test B2</b>			.37	.00
<b>Gap-fill test A1</b>	Class 2	25	8.04	12.00
<b>Gap-fill test A2</b>			8.80	12.00
<b>Gap-fill test B1</b>	Class 2	25	8.04	8.00
<b>Gap-fill test B2</b>			8.24	8.00

Class 1= grade 6; Class 2=grade 8

Table 3.10

*The Wilcoxon signed-rank results for the rest-retest reliability*

Paired of assessment	Population	N	z	Sig. (2-tailed)
<b>Timed GJT1 vs. GJT2<sup>44</sup></b>	Class 1	27	-1.095	.273
<b>Timed GJT1 vs. GJT2</b>	Class 2	25	-1.620	.105
<b>Gap-fill test A1 vs. A2</b>	Class 1	27	-1.000	.317
<b>Gap-fill test B1 vs. B2</b>	Class 1	27	-1.000	.317
<b>Gap-fill test A1 vs. A2</b>	Class 2	25	-.1434	.151
<b>Gap-fill test B1 vs. B2</b>	Class 2	25	-.852	.394

Class 1= grade 6; Class 2=grade 8

With respect to the correlational results of the two-set test scores taken at different timing points, Spearman's correlation coefficients,  $r_s$ , are reported in Table 3.11. In the

<sup>44</sup> The 1 & 2 attached in the test title mean the different timings for subjects taking the test. For example, GJT 1 refers to the first time participants took this test, and GJT 2 refers to the second time a week after the first time.

timed GJT, there was a positively significant relationship observed in Class 1,  $r_s = .598$ ,  $p = .001 < .01$ , and in Class 2,  $r_s = .800$ ,  $p = .00 < .01$ . As for the gap-fill test, the results showed that both scores of A and B versions of the gap-fill test correlated well with those scored a week later in both Class 1 (A version:  $r_s = 1.00$ ; B version:  $r_s = 1.00$ ) and Class 2 (A version:  $r_s = .914$ ,  $p = .00 < .01$ ; B version:  $r_s = .949$ ,  $p = .00 < .01$ ). The reason why the correlation coefficients in Class 1 were highly positively correlated in both versions of the gap-fill test at different testing times was that the same 26 out of 27 participants in Class 1 did not score at two different testing times. Only one, and the same participant, scored in both versions at two different testing times.

Table 3.11

*The results of Spearman's correlation for the test-retest reliability*

<b>Paired of assessments</b>	<b>Population</b>	<b>N</b>	<b>Spearman's <math>r_s</math></b>	<b>Sig. (2-tailed)</b>
<b>Timed GJT1 vs. GJT2</b>	Class 1	27	.598**	.001
<b>Timed GJT1 vs. GJT2</b>	Class 2	25	.800**	.000
<b>Gap-fill test A1 vs. A2</b>	Class 1	27	1.000**	.
<b>Gap-fill test B1 vs. B2</b>	Class 1	27	1.000**	.
<b>Gap-fill test A1 vs. A2</b>	Class 2	25	.914**	.000
<b>Gap-fill test B1 vs. B2</b>	Class 2	25	.949**	.000

Class 1= grade 6; Class 2=grade 8

To sum up, the non-parametric Wilcoxon signed-rank test showed that no statistically significant differences between paired test scores were found in the timed GJT and A & B versions of the gap-fill test, completed at two different testing times. In addition, the test scores obtained at two different testing times were highly positively correlated. The correlation coefficients of both versions of the gap-fill test were greater than the value of .8, suggested by Kline (2000) as a criterion for the test-retest reliability. Although the

correlational result of the timed GJT in Class 1 was below the threshold value of .8,  $r_s = .598$ , this might be due to the fact that participants had not learnt the target form, so they might have employed the strategy of guessing, possibly leading to the inconsistent performance in the timed GJT. However, the performances of Class 2 in the timed GJT were considered as consistent. As a result, the test-retest reliability is encouraging. The results suggested that the assessments were fairly reliable in terms of the timed GJT and the gap-fill test.

### 3.5.2.2 Cronbach's alpha ( $\alpha$ )

The internal consistency reliability of the tests was estimated by means of Cronbach's alpha ( $\alpha$ ). Cronbach's  $\alpha$ , is generally used to test a specific construct in a measure, based on the computation of correlations between each individual test item and the overall test items. The value of Cronbach's  $\alpha$  ranges from 0.0 to 1.0. The higher  $\alpha$  is, the more reliable a test is. With regard to the criterion of an acceptable Cronbach's  $\alpha$ , Carmines & Zeller (1979, p. 51) suggested the internal consistency reliability for a scale should not be below .8. Kline (2000, p.15) noted a value of .7 as a minimum indicator for a good consistent test. In general, Cronbach's  $\alpha$  value of above .7 or .8 is acceptable to indicate a reliable test (Field, 2005, p.668). Pallant (2007, p. 98) suggested that a value of about .7 is considered acceptable, and a value of above .8 is preferable.

As Cronbach's  $\alpha$  requires only the administration of a single test, the participants' performances in the main study on the timed GJT, the gap-fill test, the vocabulary test, and the picture-narration test at the post-test were analysed to estimate the internal reliability of these tests used for the current study<sup>45</sup>. The results of internal reliability,

---

<sup>45</sup> The Cronbach's alphas for the timed GJT, the gap-fill test, and the vocabulary test were computed on the basis of a total of 99 instructional participants, including the 9 outliers who were excluded for further analyses in Chapter 4 & 5. The Cronbach's alpha for the picture narration was based on 36 instructional

descriptive statistics, and the number of test items of each individual achievement test are tabulated in Table 3.12. Note that among the achievement assessments, the internal consistency reliability of the structured conversation was not ascertained, given that the structured conversation was scored in the format of obligatory context, in which is not possible to compute Cronbach's  $\alpha$ .

Cronbach's alpha obtained for the timed GJT (20 target items) was .9233, which was considered fairly reliable, suggesting that the timed GJT used in this study was highly internally consistent. In terms of the gap-fill test, Cronbach's alpha estimated for both A and B versions was .9678, and .9618 respectively, both over .9. Thus, both versions of the gap-fill test used in this study were considered to be very reliable in terms of internal consistency. Cronbach's  $\alpha$  estimated for the picture-based narration test (8 targeted items) was .9279 for the A version and .7421 for the B version. Although the Cronbach's  $\alpha$  in the B version was not as high as that in the A version, it is considered to be acceptable. As for the vocabulary test, Cronbach's  $\alpha$  estimated for the A version was .7222, and .8344 for the B version. Therefore, the internal consistency reliability of both A & B versions of the vocabulary tests is considered acceptable.

To sum up, Cronbach's alphas computed for the timed GJT, the gap-fill test, the picture-based narration test, and the vocabulary test are regarded as reliable and consistent, given that all of Cronbach's alphas were greater than .7. In addition, by observing the standard deviations (SD) of both versions of the gap-fill test and the picture-based narration test, the internal consistency reliability results in this study appeared to lend support to Brown's (1996) claim that a measure is more reliable if it is delivered to a group of a wider range of abilities than if delivered to that of a more narrow range of

---

participants, including the 8 outliers who were excluded for analyses in Chapter 4 & 5.



abilities.

Table 3.12

*The internal reliability results and descriptive statistics on achievement tests*

<i>Test</i>	<i>N of items</i>	<i>Cronbach's <math>\alpha</math></i>	<i>M</i>	<i>SD</i>
<b>Timed GJT</b>	20	.9233	20.65	11.43
<b>The gap-fill test</b>				
Version A	8	.9678	2.22	3.27
Version B	8	.9618	1.88	3.03
<b>The picture narration</b>				
Version A	8	.9279	1.08	2.26
Version B	8	.7421	2.50	2.11
<b>The vocabulary test</b>				
Version A	10	.7222	2.46	1.80
Version B	10	.8344	1.88	2.30

### ***3.5.3 The comparability of the two versions of the achievement assessments***

The comparability of the gap-fill test, the picture narration test and the vocabulary test used in this study was achieved and reported in this section. Due to the fact that the only discrepancy between the two versions of the timed GJT was the order of test items, the equivalence of the two versions of the timed GJT was reasonably presumed to be identical. Accordingly, the comparability of the timed GJT in the A & B versions was not examined.

The two versions of the gap-fill test were conducted in Class 1 and Class 2, which also took part in assessing the validity and test-retest reliability. Both Class 1 and Class 2 received both A and B versions of the gap-fill test together in a session. Note that the inclusion of two classes at different developmental stages was mainly intended to assess

the validity of the achievement assessments (see section 3.5.1), which was not directly related to the examination of comparability of assessments, but the difference does not affect the reliability tests as they should be reliable regardless of the proficiency of the participants. Note that the comparability of the two versions of the picture-based narration test was not examined in Class 1 and Class 2 due to the limitation on time and difficulty in recruiting the participants. However, this was achieved by comparing the A and B versions delivered to the participants in the main study at the pre-test.

The participants in the control group for the main study were included in the vocabulary tests in order to test the comparability of the two versions. Each participant in the control group was required to take both A and B vocabulary tests during the pre-test phase. The reason why the participants in the control group were recruited rather than those in Class 1 and Class 2 is as follows. In Taiwan, the schools have been entitled to make their own decision on the adoption of a specific English textbook. There are a variety of English textbooks published by various publishers for schools to choose which textbook better suits their needs for English courses. It is noted that the syllabus of the various textbooks at the same grade is prescribed according to the guidelines drawn up by the Taiwanese Ministry of Education. Although the syllabus is essentially the same, the use of vocabulary in different textbooks may vary to some extent. This could affect the results since the participants of Class 1 and Class 2 used different English textbooks from those participating in the main study.

Given that the four classes (3 instructional groups and one control group) in the main study came from the same English learning setting (they were taught by the same English teacher, and used the same English textbook at the same school), the results could be more convincing if the participants of the control group, as opposed to Class 1

and Class 2 in different schools from the main study, were recruited to substantiate the comparability of the two versions of the vocabulary tests. Based on the aforementioned viewpoints, the control group was invited to test the equivalence of the two versions of the vocabulary test. Since the participants in the control group received both versions of the vocabulary test at the pre-test, they were not invited to take part in the post-test and the delayed post-test. However, one disadvantage of excluding the control group in the vocabulary post-tests was that the investigation of relative effectiveness of the three interventions would be restricted to some extent without the comparison with a control group.

#### *3.5.3.1 The comparability results of the two versions of the achievement assessments*

##### *i) Can parametric tests be used?*

The results of the K-S test for the A & B versions of the achievement assessment with regard to its normality of distributions are provided in Appendix 29. The K-S test results showed that the two versions of the gap-fill test, the picture narration test, and the vocabulary test were all non-normal distributions. Due to the fact that the parametric assumption of normality was severely violated, the parametric test (i.e. the dependent *t*-test) was less powerful in examining the comparability of different versions assessments. As a result, the non-parametric equivalent of the dependent *t*-test, the Wilcoxon signed-rank test, was performed to assure the equivalence of these tests. The results of parametric test are given in Appendix 39 in order to maintain the parity of other studies. The pattern of statistically significant results did not differ between the nonparametric and parametric tests.

##### *ii) The results of the non-parametric Wilcoxon signed-rank test for the comparability of the two versions of the achievement assessments.*

Table 3.13 summarises the mean scores and the median for the comparability of the two versions of the achievement assessments.

Table 3.13

*Descriptive statistics for the comparability of the two versions of the tests*

<b>Assessment</b>	<b>Population</b>	<b>N</b>	<b>Mean</b>	<b>Median (Md)</b>
<b>Gap-fill test A</b>	Class 1	27	.33	.00
<b>Gap-fill test B</b>			.30	.00
<b>Gap-fill test A</b>	Class 2	25	8.04	12.00
<b>Gap-fill test B</b>			8.04	8.00
<b>Picture test A</b>	Subjects at the pre-test <sup>46</sup>	21	.62	.00
<b>Picture test B</b>		27	.93	1.00
<b>Vocabulary test A</b>	Control group <sup>47</sup>	34	1.59	1.00
<b>Vocabulary test B</b>			1.47	1.00

The results of the non-parametric tests for assessing comparability of the two versions of the achievement assessments are summarised in Table 3.14. The results of the Wilcoxon signed-rank test showed that statistically significant differences between both A and B versions of the gap-fill test were not found in either Class 1 ( $z=-1.000$ ,  $p=.317>.05$ ) or Class 2 ( $z=-.810$ ,  $p=.418>.05$ ). No statistically significant difference was found in the two versions of the picture narration test ( $W_s=494.00$ ,  $p=.637>.05$ ) by the Wilcoxon rank-sum test. In terms of the vocabulary test, the result of the Wilcoxon signed-rank test showed that both versions of the vocabulary test were not found to be significantly different ( $z=-.853$ ,  $p=.394>.05$ ). Due to the fact that no significant difference was observed when testing the equivalence of tests, any differences observed across the timing of testing (the pre-test, the post-test, and the delayed post-test), or across the groups (different interventions) in later analyses should not be attributed to

<sup>46</sup> The subjects included those outliers who were excluded in the analyses in Chapter 4 & 5.

<sup>47</sup> The subjects included those outliers who were excluded in the analyses in Chapter 4 & 5.

the non-equivalent versions of assessments in terms of the gap-fill test, the picture narration test, and the vocabulary test.

Table 3.14

*The results of the comparability of the two versions of the tests*

<b>Pairs of Assessments</b>	<b>Classes</b>	<b>N</b>	<b>z or Ws</b>	<b>Sig. (2-tailed)</b>
<b>Gap-fill test A vs. B</b>	Class 1	27	$z=-1.000$	.317
<b>Gap-fill test A vs. B</b>	Class 2	25	$z=-.810$	.418
<b>Picture test A vs. B</b>	Subjects in main study at the pre-test	48	$W_s=494.000$	.637
<b>Vocabulary test A vs. B</b>	control group	34	$z=-.853$	.394

To sum up, the results of the non-parametric test suggested that the A & B versions of the gap-fill test, the picture narration test and the vocabulary test were considered as being equivalent in terms of the level of difficulty. Consequently, the results showed convincing evidence for the claim that any differences observed on these achievement assessments in the main study at the pre-test, the post-test, and the delayed post-test should not be ascribed to the use of the different versions of the tests.

Furthermore, the delivery of the achievement assessments was a split-block design and each achievement assessment had two versions so as to reduce the likelihood of a test effect to some extent (Marsden, 2006, p.527; Toth, 2006, p.343). Since the two versions of the measures has been demonstrated to be equivalent (see Table 3.14), and the results of test-retest conditions one week apart did not reveal any significant improvement either in Class 1 or in Class 2 (see Table 3.9), any differences found in the achievement assessments at the post-tests in Chapter 4 & 5 should not be attributed to the test effect.

### ***3.6 The limitations of the achievement assessments:***

Some limitations of the achievement assessments used in this study are acknowledged in this section. It is worth noting that the acknowledgement of these limitations does not signify that the assessments are ineffective. However, they will be taken into consideration when interpreting the results of this study.

### ***3.6.1 The elicitation tests for measuring implicit and explicit knowledge***

Although it has been mentioned elsewhere in this thesis, it is worth repeating here regarding the measurements of implicit and explicit knowledge. Developing *pure* measures to assess both implicit and explicit knowledge is difficult (R. Ellis, 2005; DeKeyser, 2003). DeKeyser (2003, p.320) pointed out that even setting a given time constraint in a specific test does not guarantee a pure measure of implicit knowledge, though a speeded up test is more effortful for the retrieval of explicit than implicit knowledge. On the other hand, even though a test is carried out without time constraint, it is likely to elicit participants' implicit knowledge. R. Ellis (2005) stated that the implicit and explicit measures were expected to "predispose learners to access one or the other type of knowledge only probabilistically" (p. 153). Therefore, it is important to stress here that the elicitation of implicit and explicit knowledge is probabilistic. The adoption of measures to delve into explicit and implicit knowledge of language for the current study does not mean that these measures were used without reservations. The reason for adopting them is that they are recommended and considered as being *relatively* separate measures of implicit and explicit knowledge.

In addition, the implicit measures used in this study were implemented with time constraint. The potential problem derives from placing participants under time pressure to complete the task, in that it was likely to provoke participants' anxiety and unfavourably affect their performance on the assessments (Purpura, 2004). However the

participants' anxiety about the speeded up tests was not examined. In this study, the issue concerning whether the implicit measures gave rise to anxiety on the part of participants and then affected their performance was not clear.

### ***3.6.2 Reservations concerning the post-task retrospective self-report***

The premise for using verbal reports in the structured interview and a written post-task questionnaire after the expected implicit measures is that explicit knowledge is potentially verbalisable. However, some SLA researchers have cautioned against exploring explicit knowledge by means of having learners verbalise rules due to learners' forgetfulness and being unable to verbalise them (Bialystock, 1979; R. Ellis, 2004, 2005; Hu, 2002;) (see Section 2.3.2.3.). Despite these risks, Hu (2002) stated that "in the SLA literature, there is general support for verbal reporting as a test of explicit knowledge" (p.360). The threats to the validity of adopting the retrospective self-report are acknowledged in order to make it clear that the researcher/I do not use this approach without reservation.

### ***3.6.3 The validity of the achievement assessments***

As reported earlier, the validity of the oral test and the vocabulary test were not examined. In addition, the validity results for achievement assessments reported in 3.5.1 primarily concern the internal validity (i.e. whether the assessments can be used to detect the impact of the interventions). Nevertheless, the construct validity<sup>48</sup> of an assessment regarding the measure of implicit and explicit knowledge was not examined. The operationalisation of the construct validity for measuring implicit and explicit knowledge in this study followed R. Ellis' claims (2005) that the timed GJT and oral

---

<sup>48</sup> According to Trochim (2001), construct validity concerns "how well your actual programs or measures reflect your ideas or theories" (p.69).

tests can be appropriate for measuring implicit knowledge, whereas a test without time constraint tends to elicit explicit knowledge. It is acknowledged that failing to demonstrate the construct validity (i.e. explicit and implicit knowledge) of measurement would result in a risk to the interpretation of the results.



## Chapter 4 The results of the achievement assessments

### Introduction

This chapter consists of two sections. In order to examine the impact of each specific intervention, the first section focuses on presenting the performance of the participants in each individual achievement assessment by comparing their mean scores. The second section aims to explore the relationship between the achievement assessments in order to investigate what type of knowledge was derived from PI activities.

In the first section, the results of the individual assessments are discussed in the following order: the timed GJT, the gap-fill test, the picture-based narration test, the structured conversation, and, finally, the vocabulary test. The layout of the presentation of the results for each achievement test is as follows (except for the presentation of the structured conversation, given that its results were obtained by the calculation of the mean percentage of suppliance in obligatory contexts instead of a comparison of the mean scores between the groups):

- 1) justifying the application of non-parametric or parametric tests for further analysis based on the results of the K-S test and Levene's test;
- 2) presenting the results of the pre-test to ensure that each group is homogenous in relation to the participants' knowledge of the target feature at the outset;
- 3) reporting the results obtained from the non-parametric or parametric test on the achievement assessment. Note that the descriptive statistics (i.e. mean scores and standard deviation) are also reported, even though a parametric test is not tenable, given that the computation of effect size requires the mean scores and standard deviations;

4) reporting the effect size (Cohen's  $d$ ) to examine the magnitude of interventions by comparing the instructional groups with the control group, and by the scores at the pre-test with those at the post-tests (Norris & Ortega, 2000). In addition, a meta-analysis of effect size on previous PI studies using the similar achievement test is also reported.

#### **4.1 Comparison of the mean scores to explore the effectiveness of the intervention**

##### ***4.1.1 Analysis of the timed GJT***

###### *4.1.1.1 Can parametric tests be used for further analysis of the timed GJT?*

The K-S test showed that most of the scores of the groups in the timed GJT at the pre-test, post-test and delayed post-test were not normally distributed in that they were statistically significantly deviant from a normal distribution, except for the RA group at the pre-test,  $D(31)=.101$ ,  $p=.200 >.05$ ; the R group at the pre-test,  $D(29)=.131$ ,  $p=.200 >.05$ ; the A group at the post-test,  $D(30)=.126$ ,  $p=.200 >.05$ ; the control group at the post-test,  $D(30)=.147$ ,  $p=.099 >.05$  (see Appendix 31 for a tabular summary). In terms of Levene's test, only the scores at the pre-test were non-significant ( $p=.994 >.05$ ). The scores at both the post-test and the delayed post-test were significant ( $p=.000 <.01$  in both post-tests), suggesting that the variances for the four groups (the RA, R, A, and the control groups) at these post-tests were not the same (see Appendix 32). As the normality and homogeneity of variance assumptions were severely violated in the majority of the data, non-parametric tests were carried out for further statistical analyses on the timed GJTs. In addition, due to the fact that there is no non-parametric alternative for the mixed between-within subjects ANOVA in the SPSS, the Friedman's test, an alternative to a one-way repeated-measures ANOVA, was performed to investigate the impact of the interventions. In order to maintain parity with other studies, the results of applying mixed design ANOVA to the timed GJT are provided in Appendix 33, using

conditions as a between-group variable and time of test as a within-group variable, followed by planned contrasts once the ANOVA detected differences. Note that the results of the parametric test did not differ from those obtained by the non-parametric tests.

#### *4.1.1.2 The pre-test scores of the timed GJT*

The Kruskal-Wallis test (i.e. the non-parametric counterpart of one-way ANOVA) was carried out to analyse the raw scores of the timed GJT between the four groups at the pre-test. The results of the statistical analysis indicated that there was no significant difference in the timed GJT between the four groups prior to the interventions,  $H(3) = .376, p = .945 > .05$ . Consequently, the participants in the four groups, statistically speaking, had about the same knowledge of the target feature (i.e. the English regular past tense) in terms of the timed GJT before the intervention. Any differences observed between the instructional groups at the post-test and the delayed post-test should not therefore be attributed to any imparity between the groups at the beginning stage.

#### *4.1.1.3 The results of the timed GJT*

Table 4.1 and Figure 4.1 show the timed GJT mean scores of the three instructional groups and the control group at the pre-test, the post-test, and the delayed post-test. From Table 4.1 and Figure 4.1, it can be observed that the mean scores of both the RA and the R groups increased from the pre-test to the post-tests, but the same result was not found in the A group and the control group. Furthermore, the RA and R groups appeared to maintain their learning gains in the delayed post-test.

Table 4.1

*Descriptive statistics for the GJTs*

GROUP	N	The pre-test		The post-test		The dp test	
		Mean	SD	Mean	SD	Mean	SD
RA	31	11.16	4.52	21.90	10.82	23.32	10.42
R	29	11.38	4.79	21.17	11.73	20.10	11.13
A	30	11.87	4.82	13.70	6.59	14.43	6.33
C	30	12.10	5.02	12.57	4.84	12.77	3.63

\*The full score in the timed GJT was 40

\* RA=group of referential + affective activities; R=group of referential activities only;

A=group of affective activities only; C= control group

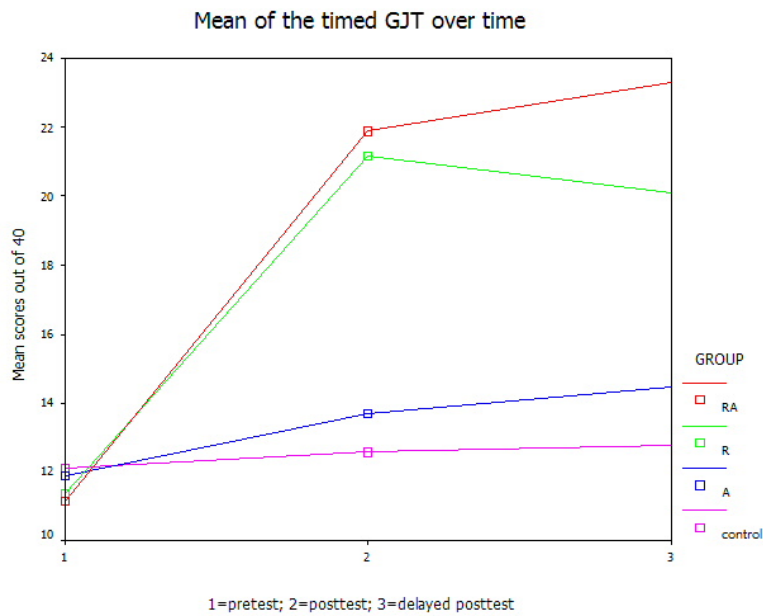


Figure 4.1 Scores on the timed GJTs over time

Table 4.2 summarises the results of the Friedman test for each group in the timed GJTs.

The results indicate that a statistically-significant difference was found in the timed GJTs in the RA group ( $\chi^2(2) = 31.776, p=.000 < .01$ ) and in the R group ( $\chi^2(2) = 17.568, p=.000 < .01$ ), suggesting that the test scores from both groups obtained from the three testing phases were, statistically speaking, different. On the other hand, no difference was found in the A group ( $\chi^2(2) = 3.431, p=.180 > .05$ ) and the control

group ( $\chi^2(2) = 0.299, p=.861 >.5$ ), suggesting that participants' performance on the timed GJT did not change over time.

Table 4.2

*The results of the Friedman test in the timed GJTs*

<b>Group</b>	<b>N</b>	<b>Chi-square (<math>\chi^2</math>)</b>	<b>Degree of Freedom (DF)</b>	<b>Sig. (2-tailed)</b>
<b>RA</b>	31	31.776	2	.000*
<b>R</b>	29	17.568	2	.000*
<b>A</b>	30	3.431	2	.180
<b>Control</b>	30	.299	2	.861

However, the Friedman test did not indicate where the differences were in terms of the repeated dependent variable (i.e. between the pre-test and the post-test, the pre-test and the delayed post-test, and/or the post-test and the delayed post-test). A *post-hoc* test for Friedman's test, the Wilcoxon signed-rank test (using a Bonferonni adjusted alpha value to control for Type I error<sup>49</sup>) (see Field, 2005, p.563; Pallant, 2007, p. 230-231), was carried out to identify where the differences were in both the RA and the R groups. Due to the fact that three comparisons were conducted (i.e. the pre-test vs. the post-test; the pre-test vs. the delayed post-test; the post-test vs. the delayed post-test), the revised alpha value for determining statistical significance was .0176 (i.e.,  $.05/3 = .0167$ ).

The results of the *post-hoc* test (i.e. the Wilcoxon signed-rank test) for the Friedman test are summarised in Table 4.3. The results reveal that in the RA group the differences observed were between the pre-test and the post-test ( $z=-4.446, p=.000<.0167$ ), and between the pre-test and the delayed post-test ( $z=-4.497, p=.000<.0167$ ). No difference

---

<sup>49</sup> Type I error refers to wrongly accepting that there is a genuine effect between the groups observed, when in fact there is not.

was found between the post-test and the delayed post-test ( $z=-1.650$ ,  $p=.099>.0167$ ). For the R group, it was found that the differences lay between the pre-test and the post-test ( $z=-3.783$ ,  $p=.000 <.0167$ ), and between the pre-test and the delayed post-test ( $z=-3.571$ ,  $p=.000 <.0167$ ), instead of the post-test and the delayed post-test ( $z=-1.797$ ,  $p=.072 >.0167$ ). Furthermore, a Wilcoxon rank-sum was carried out to examine the learning gains (i.e. to subtract each participant's pre-test score from that achieved at a post-test) in the RA and the R groups at the post-tests. The results reveal that no significant differences in learning gains were observed between the RA and the R groups at the post-test ( $W_s=859.000$ ,  $p=.706 >.05$ ), and at the delayed post-test ( $W_s=777.000$ ,  $p=.112 >.05$ ).

Table 4.3

*The results of the post-hoc test for the Friedman test in the timed GJTs*

<b>Group</b>	<b>Contrasting</b>	<b>z</b>	<b>Sig. (2-tailed)</b>
<b>RA</b>	pre- vs. post- tests	-4.446	.000*
	pre- vs. dp- tests	-4.497	.000*
	post- vs. dp- tests	-1.650	.099
<b>R</b>	pre- vs. post- tests	-3.783	.000*
	pre- vs. dp- tests	-3.571	.000*
	post- vs. dp- tests	-1.797	.072

According to the examination of the mean ranks and median values (see Table 4.4), the differences observed in Table 4.3 were due to the fact that the scores of both the RA and the R groups at the post-tests were significantly higher than those at the pre-test. In the RA group, the mean rank showed an increase from the pre-test (MR=1.21) to the post-test (MR=2.29), and from the pre-test (MR=1.21) to the delayed post-test (MR=2.50). With respect to the R group, the mean rank increased from the pre-test (MR=1.40) to the post-test (MR=2.43), and from the pre-test (MR=1.40) to the delayed post-test

(MR=2.17). These results suggest that the interventions of the RA and R groups significantly improved learners' performance in undertaking the timed GJT, and that the learning gains were sustained six weeks after the intervention had been completed. On the other hand, the fact that the performance of learners in both the A and the control groups did not significantly change over time reveals that the instruction of the A group did not significantly assist learners in undertaking the timed GJT.

Table 4.4

*Descriptive statistics for non-parametric tests in the timed GJTs*

Group	N	The pretest		The post-test		The dp test	
		mean rank (MR)	Median (Md)	mean rank (MR)	Median (Md)	mean rank (MR)	Median (Md)
RA	31	1.21	11.00	2.29	16.00	2.50	22.00
R	29	1.40	10.00	2.43	17.00	2.17	15.00
A	30	1.78	11.00	1.97	12.50	2.25	14.00
C	30	1.93	11.50	2.07	11.50	2.00	13.50

#### 4.1.1.4 The effect size of the timed GJTs

The effect size reported here was computed by means of Cohen's  $d$  (see Section 3.4.5 for the formulation). Note that the effect size was calculated by the contrasts between the pre-test and the post-tests, and the contrasts between the experimental groups and the control group. Table 4.5 summarises the magnitudes of effect for the effectiveness of the intervention by using the control group as a comparison group. On average, both the RA group and the R group were observed to have large effect sizes at the post-tests as Cohen's  $d$  were all larger than .8. The A group was found to have the smallest effect sizes at the timed GJT post-tests, and the effect observed for the intervention of the A group would be considered a small effect ( $.2 < d < .5$ ).

Table 4.6 displays the magnitudes of change from the pre-test to the post-tests in terms

of the timed GJT. Once again, a large effect size of change from the pre-test to the post-tests was observed for the intervention of both the RA and the R groups ( $d > .8$ ), and only a small effect size was observed for the A group intervention ( $.2 < d < .5$ ). Table 4.6 shows that the effectiveness of the intervention for both the RA group and the R group was substantial at the post-tests. Although the magnitude of the R group interventional effect dropped at the delayed post-test when compared with the change observed in the post-test, its effect size was still considered a large effect. According to Table 4.6, small effect sizes were found in the A group, and the effect sizes of the control group were negligible. It is noted that the effect sizes observed in the RA and R groups in Tables 4.5 and 4.6 were all greater than the mean effect size of the meta-linguistic judgment (mean  $d = .82$ ) reported by Norris & Ortega<sup>50</sup> (2000, p.471).

Table 4.5

*The magnitudes of instructional effect on the timed GJT*

<b>Groups</b>	<b><i>D</i> at the post-test</b>	<b><i>d</i> at delayed post-test</b>
<b>RA vs. C</b>	1.18	1.49
<b>R vs. C</b>	1.05	1.00
<b>A vs. C</b>	0.20	0.33

Table 4.6

*The magnitudes of change from the pre-test to the post-tests on the timed GJT*

<b>Group</b>	<b>Pre to post-test</b>	<b>Pre to delayed post-test</b>
<b>RA</b>	1.40	1.63
<b>R</b>	1.19	1.10
<b>A</b>	0.32	0.46
<b>C</b>	0.10	0.15

In addition, a meta-analysis was carried out in order to investigate the relative instructional magnitude of previous PI studies to this study. Apart from the current

<sup>50</sup> Norris & Ortega (2000) investigated the experimental and quasi-experimental studies' effectiveness of L2 instruction by comparing the effect sizes. They coded metalinguistic judgments as those that require participants to "evaluate the appropriacy or grammaticality of L2 target structures as used in item prompts (e.g., grammaticality judgment tasks)" (p.440).



study, only one study, so far, has been conducted and used the GJT to examine the instructional impact (Toth, 2006). A meta-analysis of Toth's study regarding its effect sizes is given in Table 4.7.

Table 4.7

*The meta-analysis of previous PI studies on the GJT*

<b>Toth (2006)</b>	<b>PI vs. C at pt</b>	<b>PI vs. C at dp</b>	<b>pre- to pt</b>	<b>pre- to dp</b>
<b>effect size</b>	0.88	0.62	1.39	1.13

\*C= control group; pre= pre-test; pt= post-test; dp= delayed post-test

The results reveal that a large effect size was observed on the GJT of Toth's PI group except for that at the delayed post-test when PI is compared with the control group ( $d=.62$ , a medium effect size). In terms of the relative instructional magnitude, it appeared, as a whole, that the instructional magnitude of the RA and the R effect was larger than that of the PI group in Toth's study. However, the magnitude of the effect of the A group in the current study was smaller than that of the PI group in Toth's study.

#### ***4.1.2 Analysis of the gap-fill test***

##### ***4.1.2.1 Can parametric tests be used for further analysis of the gap-fill test?***

The results of the K-S test for the gap-fill test were consistent (see Appendix 31). The results indicate that all the scores of the groups at the gap-fill pre-test and post-tests were abnormally distributed due to the fact that they were statistically significantly different from a normal distribution. Levene's test showed that none of the scores in the gap-fill test at different testing times was non-significant, suggesting that the variances for the four groups were not the same (see Appendix 32). As the assumptions of normality and homogeneity of variance for running parametric tests were severely violated in all the data from the gap-fill tests, parametric tests were not viable for the

analysis of the gap-fill tests. In this circumstance, a non-parametric test, the Friedman test, was used to assess the impact of the interventions. In order to be consistent with other studies, the results of applying mixed design ANOVA to the gap-fill test are provided in Appendix 34, with conditions as a between-group variable and time of test as a within-group variable, followed by planned contrasts. Again, the results of the parametric test did not differ from those obtained by the non-parametric tests.

#### 4.1.2.2 The pre-test scores of the gap-fill test

The Kruskal-Wallis test was performed to analyse the raw scores of the gap-fill test between the four groups at the pre-test. The results reveal that no significant differences were found in the gap-fill test between the four groups prior to the interventions,  $H(3) = 2.871, p = .412 > .05$ . As a result, any differences between the four groups at the gap-fill post-tests should not be ascribed to any imparity at baseline.

#### 4.1.2.3 The results of the gap-fill test

Table 4.8 and Figure 4.2 display the mean scores for the gap-fill test of the three instructional groups and the control group at the pre-test and the two post-tests.

Table 4.8

#### *Descriptive statistics for the gap-fill test*

GROUP	N	The pre-test		The post-test		The dp test	
		Mean	SD	Mean	SD	Mean	SD
RA	31	0.03	.18	2.32	3.19	2.65	3.45
R	29	.00	.00	1.86	2.81	2.10	3.19
A	30	.00	.00	0.07	.37	.30	1.21
Control	30	.00	.00	.00	.00	.00	.00

\* The total possible score in the gap-fill test was 8

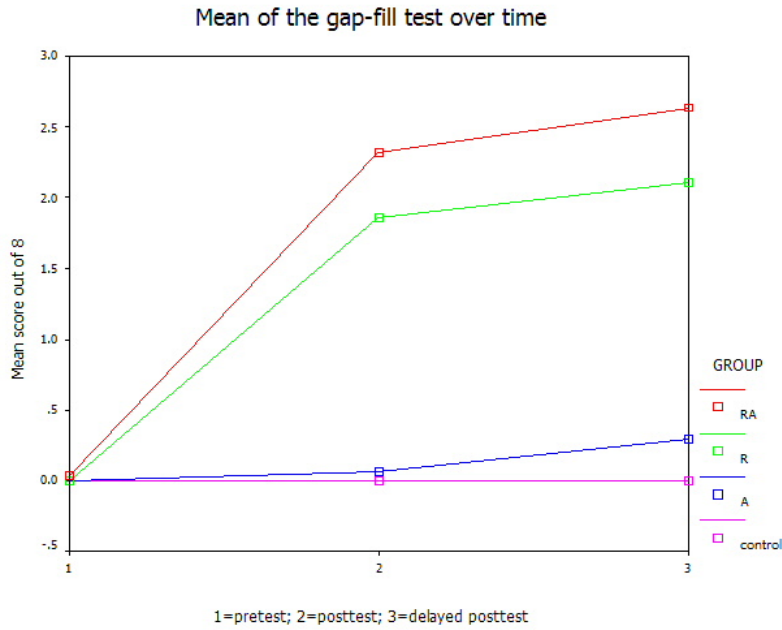


Figure 4.2 Scores on the gap-fill tests over time

Table 4.8 shows that the mean scores of both the RA and the R groups increased from the pre-test to the post-test, and from the pre-test to the delayed post-test. However, the mean scores of the A group and the control group were maintained consistently across the three different testing phases. Figure 4.2 shows that the patterns of the performances of the RA and the R groups in the gap-fill test were almost parallel, given that the test scores of both the RA and the R groups increased from the pre-test to the post-test. Furthermore, the increased test scores observed at the post-test did not drop at the delayed post-test. On the other hand, Figure 4.2 shows that the A and the control groups both remained the same from the pre-test to the delayed post-test test. Even though the participants in the A group had received some treatment, their mean scores were not different from those of the control group, in which the participants received no treatment.

Table 4.9 summarises the results of the Friedman test on the gap-fill tests. The results

indicate that significant differences in the test scores taken at three different timings were found in the RA group ( $\chi^2(2) = 18.00, p=.00 < .01$ ) and the R group ( $\chi^2(2) = 15.95, p=.00 < .01$ ), which suggests that the performances of the participants in both groups on the gap-fill test had changed over time. In contrast, no statistically significant difference was observed in the A group ( $\chi^2(2) = 2.00, p=.368 > .05$ ) and the control group ( $\chi^2(2) = .00, p=1.00 > .05$ ), suggesting that there was no significant change in the three different testing phases in either the A or the control groups.

Table 4.9

*The results of the Friedman test on the gap-fill tests*

<b>Group</b>	<b>N</b>	<b>Chi-square (<math>\chi^2</math>)</b>	<b>Degree of Freedom (DF)</b>	<b>Sig. (2-tailed)</b>
<b>RA</b>	31	18.000	2	.000*
<b>R</b>	29	15.953	2	.000*
<b>A</b>	30	2.000	2	.368
<b>Control</b>	30	.000	2	1.000

As statistically-significant differences were identified in both the RA and the R groups, a *post-hoc* test, using the Wilcoxon signed-rank test, was administered to detect where the differences were in both the RA and the R groups. In order to control for Type I error, note that the adjusted alpha value was .0167 instead of .05 due to three contrasts being compared. Table 4.10 shows the results of the *post-hoc* test on the gap-fill tests. The results indicate that in the RA group differences were found between the pre-test and the post-test ( $z = -3.071, p = .002 < .0167$ ), and between the pre-test and the delayed post-test ( $z = -3.088, p = .002 < .0167$ ). No difference was found between the post-test and the delayed post-test in the RA group ( $z = -.671, p = .502 > .0167$ ). Regarding the R group, the *post-hoc* test results reveal that the significant differences observed were between

the pre-test and the post-test ( $z = -2.944, p = .003 < .0167$ ), and between the pre-test and the delayed post-test ( $z = -2.816, p = .005 < .0167$ ), and not between the post-test and the delayed post-test ( $z = -1.512, p = .131 > .0167$ ). Furthermore, a Wilcoxon rank-sum was performed to examine whether any differences existed in learning gains between the RA and the R groups. The results showed that *no* significant differences were found between the RA and the R group in learning gains at the post-test ( $W_s = 865.00, p = .741 > .05$ ), and at the delayed post-test ( $W_s = 857.00, p = .843 > .05$ ).

Table 4.10

*The results of the post-hoc test for the Friedman test on the gap-fill tests*

<b>Group</b>	<b>Contrasting</b>	<b>Z</b>	<b>Sig. (2-tailed)</b>
<b>RA</b>	pre- vs. post- tests	-3.071	.002*
	pre- vs. dp- tests	-3.088	.002*
	post- vs. dp- tests	-.671	.502
<b>R</b>	pre- vs. post- tests	-2.944	.003*
	pre- vs. dp- tests	-2.816	.005*
	post- vs. dp- tests	-1.512	.131

According to the examination of the mean ranks and median values (see Table 4.11), it was found that the differences detected in Table 4.10 were due to the fact that the participants in both the RA and the R groups performed significantly better at the post-tests than at the pre-test. In the RA group, the mean rank increased from the pre-test (MR=1.61) to the post-test (MR=2.15), and from the pre-test (MR=1.61) to the delayed post-test (MR=2.24). In the R group, the mean rank rose from the pre-test (MR = 1.64) to the post-test (MR=2.12), and from the pre-test (MR=1.64) to the delayed post-test (MR=2.24). These results suggest that the interventions of the RA and R groups significantly improved learners' performance on taking the gap-fill test, and the learning gains appeared to be sustained six weeks after the intervention had finished.

It should be noted that although the statistical results reported above suggest that the intervention given to the RA and the R groups helps them undertake the gap-fill test, the result did not amount in most cases to substantial learning gains. Table 4.8 shows that both interventions merely led to learners completing on average two correct insertions.

Table 4.11

*Descriptive statistics for non-parametric tests on the gap-fill tests*

Group	N	The pre-test		The post-test		The dp test	
		mean rank (MR)	Median (Md)	Mean rank (MR)	median (Md)	mean rank (MR)	median (Md)
RA	31	1.61	.00	2.15	.00	2.24	.00
R	29	1.64	.00	2.12	.00	2.24	.00
A	30	1.95	.00	2.00	.00	2.05	.00
C	30	2.00	.00	2.00	.00	2.00	.00

#### 4.1.2.4 The effect size of the gap-fill test

Table 4.12 summarises the magnitudes of effect for the intervention in comparison with the control group. Overall, large effect sizes ( $d > .8$ ) were found in the RA and the R groups, and the intervention of the RA group had slightly larger effect sizes than those of the R group at the post-test and the delayed post-test. In addition, the intervention of both the RA group and the R group had substantially larger effect sizes than those of the A group.

Table 4.12

*The magnitudes of instructional effect on the gap-fill test*

Group	<i>D</i> at the post-test	<i>d</i> at the delayed post-test
RA vs C	1.43	1.51
R vs C	1.35	1.34
A vs C	0.36	0.50

Table 4.13 shows the magnitudes of change on the gap-fill test from the pre-test to the

post-tests. Once again, it can be observed that both the RA and the R interventions were associated with greater change than the A intervention from the pre-test to the post-tests. Both the RA group's and the R group's effect sizes of the change on the gap-fill test were considered to be a large effect ( $d > .8$ ), and the effect of the change in the A group was regarded as a small effect ( $.2 < d < .5$ ). In addition, the strength of the effect sizes in the RA and the R groups reported in Tables 4.12 and 4.13 was larger than the mean effect size of the constrained constructed response measures (mean  $d=1.20$ ) reported by Norris & Ortega<sup>51</sup> (2000, p.471). It should be noted that the computation of the control group was not performed, given that none of the participants scored on the gap-fill test at the post-tests.

Table 4.13

*The magnitudes of change from the pre-test to the post-tests on the gap-fill test*

<b>Group</b>	<b>Pre to post-test</b>	<b>Pre to delayed post-test</b>
<b>RA</b>	1.36	1.44
<b>R</b>	1.32	1.32
<b>A</b>	0.36	0.50
<b>C</b>	----	----

The results of the meta-analysis on effect size of previous PI studies are reported in Table 4.14. It is noted that the nature of the assessments (i.e. the sentence completion and blank-fill in text) used by the pooled PI studies were compatible with the gap-fill test used in the current study. The effect sizes were calculated by contrasting the pre-test with the post-tests, because most pooled PI studies in Table 4.14 did not include a control group, except for Wong's (2004b). To the best of the author's knowledge, so far only five studies have been conducted to compare the effectiveness of PI's components

---

<sup>51</sup> Norris & Ortega (2000) coded the constrained constructed response measures as those that require participants to "produce the target form(s) under highly regulated circumstances, where the use of the appropriate form was essential for grammatical accuracy to occur ... ranging in length from a single word up to a full sentence" (p.440).

(Benati, 2004a, 2004b; Farley, 2004b; VanPatten & Oikkenon, 1996; Wong, 2004b). Note that these five studies mainly set out to compare component one (i.e. explicit grammar explanation) with component two *plus* component three (i.e. the structured input activities), and that none of them intended to separate the structured input activities. As the PI activities delivered to the participants in this study were under the condition that component one (explicit grammar explanation) was removed, the effect sizes of pooled prior PI studies reported here were computed based on those of SIA-only groups (similar to the RA group operationalised in the current study), instead of those receiving the full 'PI package'.

Also, it is noted that Sanz & Morgan-Short's study (2004) did not aim to compare the effectiveness of the PI components. They aimed to investigate the effectiveness of explicit information given *before* and *during* the intervention, but their intervention was delivered *via* PI activities. The (-E, -F) group reported in their study was rather parallel to the RA group of the current study. In their (-E, -F) group, no explicit grammar explanation was given before or during the intervention. During the instructional phases, the indication of the correctness of the participants' responses to referential activities was along the lines of 'OK' or 'Sorry, try again!' (p. 56). In this regard, Sanz & Morgan-Short's study was included in the meta-analysis.

The meta-analysis results on the effect sizes found in previous PI studies show that the SIA-only groups, on average, had more instructional impact on the written production test at the post-test than that of the interventions of this study, because the mean  $d$  of the SIA-only groups ( $d=1.78$ ) is greater than those of the current study ( $d=1.36$  in the RA group;  $d=1.32$  in the R group; and  $d=.36$  in the A group). Larger effect sizes were also observed at the delayed post-test in prior PI studies. However, the effect size of both the



RA and the R groups at the post-test was greater than that of four of the six studies reported in Table 4.14, the exceptions being Benati's two studies (2004a & 2004b).

