Developing and Validating an Internet-based Battery of *Tests of Aptitude for Language Learning* (TALL)

Junlan Pan

PhD

University of York

Education

September 2023

Abstract

Measuring language learning aptitude faces theoretical and methodological challenges. First, the development of aptitude batteries has not generally kept pace with changes in theoretical frameworks that conceptualise aptitude constructs pertaining to second language (L2) learning. Second, it is crucial to ascertain the reliability and validity of aptitude batteries prior to conducting aptitude-learning research, yet this step has been surprisingly neglected to date (cf. Bokander & Bylund, 2020). In response to these gaps, an internet-based battery, *Tests of Aptitude for Language Learning* (TALL), has been developed, informed by the theoretical frameworks of the Stages Approach (Skehan, 2016) and the Phonological/Executive (P/E) Model (Wen, 2016), as well as major existing aptitude batteries. TALL measures four facets of aptitude that represent cognitive abilities involved in the early stages of L2 learning and development: associative memory, phonetic coding ability, language analytic ability, and working memory (specifically, phonological short-term memory and executive control capacity). These abilities are measured by five subtests, i.e., Vocabulary Learning, Sound Discrimination, Language Analysis, Serial Nonwords Recall, and Complex Span Tasks, respectively. TALL employs domain-specific verbal stimuli and has two separate test suites to differentiate the modalities (aural and written) of test items.

Initial validation checks were conducted with 165 participants (L1 Chinese undergraduates with L2 English) taking two sessions of tests with items counterbalanced across modality and test session. Results of analyses at the subtest, item, and battery levels suggested that, in general, TALL displayed satisfactory reliability and internal validity for measuring aptitude conceptualised in the theoretical frameworks. Linear mixed-effects modelling analyses revealed significant effects of modality on test results. Multiple regression analyses revealed that aural and written suite of TALL could explain 16% and 19% variance, respectively, of the self-reported L2 proficiency scores. Implications for battery refinement, further scrutiny of validation, utilisation of TALL in aptitude related research, and the potential of TALL as an open research tool are discussed.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I have learned that completing a PhD thesis is very much like running a marathon – a long process with a definite end. I would like to express my heartfelt gratitude to everyone who cheered me on.

I am profoundly grateful for the invaluable guidance from my supervisor, Prof. Emma Marsden. My connection with her dates back to my master's degree when I first encountered her module on Research Methods, which played an important role in shaping my academic journey. Her unwavering commitment to methodological rigour and open research practices have deeply influenced my dedication to these principles. Without her encouragement, her patience in allowing me to navigate my interests and formulate answers to the profound questions she raised during the supervision, and her belief in my ability to pursue a PhD, I would never have reached this point. I would also like to express my deep appreciation to my thesis advisor, Dr Guilia Bovolenta, for posting the right questions, offering invaluable feedback, and suggesting relevant references.

I am indebted to the School of Foreign Languages and Cultures at Chongqing University for providing funding to support the development of *Tests of Aptitude for Language Learning* (TALL). I am grateful to my colleagues in China for their assistance in recruiting participants, and to all the students who participated in my research–without you, none of this would have been possible. Special thanks to Daiva Judges for creating and recording stimuli adapted from Lithuanian, Hanlin Chen for her invaluable assistance with scoring and coding. I have been fortunate to work with wonderful colleagues in the PhD office at York, and I deeply appreciate their encouragement.

I dedicate this thesis to my dear parents. Thank you, 妈妈爸爸, for your love and support, for always being there. I extend my gratitude to my dear sister, who cared for our parents while I have been away for more than four years. Lastly, to my husband, thank you for being by my side, cooking to satisfy my taste buds, and keeping homesickness at bay. Life would be unbearably challenging without you.

Author's Declaration

I declare that this thesis, including all data presented in it, is original work and that I am the sole author. This work has not previously been presented for an award at this, or any other, University. The website of *Test of Aptitude for Language Learning* (TALL) was designed by me, with the testing interface and the back-end database being developed by outsourced developers. All data collection, test scoring and data analyses were carried out by me, with another Chinese native speaker as the second marker who scored responses in the subtest of Serial Nonwords Recall. All sources used are acknowledged as References.