Table 4.14

*The meta-analysis of previous PI studies on the written production test*

<b>PI studies (SIA only)</b>	<b>Assessment</b>	<b><i>d</i> at pre- to post-test</b>	<b><i>d</i> at pre- to dp</b>
<b>VanPatten &amp; Oikkenon (1996)</b>	Sentence completion	0.65	x
<b>Benati (2004 a)</b>	Blank-fill in text	3.88	3.19
<b>Benati (2004 b)</b>	Gap-fill test	3.00	x
<b>Farley (2004 b)</b>	Sentence completion	1.28	1.23
<b>Wong (2004 b)</b>	Sentence completion	1.07	x
<b>Sanz &amp; Morgan-Short (2004)<sup>52</sup></b>	Sentence completion	0.77	x
<b>Mean <i>d</i></b>		<b>1.775</b>	<b>2.21</b>

\* x=N/A

### **4.1.3 Analysis of the picture-based narration test**

#### **4.1.3.1 Can parametric tests be used for further analysis on the picture narration test?**

The results of the K-S test for the picture-based narration test show that the majority of the data did not meet the normality assumption for performing parametric tests, except for the RA group at the post-test,  $D(10) = 0.168$ ,  $p = .200 > .05$ , the R group at the pre-test,  $D(9) = .272$ ,  $p = .054$ , and the control group at the post-test,  $D(9) = .272$ ,  $p = .054 > .05$  (see Appendix 31). Levene's test for homogeneity of variance assumption was not satisfied at except for the pre-test ( $p = .133 > .05$ ) (see Appendix 32). Thus, parametric tests were not used for statistical analyses on the picture-based narration test. As a result, Friedman's ANOVA was conducted to investigate the impact of the intervention

<sup>52</sup> Sanz & Morgan-Short used two types of the written production test, namely a sentence-completion test and a video-retelling test. The effect size was computed *via* the sentence-completion test as it is more like the gap-fill test used in the current study.

on the picture-based narration test. The results of applying mixed design ANOVA to the picture-based narration test are provided in Appendix 35 for comparison with other studies. However, the results of the parametric test did not entirely back up those produced by the non-parametric tests. Nevertheless, as the assumptions for conducting parametric tests were not upheld, the use of the non-parametric tests was more trustworthy than parametric tests.

#### 4.1.3.2 *The pre-test scores of the picture-based narration test*

The Kruskal-Wallis test was performed to analyse the scores of the picture-based narration test between the four groups at the outset. The results indicate that no significant differences in the picture-based narration test were observed between the four groups prior to the interventions,  $H(3) = 9.299, p = .098 > .05$ . Therefore, any differences between the four groups at the post-test and the delayed post-test should not be attributed to any initial imparity between the groups.

#### 4.1.3.3 *The results of the picture-based narration test*

Table 4.15 and Figure 4.3 show the mean scores for the picture-based narration test of the four groups at the pre-test, the post-test and the delayed post-test.

Table 4.15

*Descriptive statistics for the picture-based narration test*

Group	N	The pre-test		The post-test		The dp test	
		Mean	SD	Mean	SD	Mean	SD
<b>RA</b>	10	.80	1.03	1.60	1.43	2.10	2.33
<b>R</b>	9	.67	.71	1.56	2.40	.11	.33
<b>A</b>	9	.33	.50	.22	.44	.11	.33
<b>Control</b>	9	.33	.50	.67	.71	.78	1.72

*\*The total possible score in the picture-based narration test was 8*

Table 4.15 and Figure 4.3 shows that the mean scores of both the RA and the R groups and the control group appear to increase from the pre-test to the post-test. The mean score of the A group decreased by .11. However, it is noted that the learning gains of each group were marginal, with the RA group increasing by .80, the R group by .89, and the control group by .34 from the pre-test to the post-test. In addition, only the RA group among the instructional groups maintained its learning gains from the post-test to the delayed post-test.

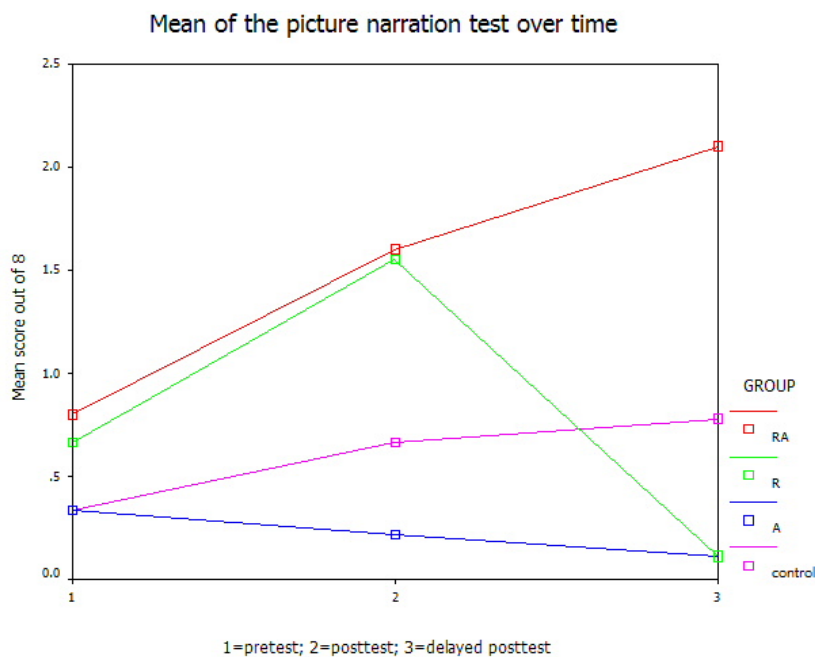


Figure 4.3 Scores on the picture-based narration tests over time

The results of the Friedman test regarding the picture-based narration test are summarised in Table 4.16. The mean ranks and median values of the picture-based narration test are presented in Table 4.17. The results of the Friedman test reveal that there was no statistically-significant difference found in any of the four groups (RA:  $\chi^2(2) = 3.000, p = .223 > .05$ ; R:  $\chi^2(2) = 4.455, p = .108 > .05$ ; A:  $\chi^2(2) = 1.500, p = .472 > .05$ ; control:  $\chi^2(2) = 1.652, p = .438 > .05$ ). As no significant differences were observed in any of the groups, *post hoc* tests, therefore, were not carried out for further

statistical analyses. Consequently, the results of the Friedman test suggest that the interventions did not induce significant improvement to the oral production of the English past tense ‘-ed’ feature in the picture-based narration test.

Table 4.16

*The results of the Friedman test on the picture-based narration test*

<b>Group</b>	<b>N</b>	<b>Chi-square (<math>\chi^2</math>)</b>	<b>DF</b>	<b>Sig. (2-tailed)</b>
<b>RA</b>	10	3.000	2	.223
<b>R</b>	9	4.455	2	.108
<b>A</b>	9	1.500	2	.472
<b>Control</b>	9	1.652	2	.438

Table 4.17

*Descriptive statistics for non-parametric tests on the picture-based narration test*

<b>Group</b>	<b>N</b>	<b>The pre-test</b>		<b>The post-test</b>		<b>The dp test</b>	
		<b>mean rank (MR)</b>	<b>Median (Md)</b>	<b>mean rank (MR)</b>	<b>median (Md)</b>	<b>mean rank (MR)</b>	<b>Median (Md)</b>
<b>RA</b>	10	1.75	.50	1.90	1.50	2.35	1.00
<b>R</b>	9	2.17	1.00	2.28	1.00	1.56	.00
<b>A</b>	9	2.17	.00	2.00	.00	1.83	.00
<b>C</b>	9	1.89	.00	2.28	1.00	1.83	.00

#### *4.1.3.4 The effect size of the picture-based narration test*

Although the examination of effect size on non-significant findings may seem odd and cause some debate, it is appropriate to report effect size, since effect sizes and inferential statistics yield two distinct kind of information (Ortega, 2009, private communication). Effect sizes tell us about an estimate of the magnitude of an observed effect; inferential statistics tell us about the probabilities, namely whether the observations are generalisable to new samples from the same population or whether they are likely to represent just luck/chance. Given that the sample size of the oral tests

in this study was rather small, ignoring the effect sizes might increase the chance of making a type two error, because statistical inference is sensitive to sample size. In the light of this consideration, effect sizes were calculated and are reported.

Table 4.18 summarises the magnitudes of instructional effect in comparison with the control group. According to Table 4.18, it is observed that the RA intervention had a large effect size ( $d=.85$ ) at the post-test and a medium effect size ( $d=.65$ ) at the delayed post-test. A medium effect size was also found for the intervention of the R group at the post-test ( $d=.57$ ).

Table 4.18

*The magnitudes of effect for the intervention on the picture narration test*

<b>Group</b>	<b><i>d</i> at the post-test</b>	<b><i>d</i> at the delayed post-test</b>
<b>RA vs. C</b>	0.85	0.65
<b>R vs. C</b>	0.57	-0.65
<b>A vs. C</b>	-0.78	-0.65

Table 4.19 shows the magnitudes of change from the pre-test to the post-tests on the picture-based narration test. Though no significant difference was found in any of the groups' performances on the picture narration test according to the Friedman test (see Table 4.16), medium effect sizes of change from the pre-test to the post-test were observed for the RA group ( $d=.65$ ), the R group ( $d=.57$ ), and the control group ( $d=0.56$ ). Also, medium effect size of change from the pre-test to the delayed post-test was observed for the RA intervention ( $d=.77$ ). In general, the medium effect sizes observed in both the RA and the R groups were all greater than the average effect size of the free constructed response measures<sup>53</sup> (mean  $d=.55$ ) reported by Norris & Ortega (2000, p.

<sup>53</sup> Norris & Ortega (2000) coded the free constructed response measures as those that "required participants to produce language with relatively few constraints and with meaningful communication as the goal for L2 production" (p.440).

471), although it is acknowledged that the picture-based narration test used in this study was possible more controlled than those analysed by Norris & Ortega.

Table 4.19

*The magnitudes of change from the pre-test to the post-tests on the picture narration test*

<b>Group</b>	<b>Pre to post-test</b>	<b>Pre to delayed post-test</b>
<b>RA</b>	0.65	0.77
<b>R</b>	0.57	-1.08
<b>A</b>	-0.23	-0.53
<b>C</b>	0.56	0.41

The results of a meta-analysis on the effect sizes of previous PI studies are reported in Table 4.20. To the best of the author's knowledge, so far only six studies have set out to explore the impact of PI on learners' oral performance. However, only four studies are reported in Table 4.20 due to insufficient information provided to calculate the effect size in Salaberry's (1997) and Benati's (2001) studies. Also, it is stressed here that three of the four PI studies presented in Table 4.20 received the full PI package (i.e. the explicit grammar explanation plus structured input activities) except for Benati's (2004b) study.

Table 4.20

*The meta-analysis of previous PI studies on the oral test*

<b>PI studies</b>		<b>Assessment</b>	<b><i>d</i> at pre- to post-test</b>	<b><i>d</i> at pre- to dp</b>
<b>Benati (2004b)</b>		Pictured narration	4.40	x
<b>VanPatten &amp; Sanz (1995)</b>		Video narration	.46	x
<b>Marsden (2006)</b>	<b>School 1</b>	Picture narration	.87	.98
	<b>School 2</b>	Picture narration	.26	x
<b>Erlam (2003)</b>		Picture narration	.44	.14
<b>Mean <i>d</i></b>			<b>1.29</b>	<b>.56</b>

\*  $x=N/A$

According to Table 4.20, it can be seen that the mean effect size of prior PI studies ( $d=1.29$ ) at the post-test was larger than all of the effect sizes, which were produced on the picture-based narration in the current study. However, the effect sizes of the RA( $d=.65$ ) and the R( $d=.57$ ) groups were greater than the effect sizes produced by VanPatten & Sanz's, Marsden's (school 2), and Erlam's studies. With respect to the delayed post-test, the effect size of the RA group ( $d=.77$ ) was greater than the average effect size found in previous PI studies (mean  $d=.56$ ), but the effect sizes produced by both the R and the A group were negligible.

#### ***4.1.4 Analysis of the structured conversation test***

Due to the fact that it was unpredictable how many target verb stems the participants would produce in this test, the mean of rate of suppliance in obligatory contexts in the structured conversation test in each group was calculated. Table 4.21 presents each group's performance in the structured conversation.

Table 4.21

*Mean rate of suppliance in obligatory contexts for structured conversation*

GROUP	N	Pre-test		Post-test		Delayed post-test	
		M %	proportions in group	M %	proportions in group	M %	proportions in group
RA	10	10.0	1/33	10.0	1/33	13.4	4/41
R	9	10.0	3/33	3.7	2/33	0.0	0/36
A	9	5.6	1/34	0.0	0/19	3.7	1/29
Control	9	1.9	1/55	5.3	2/47	0.0	0/49

The mean rate of the obligatory context was calculated using the following procedures. First, each student's ratio of producing the '-ed' tokens in each group was added up. For example, if in one group a student produced three verbs and one of these three verbs

correctly attached the ‘-ed’ token, this student’s ratio of producing the ‘-ed’ token would be 1/3. By adding up each student’s ratio, a total value K could be obtained. Then, K was divided by the total number of participants in each group; then the mean rate of appliance in obligatory contexts was obtained. In Table 4.21, the proportions (A/B) in each cell signifies:

1. A is the total number of the ‘-ed’ tokens the participants in a group produced;
2. B is the total obligatory contexts the participants in a group produced.

From Table 4.21, it can be seen that the participants’ use of regular past tense in this test was negligible. However, it is likely that the participants had some knowledge of the ‘-ed’ rule, but they encountered problems producing it during this test, given that a significant improvement was found in the gap-fill test. More detailed discussion about why learners did not do this test well can be found in Section 2.3.2.4 & 6.3.4. Although the RA group seems to improve slightly from the pre-test (10%) to the delayed post-test (13.4%), the result of either a non-parametric test (Wilcoxon signed-rank test) or a parametric test (dependent *t*-test) did not detect any significant differences in the RA group between the percentage of producing the ‘-ed’ feature at the pre-test and at the delayed post-test. The result of the Wilcoxon signed-rank test was  $z = -.552$ ,  $p = .581 > .05$ ; the result of the dependent *t*-test was  $t(9) = -.227$ ,  $p = .825 > .05$ .

#### ***4.1.5 Analysis of the vocabulary test***

##### *4.1.5.1 Can parametric tests be used for further analysis on the vocabulary test?*

The results of the K-S test for the vocabulary test (see Appendix 31) show that most test scores achieved by the groups at the vocabulary test at three different testing times violated the assumption of normality, except for the RA group at the delayed post-test,  $D(31) = .151$ ,  $p = .071 > .05$ . Nevertheless, Levene’s test did not detect any significant



differences between groups at different testing times (see Appendix 32). However, as the normality assumption of the parametric test was seriously violated in most of the data, there was no justification for performing parametric tests on the vocabulary test data. Instead, the Friedman test was carried out to investigate the impact of the interventions on vocabulary learning. The results of mixed design ANOVA to the vocabulary test are provided in Appendix 36. However, the results of the parametric test did not entirely agree with those observed in the non-parametric tests. Nevertheless, as the assumptions for conducting parametric tests were severely broken, the use of the non-parametric tests was more justifiable than using parametric tests.

#### 4.1.5.2 *The pre-test scores of the vocabulary test*

The results of the Kruskal-Wallis test showed no significant differences in the vocabulary test between the three groups,  $H(2) = .846, p = .655 > .05$ . Any differences found between the groups at the post-test and the delayed post-test should therefore not be ascribed to baseline imparity.

#### 4.1.5.3 *The results of the vocabulary test*

The mean scores for the vocabulary test of the three experimental groups at the pre-test and the post-tests are presented in Table 4.22 and in Figure 4.4.

Table 4.22

#### *Descriptive statistics for the vocabulary test*

Group	N	The pre-test		The post-test		The dp test	
		Mean	SD	Mean	SD	Mean	SD
RA	31	1.10	1.27	2.35	1.96	2.23	1.71
R	29	1.24	1.92	1.59	1.74	1.59	1.88
A	30	1.30	1.60	2.20	1.99	1.60	1.40

*\*The total possible score in the picture-based narration test is 10*

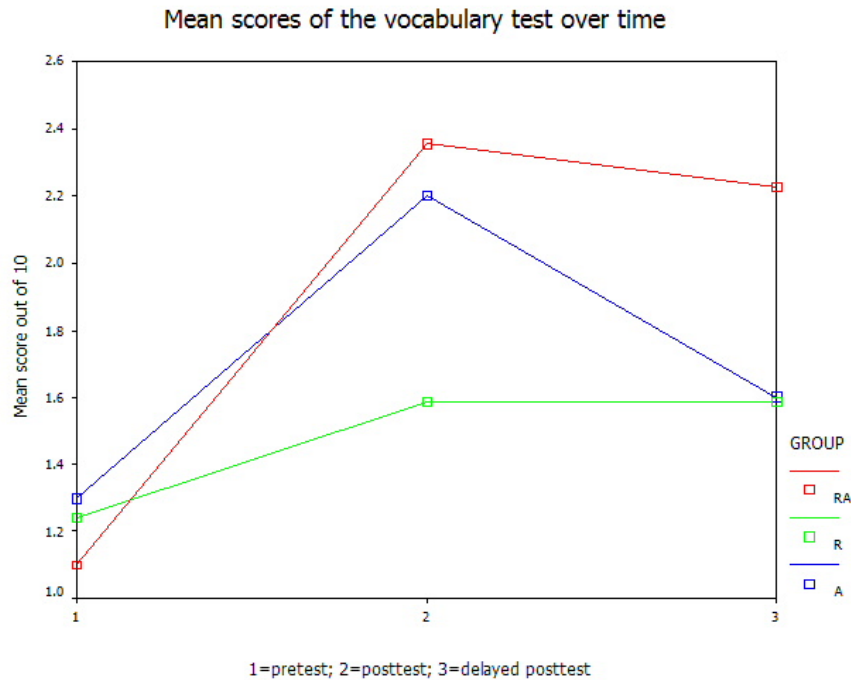


Figure 4.4 Scores on the vocabulary tests over time

Table 4.22 shows that the mean scores of the RA, R and A groups increased from the pre-test to the post-test by 1.25, 0.35 and 0.90 respectively. Based on the mean differences, it is clear that the R group exhibited less improvement in the vocabulary test than the other two instructional groups. Although all three groups made an improvement between the pre-test and the post-test, it can be seen that the mean differences (i.e. the learning gains) were quite small. Figure 4.4 shows that the performances of the RA group and the A group were similar from the pre-test to the post-test, as the mean score of both groups increased, but only the RA group maintained the learning gains in the delayed post-test, whereas the A group did not. The mean score of the A group dropped from the post-test to the delayed post-test by .60. The data displayed in Table 4.22 and in Figure 4.4 could suggest that the intervention of the R group was less conducive to learning vocabulary, compared with the other two interventions. The most promising intervention among the three was that of the RA group, because the mean score of RA group increased most between the pre-test and the

post-test. Furthermore, retention was observed six weeks after the intervention had been completed. The intervention of the A group also made a contribution to vocabulary learning after the accomplishment of the intervention, but retention was not sustained six weeks later.

Table 4.23

*The results of the Friedman test on the vocabulary test*

<b>Group</b>	<b>N</b>	<b>Chi-square (<math>\chi^2</math>)</b>	<b>Degree of Freedom</b>	<b>Sig. (2-tailed)</b>
<b>RA</b>	31	16.178	2	.000*
<b>R</b>	29	2.909	2	.234
<b>A</b>	30	12.881	2	.002*

According to Table 4.23, the results of the Friedman test demonstrate that there was a significant effect in the RA group ( $\chi^2(2) = 16.178$ ,  $p = .000 < .01$ ) and in the A group ( $\chi^2(2) = 12.881$ ,  $p = .002 < .05$ ) in the vocabulary tests. No significant difference was found in the R group ( $\chi^2(2) = 2.909$ ,  $p = .234 > .05$ ). These results suggest that the intervention of the R group did not lead to any improvement in learning vocabulary as the performance of the learners did not significantly improve over time. In contrast, the significant differences observed in both the RA and the A groups suggest that the performance of these two groups in the vocabulary test had changed over time.

The *post-hoc* test, the Wilcoxon signed-rank test, was used to determine where the difference observed in Table 4.23 was. Note that the revised alpha value for determining statistical significance was .0167 instead of .05. Table 4.24 presents the *post-hoc* results for the Friedman test. The results show that the difference found in the RA group was between the pre-test and the post-test ( $z = -3.259$ ,  $p = .001 < .0167$ ), and between the pre-

test and the delayed post-test ( $z=-3.537$ ,  $p=.000<.0167$ ). No difference was found between the post-test and the delayed post-test ( $z=-.274$ ,  $p=.784 >.0167$ ). With respect to the A group, a significant difference was observed between the pre-test and the post-test ( $z=-2.502$ ,  $p=.012 <.0167$ ). However, no significant difference was detected in the post-test or the delayed post-test ( $z=-1.733$ ,  $p=.083 >.0167$ ), nor observed in the pre-test and the delayed post-test ( $z=-2.183$ ,  $p=.029 >.0167$ ). Furthermore, Wilcoxon rank-sum was performed to examine the learning gains between the RA and the A groups. The results reveal that there were no significant differences between the RA and the A groups in vocabulary learning gains at the post-test ( $W_s=893.00$ ,  $p=.586 >.05$ ). However, a significant difference was observed between the RA and the A groups in vocabulary learning gains at the delayed post-test ( $W_s=749.00$ ,  $p=.006 <.05$ ).

Table 4.24

*The results of the post-hoc test for the Friedman test on the vocabulary test*

<b>Group</b>	<b>Contrasting</b>	<b>z</b>	<b>Sig. (2-tailed)</b>
<b>RA</b>	pre- vs. post- tests	-3.259	.001*
	pre- vs. dp- tests	-3.537	.000*
	post- vs. dp- tests	-.274	.784
<b>A</b>	pre- vs. post- tests	-2.502	.012*
	pre- vs. dp- tests	-2.183	.029
	post- vs. dp- tests	-1.733	.083

According to the examination of the mean ranks and median values on the vocabulary test (see Table 4.25), the differences observed in Table 4.24 were due to the increased scores from the pre-test to the post-test in both the RA and the A groups. In the RA group, the mean rank increased from the pre-test (MR=1.47) to the post-test (MR=2.27), and this increase was retained from the post-test (MR=2.27) to the delayed post-test (MR=2.26). In the A group, the mean rank increased from the pre-test (MR = 1.65) to

the post-test (MR=2.42), but it dropped from the post-test (MR=2.42) to the delayed post-test (MR=1.93). These results suggest that the interventions of the RA and A groups made a significant contribution to learning vocabulary, but the retention was only observed in the RA group six weeks after the completion of the intervention. Although the R group seemed to make some improvement from the pre-test to the post-tests according to Table 4.25, the fact that no significant difference was observed in the R group according to the Friedman test (see Table 4.23) suggests that the intervention of the R group did not have any significant instructional impact on learning vocabulary.

Table 4.25

*Descriptive statistics for non-parametric tests on the vocabulary test*

Group	N	The pre-test		The post-test		The dp test	
		mean rank (MR)	Median (Md)	mean rank (MR)	Median (Md)	mean rank (MR)	Median (Md)
RA	31	1.47	1.00	2.27	3.00	2.26	2.00
R	29	1.79	.00	2.07	1.00	2.14	1.00
A	30	1.65	1.00	2.42	1.50	1.93	2.00

#### *4.1.5.4 The effect size of the vocabulary test*

As no control group took the vocabulary test, no effect size with respect to the contrasts between the instructional and control groups was calculated. Table 4.26 shows the magnitudes of change from the pre-test to the post-tests in the vocabulary test. It was found that the strength of the effect size in the RA group was considered to be medium ( $.5 < d < .8$ ), and that of the A group was small ( $.2 < d < .5$ ). The strength of the effect size in the R group was negligible.

.Table 4.26

*The magnitudes of change from the pre-test to the post-tests on the vocabulary test*

<b>Group</b>	<b>Pre to post-test</b>	<b>Pre to delayed post-test</b>
<b>RA</b>	0.77	0.76
<b>R</b>	0.19	0.18
<b>A</b>	0.50	0.20

#### ***4.1.6 Summaries of the results from 4.1.1 to 4.1.5***

The results pattern obtained from the timed GJT and the gap-fill test was similar.

According to significance tests, significant differences were found in both the RA and the R groups. The *post-hoc* tests reveal that the difference is between the pre-test and the post-test, and between the pre-test and the delayed post-test. This is because participants' test score at the post-tests was significantly higher than that at the pre-test. Furthermore, no significant difference was found in learning gains between the performance of the RA and R groups at either the post-test or the delayed post-test. On the other hand, no significant difference was observed in both the A and control groups. Both the RA and the R groups produced large effect sizes between the pre-test and the post-tests and between the instructional group and the control group. Only small or negligible effect sizes were observed in the A group.

No significant difference was found in any of the groups either in the picture-based narration test or in the structured conversation. Although some instructional impact was observed in both the RA and the R groups based on the effect size, medium and small effect sizes were observed in the control group.

As for the vocabulary test, significance tests indicate that the test score of the RA and the A groups at the post-test was larger than that at the pre-test, and that the test score of

the RA group at the delayed post-test was larger than that of the pre-test. On the other hand, no significance was found in the R group. Medium effect size between the pre-test and post-tests was observed in the RA group at the post-tests. The A group yielded a small effect size. The R group produced negligible effect sizes.

#### **4.2 The relationships between the achievement assessments**

This section reports the results of the relationships between the achievement assessments by means of statistical tests such as correlations and principal component analysis in order to investigate the research question regarding the type of knowledge being derived from the intervention. As mentioned in Chapter 3.3, assessments with a time constraint (the GJT, the picture-based narration test, and the structured conversation) were expected to elicit the participants' implicit knowledge. It was therefore thought that these assessments would load on the same factor by the principal component analysis. On the other hand, the gap-fill test without a time constraint would load on the other factor, namely explicit knowledge. Therefore, four achievement assessments (the timed GJT, the gap-fill test, and two oral tests) were used to explore what type of knowledge was promoted by the interventions.

It is noted that only 37 participants were selected to take the oral test. Due to the fact that the principal component analysis was undertaken to investigate the different knowledge induced by the interventions, nine participants in the control group, receiving no instruction, were removed from the data pool. Thus, the final data pool for the administration of the principal component analysis was comprised of 28 participants (RA=10; R=9; A=9), and it is acknowledged here that the sample size for carrying out the principal component analysis was small. As the maximum score of each test was different and the computation of the structured conversation was done by the number of

verbs with ‘-ed’ ending being divided by the total number of verb stems produced by each participant, the analysis in this section was based on the percentages instead of raw test scores. The layout of this section is as follows:

- 1) presenting the relationship between the tests at the post-test in each instructional group using the correlation and the principal component analysis;
- 2) presenting the relationship between the tests at the delayed post-test in each instructional group using the correlation and the principal component analysis;
- 3) presenting the results of participants’ self-reports.

It should be noted that the SPSS software is defaulted to use Pearson correlation to perform a principal component analysis. Thus, the results of correlation reported in section 4.2.1 and 4.2.2 were based on Pearson correlation instead of Spearman correlation.

#### ***4.2.1 The relationship between the tests at the post-test***

##### *4.2.1.1 The results of the RA group*

Descriptive statistics for the RA group in the four tests at the post-test are presented in Table 4.27.

Table 4.27

*Descriptive statistics for the RA group at the post-test*

<b>Test</b>	<b>Mean %</b>	<b>SD</b>
<b>Timed GJT</b>	74.00	23.90
<b>Gap-fill test</b>	56.25	41.77
<b>Picture-based narration</b>	20.00	17.87
<b>Structured conversation</b>	10.00	31.62

*N=10*



The results of Pearson correlation are reported in Table 4.28. Table 4.28 shows that the timed GJT was significantly positively correlated to the gap-fill test,  $r=.779$ ,  $p=.008 < .01$ . No significant association was observed between other tests.

Table 4.28

*Pearson correlation matrix for the tests of the RA group at the post-test*

<b>Test</b>	<b>GJT</b>	<b>GAP</b>	<b>Picture-based narration</b>	<b>Structured conversation</b>
<b>Timed GJT</b>	---	.779**	.280	.088
<b>Gap-fill test</b>		---	.140	-.053
<b>Picture-based narration</b>			---	.344
<b>Structured conversation</b>				---

*N=10 \* $p < .05$  \*\* $p < .01$*

The results of the principal component analysis for the four tests in the RA group at the post-test are given in Table 4.29. These results reveal that two components were extracted in the RA group at the post-test. The eigenvalues of the two components after extraction were 1.903 and 1.272, respectively. Both the eigenvalues of the extracted components were greater than Kaiser's eigenvalue-greater-than-one rule, which is the criterion for deciding the substantive importance of the eigenvalue. Overall, the two components accounted for 79.4% of the total variance.

Table 4.30 presents the results of the principal component factor analysis after the rotation of the RA participants' test scores. The results indicate that the timed GJT and the gap-fill test loaded heavily (higher than .9) on component 1. The two oral tests, the picture-based narration and the structured conversation, loaded heavily (higher than .7) on component 2. On the basis of these results, it is inferred that the timed GJT and the gap-fill test elicited the same type of knowledge; on the other hand, the two oral tests

tapped the same type of knowledge.

Table 4.29

*Principal component analysis of the RA group at the post-test<sup>54</sup>*

<b>Component</b>	<b>Eigenvalue</b>	<b>% of variance</b>	<b>% of cumulative</b>
<b>1</b>	1.903	47.578	47.578
<b>2</b>	1.272	31.791	79.369

Table 4.30

*Loadings after the oblique rotation of the RA at the post-test*

<b>Test</b>	<b>Component 1</b>	<b>Component 2</b>
<b>Timed GJT</b>	.922	
<b>Gap-fill test</b>	.952	
<b>Picture-based narration</b>		.773
<b>Structured conversation</b>		.859

#### 4.2.1.2 The results of the R group

Table 4.31 presents the descriptive statistics for the R group in the four achievement assessments at the post-test.

Table 4.31

*Descriptive statistics for the R group at the post-test*

<b>Test</b>	<b>Mean %</b>	<b>SD</b>
<b>Timed GJT</b>	83.06	23.34
<b>Gap-fill test</b>	58.33	34.23
<b>Picture-based narration</b>	19.44	30.05
<b>Structured conversation</b>	3.70	11.11

N=9

<sup>54</sup> Regarding the factorability of the principal component analysis, the value of KMO measure for sampling adequacy was .508, which was barely acceptable; the determinant of the R-matrix was .306, which was safe to perform the principal component analysis. However, Bartlett's test of sphericity was violated ( $p = .231 > .05$ ). Thus, the factorability of the RA group at the post-test was challenged.

Table 4.32 shows the correlation coefficients between the tests in the R group at the post-test. The results of correlation reveal that the timed GJT was significantly positively associated with the gap-fill test,  $r=.805$ ,  $p=.009<.01$ . A significant positive correlation was also observed between the picture narration and the structured conversation,  $r=.849$ ,  $p=.004<.01$ .

Table 4.32

*Pearson correlation matrix for the tests of the R group at the post-test*

<b>Test</b>	<b>GJT</b>	<b>GAP</b>	<b>Picture-based narration</b>	<b>Structured conversation</b>
<b>Timed GJT</b>	---	.805**	.334	.272
<b>Gap-fill test</b>		---	.392	.456
<b>Picture-based narration</b>			---	.849**
<b>Structured conversation</b>				---

$N=9$  \*\* $p < .01$

The results of the principal component analysis for the assessments in the R group at the post-test are given in Table 4.33. These results reveal that two components were obtained in the R group at the post-test. The eigenvalues of the two extracted components were 1.107 and 2.558. Both of these were greater than Kaiser's eigenvalue-greater-than-one rule. Overall, the two extracted components accounted for 91.6% of the total variance, which was substantial. The results of the principal component factor analysis after the rotation of the R participants' test scores are given in Table 4.34.

These results reveal that the timed GJT and the gap-fill test loaded heavily (higher than .9) on component 1. Also, the two oral tests (the picture-based narration and the structured conversation) loaded strongly on component 2. On the basis of these results, it is suggested that the timed GJT and the gap-fill test tapped the same type of knowledge, and that the oral tests tapped another type of knowledge.

Table 4.33

*Principal component analysis of the R group at the post-test<sup>55</sup>*

<b>Component</b>	<b>Eigenvalue</b>	<b>% of variance</b>	<b>% of cumulative</b>
<b>1</b>	1.107	27.683	27.683
<b>2</b>	2.558	63.950	91.634

N=9

Table 4.34

*Loadings after the oblique rotation of the R group at the post-test*

<b>Test</b>	<b>Component 1</b>	<b>Component 2</b>
<b>Timed GJT</b>	.984	
<b>Gap-fill test</b>	.907	
<b>Picture-based narration</b>		.956
<b>Structured conversation</b>		.946

#### 4.2.1.3 The results of the A group

The descriptive statistics for the A group in the assessments at the post-test are displayed in Table 4.35. According to Table 4.35, none of participants scored in the gap-fill test or the structured conversation. Therefore, principal component analysis could not be conducted for the A group.

Table 4.35

*Descriptive statistics for the A group in the post-test*

<b>Test</b>	<b>Mean %</b>	<b>SD</b>
<b>Timed GJT</b>	39.72	21.34
<b>Gap-fill test</b>	.00	.00
<b>Picture-based narration</b>	2.78	5.51
<b>Structured conversation</b>	.00	.00

N=9

<sup>55</sup> The performance of the principal component analysis in the R group at the post-test was suitable due to the following results: KMO=.510; the *p* value of Bartlett's test of sphericity was .015, *p*<.05; the determinant was .06748 >.00001

Table 4.36 presents the correlation coefficients of the tests in the A group in the post-test. It can be seen that there was a significant positive correlation between the timed GJT and the picture-based narration test,  $r=.738$ ,  $p=.023 < .05$ . No other correlation coefficients could be obtained due to the fact that none of the participants scored in the gap-fill test or the structured conversation.

Table 4.36

*Pearson correlation matrix for the tests of the A group in the post-test*

<b>Test</b>	<b>GJT</b>	<b>GAP</b>	<b>Picture-based narration</b>	<b>Structured conversation</b>
<b>Timed GJT</b>	---	---	.738*	---
<b>Gap-fill test</b>		---	---	---
<b>Picture-based narration</b>			---	---
<b>Structured conversation</b>				---

$N=9$  \* $p < .05$

#### *4.2.1.4 Summaries of the principal component analysis results at the post-test*

Based on the results described above, the components extracted from the instructional groups in the post-test are summarised in Table 4.37. Overall, the results indicate a two-factor solution. It was found that the results of both the RA and the R groups were in a similar pattern based on the fact that two components were extracted, and the timed GJT and the gap-fill test loaded on the same component, and the two oral tests loaded on the other component. These results suggest that the participants in both groups demonstrated similar performances in the timed GJT and the gap-fill test, and in the picture narration test and the structured conversation. Note that participants' performances in the timed GJT turned out to be different from those in the oral tests, contrary to expectations. It is acknowledged here that the results obtained from the principal component analysis in the RA group could be challenged due to the suitability

of performing the principal component analysis being broken (i.e. Bartlett's test of sphericity).

Table 4.37

*Summaries of the principal component analysis results at the post-test*

<b>Instructional group</b>	<b>N of components</b>	<b>Tests loaded on Component 1</b>	<b>Tests loaded on Component 2</b>
<b>RA</b>	2	GJT + GAP	2 oral tests
<b>R</b>	2	GJT + GAP	2 oral tests
<b>A</b>	---	---	---

#### **4.2.2 The relationship between the tests at the delayed post-test**

##### **4.2.2.1 The results of the RA group**

Table 4.38 summarises the descriptive statistics for the performance of the RA group in the four achievement assessments at the delayed post-test.

Table 4.38

*Descriptive statistics for the RA group at the delayed post-test*

<b>Test</b>	<b>Mean %</b>	<b>SD</b>
<b>Timed GJT</b>	74.50	21.43
<b>Gap-fill test</b>	71.25	39.55
<b>Picture-based narration</b>	26.25	29.13
<b>Structured conversation</b>	13.43	31.27

*N=10*

Table 4.39 shows the correlation coefficients across the four tests in the RA group at the delayed post-test. These results reveal that the timed GJT had a significantly positive association with the gap-fill test,  $r=.862$ ,  $p=.001 < .01$ . No significant correlation was found between the other tests in the RA group at the delayed post-test.

Table 4.39

*Pearson correlation matrix for the RA group at the delayed post-test*

<b>Test</b>	<b>GJT</b>	<b>GAP</b>	<b>Picture-based narration</b>	<b>Structured conversation</b>
<b>Timed GJT</b>	---	.862**	.474	.241
<b>Gap-fill test</b>		---	.472	.235
<b>Picture-based narration</b>			---	.594
<b>Structured conversation</b>				---

*N=10 \*\*= $p < .01$*

Table 4.40 presents the results of the principal component analysis in the RA group in the delayed post-test. These results reveal that two components were obtained from the tests in the RA group. The eigenvalues of the two extracted components were 2.467 and 1.039. Both the eigenvalues of the extracted components were greater than Kaiser's eigenvalue-greater-than-one rule, which is the criterion for deciding the substantive importance of the eigenvalue. Overall, the two components accounted for 87.6% of the total variance.

Table 4.40

*The principal component analysis of the RA group at the delayed post-test<sup>56</sup>*

<b>Component</b>	<b>Eigenvalue</b>	<b>% of variance</b>	<b>% of cumulative</b>
<b>1</b>	2.467	61.664	61.664
<b>2</b>	1.039	25.978	87.642

Table 4.41 summarises the results of the principal component factor analysis after the oblique rotation of the RA participants' performances in the tests. The results reveal that both the timed GJT and the gap-fill test loaded heavily (higher than .9) on component 1.

<sup>56</sup> The administration of the principal component analysis in the RA group at the delayed post-test was satisfactory due to the following results: KMO=.618; the  $p$  value of Bartlett's test of sphericity was .028<.05; the determinant was .126 >.00001

The two oral tests loaded on component 2. On the basis of these results, it is suggested that the performance of the RA participants in the timed GJT and in the gap-fill test were similar, and their performances in the two oral tests also resembled one another.

Table 4.41

*Loadings after the oblique rotation of the RA group at the delayed post-test*

<b>Tests</b>	<b>Component 1</b>	<b>Component 2</b>
<b>Timed GJT</b>	.958	
<b>Gap-fill test</b>	.961	
<b>Picture-based narration</b>		.760
<b>Structured conversation</b>		.971

#### 4.2.2.2 The results of the R group

Table 4.42 summarises the descriptive statistics for the R group's performances on the assessments at the delayed post-test. It can be observed that the structured conversation had zero variance. As a result, no further principal component analysis could be performed for the R group.

Table 4.42

*Descriptive statistics for the R group at the delayed post-test*

<b>Test</b>	<b>Mean %</b>	<b>SD</b>
<b>Timed GJT</b>	75.56	26.00
<b>Gap-fill test</b>	62.50	38.02
<b>Picture-based narration</b>	1.39	4.17
<b>Structured conversation</b>	.00	.00

*N*=9

The results of the correlation coefficients between the four tests in the R group at the delayed post-test are presented in Table 4.43. The results reveal that there was a



significant positive correlation between the timed GJT and the gap-fill test in the R group at the delayed post-test,  $r=.778$ ,  $p=.013 < .05$ .

Table 4.43

*Pearson correlation matrix for the tests of the R group at the delayed post-test*

Test	GJT	GAP	Picture-based narration	Structured conversation
Timed GJT	---	.778*	-.260	---
Gap-fill test		---	-.616	---
Picture-based narration			---	---
Structured conversation				---

$N=9$ ; \*  $p < .05$

#### 4.2.2.3 The results of the A group

Table 4.44 provides the descriptive statistics for the A group's performance in the delayed post-test. Table 4.45 summarises the correlation coefficients between the four tests for the A group in the delayed post-test. It will be noted that the correlations are generally fairly low. It was speculated that higher correlations might have been observed between the timed GJT and the gap-fill test, if a larger sample had been used.

Table 4.44

*Descriptive statistics for the A group at the delayed post-test*

Test	Mean %	SD
Timed GJT	36.39	16.54
Gap-fill test	4.17	12.50
Picture-based narration	1.39	4.17
Structured conversation	3.70	11.11

$N=9$

Table 4.45

*Pearson correlation matrix for the A group at the delayed post-test*

Tests	GJT	GAP	Picture-based narration	Structured conversation
Timed GJT	---	.309	-.031	-.202
GAP		---	-.125	-.125
Picture-based narration			---	-.125
Structured conversation				---

\*  $p < .05$  \*\*  $p < .01$

The results of the principal component analysis indicate that two components were retained in the A group (see Table 4.46). The eigenvalues of the two extracted components were 1.436 and 1.125. Both eigenvalues of the extracted components were greater than 1, and both components together accounted for a total of 64% of the variance.

Table 4.46

*The principal component analysis of the A group at the delayed post-test<sup>57</sup>*

<b>Component</b>	<b>Eigenvalue</b>	<b>% of variance</b>	<b>% of cumulative</b>
<b>1</b>	1.436	35.906	35.906
<b>2</b>	1.125	28.125	64.031

Table 4.47 summarises the results of the principal component factor analysis after the oblique rotation of the A group participants' test scores. The results reveal that both the timed GJT and the gap-fill test loaded heavily on component 1. Two oral tests (the picture-based narration and the structured conversation) loaded on component 2. According to these results, it appears that the timed GJT and the gap-fill test elicited similar patterns within the tests, and that the oral tests had other parallel patterns.

Table 4.47

*Loadings after the oblique rotation of the A group at the delayed post-test*

<b>Test</b>	<b>Component 1</b>	<b>Component 2</b>
<b>Timed GJT</b>	.774	
<b>Gap-fill test</b>	.733	
<b>Picture-based narration</b>		-.850
<b>Structured conversation</b>		.567

<sup>57</sup> Concerning the factorability of principal component analysis in the A group at the delayed post-test, the results that the KMO value was .538 and the determinant was .833 suggest the suitability of running the principal component analysis. However, Bartlett's test of sphericity was not met ( $p=.983 >.05$ ).

#### *4.2.2.4 Summaries of the results of the principal component analysis at the delayed post-test*

According to the results described above, the components extracted from the assessments in each instructional group are given in Table 4.48. Inspection of Table 4.48 reveals that the results of both the RA and the A groups were similar: a) two components being retained; b) the timed GJT and the gap-fill test being loading on the same component; and c) two oral tests being loading on the other component. Overall, the results specify a two-factor solution in both the RA and the A groups, suggesting that the participants' reactions to the timed GJT and the gap-fill test were not the same as the way that they reacted to the oral tests at the delayed post-test. However, it is acknowledged here that the results acquired from the principal component analysis in the A group, although in line with the results from the RA group, could be questioned due to the assumption of factorability being broken (i.e. Bartlett's test of sphericity).

Table 4.48

#### *Summaries of the principal component analysis results at the delayed post-test*

<b>Instructional Group</b>	<b>N of components</b>	<b>Tests loaded on Component 1</b>	<b>Tests loaded on Component 2</b>
<b>RA</b>	2	GJT + GAP	2 oral tests
<b>R</b>	---	---	---
<b>A</b>	2	GJT + GAP	2 oral tests

### ***4.2.3 Analysis of the participants' self-reports***

#### *4.2.3.1 Analysis of the post-task questionnaire following the timed GJT*

The post-task questionnaire was designed to explore whether or not the participants drew on explicit knowledge when undertaking the timed GJT. The questionnaire was distributed immediately after the completion of the timed GJT. The participants had to reflect on whether or not they had used a grammar rule. Note that only those participants who reported that they had used or added '-ed', and/or gave examples with

the targeted feature, were categorised as rule-users. For example, if a participant merely wrote down ‘-ed’, or provided an incorrect example such as ‘I *goed* out for a meal’, they were classified as a rule-user. However, if a participant wrote down ‘I watch TV last week’, s/he was not classified as a rule-user. It should be noted that the post-task questionnaire was also administered to the participants in the control group. Given that *none of the participants reported that they had used the rule at the post-tests*, the self-report of the control group is not reported in Table 4.49 and Table 4.50.

Table 4.49 shows the responses of the participants in the instructional groups to the post-task questionnaire at the post-test. According to Table 4.49, the participants, overall, expressed that they had not used the targeted grammatical rule (i.e. the English regular past tense) whilst taking the timed GJT. Seventy-six out of 90 participants reported that they did not resort to the targeted rule at the post-tests. Pearson’s chi-square was performed to examine whether there was any difference between participants’ self-reports across the three instructional groups. The result showed that a significant difference between groups was observed in participants’ self-reports,  $\chi^2(2) = 7.766, p = .021 < .05$ . Examination of Table 4.49 shows that the difference was due to the fact that the RA group had the higher proportion (9 out of 14 (64%)) of reporting use of the targeted rule in the timed GJT than the A group (1 out of 14 (7.1%)).

Table 4.50 shows the responses of the participants in the instructional groups to the post-task questionnaire at the delayed post-test. On the whole, the total instances of rule-use and non-rule-use were identical to those at the post-test (76 reported rule-use, and 14 non-rule-use). Pearson’s chi-square was carried out to examine whether there was any difference between the participants’ self-reports across the three groups. The result reveal that no significant difference was found between the groups,  $\chi^2(2) =$

3.784,  $p = .151 > .05$ .

Table 4.49

*Cross-tabulation of Group \* post-task questionnaire at the post-test<sup>58</sup>*

Source		Group			Total
		RA	R	A	
<b>Using rules? No</b>	Count	22	25	29	76
	Expected count	26.2	25.4	25.3	76.0
	% within report	28.9%	32.9%	38.2%	100.0%
<b>Yes</b>	Count	9	4	1	14
	Expected count	4.8	4.5	4.7	14.0
	% within report	64.3%	28.6%	7.1%	100%
<b>Total</b>	Count	31	29	30	90
	% within report	34.4%	32.2%	33.3%	100.0%

\* 'Yes' = using the targeted grammar rule while taking the timed GJT;

\* 'No' = not using the targeted grammar rule while taking the timed GJT.

Table 4.50

*Cross-tabulation of Group \* post-task questionnaire at the delayed post-test<sup>59</sup>*

Source		Group			Total
		RA	R	A	
<b>using rules? No</b>	Count	23	26	27	76
	Expected count	26.2	24.5	25.3	76.0
	% within report	30.3%	34.2%	35.5%	100.0%
<b>Yes</b>	Count	8	3	3	14
	Expected count	4.8	4.5	4.7	14.0
	% within report	57.1%	21.4%	21.4%	100%
<b>Total</b>	Count	31	29	30	90
	% within report	34.4%	32.2%	33.3%	100.0%

<sup>58</sup> Three cells (50%) have expected counts of less than .5. Therefore, the assumption of performing the chi-square test was not satisfied.

<sup>59</sup> Three cells (50%) have expected counts of less than .5. As a result, the assumption of performing the chi-square test was not satisfied.

Furthermore, a biserial correlation<sup>60</sup> was carried out to examine the association between the participants' self-reports and their timed GJT test scores. The results are summarised in Table 4.51. In terms of the post-test, the results of the biserial correlation reveal that there was a significantly positive relationship between the timed GJT scores and the participants' self-reports in the RA group,  $r_b = .794$ ,  $p = .000 < .01$ , and in the R group  $r_b = .818$ ,  $p = .004 < .01$ . The significantly positive associated results suggest that the higher the scores a participant achieved in the timed GJT, the more likely s/he would be to report using the targeted grammatical rule to undertake the test. With respect to the delayed post-test, the timed GJT scores were significantly positively related to the participants' self-reports in all the groups (the RA group,  $r_b = .754$ ,  $p = .001 < .01$ ; the R group,  $r_b = .840$ ,  $p = .006 < .01$ ; the A group,  $r_b = .947$ ,  $p = .002 < .01$ ). Overall, the results of the analysis of the post-task questionnaire for the timed GJT reveal that the participants' performances in the timed GJT were positively related to whether they were resorting to the targeted grammatical rule to carry out the test. These results imply that the timed GJT used in this current study appeared to have elicited the participants' explicit knowledge.

Table 4.51

*The biserial correlation between post-task questionnaire and their test scores in the timed GJT*

Group	N	Self-report vs Timed GJT	
		Post-test	Delayed post-test
RA	31	.794**	.754**
R	29	.818**	.840**
A	30	.043	.947**

<sup>60</sup> A biserial correlation is suggested for use when one of the two variables is dichotomous and 'continuous' (see Field, 2005, Chapter 4). Being 'continuous' means that there is continuum between the two variables (i.e. 'using the rule' vs 'not using the rule') as the degree of participants using the rule while taking the test varied individually.

#### 4.2.3.2 Analysis of the post-task interview following the oral tests

The post-task interview was conducted to explore whether or not participants used explicit knowledge when undertaking the oral tests. The interview was administered immediately after the completion of the picture-based narration and the structured conversation. During the interview, participants had to express whether or not they had used a grammar rule to undertake the oral tests. As with the post-task questionnaire, only those participants who reported using the rule, with either the verbalisation of the rule or the provision of an example containing the targeted feature, were categorised as rule-users. Given that *none of the control group participants reported using the rule at the post-tests*, the self-report of the control group is not reported in this section.

Table 4.52 presents the participants' responses in the interview at the post-test.

Examination of Table 4.52 shows that most of the participants expressed that they had not used the targeted grammatical rule whilst taking the oral tests (24 out of 28), only four participants reported using the rule. Pearson's chi-square was carried out to examine whether any difference existed between the participants' self-reports across the groups. No significant difference was observed in the participants' self-reports,  $\chi^2(2) = 2.230$ ,  $p = .328 > .05$ .

Table 4.53 shows the participants' responses in the interview at the delayed post-test. Pearson's chi-square was performed to investigate whether there was any difference between the participants' self-reports across the groups. No significant difference was observed in the participants' self-reports,  $\chi^2(2) = 2.733$ ,  $p = .255 > .05$ .

Table 4.52

*Cross-tabulation of Group \* post-task interview at the post-test<sup>61</sup>*

Source		Group			Total	
		RA	R	A		
Using rules?	No	Count	8	7	9	24
		Expected count	8.6	7.7	7.7	24.0
		% within report	33.3%	29.2%	37.5%	100.0%
	Yes	Count	2	2	0	4
		Expected count	1.4	1.3	1.3	4.0
		% within report	50.0%	50.0%	.0%	100%
<b>Total</b>		Count	10	9	9	28
		% within report	35.7%	32.1%	32.1%	100.0%

\* 'Yes' = using the targeted grammar rule while taking the oral tests;

\* 'No' = not using the targeted grammar rule while taking the oral tests.

Table 4.53

*Cross-tabulation of Group \* post-task interview at the delayed post-test<sup>62</sup>*

Source		Group			Total	
		RA	R	A		
Using rules?	No	Count	6	5	8	19
		Expected count	6.8	6.1	6.1	19.0
		% within report	31.6%	26.3%	42.1%	100.0%
	Yes	Count	4	1	1	9
		Expected count	3.2	2.9	2.9	9.0
		% within report	44.4%	44.4%	11.1%	100%
<b>Total</b>		Count	10	9	9	28
		% within report	35.7%	32.1%	32.1%	100.0%

\* 'Yes' = using the targeted grammar rule while taking the oral tests;

\* 'No' = not using the targeted grammar rule while taking the oral tests.

A point-biserial correlation<sup>63</sup> was then performed to investigate the association between the participants' self-reports and their merged oral test scores<sup>64</sup>. Given that the interview was conducted after the completion of the two oral tests together, the merged oral test

<sup>61</sup> Three cells (50%) have expected counts of less than .5. Therefore, the assumption of performing the chi-square test was not satisfied.

<sup>62</sup> Three cells (50%) have expected count less than .5. As a result, the assumption of performing the chi-square test was violated.

<sup>63</sup> As two correlation coefficients obtained by means of the biserial correlation were larger than 1 (RA group at the post-test and the A group at the delayed post-test), which is beyond normal correlation values between -1 and 1), the point-biserial correlation was used and is reported in Table 4.54.

<sup>64</sup> As the self-report referred to both oral tasks, the oral merged test scores was the overall percentage scores of both oral tasks i.e. total correct *-ed* use / (obligatory contexts + 8).



scores were used to run the correlation. The results are summarised in Table 4.54. For the post-test, only a significant positive correlation was found in the R group,  $r_{pb} = .882$ ,  $p = .002 < .05$ , suggesting that the higher the scores a participant achieved, the more likely s/he was to report using the targeted grammatical rule to take the oral tests. No correlation coefficient was obtained in the A group, because none of the participants expressed using the rule during the oral tests. In terms of the delayed post-test, a significantly positive association was only observed in the RA group,  $r_{pb} = .656$ ,  $p = .039 < .05$ .

Table 4.54

*The point-biserial correlation between post-task interviews and merged oral scores*

Group	N	Self-report vs Timed GJT	
		Post-test	Delayed post-test
RA	10	.402	.656*
R	9	.882**	.395
A	9	---	.661

\* $p < .05$ ; \*\* $p < .01$

## **Chapter 5**

### **The results of the questionnaires regarding the participants' bio-data and their attitudes towards the interventions, and from the ANCOVA**

#### **Introduction**

This chapter examines whether any confounding factors existed which might have potentially affected the results reported in Chapter 4. This chapter consists of three sections. The first and second sections report the results obtained from the data which was collected through the questionnaires. The questionnaire regarding participants' bio-data and English learning background is given as Appendix 9. The attitudinal questionnaire is Appendix 10. Section 1 is concerned with the participants' bio-data and their English learning backgrounds. Section 2 presents the results obtained from the ANCOVA, using the confounding variables identified in Section 1 as a covariate. The final section explores participants' attitude towards the interventions.

#### **5.1 Analysis of the participants' bio-data and English learning backgrounds**

The overall focus of this section concerns the relationship between some potentially confounding variables and the scores of achievements tests (the timed GJT, the gap-fill test, and the vocabulary test). Note that the oral data were excluded from the analysis in this section for the reason that no significant oral improvement was observed in any of the groups after receiving the intervention (see Sections 4.1.3.3 & 4.1.4). In addition, the control group did not take the vocabulary test at the post-tests, so a correlation between the control group and the vocabulary test was not possible. Also, it is noted that no participant in the control group scored on the gap-fill test at the post-tests, so that the computation of a correlation coefficient between this test and the potential confounding factors was not achieved.

The five potential confounding factors are reported in the following order: experience of travelling to English-speaking countries, length of English learning experience, extra exposure to English outside school, and experience of contact with English native speakers outside the classroom.

### 5.1.1 Experience of travelling to English-speaking countries

Table 5.1 shows participants' experience of travelling to English-speaking countries.

Table 5.1

*The cross-tabulation of whether participants had travel experience in English-speaking countries \* by group<sup>65</sup>*

Source			GROUP				Total
			RA	R	A	C	
<b>Travel?</b>	No	Count	28	25	28	28	109
		Expected count	28.2	26.3	27.3	27.3	109.0
		% within group	90.3%	86.2%	93.3%	93.3%	90.8%
	Yes	Count	3	4	2	2	11
		Expected count	2.8	2.7	2.8	2.8	11.0
		% within group	9.7%	13.8%	6.7%	6.7%	9.2%
<b>Total</b>	Count	31	29	30	30	120	
	% within group	100.0%	100.0%	100.0%	100.0%	100.0%	

\* yes = have experience of travelling in English-speaking countries

no = have no experience of travelling in English-speaking countries

Table 5.1 shows that only a minority of the participants (9.2%) had travelled to English-speaking countries such as the USA, UK, Australia, Canada and Singapore prior to the intervention. None of the participants had lived in an English-speaking country on a long-term basis, and the duration of stay in an English-speaking country was between

<sup>65</sup> Four cells (50%) have expected counts of less than .5; therefore, the assumption of performing the chi-square test was broken.

one and four weeks. No distinct difference was found between the groups. As the participants responded to this question by expressing ‘yes’ or ‘no’, Pearson’s chi-square was performed to examine whether there was any difference in participants’ responses to whether they had travel experience in English-speaking countries across the groups. The result showed that no significant difference between groups was observed in participants’ responses regarding their experience of travel to English-speaking countries,  $\chi^2(3) = 1.206, p = .752 > .05$ .

Table 5.2 summarises the results regarding the relationships between experience of travelling in English-speaking countries and the tests scores over time.

Table 5.2

*The point-biserial correlations between experience of travel in an English-speaking country and scores in the achievements tests*

Group	N	Post-test			Delayed post-test		
		GJT	Gap	Voc	GJT	Gap	Voc
RA	31	.146	.280	.336	.117	.259	.346
R	29	.150	.092	.097	.152	.178	.306
A	30	.322	.695**	.246	.196	-.068	.271
C	30	.249	---	---	.317	----	---

\* correlation is significant at the .05 level (2-tailed)

\*\* correlation is significant at the .01 level (2-tailed)

The results of Table 5.2 indicate that no significant relationships, in general, were observed. There was only a positive relationship observed in the A group on the gap-fill test at the post-test,  $r_{pb} = .695, p = .000 < .01$ . It is noted that most of the correlation coefficients reported in Table 5.2 were non-significant, and the results of Pearson’s chi-square test did not reveal any statistical significant differences in the travel experience of each group. Furthermore, a significant correlation was only found in the A group, and the results reported in Chapter 4 suggest that the A group, statistically speaking, did not

make any improved performance on the gap-fill test. Based on these viewpoints, participants' experience of travelling in English-speaking countries should not be considered as a confounding variable, and any differences observed in their performances on the achievement tests should not be attributed to it.

### ***5.1.2 Length of English learning experience***

Table 5.3 presents the descriptive statistics concerning the length of the participants' English learning experience. On the whole, the participants had been learning English for an average of 4.6 years. The average number of years of learning English was 4.63 in the RA group, 4.79 in the R group (the highest of the four groups), 4.40 in the A group (the lowest), and 4.53 in the control group. It appeared that the participants in the R group had the longest English learning experience (average 4.79 years), and the A group had the shortest (average 4.4 years). Note that the participants' formal English lessons at school commenced at grade 3, which was about 3.5 years prior to their participation in the current study. However, the average number of years of learning English in all four groups was greater than 3.5 years, suggesting that the participants had had extra exposure to English before their formal school English lessons started. It was therefore essential to examine whether the length of their English learning experience was a potential factor which might interfere with the current study.

As the K-S test results showed that none of the four groups fulfilled the normality assumption (RA group:  $D(31)=.286$ ,  $p=.000<.01$ ; R group:  $D(29)=.240$ ,  $p=.000<.01$ ; A group:  $D(30)=.262$ ,  $p=.000<.01$ ; Control group:  $D(30)=.292$ ,  $p=.000<.01$ ), the Kruskal-Wallis test, a non-parametric test equivalent to a one-way ANOVA, was conducted to compare the mean differences between the groups. The results showed that no significant difference was observed between the groups in terms of their English

learning length,  $H(3) = 1.153$ ,  $p = .764 > .05$ .

Table 5.3

*Descriptive statistics of participants' English learning length*

<b>Group</b>	<b>N</b>	<b>Mean</b>	<b>SD</b>	<b>Min</b>	<b>Max</b>
<b>RA</b>	31	4.629	1.915	3.0	10.0
<b>R</b>	29	4.793	1.656	3.0	8.0
<b>A</b>	30	4.400	1.694	3.0	9.0
<b>C</b>	30	4.533	1.889	3.0	9.0
<b>Total</b>	120	4.587	1.777	3.0	10.0

*NB: The mean was calculated by years.*

Spearman's correlation ( $r_s$ )<sup>66</sup> was carried out to examine the relationship between participants' English learning length and their test scores at the post-tests. The results of the correlation, summarised in Table 5.4, reveal that a significant relationship was found in the RA group on the vocabulary test at the post-test ( $r_s = .357$ ,  $p = .048 < .05$ ) and at the delayed post-test ( $r_s = .420$ ,  $p = .019 < .05$ ), in the R group on the gap-fill test at the delayed post-test ( $r_s = .443$ ,  $p = .016 < .05$ ), and in the control group on the timed GJT at the post-test ( $r_s = .396$ ,  $p = .03 < .05$ ). Note that significant language improvement was found in the RA group on the vocabulary test (see Section 4.1.5.3) and in the R group at the gap-fill test (see Section 4.1.2.3). The significant associations in the RA group and the R group can be seen in Table 5.4, suggesting that the factor of English learning length might have impacted on the vocabulary learning of the RA group, and on the gap-fill test of the R group at the delayed post-test. As a result, an ANCOVA was carried out, in which participants' English learning length as a confounding variable was controlled while analysing these tests. The results of the ANCOVA will be reported in Section 5.2.1.

<sup>66</sup> Most of the test scores on the timed GJT and gap-fill test in each group violated the assumption of normality (see Appendix 31), so Spearman's correlation was applied instead of Pearson correlation.

Table 5.4

*Spearman's correlation between English learning length and test scores at the post-tests*

Group	N	Post-test			Delayed post-test		
		GJT	Gap	Voc	GJT	Gap	Voc
RA	31	.195	.264	.357*	.238	.264	.420*
R	29	.276	.337	.329	.297	.443*	.209
A	30	.248	.023	.102	-.045	.102	.194
C	30	.396*	---	---	.336	---	---

\* correlation is significant at the .05 level (2-tailed)

\*\* correlation is significant at the .01 level (2-tailed)

### 5.1.3 Extra exposure to English<sup>67</sup> outside school: attending extra English lessons

This section investigates whether or not participants' extra exposure to English outside school during the instructional phases of the current study was a possible variable interfering with their performance in the achievement tests.

Table 5.5

*The cross-tabulation of extra English exposure \* Group<sup>68</sup>*

Source		Group				Total	
		RA	R	A	C		
Extra exposure ?	No	Count	6	7	8	6	27
		Expected count	7.0	6.5	6.8	6.8	27.0
		% within group	19.4%	24.1%	26.7%	20.0%	22.5%
	Yes	Count	25	22	22	24	93
		Expected count	24.0	22.5	23.3	23.3	93.0
		% within group	80.6%	75.9%	73.3%	80.0%	77.5%
Total	Count	31	29	30	30	120	
	% within group	100.0%	100.0%	100.0%	100.0%	100.0%	

\* yes = had extra English exposure; no = no extra English exposure

Table 5.5 displays the proportion of the participants' English exposure outside school between the groups during the administration of the current study. Overall, ninety-three

<sup>67</sup> The extra English exposure outside school refers to attendance at an English cram school or an English language institution, or with a private English tutor.

<sup>68</sup> No cells (0%) have expected counts of less than .5. Therefore, the assumption of performing the chi-square test was satisfied.

participants (78%) had extra English exposure outside school and twenty-seven participants (23%) did not attend any English lessons after school. Pearson's chi-square was carried out to explore whether there was any difference in the extra English exposure between groups. The result showed that there was no significant difference between the groups in the participants' responses to whether or not they had had extra English exposure,  $\chi^2(3) = .627, p = .890 > .05$ .

Table 5.6 summarises the descriptive statistics of the participants' extra exposure to English outside school per week.

Table 5.6

*Descriptive statistics of participants' extra exposure to English*

<b>Group</b>	<b>N</b>	<b>Mean</b>	<b>SD</b>	<b>Min</b>	<b>Max</b>
<b>RA</b>	31	169.03	121.25	0	450.00
<b>R</b>	29	151.03	97.04	0	330.00
<b>A</b>	30	140.00	106.19	0	360.00
<b>C</b>	30	213.00	140.81	0	480.00
<b>Total</b>	120	168.42	119.50	0	480.00

\* The mean is calculated by minute(s) per week.

Overall, the average of participants' extra English exposure after school was approximately 2.8 hours (168.42 minutes) a week in a cramming school or an English institution, or with a private English tutor. Inspection of Table 5.6 shows that the control group had more extra English exposure (213 minutes per week, more than 3 hours a week) than the other groups. The A group had the least extra English exposure (less than 2.5 hours a week). The Kruskal-Wallis test<sup>69</sup> was carried out to determine whether any

<sup>69</sup> The Kruskal-Wallis test was conducted instead of one-way ANOVA as a result of the violation of the normality assumption in terms of the extra English exposure. The results of K-S test showed that only the control group reached the normality assumption,  $D(30) = .135, p = .173 > .05$ . The other three instructional



difference existed in each group's extra English exposure. The results indicated that no significant differences were detected between the groups in terms of their extra English exposure outside school,  $H(3) = 6.491, p = .09 > .05$ .

With regard to the consideration of any relationship between the duration of extra exposure to English and the test scores, the results are summarised in Table 5.7 using Spearman's correlation. These results reveal statistically significant positive correlations in the R group, except in the timed GJT at the post-test and the vocabulary test at the delayed post-test. Positive correlations were also observed in the A group at the post-test (i.e. the vocabulary test) and at the delayed post-test (i.e., timed GJT and the gap-fill test). These positive associations suggest that the extra English exposure outside school was positively related to the participants' performance in the timed GJT, the gap-fill test, and the vocabulary test, depending on the condition that they were in. This could suggest that the more extra exposure to English the participants had outside school, the more likely they were to perform better in these tests, as a function of the condition they were in. It should be noted that no significant difference between the amounts of extra English exposure was found between the different groups. However, significant associations were observed between the extra exposure and the test scores in Table 5.7. This could suggest that the factor of participants' extra exposure to English outside school could be a confounding variable which might interfere with the effectiveness of the intervention for the current study. As a result, an ANCOVA was carried out, in which participants' extra exposure to English as a confounding variable was controlled while analysing these tests. The results of the ANCOVA will be reported in Section 5.2.2.

---

groups violated the normality assumption (RA:  $D(31) = .174, p = .018 < .05$ ; R:  $D(29) = .307, p = .000 < .01$ ; A:  $D(30) = .173, p = .022 < .05$ ).

Table 5.7

*Spearman's correlation between the participants' extra exposure to English after school and the test scores*

Group	N	Post-test			Delayed post-test		
		GJT	Gap	Voc	GJT	Gap	Voc
RA	31	.197	.203	.099	.160	.022	.153
R	29	.313	.519**	.568**	.466*	.408*	.237
A	30	.247	-.055	.399*	.395*	.366*	.247
C	30	.061	---	---	.350	---	---

\* correlation is significant at the .05 level (2-tailed)

\*\* correlation is significant at the .01 level (2-tailed)

#### 5.1.4 Contact with English native speakers outside school

Table 5.8 presents participants' responses to whether or not they had had contact with English native speakers outside school.

Table 5.8

*Crosstabulation of contact with English native-speakers outside school \* Group<sup>70</sup>*

Source		GROUP				Total	
		RA	R	A	C		
Contact with English native?	No	Count	29	24	27	29	109
		Expected count	28.2	26.3	27.3	27.3	109.0
		% within group	93.5%	82.8%	90.0%	96.7%	90.8%
	Yes	Count	2	5	3	1	11
		Expected count	2.8	2.7	2.8	2.8	11.0
		% within group	6.5%	17.2%	10.0%	3.3%	9.2%
Total	Count	31	29	30	30	120	
	% within group	100.0%	100.0%	100.0%	100.0%	100.0%	

yes = had had contact with English native speakers

no = had had no contact with English native speakers

Overall, eleven out of the 120 participants (9.2% of the total) expressed the view that they had had contact with English speakers outside school by means of MSN, Email,

<sup>70</sup> Four cells (50%) have an expected count of less than .5. As a result, the assumption of performing the chi-square test was not satisfied.

Skype, or face to face, and so on. The R group had the highest numbers (n=5) contacting English native speakers, and the control group had the lowest (n=1). The frequencies with which the eleven participants had had contact with English speakers were once a week (the most), then twice a week, and then once a month. Pearson's chi-square showed that no significant difference between the groups was observed in the participants' responses to contact with English native speakers after school,  $\chi^2(3) = 3.796, p = .284 > .05$ .

Table 5.9 displays the results of the correlation between the participants' responses to whether or not they had contact with English native speakers after school and the test scores.

Table 5.9  
*Point-biserial correlations between whether or not contact was made with English native speakers and the tests*

Group	N	Post-test			Delayed post-test		
		GJT	Gap	Voc	GJT	Gap	Voc
RA	31	-.022	.099	.088	.030	.105	.277
R	29	.326	.287	.164	.313	.305	.250
A	30	-.087	-.062	.023	-.005	.477**	-.064
C	30	.329	---	---	.116	---	---

\* correlation is significant at the .05 level (2-tailed)

\*\* correlation is significant at the .01 level (2-tailed)

The results show that no significant association was found across the groups, except only in the A group on the gap-fill test at the delayed post-test,  $r_{pb} = .477, p = .008 < .01$ .

Most of the groups' test scores were not significantly related to the factor regarding contact with English native speakers, and the number of participants who had had

contact with English native speakers was very small in all groups. Last but not least, the A group did not show any significant improvement in the vocabulary test at the delayed post-test (see Section 4.1.5.3). As a result, the factor of contact with English native speakers outside school should be discarded as a confounding variable when analysing the results of the achievement tests.

### ***5.1.5 Summary of this section***

Based on this analysis of the responses to the questionnaire exploring participants' bio-data and their English learning backgrounds, it was found that the factors with respect to travel experience in English-speaking countries, and contact with English native speakers were not significantly related to the participants' performances in the timed GJT, the gap-fill test, and the vocabulary test, taken at two testing phases. These factors should therefore not be regarded as confounding variables while carrying out analyses on the achievement tests. However, the participants' English learning length and extra exposure to English outside school (such as attending English lessons after school) turned out to be associated with the test scores, depending on the condition they were in. Based on the above results, a decision was made to introduce an ANCOVA to control the effect of these confounding variables. The results obtained from the ANCOVA will be reported in the following section 5.2.

## **5.2 The results of the ANCOVA**

Given that two potential confounding variables were observed (participants' English learning length and their extra exposure to English) based on the results obtained from the correlation between test scores and the questionnaire, an ANCOVA was performed to examine whether or not these confounding variables exerted any influence on the improved performance reported in Chapter 4. Note that only those tests scores which

were observed to significantly associate with the confounding variables are analysed and reported in the following sections.

### ***5.2.1 Using participants' English learning length as a covariate***

The results reported in Section 5.1.2 suggest that the participants' English learning length should be a confounding variable, in that a significant positive relationship was found in the R group on the gap-fill tests at the delayed post-test ( $r_s = .443$ ), and in the RA group on the vocabulary test at the post-tests ( $r_s = .357$  at the post-test,  $r_s = .420$  at the delayed post-test). A one-way between-group ANCOVA was therefore conducted, controlling the confounding variable of participants' English learning length (ELL), to examine the impact of the intervention on the gap-fill test at the delayed post-test, and on the vocabulary test at the post-tests.

#### ***5.2.1.1 The results of the ANCOVA on the gap-fill test at the delayed post-test***

The ANCOVA was performed by using the instructional group (GROUP) as the independent variable (i.e., the RA, R, A, and the control groups), the gap-fill test scores at the delayed post-test as the dependent variable, and the English learning length (ELL) as the covariate. The results of the ANCOVA<sup>71</sup> on the gap-fill test at the delayed post-test are given in Table 5.10. The results reveal that the covariate, participants' English learning length, was significantly related to the effect of GROUP,  $F(1, 115) = 6.682$ ,  $p = .011 < .05$ . There was also a significant effect between instructional groups after controlling for the effect of ELL,  $F(3, 115) = 8.662$ ,  $p = .000 < .05$ .

---

<sup>71</sup> The assumption of homogeneity of regression slopes for performing the ANCOVA on the gap-fill test at the delayed post-test was satisfied,  $F(3, 112) = 1.377$ ,  $p = .254 > .05$ .

Table 5.10

*The results of the ANCOVA on the gap-fill test at the delayed post-test*

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
ELL	35.567	1	35.567	6.682	.011
GROUP	146.093	3	48.698	8.662	.000

As a significant difference between the groups was observed, planned contrasts were carried out to determine where the differences were. Table 5.11 displays the results of these planned contrasts. The results reveal that the significant differences between the groups detected by the ANCOVA were due to the differences between the RA and the control groups,  $t(115) = 4.305$ ,  $p = .000 < .05$ , and between the R and the control groups,  $t(115) = 3.269$ ,  $p = .001 < .05$ . No significant difference was observed between the A group and the control group. These results suggest that the intervention of the RA and the R groups made a contribution to the participants' improvement on the gap-fill test at the delayed post-test, while taking the effect of the participants' English learning length into account.

Table 5.11

*The results of planned contrasts on the gap-fill test at the post-test*

Contrasts	Std Error	t	Sig	95% confidence interval	
				upper bound	lower bound
RA vs C	.607	4.305	.000*	1.414	3.818
R vs C	.618	3.269	.001*	.796	3.246
A vs C	.612	.559	.577	-.871	1.555

### 5.2.1.2 The results of the ANCOVA on the vocabulary at the post-test

The result of the ANCOVA<sup>72</sup> in the vocabulary test at the post-test is provided in Table

<sup>72</sup> The assumption of homogeneity of regression slopes for performing the ANCOVA on the vocabulary test at the post-test was not violated,  $F(2,84) = .243$ ,  $p = .785 > .05$ .

5.12. The result indicates that the covariate (ELL) was significantly related to the effect of GROUP,  $F(1, 86) = 6.041, p = .016 < .05$ . When the covariate ELL was controlled, no significant effect of GROUP observed,  $F(2, 86) = 1.709, p = .187 > .05$ . Given that no significant effect was found between the instructional groups, no planned contrast test was conducted. The results of the ANCOVA suggest that no significant impact on vocabulary learning was found between the groups, whilst using participants' English learning length as a covariate.

Table 5.12

*The results of the ANCOVA on the vocabulary test at the post-test*

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
ELL	20.670	1	20.670	6.041	.016
GROUP	11.698	2	5.849	1.709	.187

### 5.2.1.3 The results of the ANCOVA on the vocabulary at the delayed post-test

Table 5.13 presents the results of the one-way between-groups ANCOVA<sup>73</sup> on the vocabulary test at the delayed post-test.

Table 5.13

*The results of the ANCOVA on the vocabulary test at the delayed post-test*

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
ELL	19.158	1	19.158	7.339	.008
GROUP	8.069	2	4.048	1.551	.218

The ANCOVA results indicate that the covariate (ELL) was significantly related to the

<sup>73</sup> The assumption of the homogeneity of regression slopes for conducting the ANCOVA on the vocabulary test at the delayed post-test was satisfied,  $F(2, 84) = .307, p = .737 > .05$ .

effect of GROUP,  $F(1, 86) = 7.339, p = .008 < .05$ . No significant effect between the groups was observed after controlling for the effect of ELL,  $F(2, 86) = 1.551, p = .218 > .05$ . Thus, no planned contrast was conducted to examine the mean differences between the groups. The fact that no significant difference was observed between the groups suggests that there was no significant effect of the intervention when taking participants' English learning length into consideration.

### ***5.2.2 Using participants' extra exposure to English as a covariate***

The results of the questionnaire regarding English learning background showed that participants' extra exposure to English outside school could possibly influence their performance on the timed GJT, the gap-fill test, and the vocabulary test, depending on the conditions that participants were in (see Section 5.1.3, and Table 5.7). So an ANCOVA was carried out by using the interventional groups (GROUP) as the independent variable, the test scores as the dependent variable, and the extra exposure to English (EEE) (i.e. the total number of minutes of exposure to English that participants had outside school per week) as the covariate.

#### ***5.2.2.1 The results of the ANCOVA on the timed GJT at the delayed post-test***

Table 5.14 presents the results of the ANCOVA<sup>74</sup> on the timed GJT at the delayed post-test. The results reveal that the covariate, the duration of participants' extra English exposure outside school per week, was significantly associated with the effect of GROUP,  $F(1, 115) = 9.400, p = .003 < .05$ . A significant effect between the groups after controlling for the effect of EEE was also found,  $F(3, 115) = 12.093, p = .000 < .05$ .

---

<sup>74</sup> The assumption of homogeneity of regression slopes for carrying out the ANCOVA on the timed GJT at the delayed post-test was not violated,  $F(3, 112) = 2.013, p = .116 > .05$ .



Table 5.14

*The results of the ANCOVA on the timed GJT at the delayed post-test*

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
EEE	624.920	1	624.920	9.400	.003
GROUP	2400.997	3	800.332	12.093	.000

Table 5.15 summarises the results of the planned contrasts. The results indicate that the significant differences observed between the groups by the ANCOVA were due to the differences between the RA and the control groups,  $t(115) = 5.421$ ,  $p = .000 < .05$ , and between the R and the control groups,  $t(115) = 3.962$ ,  $p = .000 < .05$ . No significant difference was found between the A group and the control group. These results suggest that the interventions of the RA and the R groups, as opposed to that of the A group, was significantly conducive to learners' performance on the timed GJT up to six weeks after receiving the intervention, even considering the confounding effect of participants' extra exposure to English.

Table 5.15

*The results of planned contrasts on the timed GJT at the delayed post-test*

Contrasts	Std Error	t	Sig	95% confidence interval	
				upper bound	lower bound
RA vs C	2.107	5.421	.000*	15.597	7.249
R vs C	2.160	3.962	.000*	12.838	4.279
A vs C	2.157	1.440	.153	7.379	-1.166

#### 5.2.2.2 The results of the ANCOVA on the gap-fill tests at the post-test

Table 5.16 presents the results of the one-way between-groups ANCOVA<sup>75</sup> in the gap-

<sup>75</sup> The assumption of homogeneity of regression slopes for performing the ANCOVA on the gap-fill test at the post-test was satisfied,  $F(3,112) = 2.528$ ,  $p = .061 > .05$ .

fill test at the post-test. The results reveal that the EEE was significantly related to the effect of GROUP,  $F(1, 115) = 3.974, p = .049 < .05$ , though the  $p$  value was close to the borderline. A significant effect between groups was observed,  $F(3, 115) = 10.308, p = .000 < .05$ , suggesting that the effect of GROUP, statistically speaking, was different whilst controlling the effect of EEE.

Table 5.16

*The results of the ANCOVA on the gap-fill test at the post-test*

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
EEE	17.706	1	17.706	3.974	.049
GROUP	137.781	3	45.927	10.308	.000

Table 5.17

*The results of planned contrasts on the gap-fill test at the post-test*

Contrasts	Std Error	t	Sig	95% confidence interval	
				upper bound	lower bound
RA vs C	.546	4.525	.000*	3.549	1.388
R vs C	.559	3.697	.000*	3.176	.960
A vs C	.558	.553	.581	1.415	-.797

Table 5.17 above summarises the results of the planned contrasts. It was found that the significant difference between the groups detected by the ANCOVA in Table 5.16 was due to the difference between the RA and the control groups,  $t(115) = 4.525, p = .000 < .05$ , and between the R and the control groups,  $t(115) = 3.697, p = .000 < .05$ . No significant difference was found between the A group and the control group. These results suggest that the interventions of the RA and the R groups made a significant contribution towards the participants' performance in the gap-fill test at the post-test, while controlling the effect of EEE. The fact that the A group did not significantly outperform the control group suggests that the intervention of the A group had little

impact on its learners' performance in the gap-fill test at the post-test.

### 5.2.2.3 The results of the ANCOVA on the gap-fill test at the delayed post-test

Table 5.18 displays the results of the one-way between-groups ANCOVA<sup>76</sup> on the gap-fill test at the delayed post-test. The results reveal that the EEE was not significantly related to the effect of GROUP,  $F(1, 115) = 2.873, p = .093 > .05$ . A significant effect between groups after adjusting for the effect of EEE was also obtained,  $F(3, 115) = 9.365, p = .000 < .05$ .

Table 5.18

*The results of the ANCOVA in the gap-fill test at the delayed post-test*

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
EEE	16.675	1	16.675	2.873	.093
GROUP	163.056	3	54.352	9.365	.000

Planned contrasts were performed to examine where the differences between groups were and the results are summarised in Table 5.19. Table 5.19 demonstrates that the significant difference found in Table 5.18 was the result of the differences between the RA group and the control group,  $t(115) = 4.476, p = .000 < .05$ , and between the R group and the control group,  $t(115) = 3.608, p = .000 < .05$ . The results indicate that both the RA and the R groups significantly outperformed the control group, and that the retention was sustained six weeks after the intervention had been completed, even taking participants' extra English exposure into consideration. No significant difference was detected between the A group and the control group, suggesting that the intervention of the A group did not assist learners in learning the targeted feature, in

<sup>76</sup> The assumption, the homogeneity of regression slopes, for carrying out the ANCOVA on the gap-fill test at the delayed post-test were fulfilled,  $F(3, 112) = 2.528, p = .061 > .05$ .

terms of the learning gains assessed by a gap-fill test delivered six weeks after the intervention had finished.

Table 5.19

*The results of planned contrasts on the gap-fill test at the delayed post-test*

Contrasts	Std Error	t	Sig	95% confidence interval	
				upper bound	lower bound
<b>RA vs C</b>	.623	4.476	.000	4.020	1.554
<b>R vs C</b>	.638	3.608	.000	3.567	1.039
<b>A vs C</b>	.637	.840	.403	1.798	-.727

#### 5.2.2.4 The results of the ANCOVA on the vocabulary test at the post-test

Table 5.20 summarises the result of the ANCOVA<sup>77</sup> on the vocabulary test at the post-test. The results reveal that the covariate (EEE) was significantly related to the effect of GROUP,  $F(1, 86) = 10.409$ ,  $p = .002 < .05$ . When the covariate EEE was controlled, no significant effect of GROUP was observed,  $F(2, 86) = 1.355$ ,  $p = .263 > .05$ . Due to the fact that no significant effect was found between the groups, no planned contrast test was performed. The results of the ANCOVA suggest that no significant impact on vocabulary learning was found between the groups at the post-test.

Table 5.20

*The results of the ANCOVA on the vocabulary test at the post-test*

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
<b>EEE</b>	34.003	1	34.003	10.409	.002
<b>GROUP</b>	8.853	2	4.427	1.355	.263

<sup>77</sup> The assumption of homogeneity of regression slopes for performing the ANCOVA on the vocabulary test at the post-test was not violated,  $F(2,84) = 1.958$ ,  $p = .148 > .05$ .

### ***5.2.3 Summary of the ANCOVAs which take into account English learning length and extra exposure to English as potentially confounding factors.***

#### *5.2.3.1 The results of the ANCOVA by using English learning length as a covariate*

The results of the ANCOVA showed that the R group significantly outperformed the control group. This could suggest that the improved effect observed in the R group on the gap-fill test at the delayed post-test should not be attributed to the confounding factor of English learning length. In addition, after adjusting for the effect of English learning length, there was no significant difference between the intervention groups on the vocabulary test at both the post-test and delayed post-test (see Section 5.2.1.2 and Section 5.2.1.3). The non-significant results suggest that no instructional impact was observed for the intervention groups based on the vocabulary test while statistically controlling for participants' English learning length. Thus, the improved performance of RA group on the vocabulary test, which was observed in Chapter 4, was not upheld.

#### *5.2.3.2 The results of the ANCOVA by using extra exposure to English as a covariate*

The results of the ANCOVA reveal that the R group significantly outperformed the control group on the GJT at the delayed post-test (see Section 5.2.2.1), and on the gap-fill test at the post-tests (Section 5.2.2.2 and 5.2.2.3). Although the ANCOVA did not detect any instructional impact on the vocabulary test at the post-test (see Section 5.2.2.4), the R group did not show any statistically significant improvement on the vocabulary test in any case, even without taking extra exposure to English into consideration (see Section 4.1.5.3). In sum, the improved performance of the R group on the timed GJT at the delayed post-test, and on the gap-fill test at the post-tests, which were reported in Chapter 4, should not be ascribed to the effect of the extra exposure to English.

The results of the ANCOVA reveal that no instructional impact was found in the A group on the timed GJT and the gap-fill test at the delayed post-test whilst controlling participants' extra English exposure, given that the A group did not significantly outperform the control group. In addition, the ANCOVA did not detect any significant impact between the groups on the vocabulary test at the post-test. Although the results reported in Chapter 4 (see Section 4.1.5.3 and Table 4.21) reveal that the participants in the A group made a significant improved performance on the vocabulary test from the pre-test to the post-test, the instructional impact was not convincing based on the results obtained from the ANCOVA (See Section 5.2.2.4).

### **5.3 Analysis of the attitudinal questionnaire**

An attitudinal questionnaire comprised of seven questions<sup>78</sup> was distributed at the end of the intervention phase (see Appendix 10). This questionnaire was designed to examine whether or not certain factors (the use of computers, and attitude towards the intervention) had any possible influence on the effectiveness of the intervention observed in Chapter 4. However, it is acknowledged that this attitudinal measure was fairly crude, and that the delay between the questionnaire and the actual interventions was not ideal, being delivered at the end of the delayed post-test.

The participants responded to each question by choosing either 'Yes' or 'No', and no neutral option was given to them. Point-biserial correlation was conducted to examine participants' responses across the groups. Also, as the control group did not receive any intervention, the attitudinal questionnaires were filled out only by the participants in the

---

<sup>78</sup> Note that although the questionnaire comprised seven questions, only four questions are analysed and presented in this section. Questions 1 and 2 explored the same issue concerning the operation of the computer, Questions 3 and 4 examined whether the intervention motivated the participants to different extents between the groups, and Questions 5 and 6 were concerned with the level of difficulty in the interventions. The attitudinal questions analysed and reported here are Questions 1, 3 and 5, and Question 7, which addressed participants' willingness to take part in future similar activities.

instructional groups, involving 90 participants in total.

### 5.3.1 The operation of the computer

#### ***Q1: Was it easy to operate the computer to access the intervention?***

Table 5.21 displays participants' responses to Question 1 regarding whether they had experienced any difficulty in operating the computer to access the intervention. Overall, seventy-nine participants (87.8%) responded that the operation of the computer to obtain access to the training materials was easy. On the other hand, eleven participants (12.2%) did not think that it was easy to use the computer to receive the intervention (RA (n=3), R (n=4), and A (n=4)). Table 5.21 shows that no apparently different proportion between the groups was observed.

Table 5.21

*The cross-tabulation of operation on the computer \* Group<sup>79</sup>*

Source			Group			Total
			RA	R	A	
<b>Easy to operate the computer</b>	<b>No</b>	Count	3	4	4	11
		expected count	3.8	3.5	3.7	11.0
		% within group	9.7%	13.8%	13.3%	12.2%
	<b>Yes</b>	Count	28	25	26	79
		expected count	27.2	25.5	26.3	79.0
		% within group	90.3%	86.2%	86.7%	87.8%
<b>Total</b>	Count	31	29	30	90	
	% within group	100.0%	100.0%	100.0%	100.0%	

\*'no' = not easy to operate the computer

'yes' = easy to operate the computer

Pearson's chi-square was conducted to examine whether there was any difference between the proportion of participants' responses to Question 1 between the groups. The result showed that the proportion of participants' attitudes towards operating the

<sup>79</sup> Three cells (50%) have expected counts of less than .5. Thus, the assumption of performing the chi-square test was not satisfied.

computer to access the intervention was not significantly different between the groups,  $\chi^2(2) = .288, p = .866 > .5$ . Additionally, a point-biserial correlation was performed to examine the relationship between participants' responses to Question 1 and their test scores (Table 5.22). The results reveal that no significant correlations were found. As a result, any differences found in the achievement tests between the groups should not be attributed to the level of difficulty encountered in using the computer.

Table 5.22

*The point-biserial correlation between whether or not participants felt that it was easy to operate the computer and the test scores*

Group	N	Post-test			Delayed post-test		
		GJT	Gap	Voc	GJT	Gap	Voc
RA	31	.243	.242	.174	.351	.255	.304
R	29	.171	.125	.020	.086	.109	.073
A	30	.269	.073	.241	.216	.099	.241

### 5.3.2 The motivation level of the intervention

#### ***Q3: Were the instructional materials interesting?***

Table 5.23 shows the frequencies across the instructional groups concerning whether the participants found the intervention interesting. Overall, 61 (68%) participants felt that the intervention was interesting, and 29 (32%) participants held the opposite opinion. The participants in the RA group seemed to enjoy the intervention more than those in the other two groups. On the other hand, the R intervention appeared to be the least enjoyable compared with the other two, but over half of the participants (62%) within the R group did rate it as interesting. The result of Pearson's chi-square was  $\chi^2(2) = .777, p = .678 > .05$ , suggesting that the proportion of participants' responses to whether or not the intervention was interesting was not significantly different between the groups.



Furthermore, a point-biserial correlation was administered to examine the relationship between participants' responses to Question 3 and the test scores. As can be seen from Table 5.24, no significant correlations were found, suggesting that whether or not the participants considered the intervention to be interesting or boring was not significantly related to their performances in the timed GJT, the gap-fill test, and the vocabulary test.

Table 5.23

*The cross-tabulation of attitudes towards the intervention\* Group<sup>80</sup>*

Source		Group			Total	
		RA	R	A		
<b>Interesting</b>	<b>No</b>	Count	9	11	9	29
		expected count	10.0	9.3	9.7	29.0
		% within group	29.0%	37.9%	30.0%	32.2%
	<b>Yes</b>	Count	22	18	21	61
		expected count	21.0	19.7	20.3	61.0
		% within group	71.0%	62.1%	70.3%	67.8%
<b>Total</b>		Count	31	29	30	90
		% within group	100.0%	100.0%	100.0%	100.0%

\* no = the intervention was not interesting

yes = the intervention was interesting.

Table 5.24

*The point-biserial correlation between whether or not participants felt an intervention interesting and the test scores*

Group	N	Post-test			Delayed post-test		
		GJT	Gap	Voc	GJT	Gap	Voc
RA	31	.228	.134	.155	.152	.101	.171
R	29	-.031	-.065	.060	-.071	-.042	-.021
A	30	.071	-.284	.141	.081	-.202	-.137

<sup>80</sup> No cells (0%) have an expected count of less than .5. Therefore, the assumption of performing the chi-square test was upheld.

### 5.3.3 The level of difficulty of the intervention

#### ***Q5: Were the instructional materials difficult?***

Table 5.25 summarises the attitudes of the participants towards the intervention in terms of its level of difficulty. Overall, a total of 39 (43%) participants considered the instructional material to be difficult, and 51 (57%) participants considered it to be easy. Table 5.25 shows that the A group had the highest proportion of participants who regarded the intervention as being difficult (n=15, 50% within group). On the other hand, fewer students in the R group (n=11) reported experiencing difficulty when undertaking the intervention, compared with the RA group (n=13) and the A group (n=15). The results of Pearson's chi-square test on these data indicated that no significant difference was observed across the groups in terms of the proportion of participants who perceived the intervention to be difficult,  $\chi^2(2) = .912, p = .634 > .05$ .