CHAPTER 1: INTRODUCTION

1.1 The research context

Research into second language (L2) learners' foreign language aptitude (interchangeably referred to as language aptitude or simply aptitude throughout this thesis), defined as the special cognitive abilities to learn an additional language efficiently and successfully (Carroll, 1990), has been ongoing for several decades since the inception of the concept in the 1950s (Wen et al., 2019). Over the course of sixty years, research into foreign language aptitude has experienced cycles of varying interest occasionally followed by periods of neglect. However, there has been a substantial growth of the number of publications in the first two decades of the millennium (as observed by Vuong & Wong, 2019), evidenced by the publication of significant edited volumes (Wen et al., 2019; 2023a) and thematic issues in leading journals (Doughty & Mackey, 2021; Li & DeKeyser, 2021). These outputs signify notable progress in aptitude-related research within the field of L2 learning.

Language aptitude, as a componential construct of cognitive abilities (Doughty, 2019; Doughty & Mackey, 2021), is represented by learners' performance on the tests employed to measure it. Consequently, the construct of aptitude is somewhat defined by what is measured by aptitude tests (Li & Zhao, 2021). The first comprehensive aptitude test that was influential in L2 learning research is the MLAT (the Modern Language Aptitude Test; Carroll & Sapon, 1959), followed by numerous aptitude batteries that have emerged, including the PLAB (the Pimsleur Language Aptitude Battery; Pimsleur, 1966), the CANAL-FT (the Cognitive Ability for Novelty in Acquisition of Language–Foreign Test; Grigorenko et al., 2000), the LLAMA tests (Meara, 2005; Meara & Roger, 2019), and the Hi-LAB (the High-Lever Language Aptitude Battery; Linck et al., 2013). These batteries have yielded empirical evidence that contributes to the collective knowledge of language learning aptitude. However, theoretical and methodological concerns regarding aptitude batteries used in L2 learning research have persisted.

First and foremost, as argued by Skehan (2023), aptitude batteries serve as operationalisations of various perspectives regarding the nature of L2 learning. Therefore, they ought to align with views on L2 learning processes and contribute to the evolution of theories that facilitates understanding of the nature of L2 learning. However, the most influential and widely used battery, the MLAT, which is built upon J. B. Carroll's (1962, 1973, 1981) classic four-factor model of language aptitude, has been in existence for over six decades. While it has provided a significant amount of empirical evidence, particularly before the emergence of new aptitude batteries, it has been commented as primarily focusing on the outcomes rather than the processes of L2 learning (Li, 2019). Moreover, Carroll himself (1981, 1990) noted that the MLAT lacks adequate measures of the memory system, especially working memory (WM), which can predict language learning success. In recent decades, two batteries, i.e., the LLAMA tests and the Hi-LAB, have come to dominate aptitude-related research in the field. These batteries can be seen as reflecting the advancement of aptitude theories pertaining to L2 learning. The Hi-LAB, in particular, is influenced by contemporary cognitive psychology and comprises subtests designed to measure cognitive abilities particularly relevant to achieve a high level of proficiency in L2 learning (Doughty, 2019).

However, shortcomings are associated with these two prominently used aptitude batteries. Firstly, the limited accessibility of the Hi-LAB, owing to its government-sponsored background, restricts its use to the authors of the battery and their associates and on specific learner populations. This limitation has led to only partial validation evidence, which may undermine the robustness of claims based on the battery's findings (Skehan, 2023). Secondly, despite the growing popularity of the LLAMA tests in numerous studies in recent years, attributed to its open accessibility for other researchers in the community, the lack of sufficient validation evidence for this battery raises significant concerns (Bokander & Bylund, 2020, Bokander, 2023). The absence of comprehensive aptitude research designs further compounds the limitations arising from the dominance of these two batteries, impeding a broader advancement in understanding the construct of aptitude and its role in L2 learning. As highlighted by Skehan (2023), recent progress in this realm has been characterised as "instrument-led rather than construct-led" (p. 232), revealing substantial room for further exploration.

1.2 The research aim

The primary objective of this thesis is to bridge the existing gaps by developing a comprehensive aptitude battery, *Tests of Aptitude for Language Learning* (TALL). This battery is designed to fulfil two crucial requirements: (1) to reflect a theoretical framework that underpins a componential construct of aptitude within the context of L2 learning, and (2) to undergo preliminary validation processes to ascertain the reliability and internal validity of this new battery before its use in substantive research.