Table 5.25

*Attitudes of the participants towards the level of difficulty of the intervention\* Group<sup>81</sup>*

Source		Group			Total
		RA	R	A	
<b>difficult? No</b>	Count	18	18	15	51
	expected count	17.6	16.4	17.0	51.0
	% within group	58.1%	62.1%	50.0%	56.7%
<b>Yes</b>	Count	13	11	15	39
	expected count	13.4	12.6	13.0	39.0
	% within group	41.9%	37.9%	50.0%	43.3%
<b>Total</b>	Count	31	29	30	90
	% within group	100.0%	100.0%	100.0%	100.0%

\*yes = the intervention was difficult for me

no = the intervention was not difficult for me.

According to Table 5.25, 39 out of 90 participants rated the intervention as difficult. A sub-question to Question 5 was delivered to identify which activity (reading, listening

<sup>81</sup> No cells (0%) have an expected count of less than .5. Therefore, the assumption of performing the chi-square test was satisfied.

or both) was considered to be difficult for participants. Table 5.26 summarises the frequencies of the responses of the 39 participants to Sub-question 5. According to Table 5.26, thirteen (33.3%) of the 39 participants considered that only the reading activities were difficult to undertake (five in the RA group, one in the R group, and seven in the A group). Ten (25.6% of the total) of the 39 participants experienced difficulty only in the listening activities. Sixteen (41% of the total) of the 39 participants who thought the instructional material difficult had struggled with both reading and listening activities.

Table 5.26

*Which activities were difficult?*

	Only (R)	Only (L)	Both (R&L)	Total
<b>RA</b>	5	3	5	13
<b>R</b>	1	4	6	11
<b>A</b>	7	3	5	15
<b>Total</b>	13	10	16	39

\* R=reading activities; L=listening activities

Furthermore, a point-biserial correlation was applied to examine the association between participants' responses to Question 5 and the test scores (Table 5.27).

Table 5.27

*Point-biserial correlation between whether or not participants perceived an intervention difficult and the test scores*

Group	N	Post-test			Delayed post-test		
		GJT	Gap	Voc	GJT	Gap	Voc
<b>RA</b>	31	-.373*	-.504**	-.631**	-.454*	-.528**	-.426*
<b>R</b>	29	-.184	-.295	-.309	-.092	-.275	-.094
<b>A</b>	30	.048	-.186	-.443*	-.305	-.253	-.145

\* correlation is significant at the .05 level (2-tailed)

\*\* correlation is significant at the .01 level (2-tailed)

The results of the point-biserial correlation indicate that significant negative correlations were found in the tests of the RA group over time. A significant negative association was found in the A group on the vocabulary test at the post-test. The strength of the correlational results in both the RA group and A group was either medium (i.e.,  $.3 < r_{pb} < .5$ ) or large (i.e.  $r_{pb} > .5$ ), suggesting that the less difficulty participants reported having experienced during the instructional phases, the more likely they were to score higher in the tests. Although significant negative correlation was observed in the RA and the A groups, depending on the condition that they were in, Pearson's chi-square did not detect any difference across the groups concerning the proportion of participants who rated the intervention as difficult. Therefore, the issue about the level of difficulty of the intervention would not be regarded as a confounding variable.

#### ***5.3.4 Willingness to carry out similar activities in the future***

Table 5.28 presents the responses of the participants to Question 7 concerning whether they were willing to carry out future similar activities. Overall, fifty-five participants (61%) showed their willingness, and thirty-five participants (38.9%) showed unwillingness. Pearson's chi-square was performed to investigate whether any difference existed between the proportion of participants' responses to Question 7 across the groups. The result showed that the proportion of participants' willingness to do similar activities in the future was not significantly different across the three groups,  $\chi^2(2) = .374, p = .830 > .05$ .

Table 5.28

*Willingness to carry out similar activities\* Group*<sup>82</sup>

Source	Group			Total	
	RA	R	A		
<b>Willingness No</b>	Count	13	10	12	35
	expected count	12.1	11.3	11.7	35.0
	% within group	41.9%	34.5%	40.0%	38.9%
<b>Yes</b>	Count	18	19	18	55
	expected count	18.9	17.7	18.3	55.0
	% within group	58.1%	65.5%	60.0%	61.1%
<b>Total</b>	Count	31	29	30	90
	% within group	100.0%	100.0%	100.0%	100.0%

\* no = not willing to attend similar activities; yes = willing to attend.

### 5.3.5 Summary of this section

Analysis of the responses to the attitudinal questionnaire reveals that no significant attitudinal difference was found across the groups regarding the use of the computer to gain access to the intervention. Overall, 88% of the participants thought that it was easy to operate the computer. Also, no significant attitudinal difference was observed across the groups with respect to whether they rated the instructional materials as interesting or difficult. On the whole, over half of the participants (68%) reckoned the instructional material to be interesting, and 43% of participants thought that the instructional materials were difficult. Furthermore, 61% of participants, overall, showed their willingness to participate in similar activities in the future.

Due to the fact that no significant differences were found in the responses to the attitudinal questionnaire across the three instructional groups by means of Pearson's chi-square, any further differences found in the participants' performances in the achievement tests should not be attributed to these attitudinal variables regarding the experience of difficulty in the operation of the computers, and the degree of motivation

<sup>82</sup> No cells (0%) have an expected count of less than .5. Therefore, the assumption of performing the chi-square test was upheld.

generated by instructional materials.

## Chapter 6 Discussion of the results and findings

### Introduction

This chapter pulls together the results and findings from Chapter 4 and 5, in order to answer the research questions with which this thesis is concerned. These questions explored some unverified issues in previous PI studies (i.e. the causative component in the framework of PI and PI's impact). Sub-sections 1-5 of this chapter each discuss a separate test, and link the results to previous PI studies as follows:

- 1) The findings of the timed GJT;
- 2) The findings of the gap-fill test;
- 3) The findings of the oral tests;
- 4) The findings of the vocabulary test;
- 5) The findings regarding what type of knowledge was derived from PI activities;

In sub-section 6, the factors in relative effectiveness of interventions observed are discussed based on the theoretical framework of this study.

### 6.1 Discussion of the findings of the timed GJT

#### *6.1.1 The relative impact of the interventions on the timed GJT*

According to the results obtained by Friedman's test and described in Section 4.2.2.3 (Table 4.2), it was found that both the RA and the R groups made significant improvement in the timed GJT over time. No significant difference in learning gains was detected between these two groups at the two post-tests, suggesting that the R group performed equally to the RA group. Nevertheless, the performance of the R group at the delayed post-test was significantly related to the participants' extra exposure to English (see section 5.1.3), which could potentially affect its improved performance observed. However, the results from the ANCOVA (see Section 5.2.2.1) exclude this

possibility, given that the R group significantly outperformed the control group. These results suggest that the interventions of the structured-input-activities group (the RA group) and the referential-activities-only group (the R group) were beneficial for the learners' interpretation of the English grammatical past tense marker '-ed' feature. Also, the learning gains of both groups were maintained up to six weeks after the completion of the intervention. On the other hand, no significant improvement on the timed GJT was found in the A group or the control group over time. The results suggest that the A group did not gain any significant contribution to learning the '-ed' feature.

Overall, the comparative effectiveness of the interventions observed by means of comparing the mean scores on the timed GJT was as follows:

- a) at the post-test:  $RA = R > A = C$
- b) at the delayed post-test:  $RA = R > A = C$

To sum up, the comparative results suggest that the referential activities alone did improve learners' performance on the timed GJT on the acquisition of the English '-ed' feature. However, the affective activities, delivered either alone or after the referential activities, did not contribute to any language improvement on the timed GJT.

When taking the effect size into consideration (see Section 4.1.1.4, Tables 4.5 and 4.6), the relative magnitude of the instructional effect size for the interventions on the timed GJT were as follows:

- a) comparison with the control group:
  - i) at the post-test:  $RA(1.18) > R(1.05) > A(.20)$
  - ii) at the delayed post-test:  $RA(1.49) > R(1.00) > A(.33)$



b) comparison with the pre-test score:

i) at the post-test: RA(1.40) > R(1.19) > A(.32) > C(.10)

ii) at the delayed post-test: RA(1.63) > R(1.10) > A(.46) > C(.15)

All of the effect sizes produced by both the RA and the R groups were greater than 1 and therefore considered to be large effects ( $d > .8$ ), and the effect sizes of both groups were all larger than the mean effect size of the meta-linguistic judgment (mean  $d=.82$ ) reported by Norris & Ortega (2000, p.471). It is noted that although the effect size of the RA group was larger than that of the R group, the participants in the RA group were exposed to more training items containing the target feature than those in the R group. The participants in both the R and the A groups received the same amount of targeted feature. However, only small ( $.2 > d > .5$ ) or insignificant effect sizes were observed in the A group. The results obtained from the computation of effect size suggest that structured input activities and the referential activities alone had a substantial impact on learning the English ‘-ed’ feature in terms of the timed GJT. However, affective activities alone did not have an instructional impact on learning the ‘-ed’ feature.

In terms of the relative instructional magnitude of this study in comparison with a prior PI study (Toth, 2006), the results obtained from the current study are quite promising. It was observed that the effect sizes of the RA group were all larger than those of the PI group in Toth’s study (see Section 4.1.1.4, Tables 4.5, 4.6 & 4.7). Large effect size was found in the R group and the PI group of Toth, when comparing the magnitude of change from pre-test to post-tests. However, when the control group was taken into consideration, the instructional magnitude of the R group ( $d=1.19$  at post-test and  $d=1.10$  at delayed post-test) was larger than that of the PI group in Toth’s study ( $d=.88$  at post-test and  $d=.62$  at delayed post-test). Note that the PI group in Toth’s study

received the full ‘PI package’, and the RA and the R group merely received a partial ‘PI’ package. In addition, the test items of GJT used in Toth’s study were not separately timed, and Toth’s learners were given 25 minutes overall to complete the entire test. In this sense, it is reasonable to presume that the participants in the current study encountered more time pressure compared with those in Toth’s study, which could potentially have impeded their performance in the GJT. Nevertheless, large effect sizes were found in both the RA and the R groups. These results suggest that the RA and the R group received a desirable contribution to learning the targeted ‘-ed’ feature, as measured by a timed GJT.

### ***6.1.2 Linkage of the results of the timed GJT to previous PI studies***

The intervention of the RA and the R groups in this study yielded a substantial effect in the timed GJT in terms of learning the English ‘-ed’ form. The results are in line with prior PI studies which provided empirical evidence supporting VanPatten’s lexical preference principal (Benati, 2001, 2005; Cadierno, 1995; Farley, 2001; Marsden, 2006), corroborating the claim that an alteration in learners’ default processing strategies during input processing could have a significant impact on learning, as measured by an interpretation test (Marsden, 2006; VanPatten & Cadierno, 1993a, 1993b, and others). The improved performance in both the RA and R groups are also in line with Sanz & Morgan-Short’s (2004) claim that PI could lead to learners’ language improvement by the provision of the structured input activities plus implicit feedback rather than explicit feedback. Furthermore, the performance in the RA group on the timed GJT is compatible with the claim of previous PI studies that the structured input activities (components 2 & 3) without the explicit grammar explanation (component 1) are sufficient to bring about language gains (Benati, 2004a, 2004b; Farley, 2004 b; VanPatten & Oikkenon, 1996; Wong, 2004b). However, the results obtained from the

current study suggest that this claim may need to be further refined, given that a significant impact was observed in the RA and the R groups but not in the A group, and that no significant differences in learning gains were found between the RA and the R groups. These findings suggest that different types of PI activity may have different impacts on learners' performance in an interpretation test. The lack of learning gains in the affective-activities-only group lends support to Marsden's (2006) speculation regarding the role and effectiveness of affective activities. Even though the participants in the RA group received more exemplars of the targeted feature than those in the R group, they did not outperform the R group. This implies that Wong's (2004a) claim that affective activities serve to reinforce the FMCs which occur during the referential activities is not supported.

## **6.2 Discussion of the findings of the gap-fill test**

### ***6.2.1 The relative impact of the intervention on the gap-fill test***

Based on the results displayed in Section 4.1.2.3, the resultant relative effectiveness of the interventions found by comparing the mean scores on the gap-fill test were as follows:

- a) at the post-test:  $RA = R > A = C$
- b) at the delayed post-test:  $RA = R > A = C$

The results obtained from the gap-fill test concerning the impact of the interventions are similar to those of the timed GJT. These results suggest that the interventions in the RA group and the R group led to a significant contribution to participants' performance in the gap-fill test with respect to the learning of the English '-ed' feature, given that significant differences were observed between the pre-test and the post-tests.

Furthermore, the fact that no difference was observed between the post-test and the

delayed post-test in either the RA group or the R group implies that the learning gains of both groups were upheld six weeks after the completion of the intervention. Although the performance of the R group on the gap-fill test at the post-tests was significantly related to participants' English learning length (see Section 5.1.2) and extra English exposure (see Section 5.1.3), the results from the ANCOVA (see Sections 5.2.1.1, 5.2.2.2, & 5.2.2.3) suggest that these confounding factors did not affect the interpretation of the improved performance in the R group according to Friedman's test. The R group significantly outperformed the control group when the effect of these confounding variables was controlled. Overall, the findings suggest that the interventions of both the RA and the R groups were conducive to learning the '-ed' feature as measured by the gap-fill test. On the other hand, no significant differences were observed in the A group and the control group from the pre-test to the post-tests, suggesting that the effect of the affective activities to produce the English inflection '-ed' feature in the gap-fill test was marginal.

In terms of the magnitude of the interventions, the comparative effect sizes of the interventions were as follows (see Section 4.1.2.4):

- a) comparison with the control group:
  - i) at the post-test:  $RA(1.43) > R(1.35) > A(.36)$
  - ii) at the delayed post-test:  $RA(1.51) > R(1.34) > A(.50)$
- b) comparison with the pre-test score:
  - i) at the post-test:  $RA(1.36) > R(1.32) > A(.36)$
  - ii) at the delayed post-test:  $RA(1.44) > R(1.32) > A(.50)$

Note that none of the participants in the control group scored in the gap-fill test at either

of the post-tests, so the effect size of the control group was not available. The RA group yielded the largest effect size and the A group produced the smallest effect size in comparison either with the control group or with the pre-test score. Both the effect sizes observed in the RA and the R groups are considered to be a large effect, and those of the A group are regarded as a small effect. In addition, the strength of the effect size in the RA and the R groups was greater than the mean effect size of the constrained constructed response measures (mean  $d=1.20$ ) reported by Norris & Ortega (2000, p.471), suggesting that the intervention of both the RA and the R groups was rather effective.

In addition, the results of meta-analysis on effect size of prior PI studies reveal that the structured input activities (the SIA) only groups in previous PI studies, on average, had more instructional impact on the written production test at the post-test than the interventions of this study (see Section 4.1.2.4, Tables 4.13 & 4.14). It was observed that the mean  $d$  of the SIA only groups ( $d=1.78$ ) was larger than those of the current study ( $d=1.36$  in the RA group;  $d=1.32$  in the R group; and  $d=.36$  in the A group). However, there are some reasons which could possibly explain the smaller effect size produced in the current study when compared with the findings of prior PI studies.

First, all of the post-tests in the pooled previous PI studies were administered immediately after the completion of the intervention. However, the gap-fill post-test of the current study was conducted two weeks after the intervention had been completed. This two-week interval could have attenuated the language retention to some extent. The second reason concerns the intensity of intervention. The instructional duration of pooled previous PI studies ranged from a few days and was accomplished within a week. However, the instructional duration of the current study lasted for two weeks

(four sessions in total, made up of two sessions in each of two consecutive weeks). The less intensive intervention may have produced a smaller effect size than the more intensive intervention. The fact of more delayed administration of the post-test coupled with the less intensive intervention may account for the smaller effect size observed in the current study.

In addition, the effect sizes on the gap-fill test produced by the RA and the R groups were considered to be a large effect. Although the effect size observed in both the RA and the R groups at the post-test was smaller than the mean effect size of pooled previous PI studies, the effect sizes of both groups were larger than that of four out of the six prior PI studies. This finding suggests that the intervention of the RA and the R groups produced a desirable effect size in the gap-fill tests compared with prior PI studies. Furthermore, the make-up of the SIA in previous studies was similar to that of the RA group (i.e., the PI component 2 *plus* component 3), but was not similar to the R group (only component 2). This finding suggests that referential activities alone can produce an equally beneficial effect as SIA (referential + affective) in terms of learning the ‘-ed’ feature on a gap-fill test.

### ***6.2.2 Linkage of the results of the gap-fill test to previous PI studies***

The results of the gap-fill test provide empirical evidence supporting VanPatten’s lexical preference principal. Also, the findings add credence to the claims of prior PI studies that changing the default processing strategies employed by learners during input processing leads to their language improvement in a written production test, even though the intervention was only input-based – no production was required (Benati, 2001; Farley, 2004a; VanPatten & Cadierno, 1993a, 1993b; VanPatten & Sanz, 1995, and others). Furthermore, the RA group’s performance in the gap-fill test is in line with

the findings of previous PI studies regarding the favourable role of the SIA for the effectiveness of PI (Benati, 2004a, 2004b; Farley, 2004b; Wong, 2004b; VanPatten & Oikkenon, 1996) due to the fact that no explicit grammar explanation was given to the participants in the current study. As Sanz & Morgan-Short (2004) observed, explicit information may not play an essential part in a task with task-essentialness traits, such as PI activities. However, a significant instructional impact was found in the RA and the R groups instead of in the A group, suggesting that different types of PI activity have different instructional impacts on learners' performance in a written production test. Consequently, the findings in the current study suggest that the claim in prior PI studies that the SIA was the main causative factor for the effectiveness of PI now requires refining to refer specifically to referential activities. Furthermore, the RA group did not score statistically higher than the R group, suggesting that the claim about the reinforcement of FMCs in affective activities is not sustained.

It should be noted that although the participants in both the RA and the R groups, statistically speaking, improved significantly from the pre-test to the post-tests, the instructional effect did not amount to big learning gains. Table 4.8 shows that learners' correct insertions increased on average from 0 to 2 out of 8. In addition, this test only required participants to fill a word in a gap and only 8 targeted test items were used. Thus, it is difficult to extrapolate that the same amount of learning gains would be observed if learners had been required to produce English in a less controlled written test or more test items had been employed.

### **6.3 Discussion of the findings of the oral tests**

#### ***6.3.1 The impact of interventions on the picture-based narration***

The results of the Friedman's test reveal that no significant oral improvement was

observed from the pre-test to the post-tests in any of the groups. Statistically speaking, this suggests that none of the interventions led to significant improvement of the learners' oral production of the '-ed' feature over time. However, the non-significant results might be due to the small sample pooled for the oral test. On the other hand, the results obtained from the computation of the effect size appear to suggest some instructional impact in the RA and the R groups at the post-test and the RA group at the delayed post-test (see Section 4.1.3.4) when compared with the control group and with the pre-test scores. The relative effect sizes of the interventions on the picture-based narration test were as follows:

a) comparison with the control group:

i) at the post-test: RA (.85) > R (.57) > A (-.78)

ii) at the delayed post-test : RA (.65) > R (-.65) = A (-.65)

b) comparison with the pre-test scores:

i) at the post-test: RA (.65) > R (.57) > C (.56) > A (-.23)

ii) at the delayed post-test: RA (.77) > C (.41) > A (-.53) > R (-1.08)

It should be noted that “effect sizes can be interpreted without the use of statistical significance tests” (Norris & Ortega, 2000, p.427). Even though Friedman's test did not reveal any post-instructional improvement from the pre-test to the post-tests, the effect size calculated by contrasting with the control group appear to suggest that the RA and the R groups' interventions improved the participants' oral performance in the picture narration test. The RA group produced a large effect size ( $d=.85$ ) at the post-test, and a medium effect size ( $d=.65$ ) at the delayed post-test. The R group yielded a medium effect size ( $d=.57$ ) at the post-test. On the other hand, no effect size was observed in the A group.



However, when it comes to the magnitude of change from the pre-test to the post-tests by taking the control group into consideration, the instructional impact of the intervention was not clear. Although medium effect size was observed in both the RA ( $d=.65$ ) and the R ( $d=.57$ ) groups at the post-test, medium effect was also found in the control group ( $d=.56$ ). For the delayed post-test, a medium effect was found in the RA ( $d=.77$ ) group and a small effect size was found in the control group ( $d=.41$ ). The effect size of the R and the A groups was negligible. The effect sizes observed in the control group suggest that the effect sizes observed in the RA and the R groups at the post-test(s) were not reliable. As a result, the findings obtained from the effect size are not convincing enough to justify a claim that the intervention made a contribution to the learners' oral production of the English '-ed' feature in a picture-based narration test.

With respect to the relative instructional impact of this study compared with previous PI studies, although the mean effect size of pooled previous PI studies at the post-test ( $d=1.29$ ) was greater than all of the effect sizes produced on the picture-based narration test in this study, the effect size of the delayed post-test in the RA group ( $d=.77$ ) was larger than the average effect sizes found in previous PI studies (mean  $d=.56$ ).

Furthermore, the medium effect sizes of the post-test in the RA ( $d=.65$ ) and the R ( $d=.57$ ) groups were greater than three out of the five effect sizes reported in Table 4.20 (i.e., VanPatten & Sanz ( $d=.46$ ); Marsden ( $d=.26$  at school 2); Erlam ( $d=.44$ )). Note that the participants in these three PI studies received the full PI package (i.e. the explicit grammar explanation plus structured input activities), and the participants in the current study only received a partial PI package.

### ***6.3.2 Discussion of the impact of interventions on the structured conversation***

As the structured conversation was less controlled than the picture-based narration test

in some ways, it was assumed that it would be more likely to elicit participants' implicit knowledge than the picture-based narration test. According to the mean percentages of the targeted '-ed' feature produced, the participants' performances on the structured conversation were very disappointing, given that none of the instructional groups showed significant improvement on this test (see Section 4.1.4, Table 4.21). However, was the shyness of the participants a confounding variable which affected their performance across the different instructional groups? By examination of the verb stems produced between the groups, this speculation was excluded. The Kruskal-Wallis test showed that there was no significant difference in the verb stems produced during the structured conversation across the instructional groups at the post-test,  $H(2) = 1.551$ ,  $p = .461 > .05$ , and at the delayed post-test,  $H(2) = .032$ ,  $p = .984 > .05$ <sup>83</sup>. Note that although some participants did produce the '-ed' in this test, their performance was not convincing enough to claim any significant impact of the interventions.

### ***6.3.3 Linkage of the results of the oral tests to previous PI studies***

With respect to the oral performance showed by the current study, the results (obtained either from the picture narration test or from the structured conversation) suggest that the learners' oral performance did not significantly improve after receiving the interventions. The non-significant oral performance was also reported in VanPatten & Sanz's (1995) and Marsden's (2006) studies. VanPatten & Sanz found that the PI group did not significantly outperform the control group in the structured conversation. Marsden found statistically significant positive effects of PI on an oral narration of a picture story and a semi-structured conversation in one experiment. However, in a similar experiment in a different school, Marsden found that PI had no beneficial

---

<sup>83</sup> The results of the parametric test, namely the one-way ANOVA, also backed up the findings. No significant improvement was found at the post-test,  $F(2) = 1.032$ ,  $p = .371 > .05$ , or at the delayed post-test,  $F(2) = .277$ ,  $p = .761 > .05$ .

instructional impact on the same two measures compared to a control group.

Previous PI studies have claimed that PI could be conducive to improving performance on both interpretation and production tests by specifically altering learners' processing strategies. However, this claim is not always supported by looking at the results from the current and previous studies. It appears to be that supporting evidence for PI's effectiveness in the written production test is consistent, but the evidence from the oral production test is mixed. As VanPatten & Sanz (1995) commented, different assessment tests could result in significantly different test scores and "the difference depends on whether the subjects performed the tests in the written or in the oral mode" (p.183). Similarly, the oral test results from the current study suggest that claims about PI's effectiveness need some refinement. PI could improve learners' interpretation and production of a targeted linguistic feature, though at no time are PI learners involved in output practice. However, the scope of improvement on a production test may depend on the targeted feature (e.g. simple morphological features or complex syntactic structures) (R. Ellis, 2002, p.232) and the mode of production (e.g. written or oral modes) (VanPatten & Sanz, 1995; the current study).

#### ***6.3.4 Why was the impact of the interventions not so promising in the oral tests?***

There are some reasons which may explain why the impact of the interventions on the oral tests was not as significant as that on the gap-fill test and the timed GJT. First, the identification of the sound of the '-ed' feature attached to the end of a verb is subtle. Additionally, processing the aural input is more arduous than processing the visual input (Wong, 2001). Therefore, the absence of the sound representation of the '-ed' feature might have led to the participants' poor oral performance. Furthermore, the visual FMC of the '-ed' feature is less complex than the aural one. The complexity and transparency

of a given form could affect the development of an FMC (DeKeyser, 2005). DeKeyser (2005) claimed that different forms which express the same meaning could increase the opacity and the complexity of the given form, leading to difficulty in achieving FMCs. From the visual point of view, the FMC of the ‘-ed’ feature merely requires the participants to map the meaning of pastness with a single form, and the timed GJT and the gap-fill test were administered in the visual format. On the other hand, the aural FMC of the ‘-ed’ form involves three allophones (cooked /t/, played /d/, and visited /-ɪd/), which means that participants had to connect the meaning of pastness to three forms. Moreover, the participants had never been explicitly instructed in the phonetic differences during the instructional phases. In this case, they might have experienced difficulty in identifying and then establishing the FMC of the ‘-ed’ feature in the aural mode. The failed or incomplete development of the aural FMC may have impeded their oral performance and this probably partially accounts for why a significant improvement was observed in the timed GJT and the gap-fill test, but not in the oral test.

Second, as far as the phonology is concerned, participants in the current study might have experienced difficulty in physically producing the targeted ‘-ed’ feature, as it always involves the final consonant (e.g., played /ple:d/) or final consonant clusters (e.g., cooked /cukt/). In general, final consonants and final consonant clusters are troublesome for L1 Chinese learners of L2 English as there are few final consonants in Chinese (Chang, 2001). Two participants (one in the RA group and the other in the R group) participating in the interview immediately after the oral tests actually mentioned the difficulty of producing it, stating “it was difficult to pronounce it” or “I knew I should use it, but I did not know how to pronounce it”.

However, this possible explanation was excluded, given that learners, occasionally, did demonstrate that they were capable of producing the final consonant, or consonant clusters for regular past inflections. Furthermore, they produced final consonant clusters elsewhere in their oral production, such as the plural –s and 3<sup>rd</sup> person singular –s inflections. For example, the three instructional groups produced 27 word final consonant clusters, which were not the ‘-ed’ feature, at the post-test, and they produced 29 final consonant clusters at the delayed post-test. According to the Kruskal-Wallis test, no significant difference was observed between groups in producing the final consonant clusters at the post-test ( $H(2)=3.091$ ,  $P=.213$ ) or at the delayed post-test ( $H(2)=2.305$ ,  $P=.316$ ). No significant correlations between learners’ productions of final consonant clusters and their oral merged test scores were found at the post-test (Spearman’s  $r$  for the RA group=.393, R group=.234, A group=.000) and at the delayed post-test (RA group=.089, R group=-.254, A group=.000). These results appear to suggest that the lack of gains in oral production were not related solely to speech production mechanisms.

Last but not least, the non-significant oral performance observed in the current study might be due to the fact that the interventions did not make a contribution to promoting learners’ implicit knowledge, which can be accessed readily with respect to the targeted feature during oral interaction without monitoring language production. (Although a significant improvement was found in the timed GJT, which is also a timed test, the issue regarding whether the timed GJT used in the current study did elicit participants’ implicit knowledge will be discussed in Section 6.5.2.)

## **6.4 Discussion of the findings from the vocabulary test**

### ***6.4.1 The relative impact of the interventions on the vocabulary test***

Although Friedman's test did detect significant improvement in the RA group from the pre-test to the post-tests, and in the A group from pre-test to the post-test (see Section 4.1.5.3), two confounding variables (the English learning length and the extra English exposure) were observed to be associated with the performance of these two groups (see Sections 5.1.2 & 5.1.3). After adjusting the two confounding variables by the ANCOVA, no significant difference was observed in any of the groups (see Sections 5.2.1.2, 5.2.1.3, & 5.2.2.4). Thus, the relative effectiveness of the interventions on the vocabulary test, whilst controlling for the confounding variable, either length of English learning or extra English exposure, was as follows:

- a) at the post-test: RA= R= A
- b) at the delayed post-test: RA= R= A

These results reveal that none of the interventions made any significant contribution towards the participants acquiring the vocabulary. Note that the main pedagogical purpose of PI is to assist learners in learning the grammar rather than vocabulary. The administration of the vocabulary test was to explore whether any specific intervention is more favourable for vocabulary learning than another. The pedagogical nature of the interventions may account for why the negligible impact on vocabulary learning was observed.

As far as the effect size is concerned, the comparative magnitude of the instructional effect sizes was as follows:

- a) pre-test vs post-test: RA(.77) > A(.50) > R(.19)
- b) pre-test vs delayed post-test: RA(.76) > A(.20) > R(.18)

Although the results of the ANCOVA on the vocabulary test did not show any instructional impact over time in any of the groups, the results of effect size by

contrasting the scores at the pre-test and the post-tests demonstrated some instructional impact, given that medium effect sizes were observed in the RA group at the post-test ( $d=.77$ ) and the delayed post-test ( $d=.76$ ), and in the A group at the post-test ( $d=.50$ ). The effect sizes of the R group were negligible, as they were all smaller than the ‘small effect size’ ( $.2 < d < .5$ ). These results appear to suggest that the intervention in the RA and the A groups was more advantageous for learning vocabulary at the post-test, than that in the R group, although learning vocabulary was not the focus of the intervention. Although the intervention of the RA group had the greatest impact on vocabulary learning among the three interventions, this could simply be because the RA group received more lexical exemplars of the words and more glosses of the words tested than those in the A and the R groups (see Tables 3.3 & 3.4 in Chapter 3). This could possibly explain the largest effect size observed in the RA group. However, the R group was exposed to more words and to more glosses of word tested than those of the A group during the instructional period, yet the R group had a negligible effect size and the A group had a medium effect size. This implies that affective activities were more beneficial for learning vocabulary than referential activities in terms of these effect sizes.

However, there were several factors which made differences between groups difficult to test and the results obtained from this study difficult to interpret. For example, a control group was not included for the vocabulary test. There were complex relations between the intervention materials and the vocabulary test (e.g. some groups had greater exposure to the words tested and to glosses of the words tested). Although the two confounding variables (i.e. English learning length (ELL) and the extra exposure to English (EEE)) were found to positively correlate with some test scores, no differences in reported ELL or EEE were found between the groups. Consequently, this aspect of

the research clearly requires further exploration.

#### ***6.4.2 Linkage of the results of the vocabulary test to previous PI studies***

Although Friedman's test showed that the participants in both the RA and the A groups made a significant improvement from the pre-test to the post-test, the results of the ANCOVA did not detect any instructional impacts. However, the effect size appeared to suggest some instructional impact on the A group. A medium effect size ( $d=.50$ ) was observed in the A group from the pre-test to the post-test, and a negligible effect size was found in the R group ( $d=.19$  at the post-test, and  $d=.18$  at the delayed post-test). Note that the participants in the A group received fewer lexical exemplars of the items that were tested.

So, the effect sizes and the tests of statistical significance without considering potentially extraneous variables both seem to support to a certain extent Marsden's (2004, 2006) speculation that affective activities may be more favourable for learning vocabulary than referential activities, at least in the short term. The negligible effect sizes observed in the R group suggest that the referential activities did not focus learners' attention on to the meaning of the whole sentence, somewhat out of line with PI proponents' claim that PI is a meaning-based approach to grammar pedagogy (VanPatten, 1996; Wong, 2004a). One participant in the R group spoke to the researcher after the completion of the post-test and commented that she had not paid too much attention to vocabulary during the instructional phases because there was no need for her to grasp the meaning of the text. All she needed to do was find out whether the '-ed' appeared. In this sense, referential activities may contradict the guidelines of the creation of PI activities: namely, no mechanical or non-meaningful activities (VanPatten, 1996; Wong, 2004a). However, the pattern of results in the effect sizes is



not borne out by tests of statistical significance once potentially extraneous variables are taken into account. Also, there were some factors which made the investigation and interpretation difficult as discussed above (e.g. the absence of the control group, the different amount of exposure to the word tested between groups, and so on). Thus, the interpretation of the findings from the vocabulary test is suggestive rather than conclusive.

## **6.5 Discussion of the issues regarding implicit and explicit knowledge derived from the PI activities in this study**

### ***6.5.1 Discussion of the results of the elicitation tests***

#### *6.5.1.1 The results of the principal component analysis*

According to the results produced by Principal Component Analysis (PCA), a two-component solution was specified in the RA and the R groups at the post-test (see Section 4.2.1), and in the RA and the A groups at the delayed post-test (see Section 4.2.2). Note that the PCA could not be performed for the A group at the post-test because no-one scored on the gap-fill test or the structured conversation. Nor could it be conducted for the R group at the delayed post-test, because no one scored on the structured conversation.

On the whole, the PCA results suggest that the timed GJT and the gap-fill test elicited the same construct, and the picture narration test and the structured interview elicited another. As the gap-fill test was carried out free of time pressure and the oral test has been suggested to be one technique of tapping implicit knowledge of a language (R. Ellis, 2005; Roehr, 2008), the PCA results suggest that the gap-fill test and the timed GJT used in the current study tended to draw on participants' explicit knowledge; on the other hand, the oral tests probably tended to draw on implicit knowledge. Possible

reasons why the timed GJT in this study did not tend to elicit implicit knowledge will be discussed in Section 6.5.2.

#### *6.5.1.2 The results of the participants' self-reports following the timed GJT and oral tests*

The results presented in Section 4.2.3.1 with respect to the post-task questionnaire following the timed GJT in fact suggest that the timed GJT drew on participants' explicit knowledge. A significant association was observed in both the RA and the R groups between scores and reported rule-use at the post-tests (RA:  $r_b = .794$  at post-test, and  $r_b = .754$  at delayed post-test; R:  $r_b = .818$  at post-test,  $r_b = .840$  at delayed post-test). The significant positive correlation suggests that the higher a participant scored in the timed GJT, the more possible s/he reported thinking of the targeted grammatical rule to do the test. Note that no rules were provided explicitly during the intervention and such explicit knowledge must have been induced from the referential activities and the feedback given therein. In this sense, it appears that the timed GJT *tended* to tap into participants' explicit knowledge rather than implicit knowledge, and learners had induced their explicit knowledge during the course of referential activities. It is noted that this finding should not be interpreted as the timed GJT *entirely* eliciting explicit knowledge, given that whether or not some implicit knowledge was elicited during this test is not clear, though this is possible.

In terms of participants' self-reports following the oral tests, a significant positive association was found between the merged test scores and the self-reports in the R group at the post-test ( $r_{pb} = .882$ ), and in the RA group at the delayed post-test ( $r_{pb} = .656$ ), suggesting that the higher a participant achieved in the oral tests, the more

likely s/he reported thinking of the targeted grammatical rule during the test (see Section 4.2.3.2). No significant correlation was observed in the RA group at the post-test, or in the R and the A groups at the post-test. However, the non-significant associations observed should not be interpreted as evidence for the claim that the participants drew on their implicit knowledge to do the tests, given that scores in their oral performance were very low. That is, if the participants had shown a significant improvement on the oral tests, and no significant association was found between their merged test scores and the self-reports, this would be evidence to claim that the interventions had promoted participants' implicit knowledge. However, no convincing evidence for the promotion of implicit knowledge was acquired in this study.

It is acknowledged that the validity of the self-report technique used in the current was threatened because the participants were not required to identify which oral tests (either the picture-narration test or the structured conversation) they were recalling. The reason for not conducting two separate self-reports respectively following the two oral tests was to avoid raising participants' awareness of using the rule during the structured conversation. In any case, this issue (the validity of the self-reports in terms of which task participants referred to) would only have been seriously problematic if the participants' performance in the two oral tests had been significantly different. In fact, no significant improvement was observed on either of the oral tests in any of the groups.

Furthermore, the application of the post-task self-report has come under criticism for the subjects' forgetfulness and their difficulty in verbalising a grammatical rule (R. Ellis, 2004, 2005; Bialystock, 1979). However, these reservations do not greatly alter the argument that the timed GJT drew on some explicit knowledge, because they suggest that the self-report gave a *conservative* indication about awareness, thus potentially

underestimating the strength of the relationship between the GJT scores and rule-use. Put another way, the self-reports used in the current study are unlikely to have overestimated use of explicit knowledge.

It is also acknowledged, again, that measuring implicit and explicit knowledge is probabilistic because “learners are likely to draw on whatever resources they have at their disposal irrespective of which resources are the ones suited to the task at hand” (R. Ellis, 2005, p.153). The following sections discuss why the timed GJT did not seem to draw on participants’ implicit knowledge, and to examine whether or not the oral tests in this study drew on implicit knowledge, before coming to a conclusion concerning the type of knowledge derived from PI activities.

### ***6.5.2 Why the timed GJT in this study failed to elicit implicit knowledge***

Contrary to the expectation that the timed GJT would elicit implicit knowledge as suggested by R. Ellis (2005), the PCA indicated that the timed GJT used in this study drew on participants’ explicit knowledge instead of implicit knowledge, assuming that the gap-fill test was assumed to draw on more explicit knowledge. A comparative inspection of the design and operationalisation of the timed GJT between the current study and Ellis’ study produced some explanations for this unexpected finding.

First, in Ellis’ study no intervention was delivered to the participants before or after the administration of the timed GJT. The participants only took part in the timed GJT once. On the other hand, the participants in the current study received intervention instructing a targeted feature, and they undertook the timed GJT three times in total (a pre-test and two post-tests). Second, Ellis used seventeen targeted linguistic features in his timed GJT, whereas only one linguistic feature was involved in the timed GJT of this study,

though distractors were employed. Presumably, it is easier for a participant to develop an awareness of a single targeted feature than seventeen targeted features in a test, especially when this targeted feature is instructed before and after the delivery of the test. The participant is more inclined to use the targeted linguistic feature during the testing phase if s/he knows what the targeted feature is in a test.

In addition, the operationalisation of responding to the test items might account for the discrepancy. In Ellis' study, the participants read a test item on the computer screen and they were required to press a response button within a fixed pre-estimated time, and each test item had a dichotomous option ('correct' or 'incorrect'). On the other hand, the participants in this study read the test items on the computer screen as well, but they responded to them by circling one of five options on the answer sheet within a fixed pre-estimated time. The greater number of response options (five options as opposed to two options), and the more delayed response approach (circling on a sheet of paper as opposed to pressing a response button) in this study might have increased participants' use of explicit knowledge.

Although several findings suggest that the timed GJT drew on learners' explicit knowledge (i.e. the results of the PCA, the correlations between the timed GJT and the self-reports and the correlations between the gapfill and the timed GJT), it is not certain that the participants used *solely* their explicit knowledge during these tests. The findings from the current study simply suggest tendencies and suggest that the knowledge promoted by referential activities at least partially consisted of explicit knowledge.

### **6.5.3 Did the oral tests tap into implicit knowledge?**

As mentioned above, the oral tests were designed to elicit implicit knowledge, but it is

impossible to construct a test purely to measure the two types of knowledge (R. Ellis, 2005, p.153; Roehr, 2008, p.191). In order to examine to what extent the oral tests constrained the use of explicit knowledge, Ellis (2002) suggested examining this issue by scrutinising the “quality of learners’ free production” (i.e. the occurrence of reformulation), given that the occurrence of reformulating a statement signifies the use of explicit knowledge. The number of participants who reformulated their oral output<sup>84</sup> in each group is presented in Table 6.1. Note that the number of reformulations was the same as the number of participants who reformulated, because each participant produced one reformulation.

Table 6.1

*The number of occurrences of participants’ reformulation in the oral tests*

		Picture-based narration		Structured conversation	
Group	N	Post-test	Delayed Pt	Post-test	Delayed Pt
RA	10	1	1	0	0
R	9	4	0	0	0
A	9	1	0	0	0
C	9	0	0	0	0
<b>Total</b>	37	6	1	0	0

Table 6.1 shows that reformulation only occurred in the picture-based narration instead of the structured conversation. This finding appears to suggest that the picture-based narration test did not entirely prevent participants from monitoring their oral output. The fact that no reformulation was observed in the structured conversation seems to suggest that it served better than the picture-based narration test to prevent the participants’ from monitoring their language. In addition, significant positive association between

<sup>84</sup> Only participants who reformulated the targeted feature ‘-ed’ were counted in Table 6.1. For example, a participant produced ‘walk to school’, but self-corrected it ‘walked to school’. Any reformulation of the statement irrelevant to the target feature was discarded in Table 6.1.

participants' self-reports and the merged test scores was observed in the R group at the post-test ( $r_{pb} = .882$ ), and in the RA group at the delayed post-test ( $r_{pb} = .656$ ) (see Section 4.2.3.2). It would therefore be cavalier to associate too directly the improved performance on the oral tests with implicit knowledge, especially the picture-based narration test.

Some reasons may account for why the picture narration test drew on learners' explicit knowledge to some extent. The picture-based narration test was more controlled than a spontaneous test, which would have had fewer constraints. This control may have enabled learners' oral performance to be monitored. Also, learners started the picture narration test by reading out the temporal adverbial 'yesterday' or 'last night' on the first page, in case they did not know that they were going to describe something which had happened in the past. It is possible that the procedure of 'reading out the past adverbial' might have increased the likelihood of the use of explicit knowledge.

To sum up, the findings suggest that the oral tests used in the current study did not *purely* elicit implicit knowledge based on the occurrence of reformulations (Table 6.1) and the significant associations observed (Table 4.54). However, these oral tests definitely constrained the use of explicit knowledge to some extent, given that the participants' performance was different from (worse than) that of the timed GJT and the gap-fill test.

#### ***6.5.4 What type of knowledge is derived from the different interventions?***

According to the results obtained from the significance test and the effect size, both the RA and the R groups demonstrated language improvement from the pre-test to the post-

tests in the timed GJT and the gap-fill test, suggesting that structured input activities (the RA group) and referential activities only (the R group) promoted the development of explicit knowledge of the English ‘-ed’ feature. As for the impact of the intervention of the RA and the R groups on the development of implicit knowledge, neither of the groups demonstrated it either in the structured conversation or in the picture-narration test based on the results of the significance test. Although the result of effect sizes shows some improvement in the RA group in the picture-narration test, it is not convincing enough to claim that the intervention of the RA group promoted participants’ implicit knowledge of the ‘-ed’ feature for the reasons discussed above. With respect to the A group, the participants did not significantly improve over time in any of the achievement tests, suggesting that affective activities alone, as employed in this study, did not significantly promote learners’ development of knowledge of the English ‘-ed’ feature.

These findings are in line with VanPatten & Williams’ (2007) claim that “certain morphemes that ranked low in the natural order tended to rise in rank when learners were able to monitor their production. These morphemes, such as 3<sup>rd</sup> person singular –s and *regular past tense*, were the morphemes that were more easily learned but not so easily acquired” (p.30).

#### ***6.5.5 Linkage of the results to previous PI studies***

In brief, the structured input activities and referential activities alone operationalised in this study encouraged the development of explicit knowledge. None of the interventions (RA, R, or A) appeared to significantly promote the development of implicit knowledge according to the measures used in the current study. Overall, the findings obtained from the current study concerning the effect of PI activities on implicit and explicit



knowledge lend support to de Jong's (2005) speculation "the participants in the structured-input-only group spontaneously induced explicit knowledge" (p.210), suggesting that PI activities do promote explicit knowledge. In addition, the findings substantiate DeKeyser *et al's* (2002, p.813) speculation that the PI group under the instructional condition that the explicit grammar rule is not given is an "explicit inductive group", as the learners could figure out the targeted feature by means of the yes/no feedback constantly offered to them. On the other hand, VanPatten's proposition that PI activities could assist in "building up an implicit knowledge of the language via intake facilitation (1994, p.34)" was not borne out in this study, given that the evidence for the development of implicit knowledge is not clear.

However, we cannot confidently conclude that PI can not or does not favour the development of implicit knowledge. Crucially, the little evidence on implicit knowledge in the current study might be due to the short duration of the interventions (about 2.5 hours). As Ellis (2002) stressed, "implicit knowledge can only be revealed to the learner through substantial and repeated experiences with input ... Given such an account, it is not easy to see how a few hours, several days, or perhaps even a number of weeks of FFI directed at some specific grammatical property can ensure that learners develop implicit knowledge of this feature" (p.224).

## **6.6 Discussion of the relative effectiveness of interventions for the '-ed' feature**

In terms of the effectiveness of the interventions in this study, it was found that both the RA and the R groups demonstrated a language improvement in the timed GJT and the gap-fill test from the pre-test to the post-tests. The performance of participants in the RA group and the R group did not significantly differ in the tests at the post-tests, suggesting that the intervention of R group exerted the same impact as that of the RA

group. On the other hand, the A group did not make any language improvement over time. Consequently, the findings observed in the current study indicate that referential activity is the causative factor in the effectiveness of PI, corresponding to Marsden's speculation (2006). The next section sets out to discuss in more depth why the referential activities led to greater language learning gains than the affective activities.

### ***6.6.1 From the perspective of FMCs***

As noted in Chapter 2, devising language activities to promote better FMCs is at the heart of PI. VanPatten (2002) also emphasised that task-essentialness is beneficial for grammar learning when constructing grammar tasks (see Loschky & Bley-Vroman, 1993), although task-essentialness is not restricted to input-based tasks as PI is. The vital feature of attaining task-essentialness is to require the learners to use the targeted grammatical form to achieve the tasks. However, as noted in chapter 2, only referential activities adhere to the notion of task-essentialness and FMCs (i.e. learners are pushed to interpret the targeted '-ed' feature in order to complete the task successfully), whereas affective activities do not force learners to process an FMC in order to complete the task. Learners could complete the affective activities without noticing or processing the '-ed' feature from the beginning to the end of the intervention. As the formation of an FMC has been regarded as an essential element in the effectiveness of PI, the failure to push the establishment of an FMC in the affective activities could explain the scant learning gains in the A group. These findings are in line with the assertions of previous PI studies that pushing learners to interpret the meaning of a specific form (in other words, to achieve FMCs) is requisite for L2 grammar learning (VanPatten, 2000, 2002a and elsewhere).

Furthermore, as no explicit grammar explanation was provided to participants in the

current study, the results obtained from the RA and the R groups corroborate Sanz & Morgan-Short's (2004) claim that "explicit information may not necessarily facilitate second language acquisition and that exposing learners to task-essential practice is sufficient to promote acquisition" (p.36) (though the current findings suggest that although explicit information may not be provided, learners can induce it).

In addition, the lexical preference principle in VanPatten's IP predicts that learners tend to seek out temporal adverbials to obtain the meaning of pastness, as opposed to seeking grammatical form. In this sense, it has been suggested that past tense temporal adverbials should be removed in PI activities in order to push learners to make better FMCs. However, Harrington (2004) pointed out that the grammatical form and the redundant lexical item may complement one another. Furthermore, Batstone (2002) argued that co-textual cues (i.e. lexical or linguistic cues within the input itself such as the temporal adverbial verbs) could be used as an 'anchor' at times in conjunction with the development of FMCs. Batstone's argument indicates the possible problem with affective activities demonstrated in this study, given that affective activities merely provide the con-textual cues rather than the co-textual cues<sup>85</sup> to signify the pastness. Here I refer to the *initial* intake of a *new* linguistic feature. An FMC is unlikely to be established by learners being exposed to the target form without any other cues in the input to indicate its meaning or function. This is perhaps even less likely when an explicit grammar explanation is not provided.

Taking one training item of referential activities as an example (" I walked to school

---

<sup>85</sup> 'Con-textual cues' refer to cues from a situation in which an utterance is heard or seen. For example, a learner hears an utterance from a man with a map on a street, and he/she comes to understand that he is asking for directions. 'Co-textual cues' refer to the meaning of lexical or linguistic cues within the input itself.

\_\_\_\_\_”), learners were required to choose one of two options: a) yesterday; b) tomorrow. The temporal adverbial was not given in the main sentence but appeared in the two corresponding options. The meaning of past-ness could be anchored by the appearance of temporal adverbials in corresponding options in referential activities. The feedback indicated whether a participant’s choice is right or wrong, although learners were not told why. VanPatten (2002b, p.254) suggested that this process corresponds to Batstone’s con-textual cues (i.e., the feedback provided). In this sense, both the co-textual (i.e., the temporal adverbials) and con-textual (i.e., the feedback provided) cues were given in referential activities.

On the other hand, no temporal adverbial (i.e., the co-textual cue) was provided in the affective activities except for the introduction of an affective activity (e.g., “did you do the same things *last night?*”). The learners had to choose from options such as a) listened to music; b) watched the telly and so on. The learners were directed to express their own opinions or feelings. Note that no explicit grammar explanation of the ‘-ed’ feature was given to the participants in the current study. They had no notion of what the past tense did and what it looked like. In the absence of any co-textual cues (i.e., the past-tense adverbials) and explicit information to indicate the pastness, the participants were unlikely to make FMCs, and subsequent processing, strengthening and accessing of the target form was unlikely to happen.

### ***6.6.2 From the perspective of attention***

Note that when a participant self-reported that s/he had been thinking of a grammatical rule during the testing phases, s/he was also required to provide the grammatical rules that had been used or to give examples. Only those who provided the correct grammatical rule or examples of the ‘-ed’ feature were categorised into the ‘rule-use’

group. VanPatten (2002b, 2007) claimed that ‘processing a form’ is similar to ‘detecting a form’ (Tomlin & Villa, 1994) (connecting it to a meaning), but not to ‘noticing a form’ (Schmidt, 1990 and elsewhere) (simply attending to the form itself). VanPatten argues that ‘detecting a form’ is more effective at helping learners to master the grammatical rule than ‘noticing a form’ under the condition that explicit grammar explanation is not provided. Two findings in the current study suggest that the referential activities are more conducive to ‘detecting’ a form than affective activities.

The first piece of evidence emerges from the accumulated tallies of the participants’ responses to the post-task questionnaire following the timed GJT at the post-test (see Section 4.2.3, Table 4.49). The results reveal that the RA group showed the highest percentage (64% within the responses who reported using the rule) of reported rule-users; the R group exhibited 29%, and the A group showed the lowest percentage (7%). Pearson’s chi-square test indicated a significant difference between the groups in the participants’ self-reports,  $\chi^2(2) = 7.766$ ,  $p = .021 < .05$ . It is not surprising to observe that the RA group had the highest percentage, because the RA group was exposed to more training items containing the ‘-ed’ feature. However, the R and the A groups received the same number of instances of the target feature, and yet, critically, the R group showed a higher percentage of reported rule-use than the A group. This seems to suggest that the referential activities were more likely to help detect the ‘-ed’ feature than the affective activities.

The other piece of evidence is based on the significant positive correlation observed between participants’ test scores on the timed GJT and their responses to the post-task questionnaire in the RA and the R groups at the post-test, but not in the A group (see Section 4.2.3.1, Table 4.51). This positive correlation suggests that the higher the score

a participant achieved, the more likely it is that s/he could provide the grammatical rule or examples correctly illustrating the ‘-ed’ feature. The participants’ correct provision of grammatical rules or of correct examples is in line with Schmidt’s (1990, 1995) awareness at the level of understanding (e.g. the hypothesis and rule formulations) (see Section 2.2.3.2). Though rule awareness and explicit knowledge are not the same thing, the rule knowledge is one manifestation of explicit knowledge. Greater rule awareness was generated in the R group than in the A group, suggesting that referential groups were more likely to channel participants’ attention to detect the targeted form than that of the A group.

It is noted that the above discussion of the findings is suggestive, because investigating the attentional issues was not the aim of the current study. The current study was not specifically designed to operationalise the ‘detecting’ or ‘noticing’ of a form, nor was a con-current measure (e.g., on-line think aloud) included. However, based on the above discussion, I speculate that the nature of both the referential and the affective activities required participants’ attention to complete the tasks, but that different types of activity channelled their attention to different foci. The results from the current study suggest that referential activities tend to direct participants’ attention towards a specific form, and affective activities tend to direct it towards the ‘meaning’ of lexical items.

### ***6.6.3 Further explanations for the relative effectiveness of interventions***

This section attempts to elaborate how and why referential activities are more effective than affective activities at promoting learning of the ‘-ed’ feature based on the different traits observed in these PI activities.

First, the target feature is juxtaposed with a contrasting object (i.e. the targeted feature

and another similar form) and this might be favourable toward generating a hypothesis regarding the rule governing the targeted feature. In addition, the clear-cut and systematic feedback ('yes' or 'no' to indicate correctness) provided in referential activities was more likely to offer learners an opportunity to test their hypotheses (Loschky & Bley-Vroman, 1993, p.142) and establish an FMC. In brief, learners could engage in the 'failure-driven process' described in Section 2.1.1.1 of the literature review (Carroll, 1999). Furthermore, Tomasello & Herron (1989) stated that a task which effectively generates a learner's hypothesis testing and cognitive comparison would produce superior learning. Compared with referential activities, affective activities create fewer opportunities for learners to form a hypothesis regarding the meaning or function of a form due to the absence of contrasting forms. Though timely feedback is also offered in affective activities, it relates to the semantics of their response, which does not make the target form essential. This feedback cannot be used to generate or test a hypothesis regarding form.

In addition, the likelihood of the induction of the failure-driven process by means of contrasting objects and the provision of feedback in referential activities may be also a factor in the superior effect which referential activities have over affective activities. As Loschky & Bley-Vroman (1993) commented, "it is important to create chances for learners to make errors and to receive feedback on them" (p.149). VanPatten (2002a) also stressed that "the aim of PI is in line with claims of those researchers who assert that acquisition is a failure-driven process (e.g. Carroll, 1999). ... PI is designed to cause failure in interpretation at the beginning stages of activities so that the processors can begin to 'readjust'" (p.768). Furthermore, an important function of feedback in PI was also stressed by VanPatten (2002b). VanPatten stated that "This feedback is critical and is what makes the processors understand they have failed ... feedback as to

rightness or wrongness of answer selection during activities is critical since this is what lets the learner's processing mechanism know there is a failure" (p.248). Based on this argument, although PI may be in line with the failure-driven process, this is only the case for referential activities and not affective activities, since learners involved in affective activities barely experience the failure-driven process.

### **6.7 The purpose of the affective activities**

Although the findings suggest that affective activities did not contribute to learning the '-ed' feature, I do not suggest that affective activities are superfluous and should be abandoned in the framework of PI. I support the argument that affective activities may help PI to fit into the broad outline of FonF and existing communicative approaches (VanPatten, 1993, p.439; Wong, 2004a), given that the nature of referential activities is not so communicative and learner-centred. As VanPatten (1993) stated, "Work with processing instruction and structured input should not lose sight of one of the very important tenets of communicative language teaching: a focus on the learner. Thus, processing instruction includes activities that are affective in nature, e.g., activities that ask for an opinion, a personal response, tap the student's own world, and so on" (p. 439).

In addition, although affective activities did not show conclusive evidence for learning vocabulary in this study when taking some confounding variables into consideration, the hypothesis for affective activities being more favourable to the learning of vocabulary could not be rejected according to the findings of this study. For example, the fact that the A group received fewer words and glosses of words tested than those in the R group, and that the results of effect size and Friedman's test showed some instructional impact of the A group, suggest that this hypothesis is likely to be upheld.



Clearly, however, this needs further study to verify or falsify it. If this hypothesis is substantiated, this would give further credit to the pedagogical value of PI.

Finally, the specific sequence of structured input activities (referential activities followed by affective activities) may be critical (VanPatten, 1993; Marsden, 2006). On a pedagogical level, referential activities may enable instructors to ascertain whether or not the learners have established the FMCs. Marsden (2006) speculated that “during affective activities, learners use and establish, possibly implicitly, these new form-meaning connections” made in referential activities (p.549). Also, Wong (2004a, p.44) claimed that the purpose of affective activities is to reinforce the representation of a form which has been established during referential activities by providing learners with more opportunities to hear and see the targeted feature in a meaningful context. Although this possibility was not validated by the current study as the RA group did not outperform the R-only group in any of the tests, it was not entirely excluded. It is possible that the speculation about affective activities implicitly reinforce FMCs that are made in referential activities could be corroborated in a longer and/or more intensive instructional intervention.

## Chapter 7 Summary and Conclusion

### 7.1 Summary of the current study

This thesis presents a classroom-based, quasi-experimental study involving one hundred and twenty L1 Chinese learners of English from four classes at one primary school in Taiwan. The study set out to compare the different types of Structured Input Activity (SIA) in the framework of Processing Instruction (PI), namely referential activities and affective activities, and then to explore the impact of these. The participants were allocated into three instructional groups (the RA group (structured input activities), the R group (referential only activities), and the A group (affective only activities) on the basis of their pre-test scores on a timed GJT and a gap-fill test. A non-active, test-only Control group was also used. The instructional duration was about 2.5 hours over two consecutive weeks.

The targeted linguistic feature was an English verb inflection, the past tense ‘-ed’. This target feature was chosen and the instructional materials were created following Input Processing (IP) theory and the nature of PI activities. Based on IP theory, learners tend to encounter problems in processing this targeted verb inflection, given that it often occurs along with a temporal adverb indicating the same meaning of ‘pastness’.

Learners undertaking referential activities were required to use the verb inflection to complete the tasks. The affective activities were constructed in a way which was more meaning-bearing and would make learners understand the sentential meaning, but they included exemplars of the target form.

Measurements of the instructional impact took place twice after the completion of the intervention. Post-tests were carried out within two weeks after the completion of the

intervention. A delayed post-test was then administered six weeks after the learners had completed the intervention, and was finished eight weeks after the end of the intervention. Four types of measure were designed to assess learning gains: a timed Grammaticality Judgement Test (GJT), a gap-fill test, a picture-based narration test, and a structured conversation, and a written receptive vocabulary test. The mode of measurement varied, including an interpretation test (the timed GJT) and production tests (the gap-fill and two oral tests). The tests aimed to elicit explicit knowledge (the gap-fill test) and implicit knowledge (time constrained GJT and the two oral tests) and the extent to which explicit knowledge was used (post-GJT and oral task self-reports).

## **7.2 Justification and originality of the current study**

Unlike previous PI studies, which have aimed to compare the effectiveness of PI with other types of grammar instruction, the current study was the first one to truly isolate the two types of PI activity in order to investigate their individual instructional impact. By examining these two PI activities, it was hoped to untangle the issue with reference to which component was the key causative factor that contributed to the effectiveness of PI. In addition, the current study was the first to make an attempt to distinguish what resultant knowledge PI led to. Although measuring implicit and explicit knowledge has been a methodologically problematic issue in SLA research (e.g. R. Ellis, 2005), the current study has been, so far, the first amongst PI research to introduce a timed grammaticality judgment test in an attempt to draw on participants' implicit knowledge. Due to the fact that the production tests which previous PI studies have included were more controlled and primarily based on written rather than oral modes, one of the focuses in the current study was to investigate the extent to which learners receiving PI activities could perform in a more spontaneous task (i.e., the structured conversation), in which 'monitoring the language production' was less likely.

In addition, there were some distinctive characteristics which improved the internal validity of this quasi-experimental study compared with most previous PI studies. First, the internal validity of this study was strengthened by the computerised-delivery of the instructional materials and feedbacks to ensure the uniform delivery of the interventions and thus reduce the element of instructor bias. Second, a background questionnaire exploring learners' extra-curricular English learning, and an attitudinal questionnaire concerning their attitudes towards the intervention, were distributed to participants and then analysed to scrutinise whether any possible confounding variables existed which could interfere with the instructional interventions being assessed. Third, a self-report technique (the post-task questionnaire and interview) was adopted to investigate whether or not the 'implicit tests' actually drew on learners' explicit knowledge, and the self-report technique has never been used in previous PI studies. Lastly, the validity and reliability of the testing instruments appear to have been overlooked by previous SLA researchers (Douglas, 2001). Most of the previous PI studies did not justify the measures they used. Thus, the validity and reliability of the achievement tests in the current study were estimated and reported as well as the comparability of the two-version measures used for the different testing timings (see Section 3.5).

### **7.3 Findings and Discussion**

The findings are briefly summarised here in the light of the Research Question (RQ) and Hypotheses (H) posed in the second chapter (see Section 2.3.2.5).

#### ***7.3.1 RQ 1- 4 and H1***

RQ 1: Are referential activities more beneficial for learning the English past tense '-ed' feature than affective activities in a timed Grammaticality Judgement Test (GJT)?

Yes, this was confirmed by the results obtained from this study.

RQ 2: Are referential activities more beneficial for learning the English past tense ‘-ed’ feature than affective activities in a gap-fill test?

Yes, this was confirmed by the results obtained from this study

RQ 3: Are referential activities more beneficial for learning the English past tense ‘-ed’ feature than affective activities in a picture-based narration test?

This was not fully confirmed by the results obtained from this study. Based on tests of statistical significance, neither referential activities nor affective activities were beneficial for learners on this test. Although the effect size produced some tentative evidence, it was not convincing when taking the performance of the control group into account.

RQ 4: Are referential activities more beneficial for learning the English past tense ‘-ed’ feature than affective activities in a structured conversation?

No, this was not confirmed by the results obtained from this study. Based on tests of statistical significance, neither referential activities nor affective activities were beneficial for learners on this test.

H1: Referential activities are beneficial for twelve-year-old L1 Chinese learners’ interpretation and production of the English regular past tense.

### **Discussion about the findings of RQ1 to RQ4 and H1**

H1 suggests that PI referential activities are the key factor for learners' improved performance in interpretation and production tests. Based on the results obtained from the timed GJT and the gap-fill test, H1 was accepted. It was observed that learners in the R group made a marked improvement on these two tests; on the other hand, learners in the A group did not significantly improve their performance at all. Furthermore, the learners in the RA and the R groups, statistically speaking, performed equally on the GJT and gap-fill tests. If the affective activities had been beneficial for learning the target feature, the RA group would have outperformed the R group. As a result, the evidence from the current study demonstrates that the referential activities were more effective in helping the learners learn the '-ed' feature than the affective activities.

Affective activities either alone or following referential activities did not appear to exert any significant effect in either the timed GJT or the gap-fill test. Consequently, the findings that the R group yielded significant advantage over the A group and the control group in the timed GJT and the gap-fill test from the pre-test to the post-tests confirmed H1.

However, H1 was confirmed on the basis of the findings which emerged from RQ 1 and RQ 2, but not those from RQ 3 and RQ 4. Therefore, H1 was confirmed in so far as learners' ability was improved in reading (the GJT) and in writing (the gap-fill test), but not in speaking and listening, given that no convincing evidence for learners' improved performance in either the picture-narration test or the structured conversation was found, and learners' ability to interpret in listening was not assessed in this study. Note that although the referential activities appeared to have some instructional impact on the picture-narration test at the post-test according to the effect size ( $d=.57$ ), neither the findings from the statistical significance test (the Friedman's test) nor the findings from

the ANCOVA supported RQ 3. Furthermore, a medium effect size was observed in the control group at the post-test ( $d=.56$ ), suggesting that the effect size of the R group at the post-test is not robust enough to make claims about instructional impact on the picture-based narration test. As a result, the answer to RQ 3 was not confirmed due to absence of convincing evidence.

It was speculated that the targeted ‘-ed’ feature involves three allophones ( /t/, /d/, /ɪd/). These multiple phonemic versions of the target feature might pose an obstacle for learners. It was also speculated that the learners’ sparse learning gains on the oral tests could possibly be related to their difficulty in developing certain oral production mechanisms, given that the final consonant or final consonants are uncommon in Chinese (see Section 6.3.4). However, this possibility was ruled out, because the learners demonstrated the capability of producing final consonant clusters such as plural –s ‘friends’ or 3<sup>rd</sup> person singular ‘calls’, which were not the ‘-ed’ feature. This finding could suggest that the minimal oral improvement was not related to speech production mechanisms *per se*, but it might be related to a lack of readily accessible knowledge of the feature during oral production (including, possibly, a lack of reliable phonological representations of –ed).

### **7.3.2 RQ 5 and H2**

RQ 5: Are affective activities more conducive to learning vocabulary than referential activities?

The answer to this research question was not clear due to the fact that this study did not obtain enough evidence either to support or reject it.

H2: Affective activities lead to more vocabulary learning than referential activities in twelve-year-old L1 Chinese learners.

### **Discussion about the findings of RQ 5 and H2**

H2 concerns the possibility that affective activities are more beneficial for vocabulary learning than referential activities (Marsden, 2004 & 2006). According to the results obtained from the Friedman's test and the effect size, it seems that the affective activities did exert some influence on vocabulary learning at the post-test (see Section 4.1.5). The RA group and the affective-activities-only group were not only found to improve on the vocabulary test over time according to Friedman's test, but were also found to produce medium effect sizes at the post-test. On the other hand, the referential-activities-only group was not observed to have received any significant instructional impact on vocabulary learning. These findings seem to lend support to this hypothesis. However, when taking the confounding variables (learners' English learning length and extra English exposure after school) into consideration, the results of the ANCOVA did not reveal any instructional impact among the instructional groups at the post-tests (see Sections 5.2.1.2, 5.3.1.3, & 5.2.2.4). Nevertheless, the learners in the A group did receive fewer lexical exemplars, in the intervention, of the items that were tested than those in the R group. The contradictory results obtained and the imbalanced targeted lexical exemplars between instructional groups made the interpretation of the findings complicated. The answer to this research question is not clear and this hypothesis is still largely speculative.

### **7.3.3 RQ 6 and H3-5**

RQ 6: What kind of knowledge do the four tests (i.e. the timed GJT, the gap-fill test, the picture-based narration test, and the structured conversation) tap into and what is the



relationship between this knowledge and the intervention type that the learners received?

H3: The gap-fill test without a time constraint will elicit explicit knowledge and the other three tests with a time constraint will elicit implicit knowledge.

H4: Referential activities will promote learners' explicit knowledge of the English past tense '-ed' feature.

H5: Affective activities, either alone or following referential activities, will promote learners' implicit knowledge of the English past tense '-ed' feature.

#### **Discussion about RQ 6 and H3-5**

H3 to H5 were related to investigate what type of knowledge could be promoted by PI activities in order to understand whether PI could promote learners' implicit language knowledge. Note that PI did affect the language development system to some extent on the basis of the results obtained in the current study, given that learners were only engaged in input-based activities, but they were able to produce the targeted feature in a written gap-fill test, and to a lesser extent some learners could produce it in an oral production task. It is also worth noting here that the investigation of implicit and explicit knowledge in the current study should not be interpreted as implying that implicit knowledge is superior to explicit knowledge. This study was motivated to explore an unresolved issue in previous PI studies, namely the nature of knowledge derived from PI activities.

H3 was confirmed in the current study in terms of the gap-fill test tapping into explicit

knowledge. However, H3 was not confirmed in terms of the elicitation of implicit knowledge. The tests constructed to draw on learners' implicit knowledge were: 1) the timed GJT; 2) the picture-narration test; and 3) the structured conversation. The timed GJT did not convincingly elicit implicit knowledge as was expected. The results from the principal component analysis, from the correlations between the gap-fill test and the GJT, and from the correlation between the test scores on timed GJT and the self-report questionnaire tend to suggest that the timed GJT elicited learners' explicit knowledge (see Sections 4.2.1.4, 4.2.2.4, & 4.2.3.1). As for the oral tests, medium effect sizes were observed in the RA group at the post-tests in the picture-based narration test, suggesting that the structured input activities appeared to be beneficial for learners to undertake this test. Nevertheless, the occurrence of reformulations was observed in this test in all instructional groups, which made the claim of this test tapping learners' implicit knowledge untenable. Also, the significance test (Friedman's test) did not detect any significant post-instructional impact among any of the groups, suggesting that PI activities did not promote learners' implicit knowledge, if this test *was* to be interpreted as a measure of implicit knowledge. Furthermore, the control group was observed to have a medium effect size at the post-test ( $d=.56$ ) and a small effect size at the delayed post-test ( $d=.41$ ), suggesting that the effect sizes of the RA group are not reliable enough to make claims concerning improved oral performance. In addition, none of the PI activities (SIA, Referential-only, and Affective-only groups) had any observable improvement in the structured conversation, which was less controlled than the picture-narration oral test. This finding could also suggest that the PI activities did not make any notable contribution to developing a readily accessible knowledge of the '-ed' feature during oral interaction.

H4 was empirically substantiated in the current study, given that only the groups (i.e.,

the RA and the R groups) that had referential activities made significant learning gains at the post-tests, and the gains observed tended to have characteristics of explicit knowledge. Although this result is in line with findings from previous PI studies (Fernández, 2008; Marsden, 2006; Sanz & Morgan-Short, 2004; VanPatten & Oikkenon, 1996), suggesting that PI led to learning gains without the need to provide any explicit information about the target structure, some explicit knowledge was induced from the referential activities and then used during the tests, in line with DeKeyser *et al.*'s (2002) speculation.

H5 was not verified in the current study, given that no convincing evidence of developing implicit knowledge was observed in the groups that received affective activities (i.e., the RA and A groups). However, the lack of evidence in PI activities promoting implicit knowledge in this study might be due to the short instructional duration (2.5 hours over two consecutive weeks), in that the development of implicit knowledge probably requires considerable and repeated input (Ellis, 2002). It is speculated that if a longer and more intensive instructional period had been employed, the answer to whether PI activities promote learners' implicit knowledge could have been more clear-cut.

#### ***7.3.4 RQ 7 and H6***

RQ 7: Are PI learners' improved performances retained in a delayed post-test six weeks after the instruction?

Yes, this was confirmed by the results obtained from this study.

H6: The effect of PI on twelve-year-old L1 Chinese learners' ability to interpret and

produce the English regular past tense will be retained beyond the time of instruction.

### **Discussion about H6 and RQ 7**

H6 was substantiated as the RA and R groups' improved performances on the timed GJT and the gap-fill test was clearly demonstrated in the delayed post-test, which took place six weeks after the completion of the instructions. However, H6 was only substantiated in so far as learners' abilities were improved in reading and in writing, but not in speaking and listening, due to the fact that no significant improvement was observed in the oral tests in this study and an aural interpretation test was not conducted.

### **7.4 The contribution of the current study**

Some important findings have been yielded by the current study and have made a novel contribution to PI research. First, this study has produced some findings in line with the theoretical claim, the Lexical Preference Principle in the IP, on which PI is based, predicating that the alteration of learners' default processing strategies during input processing could have a notable impact on language learning. Second, the findings of the current study suggest that PI requires more refined exploration concerning the key component contributing to its effectiveness. It was corroborated in the current study that referential activities were effective in assisting students with learning the English '-ed' feature; affective activities, with or without referential activities, were not beneficial for learning this feature.

Third, given that no explicit grammatical focus explanation was provided during the instructional phases, the findings which emerged from the current study are compatible with what prior PI studies have stressed – that well-structured input (i.e., SIA) is

sufficient to result in language learning gains, and that providing explicit grammar explanation (i.e. component 1) may not be necessary. It is noted that this does not intend to devalue the role of explicit grammar explanation in all cases of instructions. Explicit grammar explanation may be necessary in some tasks or instruction, but it is not always essential in PI. By the same token, the finding of affective activities being less effective than referential activities should not be over-interpreted as suggesting that affective activities should be abandoned from the framework of PI. Each task has its own value and strength, and the nature of affective activities makes PI fit more comfortably with the Focus-on-Form and communicative language teaching approaches. In addition, affective activities may have a favourable impact on vocabulary learning, though this speculation needs more study. Besides, a reduction in instructional impact should not be equated with a complete absence of instructional impact. In effect, small effects (.2 ~.5) were obtained in the A group in both the GJT and the gap-fill test, by comparing learners' scores at the pre-test and the post-tests. This implies that affective activities did have some instructional impact, though not as substantial as those obtained from structured input activities (R+A) and referential-only activities.

This study therefore can make a significant contribution to pedagogy and research in the area of second/foreign language learning. First, the findings are prominent in terms of the relationship between SLA theory and pedagogy. This study underlines the important role of input in language learning, specifically by indicating that input activities alone without output activities can make a positive contribution to language learning. This does seem to depend, however, on what kind of input is provided to language learners. In addition, the observable instructional impact achieved by the RA and R groups instead of the A group appears to suggest the benefit of pushing learners to make FMC in second/foreign language research and pedagogy. Although the instructional duration

used for this study was considered short according to Norris & Ortega's (2000) categorisation, large effect sizes were observed in both groups in the timed GJT and the gap-fill test. The effect sizes produced in both these tests from this study were all greater than the mean effect size of similar measures reported in Norris & Ortega's meta-analysis study. Moreover, the instructional impact was maintained six weeks after the instruction had finished. As VanPatten (2002b) stated, "one role for instruction is to facilitate and speed up acquisition of formal features" (p. 254-254).

In addition, the findings from this study inform the debate as to whether or not grammar should be taught in primary schools or included in CLT. Affective activities are more in line with the CLT, which is encouraged by the Taiwanese Ministry of Education for English Education, than referential activities. Affective activities are more meaning-based than referential activities, given that they attempt to direct learners to react in a real situation and to tap their own experience. On the other hand, referential activities are focused on the meanings of specific forms, rather than at sentence level. This study found that affective activities did not significantly lead to students' learning of the targeted form in a short instructional period. More learning gains might have resulted from affective activities if a longer instructional period had been allowed – this remains an empirical question. However, one of the major tasks of instruction is to speed up learning in a limited instructional period, and in order to learn the grammar of a language in a finite class time, an activity similar to the referential activities might be an effective option for practitioners.

Finally, one of the concerns about the counterproductive effect of the implementation of English grammar instruction in primary schools in Taiwan is that it may decrease students' motivation for English learning (Lee, 2005). However, the results obtained

from the attitudinal questionnaire used in this study did not correspond to this concern. More than half (68%) of the learners participating in this study responded that they found the activities interesting. Hence, the findings of this study suggest the implementation of the CLT does not have to be at the cost of focus-on-form tasks or grammar instruction.

## **7.5 Limitations of the current study and implications for future research**

Various limitations of this study are acknowledged mainly relating to the research design and achievement assessments. The implications for future study which have arisen from the current study are addressed or discussed along with the limitations.

### ***7.5.1 Research design***

The duration of the intervention was relatively short (about 160 minutes), so that it did not meet the criteria discussed in Section 2.3.1.9 (the 3<sup>rd</sup> limitation, though it was longer than the six studies cited). Also, the sample size of the oral tests was rather small, with fewer than 15 participants as was discussed in Section 2.3.1.9 (the 5<sup>th</sup> limitation). Thus, the interpretation of the oral test results requires circumspection. Future studies need to have interventions of longer duration, and include more participants in oral tests in order to examine PI's instructional impact.

The allocation of the participants into groups was not fully randomised but was based on their scores at the pre-test. The control group was not allocated in the same way as the instructional groups. Instead, for practical reasons, the control group was chosen randomly from one of the four classes in the participating school. The method of allocating participants used in this study was to ensure an even proficiency among the groups, but it could not “produce equivalence over a whole range of variables” (Cohen,

Manion & Morrison, 2000, p. 216). It is important to note here that ‘instructional duration’ and ‘the amount of exposure to the targeted form’ in each intervention could not be controlled due to the nature of the interventions. Specifically, the RA group had more activities than the R and the A groups, and so the RA group received a greater amount of exposure to the targeted form. (Note also that this did not result in statistically superior learning gains in the RA group compared with the R group). Critically, however, the R and A groups were exposed to the same amount of targeted form as each other, which would not affect the interpretation of H1. One limitation on the research design is that the RA group only received half the amount of affective training practice as the A group. It is possible that the more affective activities would have likely increased the learning gains of the RA group, and that the RA group would have outperformed the R group. The increased number of affective activities perhaps would have changed the answer to H3 in the current study, and this speculation needs further study to substantiate it.

Furthermore, no additional follow-up test was conducted to assess learners’ retention because the participants graduated and went on to different junior high schools one month after they had finished taking part in this study. As a result, whether the instructional impact identified in this study could be maintained a few months after the completion of the intervention was not clear. Previous PI studies have substantiated that PI’s long-term effect could be observed up to 14-16 weeks (Marsden, 2006) and even eight months after the completion of the instruction (VanPatten & Fernández, 2004). Nevertheless, further study to administer a more delayed follow-up test is advisable.

In addition, although it has been advised to implement a device to keep track of learners’ performance such as the location of errors (Sanz, 2000, p. 30), response time



and accuracy (Fernández, 2008), this was not achieved in the current study due to the limited capability of the software. Consequently, it is difficult to observe whether or not learners' attention was focused on the tasks.

Next, the juxtaposition of a contrasting form in the referential activities, and the associated correct/incorrect feedback, may have been responsible for the greater learning gain observed in the RA and R groups than in the A group, given that that the juxtaposition might highlight the targeted feature. Further research can test whether or not the juxtaposition of a contrasting form would affect the effectiveness by comparing referential activities with a condition where there is no juxtaposition and learners just have to note the presence or absence of the form.

Last but not least, Schmidt (1990, 1995, 2001) claimed that noticing is enough to learn every aspect of a language. Rosa & O'Neill's<sup>86</sup> (1999) study concluded that learning at the level of noticing can lead to an improvement in the post-test on the acquisition of Spanish. However, VanPatten stressed the importance of FMCs in PI, and the establishment of FMCs appears to be at the level of understanding, given that PI learners are required to achieve tasks by interpreting the meaning of a targeted form. Therefore, one issue that remains unclear in PI is whether FMC is *necessary* to learning a grammar feature if noticing (i.e., 'noticing' a form without attaching any meaning to it) might suffice? Further research is needed to test whether attention to form only (i.e., learners have to attend to the form but not connect it with any meaning at all) is as effective as activities which require learners to make FMCs.

---

<sup>86</sup> Rosa & O'Neill explored different levels of awareness at a syntactic level (using a Spanish conditional sentence), assessed by think-aloud protocols during a problem-solving jigsaw puzzle task. The level of noticing was operationalised by a verbal reference to the target structure, but did not necessarily mention the rule, such as *pausing after the verb* or *making a comment on it*.

### ***7.5.2 The targeted linguistic feature***

It has been posited that different aspects of language learning require different levels or processes of attention (Gass, 2004; Gass, Svetics & Lemelin, 2003; Larsen-Freeman, 2004; Schmidt, 2001; Schwartz, 1993). Schwartz (1993) suggested that learning lexicon, morphology and syntax may require different attention and awareness. Gass *et al.* (2003) investigated English learners of L2 Italian on the acquisition of three linguistic domains: syntax, morphosyntax and lexicon, and their results suggested that learning syntax and morphology requires more attention, and that lexicon requires the least. VanPatten (1994) made a proposal that “perhaps different aspects of language are processed and stored differentially” (p.31), but he has not clearly elaborated how this might happen and what different aspects of a language actually act on input processing. Most of the grammatical focuses of PI studies have been on morphology (for example, Marsden, 2006; Benati, 2005, including the current study) and morphosyntax (for example, VanPatten & Cadierno, 1993a, 1993b). However, only a few studies have examined the effect of PI on learning syntax (Farley, 2004b; Wu, 2003). As a result, the applicability of PI in learning different aspects of grammar needs more studies to attest it.

In addition, one practical issue emerges when constructing PI activities. The emphasis is on the task design for the targeted linguistic form without inherent meaning. Since the form has no ‘meaning’, how and to what extent can the PI activities manage to facilitate the formation of FMCs? VanPatten has not specifically addressed this practical issue or offered guidelines on how to devise PI activities relating to the non-meaningful form. So far, the beneficial effect of PI on non-meaningful form has not been investigated.

### ***7.5.3 Achievement assessments***

The generalisability of this study was limited to a certain extent due to the nature of the achievement assessments. For example, the GJT used in this study only adopted visual modality rather than aural modality. Furthermore, the validity of the oral tests was not assessed. Besides this, the sample recruited for analysing the oral tests was small (9-10 participants in each instructional group). Consequently, the findings from the oral tests were more suggestive than conclusive, particularly the results obtained from the principal component analysis.

In terms of exploring implicit and explicit knowledge, only post-instruction instruments were employed. The use of a post-task questionnaire is open to criticism due to the learners' lack of awareness and their consequent lack of ability to verbalise the rule. As a result, in order to obtain a more complete picture of what knowledge learners are processing, further studies could include a concurrent self-report (e.g. a think-aloud protocol) to investigate this issue. Additionally, as Dienes & Perner (1999) claimed that the distinction between implicit knowledge and explicit knowledge should be treated as a continuum instead of a dichotomy, future studies could adopt the acceptability judgment (Sorace, 1996) rather than a grammaticality judgment, with grades of certainty.

The use of a time constraint in the achievement assessment was designed originally to decrease the likelihood that the learners were monitoring their language performance during testing phases. However, the results obtained from the implicit tests suggest that the time constraint used was not sufficient to prevent such monitoring completely. Also, it is challenging to establish an appropriate time frame for all L2 learners due to individual differences between them. Therefore, how to set a valid time frame for implicit tests to avoid the subjects using explicit knowledge but still eliciting implicit

knowledge is a pressing issue for future studies to address when it comes to devising the instruments (see Ellis *et al.*, 2009).

In addition, another implication derived from this study is the use of statistical analysis procedures for the outcome measures. It was found in this study (i.e. in the picture narration test and the vocabulary test) that the use of parametric and non-parametric tests could produce different results and would affect the interpretation of the effectiveness of an intervention. However, it is common to find in the published literature that researchers apply parametric tests without justifying the use of them (i.e. reporting whether or not the assumptions were upheld). I suggest that researchers should give a detailed report about their use of statistical tests or be rigorous in their choice of them.

Finally, this study has an important implication for future research on second/foreign language assessment as test scores were significantly different as a function of the mode of the tests which were used. It could be that the difference depended on whether learners performed the tests in a written mode or an oral mode. This could have been related to the accessibility to explicit knowledge in these different test types. Thus, this study points to the need to develop multiple assessments in language research to establish a more informative investigation on the effectiveness of different types of instruction.

**Appendix 1 The official objectives of the English curriculum and guidelines for  
constructing English teaching materials  
(<http://teach.eje.edu.tw/9CC/3-2.php>)**

一、 基本理念

隨著地球村時代的來臨，國際間政治、經濟、文化往來頻繁，英語的重要性日益突顯。從資訊、科技、工商業、乃至高等教育，英語已成為國際交流之重要溝通工具。而透過英語學習，學生可體驗不同的文化，增進其對多元文化的了解與尊重。因此，重視英語教育已成為多數現代國家的教育趨勢。自從政府推動亞太營運中心的建立以來，國人深感英語溝通能力的提昇日漸重要，在社會殷切期望下，英語教學提前至國小階段實施。國民中小學英語課程設計宗旨在奠定國人英語溝通基礎，涵泳國際觀，以期未來能增進國人對國際事務之處理能力，增強國家競爭力。

國民中小學英語課程強調營造自然、愉快的語言學習環境，以培養學生之學習興趣和基本溝通能力。上課宜採輕鬆、活潑之互動教學模式。教材內容及活動設計宜生活化、實用化及趣味化；體裁多樣化。溝通能力之培養宜透過多元教材與活動練習，讓學生藉由多方面語言接觸，及實地應用來學習英語，而非由老師單向灌輸文法結構等語言知識。為了維持學生之學習動機且不增加學習負擔，教材之份量及難易度宜適中，學生之學習興趣與吸收能力應勝於教學進度之考量。(以下略)

四、課程目標

(以下略)

句型結構: 國小、國中教材可採用之句型，應以基本常用為主，避免冷僻、抽象之文法知識的灌輸

(以下略)

## 2. 教材編纂原則

(以下略)

教材編撰時,每單元宜提供生活化之情境,並融合主題、句型結構及溝通功能加以編寫。活動之設計宜多元,並強調溝通式活動,以培養學生基本的溝通能力。每單元之活動宜環繞主題或溝通功能加以設計,字彙、片語、句型之介紹應採循序漸進,由易漸難螺旋向上之模式,並適時安排複習單元,提供學生反覆練習之機會。介紹過之主題、溝通功能或文法句型,之後仍可以較高層次的應用方式再次出現。教材的主題應與學生之生活密切配合,體裁宜隨著學生年齡及英語能力的增長呈現多元的風貌。以淺白易懂與趣味化為原則,儘量將歌謠、對話、韻文、書信、故事、短劇等融入教材之中。生活化的主題配合不同的體裁,讓學生透過教材體驗豐富多樣的語言學習經驗,以提昇學習的興趣,增進學習的效果。

(以下略)

### 1. Basic concepts

In this time of globalisation, the government, the economy and culture have become increasingly international, and the importance of English has become increasingly apparent. In the fields of information, technology, industry, commerce, and even advanced education, English has become an important communicative tool for international exchange. Through the study of English, students experience and gain an understanding and respect for many other different cultures. Due to the effect of this, English education has already become an important feature of the educational systems in a majority of developed countries. Since the Taiwan government sponsored APAC was founded, Taiwanese have

deeply felt the increasing importance of being able to communicate in English. As a result, they wish to introduce English education as early as possible into the curriculums. The design objective of the Taiwanese elementary and junior high school curriculums is to build up a foundation of English abilities in students, to develop international perspectives, increase their potential to handle international concerns and to strengthen the country's competitiveness.

Taiwanese elementary and junior high English curriculums should strongly emphasise the construction of natural and enjoyable language study environments, and use a relaxed, lively and interactive teaching model to cultivate the students' interest and basic communication abilities. The content of the teaching material and activities should be suitable for daily life's use and interest, and use many types of materials. Learners' communicative ability should be cultivated using a variety of teaching materials and practice so that students want to learn English, *and not just passively absorb grammar knowledge from the teacher*. To encourage students to study without increasing pressure on them, students' interests and abilities should be considered when determining the amount and difficulty of the teaching material.

#### **4. Goals of Implementation**

(...deleted)

Sentence structure: the sentence patterns adopted in the teaching materials for all levels must use common topics and avoid obscure and abstract forms that just illustrate grammar. Structures should be presented in a step-by-step fashion, from the simple to the complex, and let the students understand the meaning through familiar patterns.

The sentence patterns introduced at elementary and junior high levels should be coordinated.

#### **2. Principles for constructing English teaching materials**

(...deleted)

When compiling teaching materials, every unit should be suitable to life's circumstances, integrate the main ideas and include sentence structures, useful words and phrases. Activities should be diverse, but must emphasize communication activities and cultivate a student's basic communication ability. Every unit should include activities which are appropriate to the main topic, and should introduce vocabulary, phrase and sentence patterns. These should follow a step-by-step, easy-to-difficult model. There should also be timely reviews, and provide the students with opportunities to practice with the material again and again. Ideas and structures introduced at low levels should connect to and be repeated in materials at higher levels, expanding their use. The main ideas in the teaching material should be closely matched to the age and English abilities of the students, while presenting a variety of diverse features. They should integrate simple, easy and interesting resources, such as songs, dialogues, rhymes, stories, and short scripts. A variety of lively and appropriate topics in the material should put the students in touch with a variety of language study experiences, which should promote their interest in the language and benefit their studies.



## Appendix 2 The ‘first noun principle’ and its corollaries in IP

*The first noun principle:* Learners tend to process the first noun or pronoun they encounter in a sentence as the subject/agent.

### Principle a. Lexical semantics principle.

Learners may rely on lexical semantics, where possible, instead of word order to interpret sentences.

### Principle b. event probabilities principle.

Learners may rely on event probabilities, where possible, instead of word order to interpret sentences.

### Principle c. contextual constraint principle.

Learners may rely less on the First Noun Principle if preceding context constrains the possible interpretation of a clause or sentence.

**Appendix 3 A tabular summary of studies related to PI**

Study	participants	N	L1	L2	treatment	assessment	Treatment Duration	posttests
Allen (2000)	High school students	N=179	N.P.	French	1.PI (N=64) 2.TI (N=61) 3.Control (N=54) (No RM)	1.interpretation task: match aural sentences with pictures 2.production task( written narrative test)	Two days	A pretest, an immediate posttest; two delayed posttest (one week and one month later)
Benati (2001)	Uni students Second semester (beginner learners)	N = 39	English	Italian	1. PI 2. TI 3. C (RM)	1. int tests (20 aural test items) 10T(future)+10D ( present tense ) 2. pro tests 2.1 written completion task ( 5T ) 2.2 oral limited response task ( 5 pictures-based story telling )	2 consecutive days (6 hours)	1. a pretest ( 3 weeks before ) 2. imm posttest 3. delayed posttest (3 weeks after)
Benati (2004a)	undergraduate	N=38	English	Italian	1.PI group (N=14) 2.SI group (N=12) 3.EI group (N=12)	1.aural int ( 10T + 10D ) 2. written production task ( 5T )	2 3-hour sessions in 2 consecutive days	1.A pretest 2.immediate posttest 3. one-month posttest
Benati (2004b)	Undergraduate	N = 31	English	Italian	1. PI=10 2. SI=11 3. EI=10	1. an interpretation(10 T+10D) 2. two production (a gap-fill test and a picture-based oral test)	2 2-hour sessions in 2 consecutive days	1. pretest ( two weeks bf the treatment ) 2. only one posttest(imm)

Benati (2005)	Secondary school School-aged Chinese and Greek (12-13)	Chinese N=47 Greek N=30	Chinese and Greek	English	Chinese: <b>RM</b> 1. PI=15 2. TI=15 3. MOI=17 Greek: 1. PI=10 2. TI=10 3. MOI=10	1. interpretation =10T +10D 2. written production	3 consecutive days for 6 hours (2 hour per day)	1. pretests (administered 2 weeks before the instruction)  2. immediate posttests
Cadierno (1995)	Uni students	N=61	English	Spanish	1.PI 2.TI 3. C	1.Interpretation task 2. production task.	two days	Three posttest: immediately, one week and one month later
Cheng (2002)	University student ( cut off point= 60)	N = 109	English	Spanish ser and estar	1.PI 2.TI 3.C	1. aural interpretation 2. written production tests: (sentence completion and a guided composition)	2 consecutive days of instruction	1. pretest: 1 week before 2. immi posttest 3. 3-week delayed posttest
Cheng (2004)	Uni students in fourth-semester	N = 83	English	Spanish	1. PI (N=29) 2. TI (N=28) 3. C (N=26)	1. interpretation task 2. production task: (sentence completion and guided composition)	N.P.	1.pretest: 1 week before 2. imm posttest 3. 3-week delayed posttest
Collentine, (1998)	Uni student	N = 54	English	Spanish	1.PI (N=18) 2.TI (N=18) 3.C (N=18)	1.interpretation test (listening and reading) 2.production test (sentence completion test: fill the blank)	2 consecutive, 50-min classes	1. pretest 3 days before treatment 2. posttest a day afterwards.

Dekeyser & Sokalski (1996)	Uni students	N=82	English	Spanish	EX1: 1. PI group 2. TI group 3. C group	1. Interpretation task 2. production tasks (fill in blanks, translate sentences, or answer questions)	6 class period	A pretest, an immediate posttest and a one week posttest
Erlam, (2003)	School-age learners (14)	N=70	English	French	1. PI 2. MOI 3. C	1. listening tests (10T +2D) 2. reading test(8T+2D) 3. written production (10T +2D) 4. oral production tests: based on a short sequence of pictures	3 45-min lessons in a week	1. 2-week Pretest 2. one-week posttest 3. 6-week delayed posttest
Farley (2001)	Uni students 4 <sup>th</sup> -semester	N = 29	English	Spanish	1. PI (N=17) 2. MOI (N=12) <b>RM</b>	1. aural int tests (12 T+12D) 2. pro tests: sentence completion(fill in the blank) (12T + 8D)	2 days ( 90 minutes )	1. pretest: on the first day of instruction before treatment. cutoff point:60 2. posttest: 1 day after 3. delayed posttest: 1 month
Farley (2004a)	Uni students: 4 <sup>th</sup> -semester Spanish	N = 50	English	Spanish	1. PI 2. MOI	1. int 24 items ( 12T+12D ) 【 9sub+3ind+12D 】 2. pro ( 9D 【 6sub+3indi 】 +12D ) indi for avoiding the overextension	2 sessions within 2 days in a week	1. pretest ( 1-day bef treatment;cutoff points 60% ) 2. posttest 3. 2-week later delayed posttest
Farley (2004b)	Uni students 4 <sup>th</sup> -semester	N = 54	English	Spanish	1. PI (23) 2. SI (31)	1. int 24 items ( 12T+12D )	Two consecutive days; a total of	1. pretest 2. two posttests.

	Spanish					2. pro (9D +12D)	100 mins	
Keating & Farley (2008)	Uni students	N=87	English	spanish	1.PI=36 2.MOI=25 3.MDI=26	1. aural interpretation test 2. production test:	70 mins	1.pretest 2.posttests: immediate, 1-week, 1-month later.
Marsden (2006): Experiment 1	13-14	N=27	English	French	1. PI group 2. EnI group 3. C	1. interpretation task(oral and aural) 2. production task (writing and speaking)	9.5 hours over 7 weeks	A pretest, an immediate posttest and a posttest between 14-16 weeks.
Morgan-short & Bowden (2006)	First-semester Uni Spanish learners	N = 43		Spanish	1.PI (N=15) 2.MOBI (N=14) 3.C (N=14)	1.aural interpretation test 2. written production test: gap-fill test (time constraints in tests: 15s in int; 20s in pro )	3 one-hour sessions in 3 weeks. O	1.pretest: 2 weeks prior to the study 2.immi posttest 3.1-week posttest
Salaberry (1997)	Uni students	N=26	English	Spanish	1.PI group n=9 2. TI group n=10 3. C group n=7	1.Interpretation task: match aural sentences with pictures 2. production tasks: (sentence completion and free narration of a one-minute silent video clip)		A pretest, an immediate posttest and a one-month delayed posttest with only interpretation test.
Sanz & Morgan-Short(2004)	Uni students 1 <sup>st</sup> or 2 <sup>nd</sup> year of Spanish	N=69	N.P.	Spanish	1. +E+F (N=21) 2. -E-F (N=20) 3.+E-F (N=15) 4.-E+F (N=13)	1. interpretation (10T+5D) 2. production (sentence completion and written video retelling)	N.P.	1. Pretest 2. posttest

Toth (2006)	Adults beginning learners	N=80	English	Spanish	1. PI (N=27) 2. COI (N=28) 3. C (N=25)	1. production: sentence completion: 2. grammaticality (a total of 50 sentences)	7 days: each day 50 mins except 25 mins taken on Day and Day 7.(300 mins)	1. pretest: cutoff point 50 % or higher 2. imm posttest 3. 24 days after
VanPatten & Cadierno (1993a)	Uni students	N=80	English	Spanish	1.PI group 2.TI group 3.C	1.interpretation task(10T+10D) 2.production task (5T+5D)	2 days	A pretest; 3 posttest (immediate, 2-week, one month)
VanPatten & Cadierno (1993b)	Uni students	N=49	English	Spanish	1. PI 2. TI 3.C	1. interpretation task 2. production task	4 days a week	A pretest; 3 posttest (immediate, 2-week, one month)
VanPatten & Fernández (2004)	Uni students	N = 45	English	Spanish	Only PI	1. aural interpretation (10 T, 5 D) 2. production (5T, 5D): sentence completion task	Two consecutive days.	Pretest, immediate posttest, and eight-month delayed posttest
VanPatten & Oikkenon (1996)	Secondary school students	N=59	English	Spanish	1. EI (N=22) 2.SIA only (N=20 ) 3.PI group (17)	1. interpretation task 2. production task	3 days	A pretest; only one posttest
VanPatten & Sanz (1995)	Uni students	N= 44	English	Spanish	1. PI (N=27) 2. C (N=17)	1.Interpretation test (20T + 6D) 2.Production test (sentence completion test, structured	2 days	1. pretest: several weeks before treatments 2. immi posttest

						interview, and video narration		
VanPatten & Wong (2004)	University students		English	French causative	1. PI 2. TI 3. C	1.interpretation task (7T +7D): aural test 2.production task (5T+5D): sentence completion	Treatment time: 45 mins	Pretest, and a posttest
Wong (2004b)	Uni students	N = 94	English	French	1. PI (N=26) 2. EI (N=22) 3. SI (N=25) 4. C (N=21)	1. interpretation (10T+10D) 2. production (6T+6D)	One day treatment	1.pretest (2-weeks before) 2.posttest
Wu (2003)	15-16, 1 <sup>st</sup> graders in high school	N=90	Chinese	English	1. PI 2. TI	1. comprehension test (identifying test and interpretation test) 2. production (fill-in-blank and sentence-combining)	4 days 100 mins	1. no pretest 2. two posttests, including an immediate and a one-month delay-posttest
Xu (2001)	grade 7 junior high school students	N=51	Chinese	English	1.PI 2.TI 3.C	1. interpretation 2. written production test : scrambled question	Two one-hour sessions	1. pre-test 2. immi pt 3. five-month dp

NOTE: TI= Traditional instruction; PI = Processing instruction; C = control group; SI = structured input activities only instruction; EI = explicit information only instruction; MOI = meaningful-output instruction; EnI = enriched input instruction; N.P. = Not Report; int = interpretation tests; pro = production tests; +/-E = +/-explanation; +/- F = +/-explicit feedback. T=target feature; D=distractor; RM=randomly assigned participants

**Appendix 4 Summary of PI-based studies' findings on interpretation tasks**

<b>Studies</b>	<b>Target structure</b>	<b>Results of interpretation tasks at post-test(s)</b>		
Allen (2000)	French causative	(immi) <b>P=T</b> T>C P>C	(1 week) <b>P=T</b> T>C P>C	(1 month) <b>P=T</b> T>C P>C
Benati (2001)	Italian future tense	Immediate posttest PI > TI PI > C TI > C		3-week delayed posttest PI > TI PI > C TI > C
Benati (2004a)	Italian future tense	Immediate posttest PI > EI PI = SI SI > EI		1-month delayed posttest PI > EI PI = SI SI > EI
Benati (2004b)	Gender agreement in Italian	Immediate post-test: PI = SI PI > EI SI > EI		
Benati (2005)	English past simple tense	immediate posttest in <u>Chinese group</u> : PI > TI PI > MOI TI = MOI		immediate posttest in <u>Greek group</u> : PI > TI PI > MOI TI = MOI
Cadierno (1995)	Spanish simple past tense	(immiate) PI > TI PI > C TI = C	(one week) PI > TI PI > C TI = C	(one month) PI > TI PI > C TI = C
Cheng (2002)	Spanish <i>ser</i> and <i>estar</i>	Immediate posttest PI > C TI = C		3-week delayed posttest TI > C PI = C
Collentine (1998)	Spanish subjunctive	PI = TI PI > C TI > C		
Dekeyser & Sokalski (1996)	T1: Spanish direct object clitics	T1	T=P, T=C, P>C (immediate posttest)	
		T2	T=P, P=C, T>C (immediate posttest)	
	T2: Spanish conditional	T1	T=P=C (one week delayed posttest)	
		T2	T=P=C (one week delayed posttest)	
Erlam (2003)	French	One-week posttest		six-week posttest



		<u>Listening:</u> MOI>PI MOI>C <u>Reading:</u> MOI>C	<u>Listening:</u> MOI=PI=C <u>Reading:</u> MOI=PI=C	
Farley (2001)	Spanish subjunctive	One-day posttest PI > MOI	1-month delayed posttest PI > MOI	
Farley (2004a)	Spanish subjunctive	Post-test PI=MOI	2-week delayed posttest PI=MOI	
Farley (2004b)	Spanish Subjunctive	Posttest 1 PI > SI PI > SI	Posttest 2 PI > SI PI > SI	
Marsden (2006)	French verb inflections in the perfect and present tense	Posttest PI> EnI (experiment 1) PI>C EnI =C (experiment 2)	Delayed posttest PI> EnI (experiment 1)	
Morgan-Short & Bowden (2006)	Spanish preverbal direct object pronouns	Immediate posttest PI> pretest MOBI>pretest C=pretest	1-week delayed posttest PI> pretest MOBI > pretest C > pretest	
Salaberry (1997)	Spanish direct object clitics	Immediate posttest T>C P>C T =P	1-month delayed posttest T>C P>C T =P	
Sanz & Morgan-Shrot (2004)	Spanish Object Pronous	(+E, +F) = (-E, -F) = (+E, -F) = (-E, +F) all groups improved significantly from the time of the pretest to the time of the posttest(sentence completion and video-retelling).		
VanPatten & Cadierno (1993a)	Spanish direct object clitics	immediate PI > TI PI > C TI = C	2-week posttest PI > TI PI > C TI = C	1-month pt PI > TI PI > C TI = C
VanPatten & Cadierno (1993b)	Spanish direct object clitics	immediate PI > TI PI > C TI = C	2-week posttest PI > TI PI > C TI = C	1-month pt PI > TI PI > C TI = C
VanPatten & Fernández (2004)	Spanish direct object clitics	Immediate posttest (8 months) > pretest Immediate posttest > delayed posttest delayed posttest > pretest		

VanPatten & Oikkenon (1996)	Spanish direct object clitics	PI > EI SI > EI PI = SI	
VanPatten & Sanz (1995)	Direct object pronouns in Spanish	PI > C	
VanPatten & Wong	French causative	PI > C PI > TI TI > C	
Wong (2004b)	French ( the use of <i>de</i> with <i>avoir</i> in French )	PI > EI PI > C SI > EI SI > C	
Wu (2003)	English Subjunctive Mood	Immediate posttest PI > TI	1-month delayed posttest PI > TI
Xu (2001)	English wh-question	Immediate posttest PI > C PI > TI	5-month delayed posttest PI=CI=T

Note: TI= Traditional instruction; PI = Processing instruction; C = control group;

EnI = enriched input instruction; MOI = meaningful-output instruction;

SI = structured input activities only instruction; EI = explicit information only instruction

> and < means statistic significance; = means no statistic significance

+/-E = +/- explanation; +/- F = +/- explicit feedback.

## Appendix 5 Summary of PI-based studies' findings on production tasks

Studies	Target structure	Results of production tasks at post-test(s)		
Allen (2000)	French causative	(immediate) T>C P>C T>P	(1 week) T>C P>C T=P	(1 month) T>C P>C T>P
Benati (2001)	Italian future tense	Immediate posttest (both in written and oral tests) PI = TI PI > C TI > C		3-week delayed posttest (both in written and oral tests) PI = TI PI > C TI > C
Benati (2004a)	Italian future tense	Immediate posttest PI > EI PI = SI SI > EI		1-month delayed posttest PI > EI PI = SI SI > EI
Benati (2004b)	Gender agreement in Italian	Immediate posttest <u>Gap-fill test:</u> PI = SIA PI > EI SIA > EI		Immediate post-test <u>Oral production test:</u> PI = SIA PI > EI SIA > EI
Benati (2005)	English past simple tense	immediate posttest in <u>Chinese group</u> : PI = TI = MOI		immediate posttest in <u>Greek group</u> : PI = TI = MOI
Cadierno (1995)	Spanish simple past tense	(immediatt) PI > C TI > C PI = TI	(one week) PI > C TI > C PI = TI	(one month) PI > C TI > C PI = TI
Cheng (2002)	Spanish <i>ser</i> and <i>estar</i>	Posttest 1 a) Sentence production PI > C TI > C b) guided composition PI > C TI > C		Posttest2 (3weeks later) a)Sentence production PI > C TI > C b) guided composition PI > C TI > C
Collentine (1998)	Spanish subjunctive	PI = TI PI > C TI > C		

Dekeyser & Sokalski (1996)	T1: Spanish direct object clitics	T1	T=P, P=C, T>C (immediate posttest)			
		T2	T>C, P>C, T>P (immediate posttest)			
	T2: Spanish conditional	T1	T=P=C (one week delayed posttest)			
		T2	T=P=C (one week delayed posttest)			
Farley (2001)	Spanish subjunctive	One-day posttest PI = MOI	1-month delayed posttest PI = MOI			
Farley (2004a)	Spanish subjunctive	posttest PI = MOI	2-week delayed posttest PI = MOI			
Farley (2004b)	Spanish Subjunctive	Posttest 1 PI > SI PI > SI	Posttest 2 PI > SI PI > SI			
Marsden (2006)	French verb inflections in the perfect and present tense	posttest PI>EnI (experiment 1)	Delayed posttest PI> EnI (experiment 1)			
		(experiment 2) Writing: EnI = PI > C Speaking: EnI = PI = C				
Morgan-Short & Bowden (2006)	Spanish preverbal direct object pronouns	Immediate posttest PI> pretest MOBI>pretest	1-week delayed posttest PI> pretest MOBI > pretest			
Salaberry (1997)	Spanish direct object clitics	T=P=C				
Sanz & Morgan-Shrot (2004)	Spanish Object Pronous	(+E, +F) = (-E, -F) = (+E, -F) = (-E, +F) all groups improved significantly from the time of the pretest to the time of the posttest(sentence completion and video-retelling).				
VanPatten & Cadierno (1993a)	Spanish direct object clitics	immediate	2-week posttest	1-month posttest		
		PI > C	PI > C	PI > C		
		TI > C	TI > C	TI > C		
VanPatten & Cadierno (1993b)	Spanish direct object clitics	PI = TI	PI = T	PI = TI		
		PI > C	PI > C	PI > C		
		TI > C	TI > C	TI > C		
VanPatten & Fernández (2004)	Spanish direct object clitics	PI = TI	PI = T	PI = TI		
		Immediate posttest (8 months) > pretest	Immediate posttest > delayed posttest			
		Immediate posttest > delayed posttest	delayed posttest > pretest			

VanPatten & Oikkenon (1996)	Spanish direct object clitics	PI = SI PI > EI SI = EI		
VanPatten & Sanz (1995)	Direct object pronouns in Spanish	Sentence completion PI > C	Video narration test: PI > C	Structured interview test: PI = C
VanPatten & Wong	French causative	PI>C TI>C PI=TI		
Wong (2004b)	French ( the use of <i>de</i> with <i>avoir</i> in French )	PI > EI; SI = EI PI > C ; EI = C SI > C ; PI = SI		
Wu (2003)	English Subjunctive Mood	Immediate posttest 1.fill-in-blank: PI = TI 2.sentence-combining TI > PI	1-month delayed posttest 1.fill-in-blank: PI > TI 2.sentence-combining PI = TI	
Xu (2001)	English wh-question	Immediate posttest PI>C	5-month delayed posttest PI>C	

Note: TI= Traditional instruction; PI = Processing instruction; C = control group; EnI = enriched input instruction; MOI = meaningful-output instruction; SI = structured input activities only instruction; EI = explicit information only instruction; > and < means statistic significance; = means no statistic significance; +/-E = +/- explanation; +/- F = +/- explicit feedback.

## **Appendix 6 The consent of the headmaster of the participating school**

To Whom It May Concern,

This is a letter to confirm that Hsin-Ying Chen has been given permission to undertake her educational research connected with her degree studies between February and May 2007 at [...] Primary School, Taitung, Taiwan. On behalf of [...] Primary School, I can also confirm that we are delighted to be a part of her project and to have the educational benefit which it brings, as we regard her teaching materials to be a part of the curriculum. In addition, we should cooperate with Hsin-Ying to assist her in carrying out her project smoothly, including the use of school facilities and the computer laboratory. If you have any concerns or enquiries related to this, please do not hesitate to contact the school or me.

Kind regards,

[...]

### Appendix 7 The raw scores of the 13 outliers

<b>ID</b>	<b>Group</b>	<b>GJT Pre-test</b>	<b>GJT PT</b>	<b>GJT DP</b>	<b>GAP Pre-test</b>	<b>GAP PT</b>	<b>GAP DP</b>
<b>1</b>	<b>RA</b>	40	40	40	8	8	8
<b>2</b>	<b>RA</b>	36	36	37	5	8	6
<b>3</b>	<b>RA</b>	40	40	40	7	8	8
<b>4</b>	<b>R</b>	40	40	40	6	8	8
<b>5</b>	<b>R</b>	35	39	40	4	7	7
<b>6</b>	<b>R</b>	36	36	40	8	8	7
<b>7</b>	<b>A</b>	38	33	33	6	6	7
<b>8</b>	<b>A</b>	40	40	38	6	7	7
<b>9</b>	<b>A</b>	36	36	40	8	8	8
<b>10</b>	<b>C</b>	38	40	40	3	7	5
<b>11</b>	<b>C</b>	32	38	40	5	7	7
<b>12</b>	<b>C</b>	39	38	38	5	5	7
<b>13</b>	<b>C</b>	36	40	38	8	3	8

\*PT= post-test; DP= delayed post-test

\*the maximum score for the GJT is 40.









\*the maximum score fro the gap-fill test is 8.

### Appendix 8 The timetable of the current study

15.1.2007 ~ 18.1.2007	Pilot study
29.1.2007 ~ 24.2.2007	School holidays
26.2.2007	Start of the main study Administering the pre-test (the GJT, the gap-fill test, and the vocabulary test)
5.3.2007 ~ 9.3.2007	Administering the pre-test (oral tests)
12.3.2007 ~ 14.3.2007	Instructional phases: delivering teaching materials
19.3.2007 ~ 21.3.2007	Instructional phases: delivering teaching materials
26.3.2007 ~ 30.3.2007	Administering the post-test (oral tests)
2.4.2007	Administering the post-test (the GJT, the gap-fill test, and the vocabulary tests)
30.4.2007 ~ 4.5.2007	Administering the delayed post-test (oral tests)
7.5.2007	Administering the delayed post-test (the GJT, the gap-fill test, and the vocabulary test)



**Appendix 9 The questionnaire regarding subjects' English learning backgrounds  
(Chinese and English versions)**


1. 請問你大約學幾年英文了呢 (包括參加學校正式課程之外的補習班或私人家教)? \_\_\_\_\_
2. 請問你是否曾經去過英語系國家旅行或居住 (例如:美國、英國、澳洲...等)?  
是  否   
 如果有的話, 請問你去了哪個或哪些國家? \_\_\_\_\_  
 什麼時候去的呢? \_\_\_\_\_  
 大約在那裡停留了多長時間呢? \_\_\_\_\_
3. 請問你放學後有沒有參加任何英文補習班、英文課後加強班、或是有私人英文家教?  
是  否   
 請問你參加課後英文輔導的頻率如何呢 (請以一週為單位; 例如:一個禮拜一次)? 1 週 \_\_\_\_\_ 次 或 其它: \_\_\_\_\_  
 每次上課的時間多長 (請以分鐘為單位; 例如:一次 2 小時=一次 120 分鐘)?  
1 次 \_\_\_\_\_ 分鐘
4. 請問你平常學校放學後有沒有跟任何說英文的外國人聊天或練習英文?  
是  否   
 如果有的話,請問他來自哪個國家? \_\_\_\_\_  
 你是如何跟他聊天或練習英文的呢?  
面對面  用 Skype  用 MSN  e-mail  其它 \_\_\_\_\_  
 請問你多常跟他用英文聊天、寫信、或傳訊息? (例如:1 個月 2 次 或 1 週 1 次)? \_\_\_\_\_

**Translation:**


1. How many years have you been studying English? \_\_\_\_\_

2. Have you ever been to any English-speaking countries (e.g. USA, UK, etc)?

Yes  No

 If yes, where have you been? \_\_\_\_\_

 When? \_\_\_\_\_

 For how long? \_\_\_\_\_

3. Do you attend any English classes at a cram school, a language institution, or do you have private English tutor after school?

Yes  No

 How often do you attend the classes (e.g. once a week, etc)? \_\_\_\_\_

 How long does the class last each time?

\_\_\_\_\_ (minutes)


4. Do you contact with any English-native speakers after school?

Yes  No

 If yes, where is s/he from? \_\_\_\_\_

 How do you contact with him/her?


face to face  Skype  MSN  email  others \_\_\_\_\_

 How often do you talk/write to him/her in English? (e.g. once a week, once a month and so on)


\_\_\_\_\_

## Appendix 10 The attitudinal questionnaire (Chinese and English versions)

1. 我覺得在電腦 '練習' 操作上，很簡單。 是  否
2. 我覺得在電腦 '練習' 操作上，有困難。 是  否

 如果你勾是的話，請你寫下你操作上的問題及困難處 \_\_\_\_\_

3. 我覺得這些練習活動很有趣。 是  否
4. 我覺得這些練習活動很無聊。 是  否
5. 我覺得這些英文練習活動很難。 是  否

 如果你勾是的話，請你勾選你覺得很難的練習活動

閱讀練習  聽力練習  兩個都很難

6. 我覺得這些英文練習活動很簡單。 是  否
7. 如果將來有機會的話，我願意再次使用與這次類似的練習活動去學英文。  
是  否

### Translation:

1. I feel the programme is easy to operate. Yes  No
2. I have problems when practising on the computer. Yes  No

If yes, what are the problems? \_\_\_\_\_

3. I feel the activities are interesting. Yes  No
4. I feel the activities are boring. Yes  No
5. I feel the activities are difficult. Yes  No
6. I feel the activities are easy. Yes  No
7. I would like to learn English by using this type of activity in future.  
Yes  No

**Appendix 11 Handout with English temporal adverbials given to the participants  
during the instructional phases**

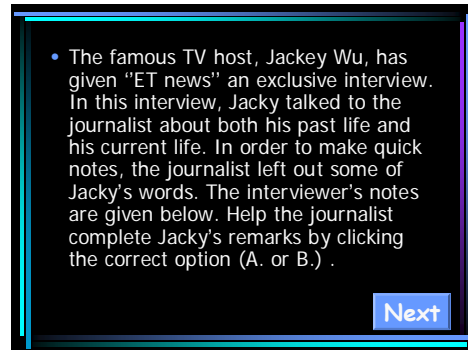
<b>過去式(past tense)</b>	<b>現在式(present tense)</b>	<b>未來式(future tense)</b>
<p align="center"><b>last</b></p> <p>examples (例子):</p> <p>last month (上個月)</p> <p>last year (去年)</p> <p>last week (上個星期)</p> <p>last weekend (上個週末)</p> <p>last Sunday (上個星期天)</p>	<p align="center"><b>every</b></p> <p>examples (例子):</p> <p>every month (每個月)</p> <p>every year (每年)</p> <p>every week (每個星期)</p> <p>every weekend (每個週末)</p> <p>every Sunday (每個星期天)</p>	<p align="center"><b>next</b></p> <p>examples (例子):</p> <p>next month (下個月)</p> <p>next year (明年)</p> <p>next week (下個星期)</p> <p>next weekend (下個週末)</p> <p>next Sunday (下個星期天)</p>
<p align="center"><b>yesterday (昨天)</b></p>	<p align="center"><b>every day (每天)</b></p>	<p align="center"><b>tomorrow (明天)</b></p>
<p align="center"><b>ago</b></p> <p>2 days ago</p> <p>2 hours ago</p>		

## Appendix 12 An example of referential activities for this study

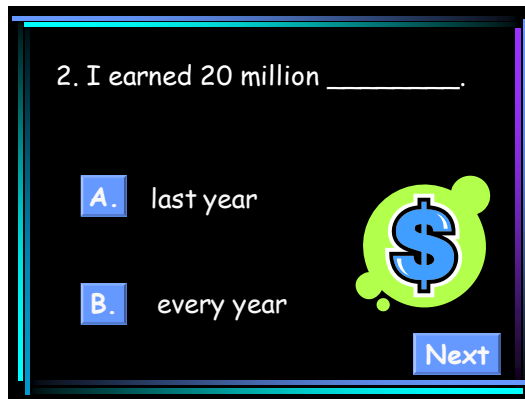
1)



2)



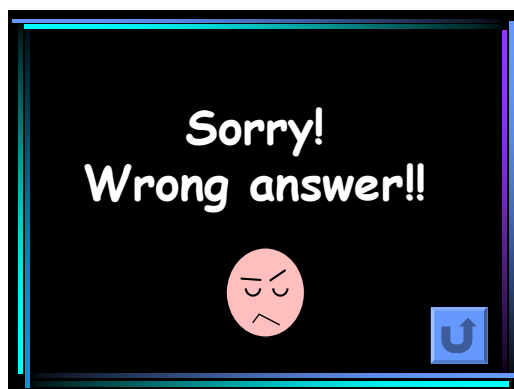
3)



4) if a participant clicked A in 3)



5) if a participant click B in 3)



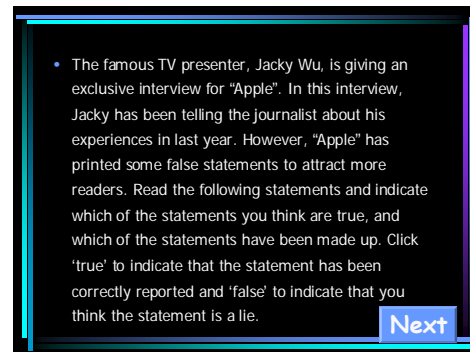
NB: Slide 2 given to the participants was written in Chinese instead of English

## Appendix 13 An example of affective activities for this study

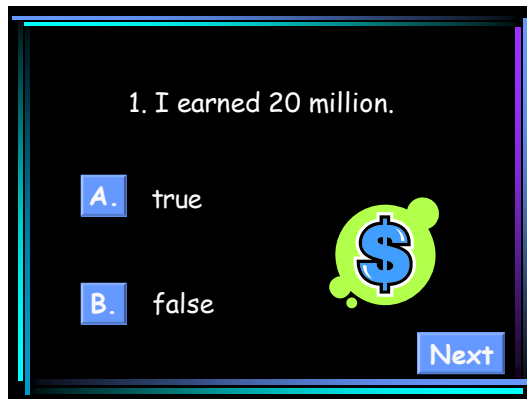
1)



2)



3)



4) if a participant clicked A in 3)



5) if a participant click B in 3)



NB: Slide 2, 4 & 5 given to the participants were written in Chinese instead of English

## Appendix 14 The test items of the timed GJT

1. My brother is work next month.
2. You watched TV last night.
3. They play badminton every Friday.
4. She is reading a book now.
5. They play baseball last weekend.
6. She worked hard last year.
7. They visited the USA last month.
8. Dad walk his dog every day.
9. You listen to the radio yesterday.
10. Are you watching DVDs tonight?
11. How old are you this year?
12. We visit Hong Kong next year.
13. They study English last night.
14. John walked to school yesterday.
15. Jessica is wearing a red skirt.
16. I visited my uncle last weekend.
17. We live in Taipei last year.
18. Is Mary visit Tainan tomorrow?
19. We watched movies two days ago.
20. I clean my room every night.
21. He visited Japan last week.
22. Ella call her friends every day.
23. I visit my grandfather last month.
24. Joey and Russ are good friends.
25. Are they are studying math now?
26. We played basketball last weekend.
27. We walk to school every day.
28. Bill listened to music last night.
29. You play computer games last night.
30. Joe and Mary is studying English.
31. My parents visit China last year.
32. Is your father a teacher?
33. We clean our house yesterday.
34. I brush my teeth last night.
35. Peter do not like apples.
36. Mary played the piano last night.

37. Do you like your math lesson?
38. We watch TV last night.
39. Mum go to London tomorrow.
40. Is he often play the guitar?



### Appendix 15 The answer sheet of the timed GJT

(English translation) Please circle a number for each test item (-2 to +2) according to your judgement of the sentence. Circling -2 means you are pretty sure (nearly 100 percent sure) that this sentence is incorrect; on the other hand, +2 means that you are pretty sure (nearly 100 percent sure) that this sentence is correct. -1 means that you are not very sure this sentence is wrong but it looks like it's incorrect to you; on the other hand, +1 means that you are not very sure this sentence is correct but it looks like it is. Circling 0 means that you have no idea about the correctness of this sentence.

	correct	possibly correct	don't know	possibly incorrect	incorrect
1.	+2	+1	0	-1	-2
2.	+2	+1	0	-1	-2
3.	+2	+1	0	-1	-2
4.	+2	+1	0	-1	-2
5.	+2	+1	0	-1	-2
-----					
6.	+2	+1	0	-1	-2
7.	+2	+1	0	-1	-2
8.	+2	+1	0	-1	-2
9.	+2	+1	0	-1	-2
10.	+2	+1	0	-1	-2
-----					
11.	+2	+1	0	-1	-2
12.	+2	+1	0	-1	-2
13.	+2	+1	0	-1	-2
14.	+2	+1	0	-1	-2
15.	+2	+1	0	-1	-2
-----					
16.	+2	+1	0	-1	-2
17.	+2	+1	0	-1	-2
18.	+2	+1	0	-1	-2
19.	+2	+1	0	-1	-2
20.	+2	+1	0	-1	-2

	correct	possibly correct	don't know	possibly incorrect	incorrect
21.	+2	+1	0	-1	-2
22.	+2	+1	0	-1	-2
23.	+2	+1	0	-1	-2
24.	+2	+1	0	-1	-2
25.	+2	+1	0	-1	-2
-----					
26.	+2	+1	0	-1	-2
27.	+2	+1	0	-1	-2
28.	+2	+1	0	-1	-2
29.	+2	+1	0	-1	-2
30.	+2	+1	0	-1	-2
-----					
31.	+2	+1	0	-1	-2
32.	+2	+1	0	-1	-2
33.	+2	+1	0	-1	-2
34.	+2	+1	0	-1	-2
35.	+2	+1	0	-1	-2
-----					
36.	+2	+1	0	-1	-2
37.	+2	+1	0	-1	-2
38.	+2	+1	0	-1	-2
39.	+2	+1	0	-1	-2
40.	+2	+1	0	-1	-2

## Appendix 16 The consent form for L1 participants

Researcher: Hsin-Ying, Chen

Email: [hc138@york.ac.uk](mailto:hc138@york.ac.uk)

Address: Department of Educational Studies  
University of York, YO10 5DD

### Consent to participate in Research

**Introduction:**

You are invited to participate in a research study. You will be asked to judge how good some English sentences are as quickly as you can. If you decide to participate, please sign and date the last line of this form. If at any point you change your mind and no longer want to participate, you can stop.

**Confidentiality:**

All of the information you give us will be treated confidentially. No one apart from the researcher will know your name. If you have any enquires or question about the research, you can contact the researcher by the telephone number or the email at the top of this form.

Thanks for your help!

Your signature \_\_\_\_\_ Date \_\_\_\_\_











**Appendix 17 Ten examples for the timed GJT**  
(which bore no relation to the targeted features)

1. I like dogs.
2. You do not play basketball.
3. Is they students?
4. John go to school every day.
5. He have a bike.
6. John and Mary are classmates.
7. Do you walk to school every day?
8. She is do her homework.
9. The trees is tall.
10. My sister is playing the piano.

## Appendix 18 The gap-fill test: Version A


姓名: 班級: 座號:


I. 填充題: 請將空格裡的單字填入 (共 15 題)


- \_\_\_\_\_ 1. Eric \_\_\_\_\_ the wall yesterday.  (擦油漆)
- \_\_\_\_\_ 2. I often take a \_\_\_\_\_ to school.  (公車)
- \_\_\_\_\_ 3. Joe \_\_\_\_\_ the guitar two days ago.  (彈吉他)
- \_\_\_\_\_ 4. My parents like eating \_\_\_\_\_.  (水果)
- \_\_\_\_\_ 5. Bill \_\_\_\_\_ English every night.  (讀書)
- \_\_\_\_\_ 6. Greg \_\_\_\_\_ his dinner last night.  (煮飯)
- \_\_\_\_\_ 7. Jerry is very \_\_\_\_\_.  (胖的)
- \_\_\_\_\_ 8. My brother \_\_\_\_\_ his room last weekend.  (打掃)
- \_\_\_\_\_ 9. Tommy \_\_\_\_\_ his teacher for help yesterday.  (要求)
- \_\_\_\_\_ 10. Janet \_\_\_\_\_ her dogs every weekend.  (溜狗)

(請翻頁繼續作答)

\_\_\_\_\_ 11. Peter \_\_\_\_\_ TV yesterday.  (看電視)

\_\_\_\_\_ 12. My mum \_\_\_\_\_ at the department store yesterday.  (購物)

\_\_\_\_\_ 13. My friends \_\_\_\_\_ the newspaper every day.  (讀報紙)


\_\_\_\_\_ 14. My uncle is working in the \_\_\_\_\_.  (醫院)

\_\_\_\_\_ 15. Johnson \_\_\_\_\_ last night.  (跳舞)


## Appendix 19 The gap-fill test: Version B


姓名: 班級: 座號:


I. 填充題: 請將空格裡的單字填入 (共 15 題)


\_\_\_\_\_ 1. I \_\_\_\_\_ my face an hour ago.  (洗臉)


\_\_\_\_\_ 2. It is very \_\_\_\_\_ outside.  (寒冷的)


\_\_\_\_\_ 3. Peter \_\_\_\_\_ the piano last weekend.  (彈鋼琴)


\_\_\_\_\_ 4. I \_\_\_\_\_ a movie two days ago.  (看電影)


\_\_\_\_\_ 5. My brother is \_\_\_\_\_.  (生病; 不舒服的)

\_\_\_\_\_ 6. My sister \_\_\_\_\_ the dog last weekend.  (溜狗)


\_\_\_\_\_ 7. Mack \_\_\_\_\_ math every night.  (讀書)


\_\_\_\_\_ 8. I \_\_\_\_\_ my friends yesterday.  (打電話)


\_\_\_\_\_ 9. I want a \_\_\_\_\_ as a birthday present.  (腳踏車)


\_\_\_\_\_ 10. I \_\_\_\_\_ to music last night.  (聽音樂)


(請翻頁繼續作答)

\_\_\_\_\_ 11. Russ is \_\_\_\_\_ for the bus.  (等待)

\_\_\_\_\_ 12. Bill \_\_\_\_\_ his pictures to me yesterday.  (展示)



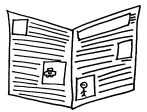












\_\_\_\_\_ 13. I read \_\_\_\_\_ every day.  (報紙)

\_\_\_\_\_ 14. It's 6 o'clock. I am \_\_\_\_\_.  (飢餓的)



\_\_\_\_\_ 15. Stephen \_\_\_\_\_ dinner for me last night.  (煮飯)







Appendix 20 Quick revision list for the gap-fill test: Version A

1.  醫院=hospital
2.  購物=shop
3.  '讀'報紙=read
4.  水果=fruit
5.  公車=bus
6.  煮飯=cook
7.  擦油漆=paint
8.  讀書=study
9.  跳舞=dance
10.  彈吉他=play guitar
11.  打掃=clean
12.  胖的=fat
13.  要求;問=ask
14.  溜狗=walk the dog
15.  看電視=watch TV

Appendix 21 Quick revision list for the gap-fill test: Version B

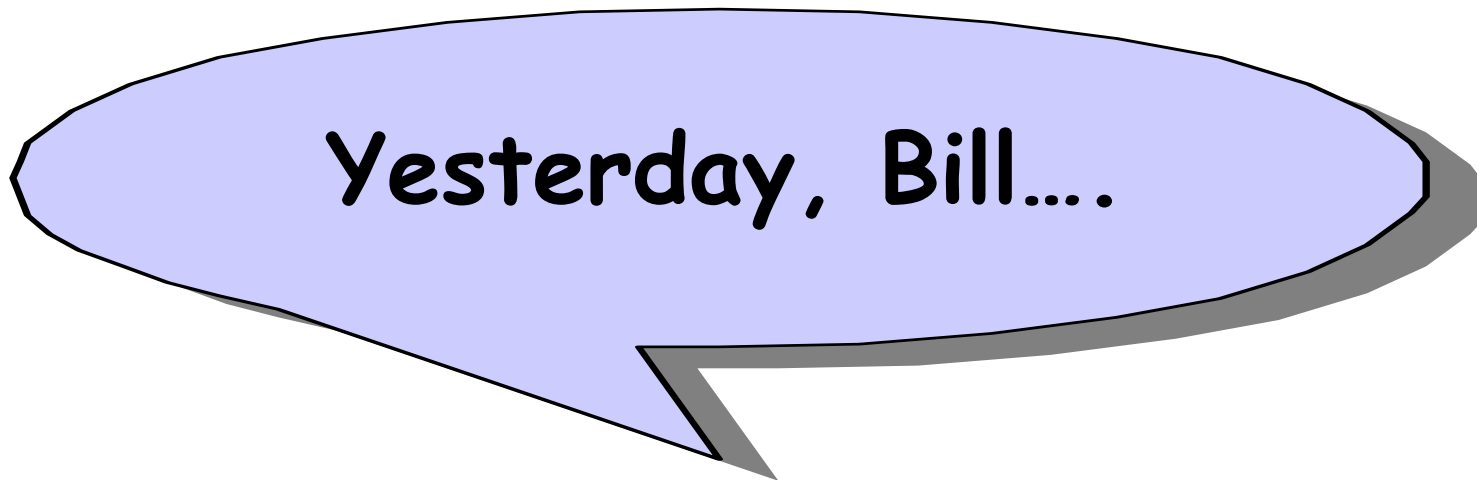
1.  聽音樂=listen to music 2.  腳踏車=bicycle 3.  讀書=study 4.  煮飯=cook

5.  報紙=newspaper 6.  寒冷的=cold 7.  打電話=call/phone 8.  飢餓的=hungry

9.  等待= wait 10.  溜狗=walk dogs 11.  展示=show 12.  彈鋼琴=play the piano

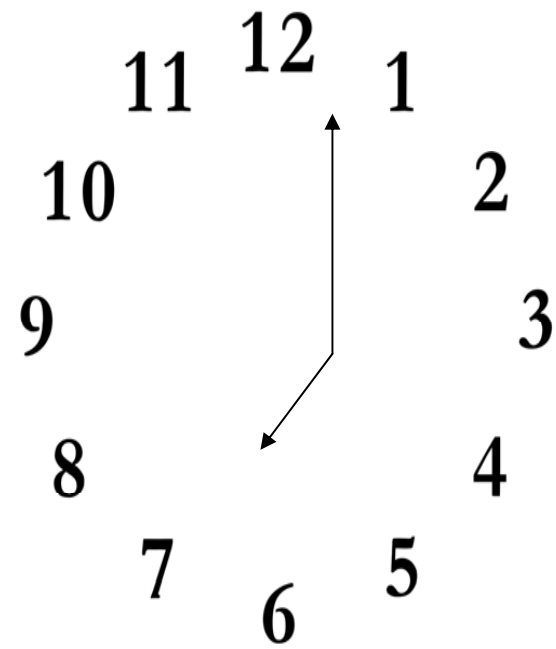
13.  洗臉=wash face 14.  看電影=watch movies 15.  生病的=sick

**Appendix 22 The picture-based narration test: Version A**



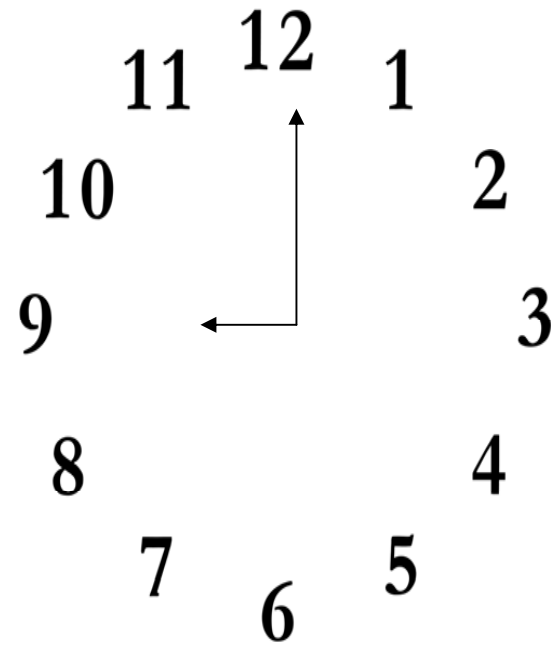


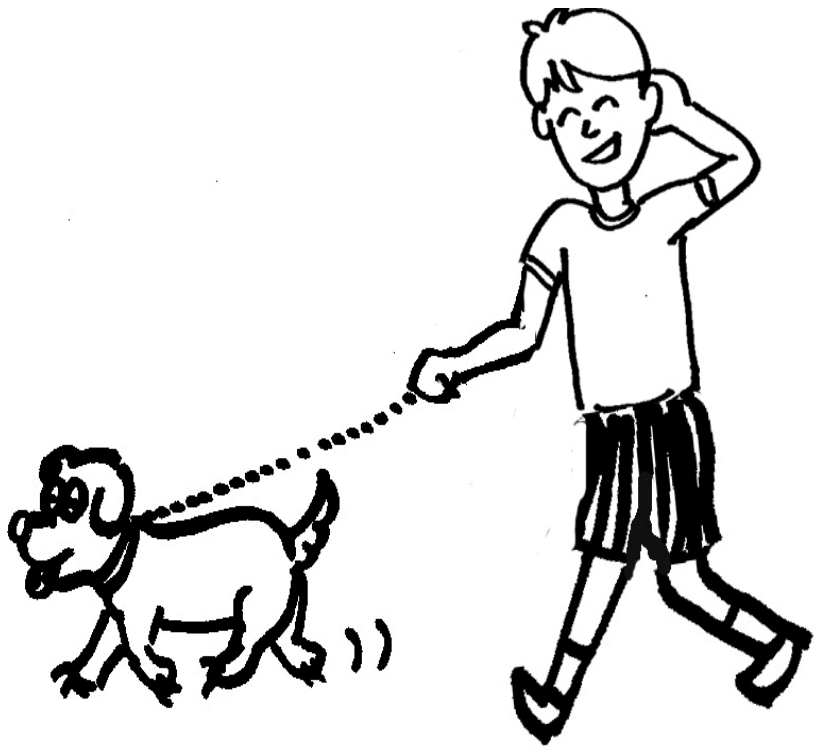
7.00 A.M.



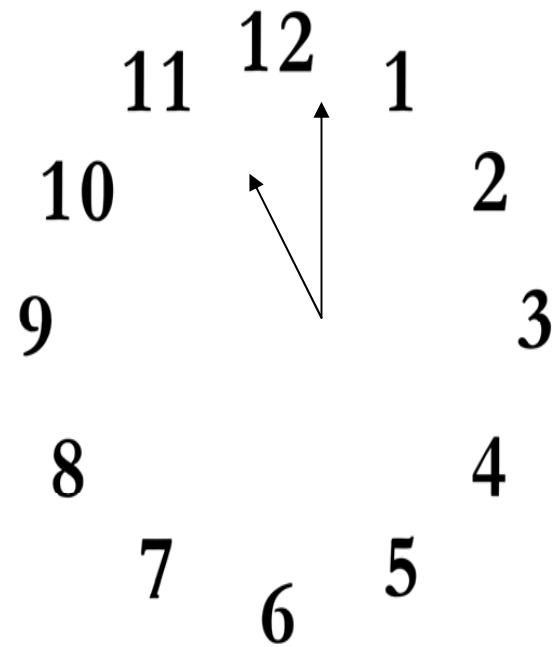


9.00 A.M.



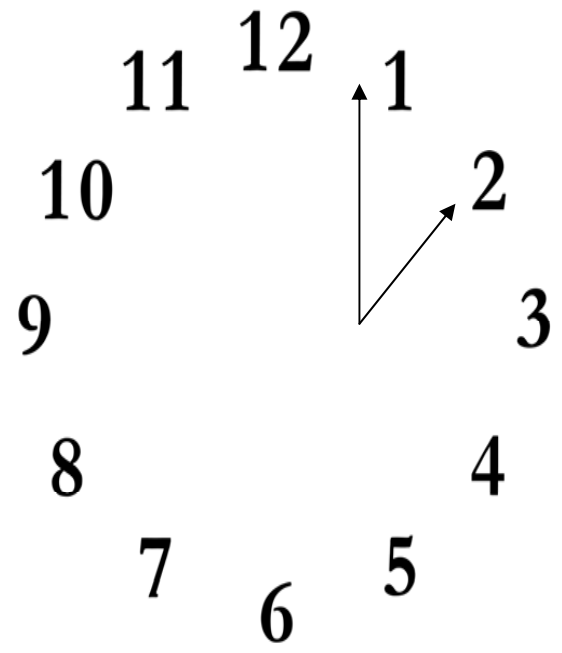


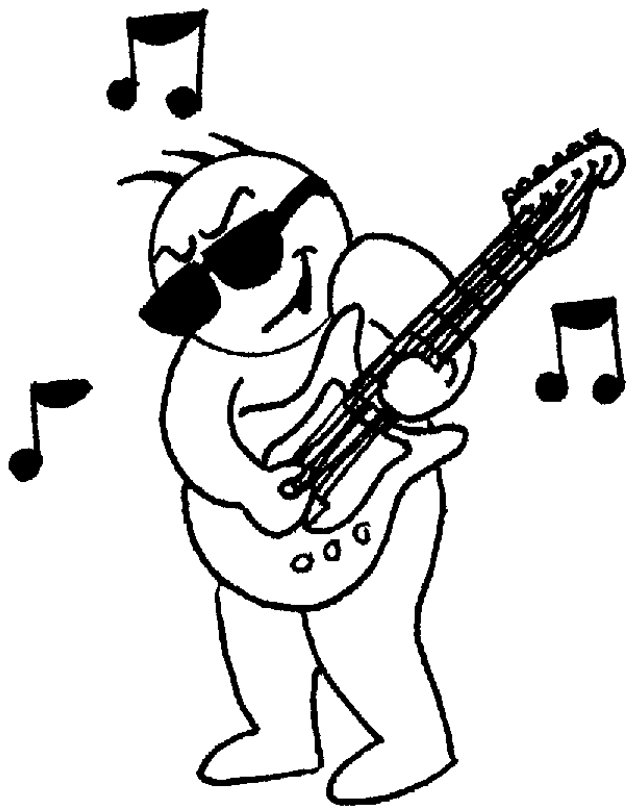
11.00 A.M.



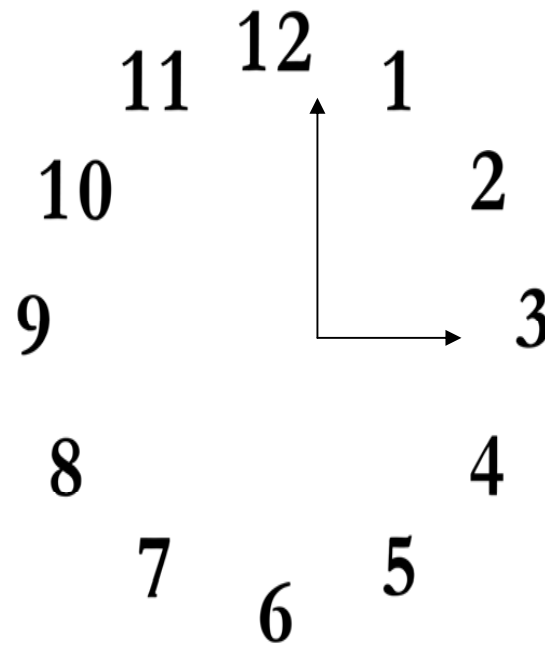


**1.00 P.M.**





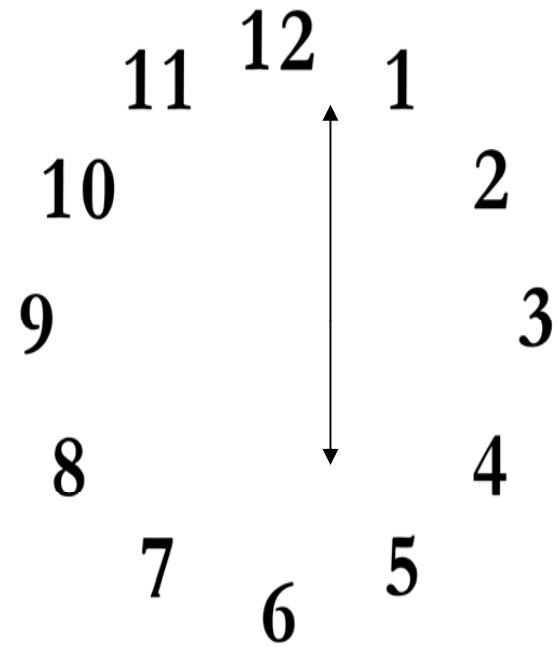
**3. 00 P.M.**





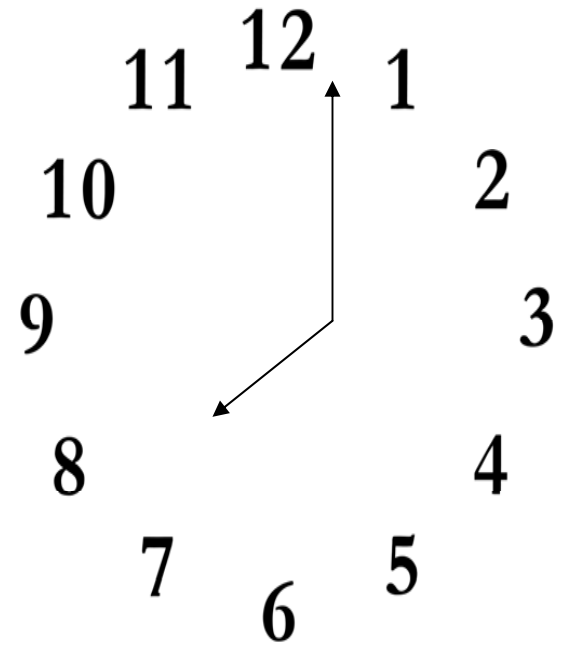


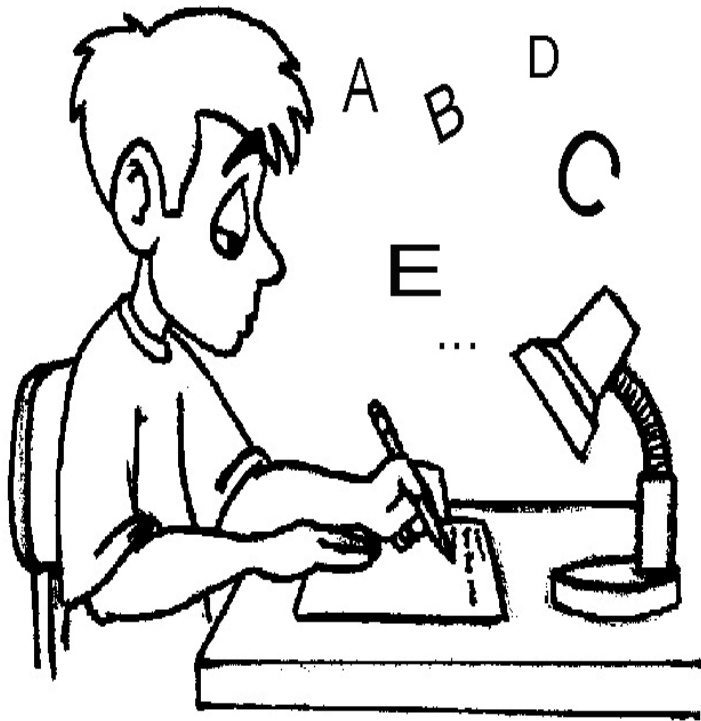
**6.00 P.M.**



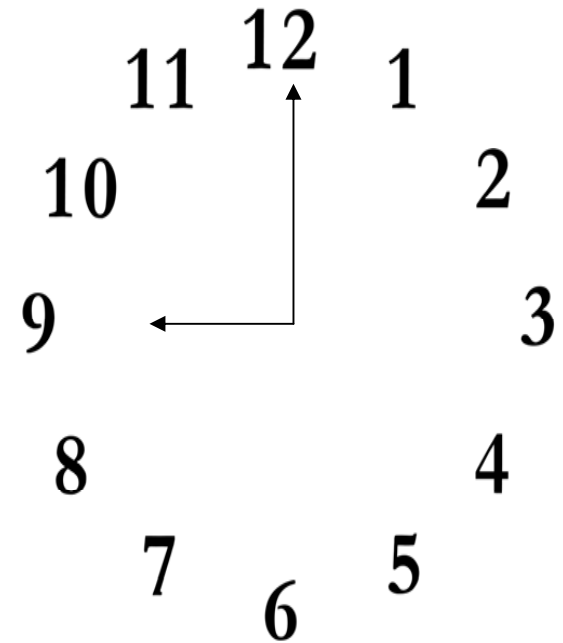


**8.00 P.M.**





**9.00 P.M.**

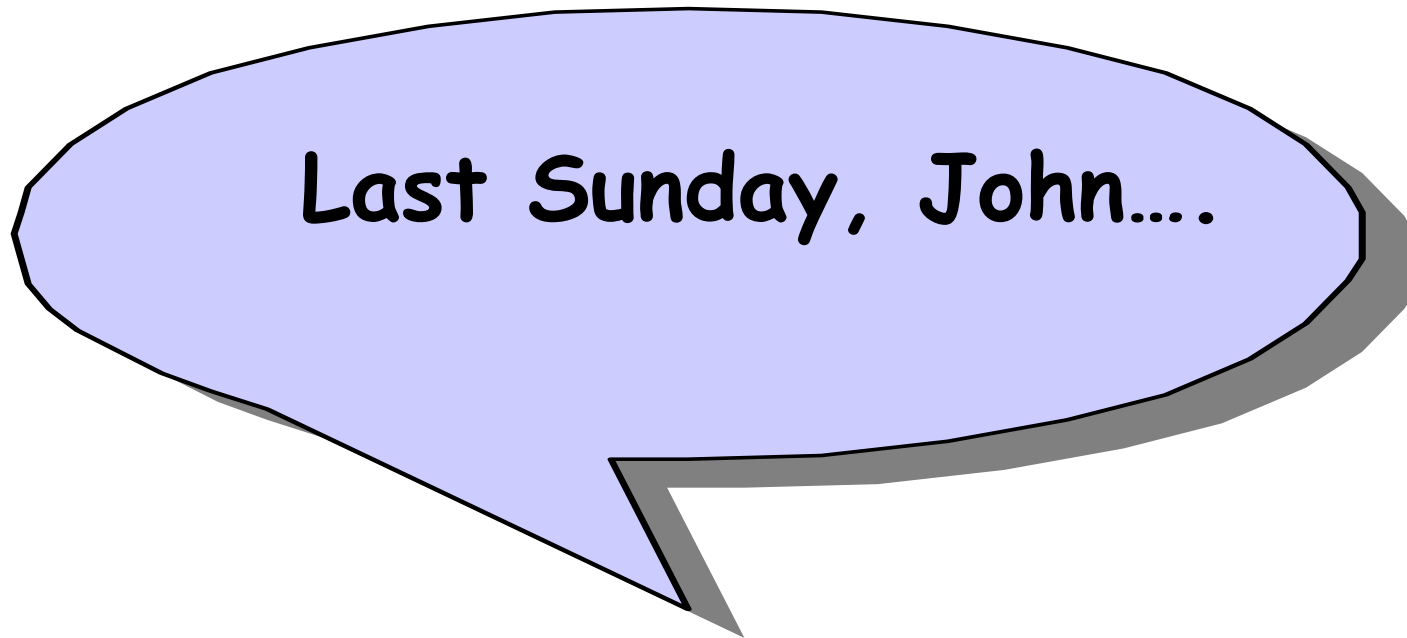




小朋友! 說完了 **Bill** 昨天做的活動, 你能不能也用英文告訴我, 你自己昨天做了哪些事呢?

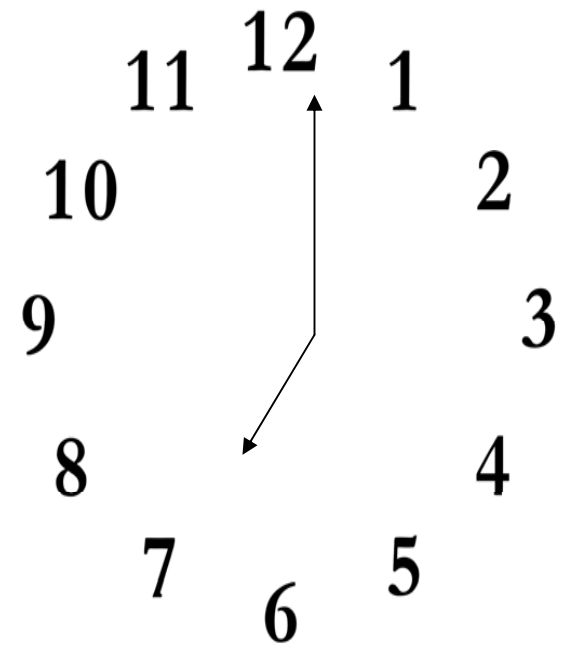
**Could you please tell me in English what you did yesterday?**

**Appendix 23 The picture-based narration test: Version B**



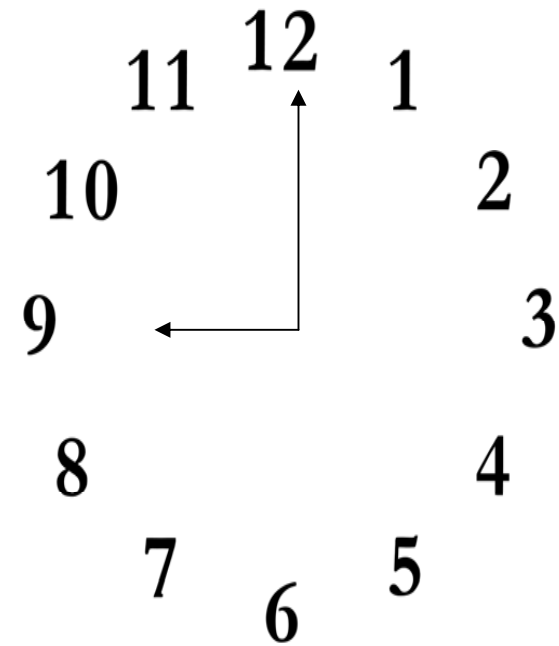


7.00 A.M.



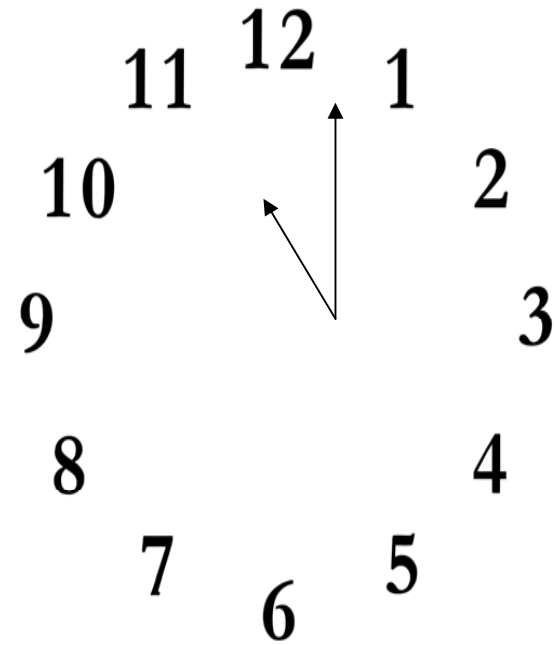


9.00 A.M.

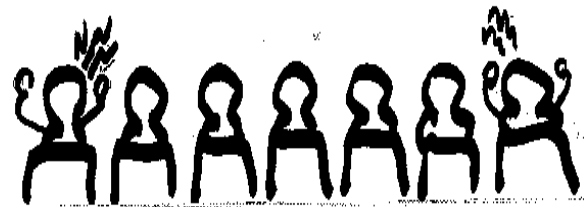
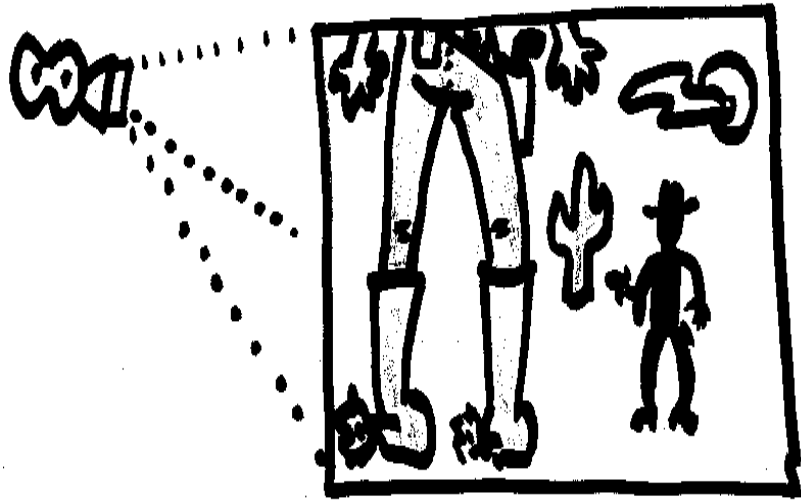




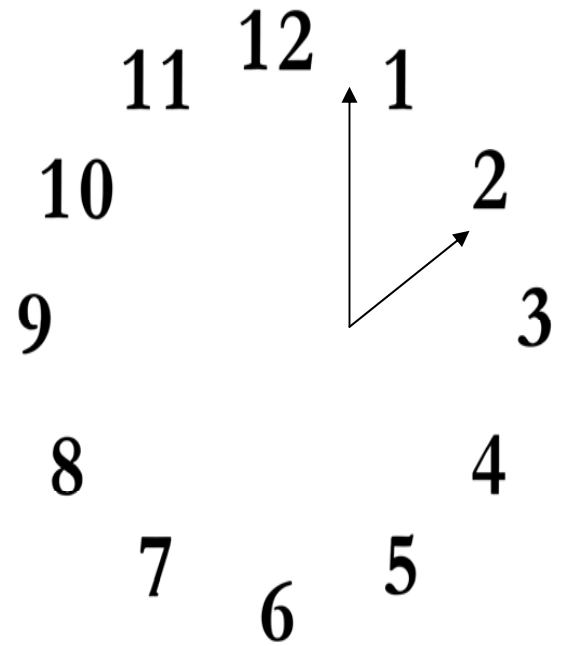
11.00 A.M.





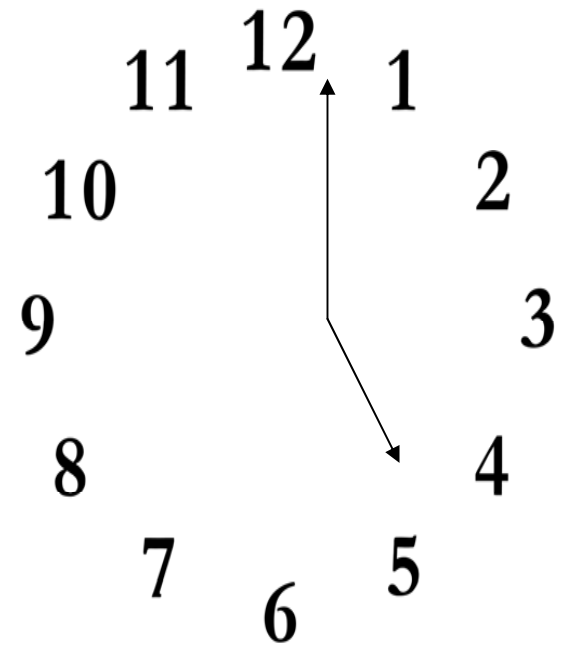


2.00 P.M.



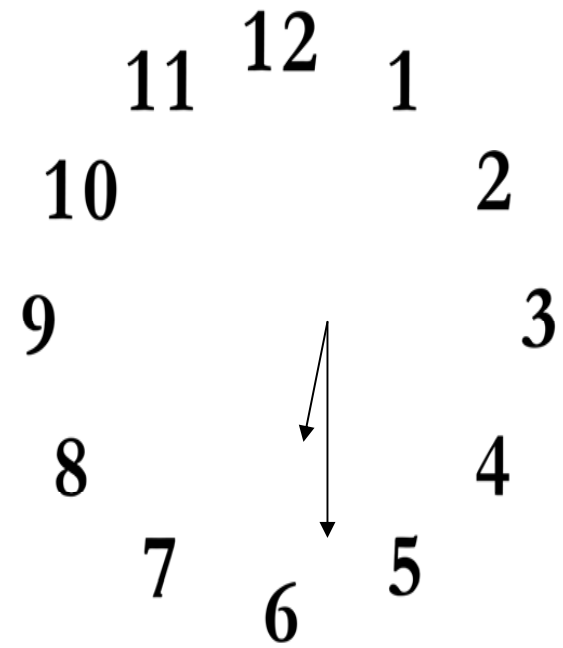


5.00 P.M.



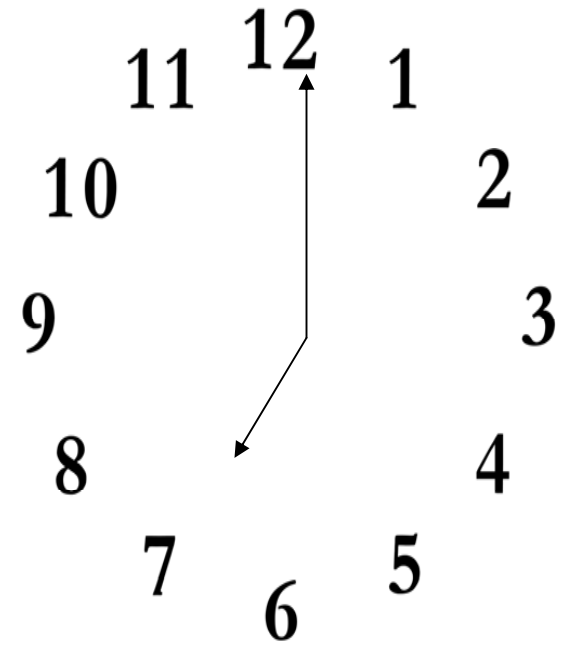


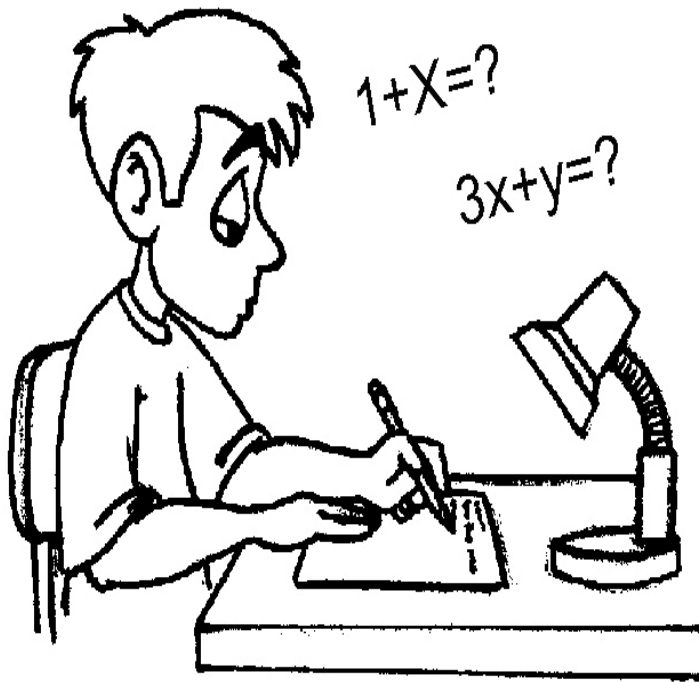
6.30 P.M.



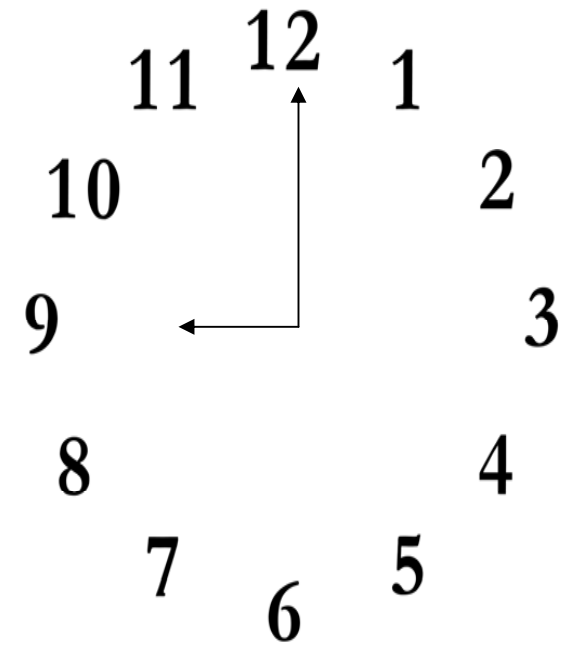


7.00 P.M.



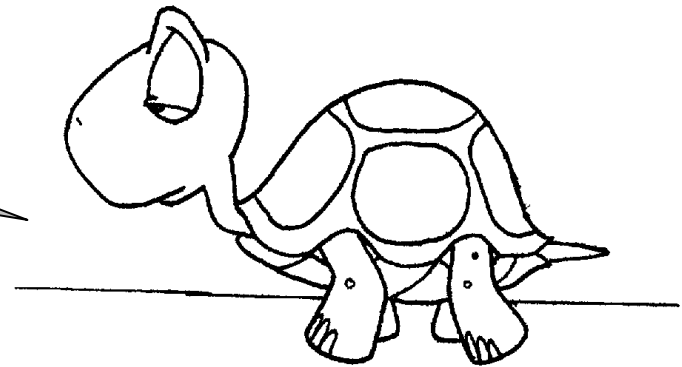


9.00 P.M.




小朋友! 說完了 **John** 上個星期天做的活動, 你能不能再用英文告訴我, 你自己上個星期天做了哪些事呢?

**Could you please tell me in English what you did last Sunday?**



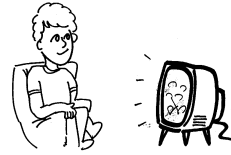
Appendix 24 Quick revision list for the picture-based narration test

1. wash hand = 洗手 
2. clean the teeth = 刷牙 
3. play the guitar = 彈吉他 
4. play basketball = 打籃球 
5. walk the dog = 溜狗 
6. listen to music = 聽音樂 
7. study math = 學習數學 
8. study English = 學習英文 
9. clean the room = 打掃房間 
10. wash dishes = 洗碗 
11. cook dinner = 煮晚餐 
12. watch movies = 看電影 

13. call = 打電話



14. watch TV = 看電視





## Appendix 25 The vocabulary test: Version A

姓名:            班級:            座號:

**※請寫下以下 10 個英文單字的中文意思：**

examples (例題):

a. happy → 快樂的

b. cat → 貓

**1. attend →**

**2. Japan →**

**3. together →**

**4. provide →**

**5. Mah-Jong →**

**6. enjoy →**

**7. pub →**

**8. plan →**

**9. invite →**

**10. discount →**

## Appendix 26 The vocabulary test: Version B

姓名:            班級:            座號:

※請寫下以下 10 個英文單字的中文意思：

*examples (例題):*

a. happy → 快樂的            b. cat → 貓

1. join →

2. Thailand →

3. a lot →

4. offer →

5. museum →

6. practice →

7. trip →

8. learn →

9. organize →

10. the Statue of Liberty →

## Appendix 27 The post-task questionnaire

請你回想一下，當你在做答剛才那 40 題時，你大多是依據什麼來判斷的？請圈選以下選項：

1. 我大部份時間，都是憑感覺做答的(覺的可能是對的，就圈可能對；覺的看起來怪怪的，就圈可能錯)。

2. 我大部份時間，都是依據文法規則去做答的。

如果你選 2 的話，你用了什麼樣的規則？（你可以寫規則，也可以舉例）

---

---

---

3. 其它（如果你是依據別的方式來做答的，就請你寫下來）

---

---

---

-----  
**Translation:**

Could you please recall how you responded to the 40 test items most of the time while taking the task? Please circle one of the following options.

1: I responded to the test items by feeling most of the time.

2: I responded to the test items by using grammar rules.

If you chose 2, what grammar rules were you thinking of? (you can write down the grammar rules or provide examples)

---

---

---

3. others

---

---

---

**Appendix 28 The interview sheet for an interviewer to note down participants' responses in the post-task interview**

姓名:

座號:

班級:

Q1: 問小朋友剛剛在講英文時, 有沒有想到一些文法規則?

A. 有

B. 沒有

Q2: 如果學生回答”有”, 問學生‘那他們想到了哪些文法規則?’ 請學生舉例子說明, 並把學生講的變化及例子寫下

Q3: 如果學生回答, 說他們想到要加 -ed, 再一次確認, ‘剛剛他們在講英文時, 真的有’想到要在動詞後加-ed 或做任何變化’?

-----  
Translation:

Q1: Please ask the student whether or not s/he was thinking of a grammatical rule during the oral test?

A. yes

B. no

Q2: If s/he said ‘yes’, please ask her/him what grammatical rules s/he was thinking of while taking the oral task. You can ask him/her to give a grammatical rule or provide an example containing the grammatical rule.

Q3: If s/he mentioned ‘-ed’, please ask her/him to **confirm** that s/he was thinking of using the ‘-ed’ while taking the oral test.

**Appendix 29 Results of the K-S test for achievement test versions**

<b>Assessments</b>	<b>Populations</b>	<b>N</b>	<b>Statistic</b>	<b>df</b>	<b>Sig.</b>
<b>GJT 1</b>	<b>Class 1</b>	27	.248	27	.000
<b>GJT 2</b>	<b>Class 1</b>	27	.168	27	.048
<b>GJT 1</b>	<b>Class 2</b>	25	.190	25	.020
<b>GJT 2</b>	<b>Class 2</b>	25	.211	25	.005
<b>GAP A1</b>	<b>Class 1</b>	27	.539	27	.000
<b>GAP A2</b>	<b>Class 1</b>	27	.539	27	.000
<b>GAP A1</b>	<b>Class 2</b>	25	.239	25	.001
<b>GAP A2</b>	<b>Class 2</b>	25	.243	25	.001
<b>GAP B1</b>	<b>Class 1</b>	27	.539	27	.000
<b>GAP B2</b>	<b>Class 1</b>	27	.539	27	.000
<b>GAP B1</b>	<b>Class 2</b>	25	.246	25	.000
<b>GAP B2</b>	<b>Class 2</b>	25	.203	25	.010
<b>Voc A</b>	<b>control group</b>	34	.209	34	.001
<b>Voc B</b>	<b>control group</b>	34	.254	34	.000
<b>Picture-narration test A</b>	<b>Subjects in main study at pre-test</b>	21	.334	21	.000
<b>Picture-narration test B</b>	<b>Subjects in main study at pre-test</b>	27	.371	27	.000

**Appendix 30 The results of Levene's test on both versions of achievement test to investigate the validity**

<b>Assessments</b>	<b>Populations</b>	<b>Levene Statistic</b>	<b>Sig.</b>
<b>GJT</b>	<b>C1 vs C2</b>	18.658	.000
<b>Gap-fill test A</b>	<b>C1 vs C2</b>	144.699	.000
<b>Gap-fill test B</b>	<b>C1 vs C2</b>	142.999	.000

**Appendix 31 Results of the K-S test for the achievement tests**

(The GJT, gap-fill, vocabulary, and picture-based narration tests)

Test	Time of test	GROUP	Statistic	df	Sig.
GJT	Pre-test	RA	.101	31	.200
		R	.131	29	.200
		A	.189	30	.008
		C	.175	30	.020
	Post-test	RA	.223	31	.000
		R	.191	29	.009
		A	.126	30	.200
		C	.147	30	.099
	Delayed post-test	RA	.158	31	.046
		R	.228	29	.000
		A	.164	30	.038
		C	.164	30	.038
Gap-fill test	Pre-test	RA	.539	31	.000
		R	----	----	----
		A	----	----	----
		C	----	----	----
	Post-test	RA	.380	31	.000
		R	.367	29	.000
		A	.539	30	.000
	Delayed post-test	RA	.391	31	.000
		R	.400	29	.000
A		.531	30	.000	
Vocabulary test	Pre-test	RA	.257	31	.000
		R	.362	29	.000
		A	.308	30	.000
	Post-test	RA	.207	31	.002
		R	.198	29	.005
		A	.227	30	.000
	Delayed post-test	RA	.151	31	.071
		R	.243	29	.000
		A	.173	30	.023
Picture narration	Pre-test	RA	.281	10	.025
		R	.272	9	.054
		A	.414	9	.000
		C	.414	9	.000
	Post-test	RA	.168	10	.200
		R	.369	9	.001
		A	.471	9	.001
		C	.272	9	.054
	Delayed posttest	RA	.282	10	.024

	R	.519	9	.000
	A	.519	9	.000
	C	.453	9	.000



**Appendix 32 The results of Levene's test on achievement tests**

<b>Assessment</b>		<b>Levene Statistic</b>	<b>df1</b>	<b>df2</b>	<b>Sig.</b>
<b>GJT</b>	Pre-test	.028	3	16	.994
	Post-test	22.121	3	16	.000
	dp	23.160	3	16	.000
<b>Gap</b>	Pre-test	4.092	3	16	.008
	Post-test	76.591	3	16	.000
	dp	78.739	3	16	.000
<b>Picture narration</b>	Pre-test	2.001	3	33	.133
	Post-test	4.991	3	33	.006
	dp	9.951	3	33	.000
<b>Vocabulary test</b>	Pre-test	2.564	2	87	.083
	Post-test	.722	2	87	.489
	dp	1.017	2	87	.366
<b>Structured conversation</b>	Pre-test	1.590	3	33	.211
	Post-test	2.562	3	33	.072
	dp	4.263	3	33	.012

### Appendix 33 The results of parametric tests on the timed GJT

Table 33.1

*Mixed design ANOVA of within-subjects effects of the GJTs*

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
TIME	2760.469	1.590	1736.097	50.379	.000
TIME * GROUP	1747.175	4.770	366.274	10.629	.000

Table 33.2

*Mixed design ANOVA of between-subjects effects of the GJTs*

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
GROUP	873.336	3	291.112	7.299	.000

Table 33.3

*Planned contrasts of within-subjects effects of GJTs*

Source	TIME	Type III Sum of Squares	df	Mean Square	F	Sig.
TIME	Pre vs. Post	3908.619	1	3908.619	54.756	.000*
	Pre vs. Dp	4360.435	1	4360.435	66.223	.000*
	Post vs. Dp	12.353	1	12.353	.455	.501
TIME * GROUP	Pre vs. Post	2544.039	3	848.013	11.880	.000*
	Pre vs. Dp	2598.572	3	866.191	13.155	.000*
	Post vs. Dp	98.915	3	32.972	1.214	.308

Table 33.4

*Planned contrasts of between-subjects effects of timed GJTs*

Contrasts	Std Error	Sig	95% confidence interval	
			upper bound	lower bound
<b>RA vs. C</b>	1.617	.000*	9.521	3.114
<b>R vs. C</b>	1.645	.003*	8.331	1.817
<b>A vs. C</b>	1.631	.601	4.085	-2.374
<b>RA vs. A</b>	1.617	.001*	-2.259	-8.666
<b>R vs. A</b>	1.645	.012*	7.476	.961
<b>RA vs. R</b>	1.632	.447	4.475	-1.987

### Appendix 34 The results of parametric tests on the gap-fill test

Table 34.1

*Mixed design ANOVA of within-subjects effects of the gap-fill test*

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
<b>TIME</b>	108.939	1.423	76.550	27.402	.000
<b>TIME * GROUP</b>	94.429	4.269	22.118	7.917	.000

Table 34.2

*Mixed design ANOVA of between-subjects effects of the gap-fill*

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
<b>GROUP</b>	64.118	3	21.373	9.865	.000

Table 34.3

*Planned contrasts of within-subjects effects of the gap-fill test*

Source	TIME	Type III Sum of Squares	df	Mean Square	F	Sig.
<b>TIME</b>	<b>Pre vs. Post</b>	133.429	1	133.429	29.555	.000*
	<b>Pre vs. Dp</b>	188.623	1	188.623	32.255	.000*
	<b>Post vs. Dp</b>	4.765	1	4.765	3.046	.084
<b>TIME * GROUP</b>	<b>Pre vs. Post</b>	128.890	3	42.963	9.516	.000*
	<b>Pre vs. Dp</b>	152.647	3	50.882	8.701	.000*
	<b>Post vs. Dp</b>	1.749	3	.583	.373	.773

Table 34.4

*Planned contrasts of between-subjects effects of the gap-fill test*

Contrasts	Std Error	Sig	95% confidence interval	
			upper bound	lower bound
<b>RA vs. C</b>	.377	.000*	2.413	.920
<b>R vs. C</b>	.383	.001*	2.081	.563
<b>A vs. C</b>	.380	.748	.875	-.631
<b>RA vs. A</b>	.377	.000*	-.798	-2.291
<b>R vs. A</b>	.383	.002*	1.959	.440
<b>RA vs. R</b>	.380	.366	1.098	-.408

## Appendix 35 The results of parametric tests on the picture-based narration test

Table 35.1

*Mixed design ANOVA of within-subjects effects of the picture-based narration tests.*

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
TIME	4.214	2	2.107	1.355	.265
TIME * GROUP	14.944	6	2.491	1.601	.161

Table 35.2

*Mixed design ANOVA of between-subjects effects of the picture-based narration test.*

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
GROUP	8.294	3	2.765	4.612	.008*

Table 35.3

*Planned contrasts of between-subjects effects of picture narration test*

Contrasts	Std Error	Sig	95% confidence interval	
			upper bound	lower bound
RA v.s. C	.356	.016	1.631	.184
R v.s. C	.365	.615	.928	-.557
A v.s. C	.365	.318	.372	-1.113
RA v.s. A	.356	.001	-.554	-2.002
R v.s. A	.365	.138	1.298	-.187
RA v.s. R	.356	.050	1.574E-03	-1.446

## Appendix 36 The results of parametric tests on the vocabulary test

Table 36.1

*Mixed design ANOVA of within-subjects effects of the vocabulary tests*

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
TIME	33.117	1.607	20.614	13.847	.000
TIME * GROUP	10.613	3.213	3.303	2.219	.084

Table 36.2

*Mixed design ANOVA of between-subjects effects of the vocabulary tests*

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
GROUP	2.661	2	1.330	.602	.550

Table 36.3

*Planned contrasts of within-subjects effects of the vocabulary tests*

TIME	Type III Sum of Squares	df	Mean Square	F	Sig.
Pre vs Post	62.598	1	62.598	22.212	.000*
Pre vs Dp	31.442	1	31.442	25.725	.000*
Post vs Dp	5.311	1	5.311	1.694	.196

**Appendix 37 The results of parametric tests for assessing the validity of achievement tests**

Table 37.1

*Descriptive statistics of Class 1 and Class 2 in assessments to investigate the validity*

<b>Assessment</b>	<b>Population</b>	<b>N</b>	<b>Mean</b>	<b>SD</b>
<b>Timed GJT</b>	Class 1	27	13.78	7.22
	Class 2	25	25.04	11.43
<b>Gap-fill test A</b>	Class 1	27	.33	1.73
	Class 2	25	8.04	6.83
<b>Gap-fill test B</b>	Class 1	27	.30	1.54
	Class 2	25	8.04	6.89

Class 1= grade 6; Class 2=grade 8

Table 37.2

*The results of independent t-test for assessing the validity of the timed GJT and two versions of the gap-fill test*

<b>Assessment</b>	<b>Paired of population</b>	<b>T</b>	<b>Sig. (2-tailed)</b>
<b>Timed GJT</b>	C1 vs. C2	-4.211	.000
<b>Gap-fill test A</b>	C1 vs. C2	-5.482	.000
<b>Gap-fill test B</b>	C1 vs. C2	-5.490	.000



**Appendix 38 The results of parametric tests for assessing the test-retest reliability of achievement tests**

Table 38.1

*Descriptive statistics of Class 1 and Class 2 for the test-retest reliability*

<b>Assessment</b>	<b>Population</b>	<b>N</b>	<b>Mean</b>	<b>SD</b>
Timed GJT1	Class 1	27	13.78	7.22
Timed GJT2			14.52	8.12
Timed GJT1	Class 2	25	25.04	11.43
Timed GJT2			26.36	12.40
Gap-fill test A1	Class 1	27	.33	1.73
Gap-fill test A2			.37	1.92
Gap-fill test B1	Class 1	27	.30	.154
Gap-fill test B2			.37	1.92
Gap-fill test A1	Class 2	25	8.04	6.83
Gap-fill test A2			8.80	6.73
Gap-fill test B1	Class 2	25	8.04	6.89
Gap-fill test B2			8.24	6.94

Table 38.2

*The results of dependent t-test for test-retest reliability*

<b>Paired of assessments</b>	<b>Population</b>	<b>t</b>	<b>Sig. (2-tailed)</b>
Timed GJT1 vs. GJT2	Class 1	-.875	.390
Timed GJT1 vs. GJT2	Class 2	-.856	.400
Gap-fill test A1 vs. A2	Class 1	-1.000	.327
Gap-fill test B1 vs. B2	Class 1	-1.000	.327
Gap-fill test A1 vs. A2	Class 2	-1.507	.145
Gap-fill test B1 vs. B2	Class 2	-.679	.503

Table 38.3

*The Pearson correlational results in assessing the rest-retest reliability*

<b>Paired of Assessments</b>	<b>Classes</b>	<b>N</b>	<b>Pearson <i>r</i></b>	<b>Sig. (2-tailed)</b>
<b>Timed GJT1 vs. GJT2</b>	Class 1	27	.842**	.000
<b>Timed GJT1 vs. GJT2</b>	Class 2	25	.794**	.000
<b>Gap-fill test A1 vs. A2</b>	Class 1	27	1.000**	.000
<b>Gap-fill test B1 vs. B2</b>	Class 1	27	1.000**	.000
<b>Gap-fill test A1 vs. A2</b>	Class 2	25	.931**	.000
<b>Gap-fill test B1 vs. B2</b>	Class 2	25	.977**	.000

**Appendix 39 The results of parametric tests for assessing the comparability of two versions of achievement tests.**

Table 39.1

*Descriptive statistics for the comparability of the two versions of assessments*

<b>Assessment</b>	<b>Population</b>	<b>N</b>	<b>Mean</b>	<b>SD</b>
<b>Gap-fill test A</b>	Class 1	27	.33	1.74
<b>Gap-fill test B</b>			.30	1.54
<b>Gap-fill test A</b>	Class 2	25	8.04	6.83
<b>Gap-fill test B</b>			8.04	6.89
<b>Picture test A</b>	Subjects at pre-test	21	.62	.86
<b>Picture test B</b>		27	.93	1.66
<b>Vocabulary test A</b>	Control group	34	1.59	1.52
<b>Vocabulary test B</b>			1.47	1.73

Table 39.2

*The results of t-test for the two versions of the achievement assessments*

<b>Paired of Assessments</b>	<b>Classes</b>	<b>N</b>	<b>t</b>	<b>Sig. (2tailed)</b>
<b>Gap-fill test A vs. B</b>	Class 1	27	1.000	.327
<b>Gap-fill test A vs. B</b>	Class 2	25	.000	1.000
<b>Picture test A vs. B</b>	Subjects at pre-test	48	-.768	.447
<b>Vocabulary test A vs. B</b>	control group	34	.849	.402

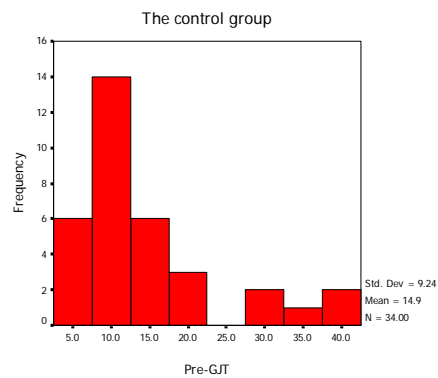
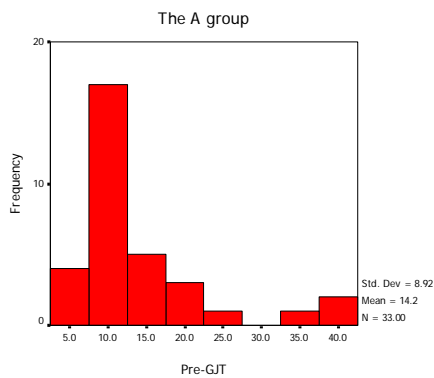
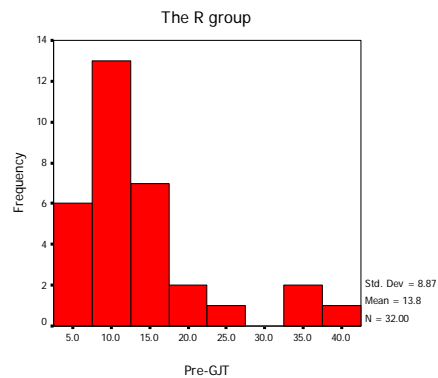
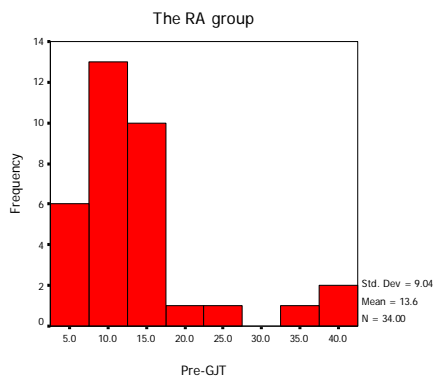
## Appendix 40 The K-S test results, the histograms, and boxplots of the timed GJT including 13 outliers

### \* The K-S test results

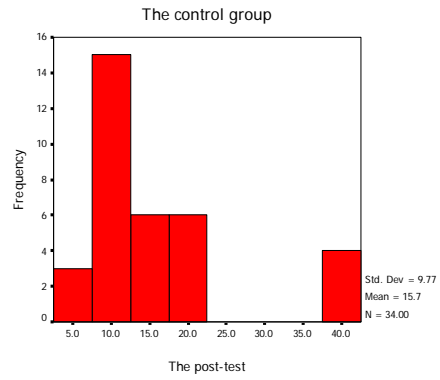
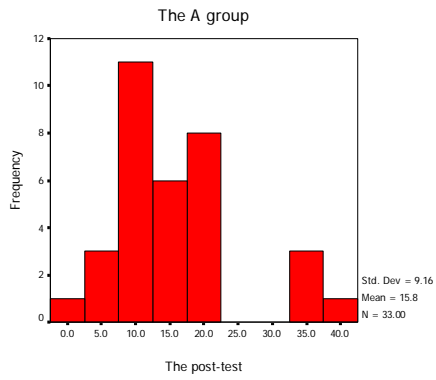
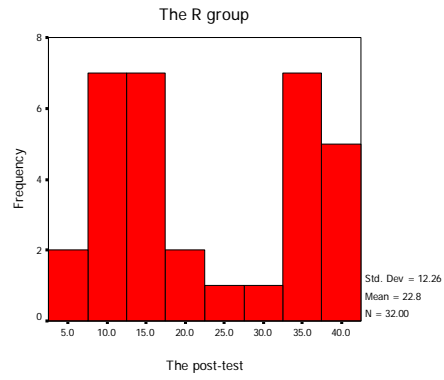
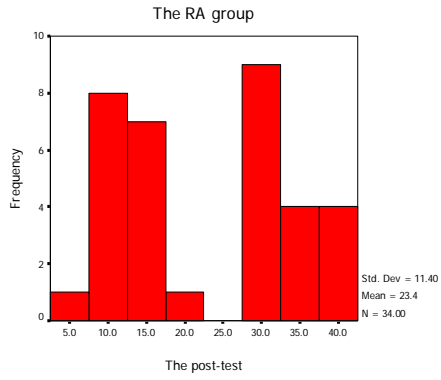
Test	Time of test	GROUP	Statistic	df	Sig.
<b>GJT</b>	<b>Pre-test</b>	RA	.126	34	.000
		R	.209	32	.001
		A	.236	33	.000
		C	.213	34	.000
	<b>Post-test</b>	RA	.212	34	.000
		R	.184	32	.007
		A	.134	33	.144
		C	.176	34	.009
	<b>Delayed post-test</b>	RA	.160	34	.028
		R	.218	32	.001
		A	.202	33	.001
		C	.261	34	.000

### \* The histograms

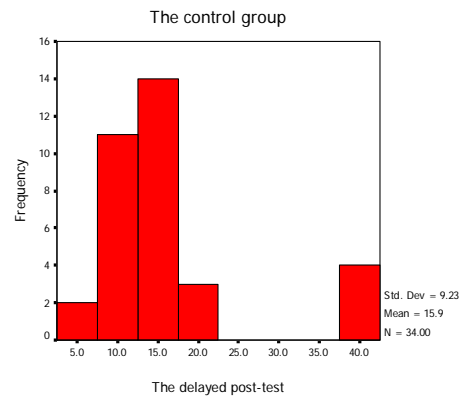
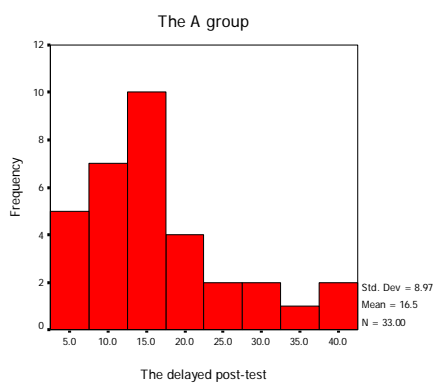
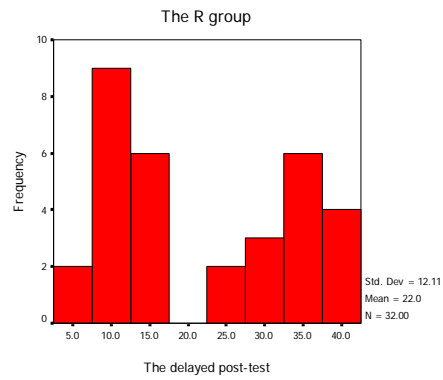
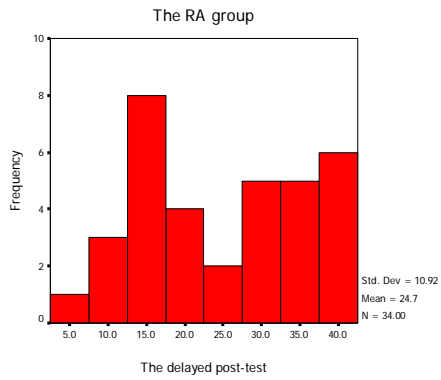
#### The pre-test



#### The post-test

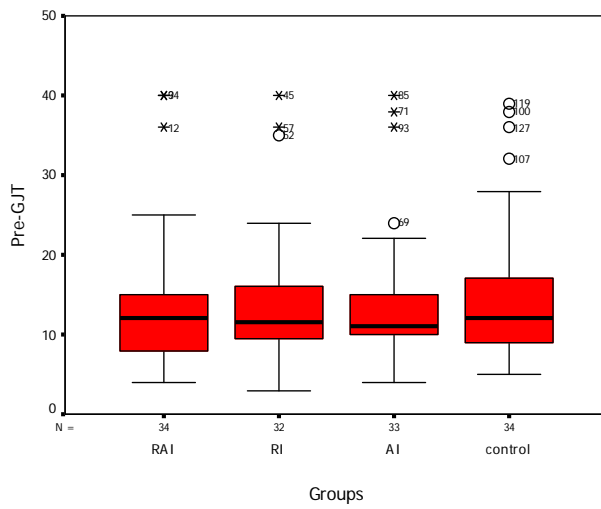


### The delayed post-test

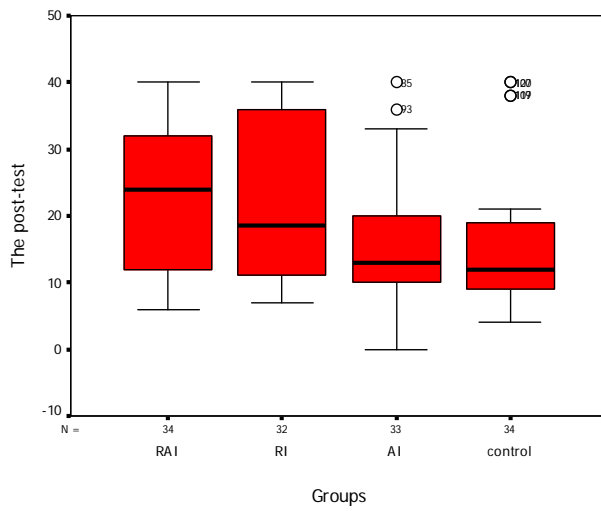


### \* The boxplots

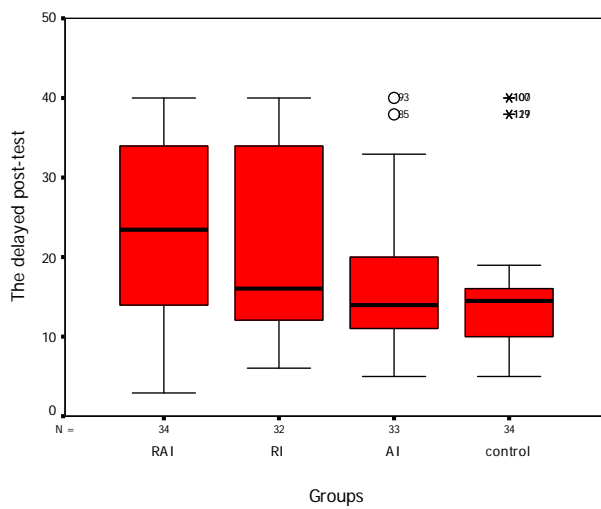
### The pre-test



### The post-test



### The delayed post-test



## Appendix 41 A sample transcription of a structured conversation

**Test administrator (T):** xxx, 你剛剛是告訴我 Bill 他昨天做的一些事情, 那你可不可以想一想, 用英文告訴我, 你昨天你自己做了哪些事情?

**Student A (S):** Yesterday, I play computer.

**T:** computer! And?

**A:** and I go running.

**T:** Go running?

**A:** Yes,

**T:** Yes, great.

**A:** and... I singing.

**T:** .... Oh, singing. 唱歌?

**A:** mm

**T:** anything else? 還有什麼其它的嗎?

**A:** sleeping

**T:** okay, 還有嗎? 沒關係, 想一想?

**A:** hmm.. Yesterday, I go to my aunt's house.

**T:** Ah! 去你伯母的家,

**A:** hmm...

**T:** 好, 還有嗎?

**A:** 沒有了

-----  
**Test administrator (T):** xxx (student A's name), you just told me about some activities that Bill did yesterday. Could you please think carefully and tell me what you did yesterday in English (in Mandarin)?

**Student A (S):** Yesterday, I play computer.

**T:** computer! And?

**A:** and (paused, cleared her throat) I go running.

**T:** Go running?

**A:** Yes,

**T:** Yes, great.

**A:** and... (paused about 2 seconds) I singing.

**T:** ... (paused about 1 second). Oh, singing, and then repeated 'singing' in Mandarin (to make sure she intended to say 'singing')?

**A:** mm

**T:** anything else? (And then repeated "anything else?" in Mandarin)

**A:** sleeping

(both T and A laughed)

**T:** okay, anything else (in Mandarin)? No hurry, please take your time to think about it (in Mandarin).

**A:** hmm.....(thinking and paused about 1 second, and cleared her throat) Yesterday (thinking and paused about 2 seconds), I go to my aunt's house.

**T:** Ah! go to your aunt's house (in Mandarin)!

**A:** mm (imply yes)...

**T:** Okay, anything else (in Mandarin)?

**A:** No (in Mandarin)

Task finished.



## References

- Allen, L.Q. (2000). Form-meaning connections and the French causative: An experiment in processing instruction. *Studies in Second Language Acquisition*, 22, 69-84.
- Anderson, J. (1993). Is a communicative approach practical for teaching English in China? Pros and cons. *System*, 21, 4, 471-480.
- Baddeley, A.D. (1986). *Working Memory*. Oxford: Oxford University Press.
- Barcroft, J. & VanPatten, B. (1997). Acoustic salience of grammatical forms: the effect of location, stress, and boundedness on Spanish L2 input processing. In W. Glass & A. Perez-Leroux (Eds), *Contemporary Perspectives on the Acquisition of Spanish, volume 2: Production Processing, and Comprehension (pp. 109-121)*. Somerville, MA: Cascadilla Press.
- Bardovi-Harlig, K. (1992). The use of adverbials and natural order in the development of temporal expression. *International Review of Applied Linguistics*, 30, 299-320.
- Batstone, R. (2002). Making sense of new language: A discourse perspective. *Language Awareness*, 11, 1, 14-29.
- Benati, A. (2001). A comparative study of the effects of processing instruction and output-based instruction on the acquisition of the Italian future tense. *Language Teaching Research*, 5, 2, 95-127.
- Benati, A. (2004a). The effects of structured input activities and explicit information on the acquisition of the Italian future tense. In B. VanPatten (Ed.), *Processing Instruction: Theory, Research, and Commentary (pp. 207-226)*. Mahwah, NJ: Erlbaum.
- Benati, A. (2004b). The effects of processing instruction and its components on the acquisition of gender agreement in Italian. *Language Awareness*, 13, 2, 67-80.
- Benati, A. (2005). The effects of processing instruction, traditional instruction and meaning-output instruction on the acquisition of the English past simple tense. *Language Teaching Research*, 9, 1, 67-93.

- Bialystok, E. (1979). Explicit and implicit judgments of L2 grammaticality. *Language Learning*, 29, 81-103.
- Bialystok, E. (1982). On the relationship between knowing and using forms. *Applied Linguistics*, 3, 181-206.
- Blau, E. (1990). The effect of syntax, speed and pauses on listening comprehension, *TESOL Quarterly*, 24, 746-753.
- Bley-Vroman, R. (1986). Hypothesis testing in second language acquisition theory, *Language Learning*, 36, 353-376.
- Bransdorfer, R. (1989). Processing function words in input: Does meaning make a difference? Paper presented at the annual meeting of the American Association of Teachers of Spanish and Portuguese, San Antonio.
- Bransdorfer, R. (1991). Chap. 2: Review of the literature. *Communicative value and linguistic knowledge in second language oral input processing*. Unpublished doctoral thesis, University of Illinois at Urbana-Champaign.
- Broadbent, D.A. (1958). *Perception and communication*. Oxford: Pergamon.
- Brown, J.D. (1996). *Testing in language programs*. Englewood Cliffs, NJ: Prentice Hall.
- Bryman, A. & Cramer, D. (2001). *Quantitative data analysis with SPSS release 10 for Windows: a guide for social scientists*. Hove: Routledge.
- Butler, Y. (2002) Second language learners' theories on the use of English articles. *Studies in Second Language Acquisition*, 24, 451-480.
- Cadierno, T., Glass, W.R., Lee, J.F., & VanPatten, B. (1991). Processing tense in second language input: Lexical cues versus grammatical cues. The University of Illinois at Urbana-Champaign.
- Cadierno, T. (1995). Formal instruction from a processing perspective: An investigation into the Spanish past tense. *Modern Language Journal*, 79, 179-193.
- Cadierno, T & Lund, K. (2004). Cognitive linguistics and second language acquisition:

- Motion events in a typological framework. In B. VanPatten (Eds.), *Form-Meaning Connections in Second Language Acquisition*. (pp. 139-154). Mahwah, NJ: Lawrence Erlbaum Associates.
- Campbell, D.T. & Stanley, J.C. (1963). *Experimental and Quasi-experimental Designs for Research*. Chicago: R. McNally.
- Canale, M. & Swain, M. (1980) Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1, 1-47.
- Carmines, E.G. & Zeller, R.A. (1979). *Reliability and Validity Assessment*. Beverly Hills: Sage Publications.
- Carroll, S.E. (1999). Putting "input" in its proper place. *Second Language Research*, 15, 4, 337-388.
- Carroll, S. & Swain M. (1993). Explicit and implicit negative feedback: An empirical study of the learning of linguistic generalizations. *Studies in Second Language Acquisition* 15, 357-86.
- Cattell, R.B., & Scheuberger, J.M. (1978). *Personality theory in action*. Champaign, IL: IPAT.
- Cheng, A.C. (2002). The effects of processing instruction on the acquisition of *ser* and *estar*. *Hispania*, 85, 308-323.
- Cheng, A.C. (2004). Processing instruction and Spanish *Ser* and *Estar*: Forms with semantic-aspectual values. In B. VanPatten (Ed.), *Processing Instruction: Theory, Research, and Commentary* (pp. 119-142). Mahwah, NJ: Erlbaum.
- Cobb, P., Confrey J., diSessa A., Lehrer R. & Schauble L. (2003). Design experiments in educational research. *Educational Researcher*. 32, 1, 9-13.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nded.). Hillsdale, NJ: Erlbaum.
- Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, 45, 1304-1312.

- Cohen, L., Manion, L. & Morrison, K. (2000). *Research Methods in Education*. (5<sup>th</sup> ed.) London: Routledge Falmer.
- Collentine, J. (1998). Processing instruction and the subjunctive. *Hispania*, 81, 3, 576-587.
- Collentine, J. (2002). On the acquisition of the subjunctive and authentic processing instruction: A response to Farley. *Hispania*, 85, 4, 879-888.
- Collentine, J. (2004). Commentary: Where PI research has been and where it should be going? In B. VanPatten (Ed.), *Processing Instruction: Theory, Research, and Commentary* (pp. 169-181). Mahwah, NJ: Erlbaum.
- Corder, S. P. (1967). The significance of learners' errors. *International Review of Applied Linguistics*, 5, 161-169.
- Dale, P.S. & Fenson, L. (1996). Lexical development norms for young children. *Behavior Research Methods, Instruments, and Computers*, 28, 125-127.
- Davies, W.D. & Kaplan, T.I. (1998). Native speaker vs L2 learners grammaticality judgments. *Applied Linguistics*, 19, 2, 183-203.
- de Graaff, R. (1997). *Differential effects of explicit instruction on second language acquisition*. The Hague: Holland Institution of Generative Linguistics.
- de Jong, N.D. (2005). Can second language grammar be learned through listening? An experiment study. *Studies in Second Language Acquisition*. 27, 2, 205-234.
- DeKeyser, R.M. (1995). Learning second language grammar rules: An experiment with a miniature linguistic system. *Studies in Second Language Acquisition*, 17, 379-410.
- DeKeyser, R.M. (2003). Implicit and explicit learning. In C. J. Doughty and M.H. Long (Eds.), *The Handbook of Second Language Acquisition* (pp. 313-348). Oxford, England: Blackwell.
- DeKeyser, R.M. (2005). What makes learning second-language grammar difficult? A review of issues. *Language Learning*, 55 (supplement), 1, 1-25.

DeKeyser, R.M., Salaberry, R., Robinson, P.J. & Harrington, M. (2002). What gets processed in processing instruction? A commentary on Bill VanPatten's "Update." *Language Learning*, 52, 4, 805-823.

DeKeyser, R.M. & Sokalski, K.J. (1996). The differential role of comprehension and production practice. *Language Learning*, 46, 613-642.

Dienes & Perner (1999). A theory of implicit and explicit knowledge. *Behavioral and Brain Sciences*, 22, 5, 745-808.

Dienes, Z., Broadbent, D. & Berry, D. (1991). Implicit and explicit knowledge bases in artificial grammar learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17, 5, 875-887.

Doughty, C. (1991). Second language instruction does make a difference: Evidence from an empirical study of L2 relativization. *Studies in Second Language Acquisition*, 13, 431-469.

Doughty, C. (2004). Commentary: When PI is focus on form it is very, very good, but when it is focus on forms.... In B. VanPatten (Ed.), *Processing instruction: Theory, research, and commentary* pp. 257-270. Mahwah, NJ: Erlbaum.

Douglas, D. (2001). Performance consistency in second language acquisition and language testing research: a conceptual gap. *Second Language Research*, 17,4, 442-456.

Dulay, H. & Burt, M. (1978). Some remarks on creativity in language acquisition. In W. C. Ritchie(Ed.), *Second language acquisition research : issues and implications*.(pp. 65-89). New York: Academic Press.

Ellis, N. (2004). The processes of second language acquisition. In B. VanPatten(Eds.), *Form-Meaning Connections in Second Language Acquisition*.(pp. 49-76). Mahwah, NJ: Lawrence Erlbaum Associates.

Ellis, N. (2005). At the interface: Dynamic interactions of explicit and implicit language knowledge. *Studies in Second Language Acquisition*, 27, 2, 305-352.

Ellis, N. (2006). Selective attention and transfer phenomena in L2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and

perceptual learning. *Applied Linguistics*, 27, 2, 164-194.

Ellis, N. (2007). The associative-cognitive CREED. In B. VanPatten & J. Williams (Eds), *Theories in Second Language Acquisition : An Introduction*. (pp. 7-95). Mahwah, NJ: Lawrence Erlbaum Associates.

Ellis, N. (2008). Temporal cognition and temporal language the first and second times around. Commentary on McCormack and Hoerl. *Language Learning*, 58:Suppl. 1, December , 115-121.

Ellis, R. (1991). Grammaticality judgments and second language acquisition. *Studies in Second Language Acquisition*, 13, 161-186.

Ellis, R. (1993). The structural syllabus and second language acquisition. *TESOL Quarterly*, 27, 1, 93-113.

Ellis, R. (2002). Does form-focused instruction affect the acquisition of implicit knowledge? *Studies in Second Language Acquisition*, 24, 2, 223-236.

Ellis, R. (2004). The definition and measurement of explicit knowledge. *Language Learning*, 54, 227-275.

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A Psychometric study. *Studies in Second Language Acquisition*, 27, 2, 141-172.

Ellis, R. (2006). Modelling learning difficulty and second language proficiency: The differential contributions of implicit and explicit knowledge. *Applied Linguistics*, 27, 3, 431-463.

Ellis, R. & Loewen, S. (2007) Confirming the operational definitions of explicit and implicit knowledge in Ellis (2005): Responding to Isemonger. *Studies in Second Language Acquisition*, 29, 119-126.

Ellis, R., Loewen, S., Elder, C., Erlam, R., Philp, J & Reinders, H. (2009). *Implicit and Explicit Knowledge in Second Language Learning, Testing and Teaching*. Bristol: Multilingual Matters.

Erlam, R. (2003). Evaluating the relative effectiveness of structured-input and output-

based instruction in foreign language learning: Results from an experimental study. *Studies in Second Language Acquisition*, 25, 4, 559-582.

Farley, A. (2001). Authentic processing instruction and the Spanish subjunctive. *Hispanis*, 84, 289-99.

Farley, A. (2002) Processing instruction, communicative value, and ecological validity: A response to Collentine's Defense. *Hispania*, 85, 4, 889-895.

Farley, A. (2004a). The relative effects of processing instruction and meaning-based output instruction. In B.VanPatten (Ed.), *Processing Instruction: Theory, Research, and Commentary* (pp. 143-168). Mahwah, NJ: Erlbaum

Farley, A. (2004b). Processing instruction and the Spanish subjunctive: Is explicit information needed? In B.VanPatten (Ed.), *Processing Instruction: Theory, Research, and Commentary* (pp. 227-240). Mahwah, NJ: Erlbaum.

Fernández, C. (2008). Reexamining the role of explicit information in processing instruction. *Studies in Second Language Acquisition*, 30, 277-305.

Field, A. (2005). *Discovering Statistics Using SPSS*. (2<sup>nd</sup> ed) London: Sage Publications.

Gass, M. (1994). The reliability of L2 grammaticality judgments. In E. Tarone, S. Gass, & A. Cohen (Eds), *Research methodology in second-language acquisition* (pp.303-322). Hillsdale, NJ: Erlbaum.

Gass, S.M. (2004). Context and second language acquisition. In B. VanPatten(Ed.), *Form-Meaning Connections in Second Language Acquisition*(pp.77-90). Mahwah, NJ: Lawrence Erlbaum Associates.

Gass, S.M., Svetics, I. & Lemelin, S. (2003). Differential effects of attention. *Language Learning*, 53, 3, 497-545.

Gathercole, S. (2008). *Working Memory and Learning: A Practical Guide for Teachers*. Los Angeles, CA; London: Sage Publications

Glass, W.R. (1994). Paper delivered at the annual meeting of the American Association for Applied Linguistics, Baltimore.

- Goldschneider, J. & DeKeyser, R. (2001). Exploring the “natural order of L2 morpheme acquisition” in English: A meta-analysis of multiple determinants. *Language Learning*, 51, 1-50.
- Gorard, S. (2002a). Can we overcome the methodological schism? Four models for combining qualitative and quantitative evidence. *Research Papers in Education*, 17, 4, 345-361.
- Gorard, S. (2002b). Political control: a way forward for educational research? *British Journal of Educational Studies*. 50, 3, 378-389.
- Gorard, S., Roberts, K. & Taylor, C. (2004). What kind of creature is a design experiment? *British Educational Research Journal*, 30, 4, 577-590.
- Goss, N., Zhang, H. & Lantolf, J.P. (1994). Two heads may be better than one: Mental activity in second-language grammaticality judgments. In E. Tarone, S. Gass, & A. Cohen (Eds), *Research methodology in second-language acquisition* (pp.263-286). Hillsdale, NJ: Erlbaum.
- Green, P. & Hecht, K. (2002) Implicit and explicit grammar: An empirical study. *Applied Linguistics*, 13, 2, 168-184.
- Hammersley, M. (2001). *Some questions about evidence-based practice in education*. Presentation to BERA, Leeds, 13-15 September, 2001.
- Han, Y. & Ellis, R. (1998). Implicit knowledge, explicit knowledge and general language proficiency. *Language Teaching Research*, 2, 1-23.
- Harley, B. (1992). Patterns of second language development in French immersion. *Journal of French Language Studies*, 2, 159-183.
- Harley, B. & Swain, M. (1984). The interlanguage of immersion students and its implications for second language teaching. In A. Davies, C. Criper, & A. Howatt (Eds.), *Interlanguage* (pp. 291-311). Edinburgh, UK: Edinburgh University Press.
- Harrington, M. (2004). Commentary: Input processing as a theory of processing input. In B. VanPatten (Ed.), *Processing Instruction: Theory, Research, and Commentary* (pp. 79-92). Mahwah, NJ: Erlbaum.



- Hiebert, J., Gallimore, R. & Stigler, J. (2002). A knowledge base for the teaching profession: What would it look like and how can we get one? *Educational Researchers*, 31, 3-15.
- Hsu, S. (2007). *High school English teachers' beliefs in grammar teaching and their classroom practices: A case study*. (Unpublished Master Dissertation). National Chung-Cheng University, Taiwan.
- Hu, G. (2002). Psychological constraints on the utility of metalinguistic knowledge in second language production. *Studies in Second Language Acquisition*, 24, 3, 347-386.
- Huang, C. (2003). *Comparing the effects of two grammar pedagogies on the learning of English grammar for junior high school students in Taiwan: Communicative focus on form and traditional grammar instruction*. (Unpublished Master Dissertation). Ming-Chuan University, Taiwan.
- Hulstijn, J.H. (1997). Second language acquisition research in the laboratory: possibilities and limitations. *Studies in Second Language Acquisition*, 19, 131-143.
- Hulstijn, J.H., Hollander, M. & Greidanus, T. (1996). Incidental vocabulary learning by advanced foreign language students: The influence of marginal glosses, dictionary use, and reoccurrence of unknown words. *The Modern Language Journal*, 80, 3, 327- 339.
- Isemonger, I.M. (2007). Operational definitions of explicit and implicit knowledge: Response to R. Ellis (2005) and some recommendations for future research in this area. *Studies in Second Language Acquisition*, 29, 101-118.
- Izumi, S. (2002). Output, input enhancement, and the noticing hypothesis. *Studies in Second Language Acquisition*, 24, 541-577.
- Jiménez, L. & Méndez, C. (1999). Which attention is needed for implicit sequence learning? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25, 236-259.
- Johnson, J.S., Shenkman, K.D., Newport, E.L., & Medin, D.L. (1996). Indeterminacy in the grammar of adult language learners. *Journal of Memory and Language*, 35, 335-352.

- Jolliffe, I.T. (1972). Discarding variables in a principal component analysis, I: artificial data. *Applied Statistics*, 21, 160-173.
- Just, M.A. & Carpenter, P.A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, 99, 1, 122-149.
- Kahneman, D. (1973). *Attention and Effort*, Englewood Cliffs, NJ: Prentice-Hall.
- Kaiser, H.F. (1974). An index of factorial simplicity. *Psychometrika*, 39, 31-36.
- Keating, G.D. & Farley, A.P. (2008). Processing instruction, meaning-based output instruction, and meaning-based drills: Impacts on classroom L2 acquisition of Spanish object pronouns. *Hispania*, 91, 3, 639-650.
- Keng, C. (2009). *Implementation of, and resistance to, communicative EFL teaching in Taiwan: an exploration of three related secondary EFL teaching contexts*. Unpublished doctoral thesis, University of York.
- Kinney, P.R. & Gray, C.D. (2000). *SPSS for Windows made simple*. Woking: Psychology P.
- Klein, W. (1986). *Second Language Acquisition*. Cambridge: Cambridge University Press.
- Kline, P. (1994). *An easy guide to factor analysis*. London: Routledge.
- Kline, P. (2000). *The handbook of psychological testing*. (2<sup>nd</sup> ed) London: Routledge.
- Krashen, S.D. (1982). *Principles and Practice in Second Language Acquisition*. Oxford, England: Pergamon.
- Krashen, S.D. (1985). *The Input Hypothesis: Issues and Implications*. London: Longman.
- Krashen, S.D. & Terrell, T.D. (1983). *The Natural Approach: Language Acquisition in the Classroom*. Hayward, CA: Alemany Press.
- Lai, S. (2004). High School English Teachers' Beliefs on Grammar Instruction in

Taiwan. (Unpublished Dissertation). National Taiwan Normal University, Taiwan.

Larsen-Freeman, D. (2004). Reflections on form-meaning connection research in second language acquisition. In B. VanPatten (Eds.), *Form-Meaning Connections in Second Language Acquisition*. (pp. 237-244). Mahwah, NJ: Lawrence Erlbaum Associates.

Lee, J.F. (1998). The relationship of verb morphology to second language reading comprehension and input processing. *The Modern Language Journal*, 82, 33-48.

Lee, J.F. (2000). Five types of input and the various relationships between form and meaning. In J. Lee & A. Valdman (Eds). *Form and Meaning: Multiple Perspectives* (pp. 43-68). Boston: Heinle & Heinle.

Lee, J.F., Cadierno, T., Glass, W.R. & VanPatten, B. (1997). The effects of lexical and grammatical cues on processing tense in second language input. *Applied Language Learning*. 8, 1-23.

Lee, J.F. & Benati, A.G. (2007a). *Second language processing: an analysis of theory, problems, and possible solutions*. London: Continuum.

Lee, J.F. & Benati, A.G. (2007b) *Delivering processing instruction in classrooms and virtual contexts: Research and practice*. London: Equinox.

Lee, J.F. & VanPatten, B. (2003) *Making communicative language teaching happen*. London: McGraw-Hill.

Lee, P. (2005). *A study of English Grammar Instruction in Elementary Schools in Taipei*. (Unpublished Master's Dissertation). National Kaohsiung First University of Science and Technology, Taiwan.

Leow, R.P. (1993). To simplify or not to simplify: A look at intake. *Studies in Second Language Acquisition*, 15, 333-355.

Leow, R. (1996). Grammaticality judgment tasks and second-language development. *Georgetown University Round Table on Languages and Linguistics*, 126-39.

Leow, R.P. (2001). Attention, awareness, and foreign language behavior. *Language*

*Learning*, 51, 1, 113-155.

Lightbown, P. M. & Spada, N. (2000). Do they know what they're doing? L2 learners' awareness of L1 influence. *Language Awareness*, 9, 4, 198-217.

Long, M. H. & Robinson, P. (1998). Focus on form: Theory, research and practice. In C. Doughty & J. Williams (Eds.), *Focus on Form in Classroom Second Language Acquisition* (pp. 15-41). Cambridge: Cambridge University Press.

Loschky, L. & Bley-Vroman, R. (1993). Grammar and task-based methodology. In G. Crookes & S. M. Gass(Eds.), *Tasks and Language Learning: Integrating Theory and Practice* (pp. 123-167). Clevedon: Multilingual Matters.

Mackey, A. & Gass, S.M. (2005). *Second Language Research: Methodology and Design*, Mahwah, NJ: Lawrence Erlbaum.

Macrory, G. & Stone, V. (2000) Pupil progress in the acquisition of the perfect tense in French: The relationship between knowledge and use. *Language Teaching Research*, 4, 55-82.

Mandell, P. (1999). On the reliability of grammaticality judgment tests in second language acquisition research. *Second Language Research*, 15, 73-99.

Mangubhai, F. (1991). The processing behaviors of adult second language learners and their relationship to second language proficiency. *Applied Linguistics*, 12, 268-297.

Marsden, E. (2004). *Teaching and learning of French verb inflections: a classroom experiment using processing instruction*, unpublished Ph.D. thesis, University of Southampton.

Marsden, E. (2006). Exploring input processing in the classroom: An experimental comparison of processing instruction and enriched input. *Language Learning*, 56, 3, 507-566.

Marsden, E. (2007). Can educational experiments both test a theory and inform practice? *British Educational Research Journal*, 33, 4, 565-588.

Matessa, M. & Anderson (2000). Modeling focused learning in role assignment.

*Language and Cognitive Processes*, 15, 3, 264-292.

Moore, L. (2002). Research design for the rigorous evaluation of complex educational interventions: lessons from health services research. *Building Research Capacity*, 1, 4-5.

Moore, L., Graham, A. & Diamond, I. (2003). On the feasibility of conducting randomised trials in education: case study of a sex education intervention. *British Educational Research Journal*, 29, 5, 673-689.

Morgan-Short, K. & Bowden, H.W. (2006). Processing instruction and meaningful output-based instruction: Effects on second language development. *Studies in Second Language Acquisition*, 28, 1, 31-65.

Munnich, E., Flynn, S., & Martohardjono, G. (1994). Elicited imitation and grammaticality judgment tasks: what they measure and how they relate to each other. In E. Tarone, S. Gass, & A. Cohen (Eds), *Research methodology in second-language acquisition* (pp.227-243). Hillsdale, NJ: Erlbaum.

Murphy, V. (1997). The effect of modality on a grammaticality judgment task. *Second Language Research*, 13, 1, 34-65.

Musumeci, D. (1989). *The ability of second language learners to assign tense at the sentence level: A cross-linguistic study*. Unpublished doctoral thesis, University of Illinois at Urbana-Champaign.

Nation, I.S.P. (1990). *Teaching and learning vocabulary*. New York: Newbury House.

Nation, P. & Waring, R. (1997). Vocabulary size, test coverage and word lists. In N. Schmitt & M. McCarthy (Eds), *Vocabulary: Description, acquisition and pedagogy*. (pp.6-19). Cambridge: Cambridge University Press.

Neumann, O. (1996). Theories of attention. In O. Neumann & A. Sanders (Eds.), *Handbook of perception and action: Vol. 3. Attention*. (pp.389-446). New York: Academic Press.

Norman, D.A. (1968). Toward a theory of memory and attention. *Psychological Review*, 75, 6, 522-536.

- Norris, J. M. & Ortega L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, 50, 3, 417-528.
- Pallant, J. (2007). *SPSS survival manual : a step by step guide to data analysis using SPSS for Windows* (3<sup>rd</sup> Ed.). Maidenhead: Open University Press.
- Paulston, C.B. (1972). The sequencing of structural pattern drills. *TESOL Quarterly*, 5, 197-208.
- Peters, A.M. (1985). Language segmentation: Operating principles for the perception and analysis of language. In D. I. Slobin (Ed.), *The cross-linguistic study of language acquisition, vol. 3: Theoretical issues* (pp. 1029-1067). Hillsdale: Erlbaum.
- Pica, T. (1985). Linguistic simplicity and learnability: Implications for syllabus design. In K. Hytlenstam & M. Pienemann (Eds.), *Modelling and Assessing Second Language Acquisition* (pp. 137-152). Clevedon: Multilingual Matters.
- Purpura, J.E. (2004). *Assessing Grammar*. Cambridge: Cambridge University Press.
- Qiu, Z.H. (2005). *Quantitative Research and Statistical Analysis in Social & Behavioral Sciences*. (2<sup>nd</sup> ed) Taipei: Wu-Nan.
- Radford, A. (1990). *Syntactic theory and the acquisition of English syntax: The nature of early child grammars of English*. Oxford: Blackwell.
- Robinson, P. (2003). Attention and memory in SLA. In C. Doughty & M. Long (Eds.), *Handbook of Second Language Acquisition*. Oxford, England: Blackwell.
- Robinson, P. (2002) Effects of individual differences in intelligence, aptitude and working memory on adult incidental SLA: A replication and extension of Reber, Walkenfield and Hernstadt, 1991. In P. Robinson (Ed), *Individual differences and instructed language learning* (pp. 211-266). Amsterdam: Benjamins.
- Robinson, P. (1995a). Review article attention, memory, and the "noticing" hypothesis. *Language Learning*, 45, 2, 283-331.
- Robinson, P. (1995b). Aptitude, awareness, and the fundamental similarity of implicit and explicit second language learning. In R. Schmidt (Ed.), *Attention and awareness in*

*foreign language learning* (pp.303-357). Hawaii: Second Language Teaching & Curriculum Center, University of Hawai'i at Manoa.

Robinson, P. & Ross, S. (1996). The development of task-based assessment in English for academic purposes programs. *Applied Linguistics*, 17, 455-476.

Roehr, K. (2008). Metalinguistic knowledge and language ability in university-level L2 learners. *Applied Linguistics*, 29, 2, 173-199.

Rosa, E.M. & O'Neill, M. (1999). Explicitness, intake and the issue of awareness: Another piece to the puzzle. *Studies in Second Language Acquisition*, 21, 511-566.

Salaberry, M.R. (1997). The role of input and output practice in second language acquisition. *The Canadian Modern Language Review*, 53, 2, 422-451.

Sanz, C. (2000). Review of implementing LIBRA for the design of experimental research in second language acquisition. *Language Learning & Technology*. 3, 2, 27-31.

Sanz, C. & Morgan-Short, K. (2004). Positive evidence versus explicit rule presentation and explicit negative feedback: A computer-assisted study. *Language Learning*, 54, 1, 35-78.

Sanz, C. & VanPatten, B. (1998). On input processing, processing instruction, and the nature of replication tasks: A response to M. Rafael Salaberry. *The Canadian Modern Language Review*, 54, 263-273.

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics*, 11, 127-158.

Schmidt, R. (1994). Deconstructing consciousness in search of useful definitions for applied linguistics. *AILA Review*, 11, 11-26.

Schmidt, R. (1995). Consciousness and foreign language learning: A tutorial on the role of attention and awareness in learning. In R. Schmidt (Ed.), *Attention and awareness in foreign language learning* (pp.1-63). Hawaii: Second Language Teaching & Curriculum Center, University of Hawai'i at Mānoa.

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and Second Language*

*Instruction*(pp.3-32). Cambridge: Cambridge University Press.

Schmidt, R.W. & Frota, S.N. (1986). Developing basic conversational ability in a second language: A case study of an adult learner of Portuguese. In R. Day (Ed.), *Talking to learn: conversation in second language acquisition*. (pp.237-326). Rowley, MA: Newbury House.

Schmitt, N. & McCarthy, M. (1997). *Vocabulary: Description, acquisition and pedagogy*. Cambridge: Cambridge University Press.

Schwartz, B. (1993). On explicit and negative data effecting and affecting competence and linguistic behavior. *Studies in Second Language Acquisition*, 15, 2, 147-163.

Scriven, M. (1998). Minimalist theory: The least theory that practice requires. *American Journal of Evaluation*. 19, 1, 57-70.

Sharwood Smith, M. (1986). Comprehension versus acquisition: two ways of processing input. *Applied Linguistics*, 7, 3, 239-256.

Sheen, R. (2007) Processing Instruction. *ELT Journal*, 61, 2, 161-163.  
<http://eltj.oxfordjournals.org/cgi/reprint/61/2/161>

Shirai, Y. (2004). A multiple-factor account for form-meaning connections in the acquisition of tense-aspect morphology. In B. VanPatten (Eds.), *Form-Meaning Connections in Second Language Acquisition*. (pp. 91-112). Mahwah, NJ: Lawrence Erlbaum Associates.

Skehan, P. (1998). *A Cognitive Approach to Language Learning*. Oxford: Oxford University Press.

Sökmen, A.J. (1997). Current trends in teaching second language vocabulary. In N. Schmitt & M. McCarthy (Eds), *Vocabulary: Description, acquisition and pedagogy*. (pp.237-257). Cambridge: Cambridge University Press.

Sorace, A. (1996). The use of acceptability judgments in second language acquisition research. In W.C. Ritchie and T.K. Bhatia (eds) *Handbook of second language acquisition* (pp. 375-409). San Diego: Academic Press.



Spada, N. & Lightbown, P. (1989). Intensive ESL programs in Quebec primary schools. *TESL Canada Journal*, 7, 11-32.

Swain, M. (1998). Focus on form through conscious reflection. In C. Doughty & J. Williams (eds) *Focus on form in classroom second language acquisition* (pp. 64-82). Cambridge: Cambridge University Press.

Terrell, T. (1991) The role of grammar instruction in a communicative approach. *The Modern Language Journal*, 75, 1, 52-63.

Tomasello, M. & Herron, C. (1988). Down the garden path: inducing and correcting overgeneralization errors in the foreign language classroom. *Applied Psycholinguistics*, 9, 3, 237-246.

Tomasello, M. & Herron, C. (1989). Feedback for language transfer errors: the garden path technique. *Studies in Second Language Acquisition*, 11, 4, 385-395.

Tomlin, R., & Villa, H. (1994). Attention in cognitive science and second language acquisition. *Studies in Second Language Acquisition*, 16, 183-203.

Torgerson, C. & Torgerson, D. (2001). The need for randomized controlled trials in educational research, *British Journal of Educational Studies*, 49, 316-329.

Torgerson, C. & Torgerson, D. (2003a). The design and conduct of randomized controlled trials in education: lessons from health care. *Oxford Review of Education*, 29, 67-80.

Torgerson, D. & Torgerson, C. (2003b). Avoiding bias in randomised controlled trials in educational research. *British Journal of Educational Studies*, 51, 1, 36-45.

Toth, P.D. (2006). Processing Instruction and a role for output in second language acquisition. *Language Learning*, 56, 2, 319-385.

Trahey, M. & White, L. (1993). Positive evidence and preemption in the second language classroom. *Studies in Second Language Acquisition*, 15, 181-204.

Treisman, A.M. (1964). The effect of irrelevant material on the efficiency of selective listening. *American Journal of Psychology*, 77, 4, 533-546.

Trochim, W.K. (1998). An evaluation of Michael Scriven's "Minimalist theory: The least theory the practice requires". *American Journal of Evaluation*, 19, 2, 243-49.

Trochim, W.K. (2001). *The Research Methods Knowledge Base* (2<sup>nd</sup> Ed.). Cincinnati, Ohio: Atomic Dog Pub.

Truscott, J. (1998). Noticing in second language acquisition: a critical review. *Second Language Research*, 14, 2, 103-135.

Tymms, P. & Fitz-Gibbon C. T. (2002). Theories, hypotheses, hunches and ignorance. *Building Research Capacity*, 2, 10-11.

VanPatten, B.(1989). Can learners attend to form and content while processing input? *Hispania*, 72, 409-417.

VanPatten, B. (1990). Attention to form and content in the input: An experiment in consciousness. *Studies in Second Language Acquisition*, 12, 287-301.

VanPatten, B. (1993) Grammar teaching for the acquisition-rich classroom. *Foreign Language Annals*, 26, 4, 435-450.

VanPatten, B. (1994). Evaluating the role of consciousness in second language acquisition: Terms, linguistic features and research methodology. *AILA Review*, 11, 27-36.

VanPatten, B. (1996). *Input Processing and Grammar Instruction in Second Language Acquisition*. Norwood, N.J.: Ablex Pub.

VanPatten, B. (2000). Processing instruction as form-meaning connections: Issues in theory and research. In J. Lee and A. Valdman (Eds) *Form and Meaning: Multiple Perspectives* (pp. 43-68). Boston: Heinle & Heinle.

VanPatten, B. (2002a). Processing Instruction: An Update. *Language Learning*, 52, 4, 755-803.

VanPatten, B. (2002b). Processing instruction, prior awareness and the nature of second language acquisition: A (partial) response to Batstone. *Language Awareness*, 11, 4, 240-258.

- VanPatten, B. (2002c). Processing the content of input-processing and processing instruction research: A response to DeKeyser, Salaberry, Robinson, and Harrington. *Language Learning*, 52, 4, 825-831.
- VanPatten, B. (2004). *Processing instruction : theory, research, and commentary*. Mahwah, NJ: Erlbaum.
- VanPatten, B. (2007). Input processing in adult second language acquisition. In B. VanPatten & J. Williams (eds.) *Theories in Second Language Acquisition: An Introduction* (pp. 115-135). Mahwah, NJ: Lawrence Erlbaum Associates.
- VanPatten, B. & Cadierno, T. (1993a). Explicit instruction and input processing. *Studies in Second Language Acquisition*, 25, 225-241.
- VanPatten, B. & Cadierno, T. (1993b). Input processing and second language acquisition: A role for instruction. *The Modern Language Journal*, 77, 45-57.
- VanPatten, B. & Fernández, C. (2004). The long-term effects of PI. In B. VanPatten (Ed.), *Processing Instruction: Theory, research, and commentary* (pp. 273-289). Mahwah, NJ: Erlbaum.
- VanPatten, B. & Oikkenon, S. (1996). Explanation versus structured input in processing instruction. *Studies in Second Language Acquisition*, 18, 495-510.
- VanPatten, B. & Sanz, C. (1995). From input to output: Processing instruction and communicative tasks. In F. Eckamn, D. Highland, P. Lee, J. Mileham and R. Rutkowski Weber (eds.) *Second Language Acquisition Theory and Pedagogy* (pp. 168-185). Mahwah, NJ: Erlbaum.
- VanPatten, B. & Wong, W. (2004). Processing instruction and the French causative: another replication. In B. VanPatten (Ed.), *Processing Instruction: Theory, Research, and Commentary* (pp. 97-118). Mahwah, NJ: Erlbaum.
- VanPatten, B., Williams, J. & Rott, S. (2004). Form-meaning connections in second language acquisition. In B. VanPatten (Eds.), *Form-Meaning Connections in Second Language Acquisition*. (pp. 1-26). Mahwah, NJ: Lawrence Erlbaum Associates.
- VanPatten, B. & Williams, J. (2007). *Theories in Second Language Acquisition: An*

*Introduction*. Mahwah, NJ: Lawrence Erlbaum Associates.

White, L. (1991). Adverb placement in second language acquisition: some effects of positive and negative evidence in the classroom. *Second Language Research*, 7, 133-161.

Wickens, C.D. (1984). Processing resources in attention. In R. Parasuraman & D. Davies (Eds.), *Varieties of attention* (pp. 63–102). New York: Wiley.

Wickens, C.D. (1989). Attention and skilled performance. In D.H. Holding (Ed.), *Human Skills*, Chichester: John Wiley.

William, J.N. (1999). Learner-generated attention to form. *Language learning*, 49, 4, 583-625.

Williams J.N. (2004). Implicit learning of form-meaning connections. In B. VanPatten(Eds.), *Form-Meaning Connections in Second Language Acquisition*.(pp. 203-218). Mahwah, NJ: Lawrence Erlbaum Associates.

Williams, J.N. (2005). Learning without awareness. *Studies in Second Language Acquisition*, 27, 269-304.

Wong, W. (2001). Modality and attention to meaning and form in the input. *Studies in Second Language Acquisition*, 23, 345-368.

Wong, W. (2004a). The nature of processing instruction. In B. VanPatten (Ed.), *Processing Instruction: Theory, Research, and Commentary* (pp. 33-64). Mahwah, NJ: Erlbaum.

Wong, W. (2004b). Processing instruction in French: The roles of explicit information and structured input. In B. VanPatten (Ed.), *Processing Instruction: Theory, Research, and Commentary* (pp. 187-206). Mahwah, NJ: Erlbaum.

Wu, C. (2003). *A Study of the Comparative Effect of Input-based Grammar Instruction and Output-based Instruction on the Acquisition of the English Subjunctive Mood*. (Unpublished Master's Dissertation). National Taiwan Normal University, Taiwan.

Xu, J. (2001). *Using Processing Instruction to Teach Wh-questions in Secondary EFL*

*Classes in Taiwan.* (Unpublished Master's Dissertation). National Tsing-Hua University, Taiwan.