In pursuit of these primary objectives, the development of a valid aptitude battery grounded in aptitude–L2 learning theories is accompanied by several considerations. First, the unresolved contrast between domain generality and specificity (Wen et al., 2017) has been tackled by devising domain-*specific* verbal stimuli for all subtests, including two for WM. This initial step is intended to pave the way for further examination of convergent validity by comparing TALL with other measures using domain-*general* stimuli, such as the Hi-LAB. Second, the design of TALL incorporates different modalities (aural and written) for the test items in three subtests, resulting in the creation of two separate test suites. This differentiation enables an investigation into the potential impact of modality on test outcomes. In addition, to mitigate the potential confounds of L2 knowledge, TALL employs the participants' first language (L1), Mandarin Chinese, as the instructional language of the entire battery and the encoding language for the stimuli in WM subtests. Furthermore, a semi-artificial language, adapted from Lithuanian, has been used to create target items in the subtests related to language learning. This approach ensures the novelty of the items for participants in the current research. Importantly, TALL has been developed into an internet-based battery, facilitating remote data collection on the test platform (https://www.tall-webtest.com), thus offering a practical solution to the restrictions imposed by the Covid-19 pandemic on lab-based research. The internet-based research capability of TALL is noteworthy, as it enables the development of TALL into an open research instrument accessible for other researchers, aligning with the spirit of open research practices (Pan & Marsden, under review). This initiative is anticipated to facilitate the accumulation of validation evidence for TALL across diverse temporal and contextual settings.

In its essence as a methodological study, this PhD research mainly provides empirical evidence to verify TALL as a reliable and valid battery for measuring language aptitude. Moreover, it seeks to explore the effects of modality in measuring aptitude and to investigate predictive validity of TALL concerning participants' L2 proficiency.

1.3 Outline of the thesis

The current chapter has established the research context and the research aims, focusing on the development and validation of an internet-based battery, TALL.

Chapter 2 will provide a comprehensive review of the relevant research literature, organised into four sections: (1) an overview of language aptitude for L2 learning, encompassing fundamental concepts of aptitude, a synthesis of research connecting aptitude and L2 learning, theories concerning language aptitude, and a dedicated review of working memory as a constituent of aptitude and a theoretical framework conceptualising WM in the context of L2 learning; (2) a synthesis of existing aptitude batteries, focusing on the theoretical inquiries and methodological challenges related to measuring the multi-faceted construct of aptitude; (3) an outline of the rationale driving the current research, with the aim of addressing empirical and methodological gaps in measuring aptitude; and (4) the formation of the research questions, which concern the internal validity of TALL as an aptitude battery, the effects of modality on the results of measuring aptitude, and the predictive validity of TALL in relation to L2 proficiency.

Chapter 3 will detail the methods employed, providing insights into the decision-making process behind the development of TALL's subtests, data collection procedure, and analysis plan.

Chapter 4 presents the outcomes of the analysis conducted for Research Question 1, followed by an accompanying discussion. This chapter concerns the internal validity of TALL as a battery for aptitude, evaluated through a validation plan that yields evidence for making (1) a generalisation inference about the representative nature of all subtests as measures for their intended constructs; (2) a scoring inference about the efficacy of the items in each subtest in assessing participants' componential abilities; and (3) an explanation inference about TALL, as a battery for aptitude, aligning with the theoretical frameworks underpinning its construction.

Chapter 5 will report the results of the analysis undertaken for Research Question 2, which examines the effects of modality in measuring aptitude, followed by a discussion of these findings.

Chapter 6 will present the findings from the analysis addressing Research Question 3, which investigates the predictive validity of subtests of TALL in explaining participants' L2-English proficiency reflected by their (self-reported) scores in the National Matriculation English Test (NMET). This chapter will also be accompanied by a discussion.

Lastly, Chapter 7 (Conclusion) will summarise the main findings and contributions of the study. It will additionally outline the limitations of the study and propose avenues for future research.

CHAPTER 2: LITERATURE REVIEW & THE CURRENT RESEARCH

This chapter starts with a review of language aptitude and L2 learning research and an outline of key theoretical frameworks that conceptualise the construct of aptitude in L2 learning. A particular focus is on two theoretical frameworks that lay the foundation of the current research. It is followed by critiques of existing major aptitude batteries, which leads to the research gaps the current research attempts to address and the research questions it aims to answer.

2.1 Aptitude for L2 learning

Foreign language aptitude, while remaining a complex and frequently contentious research area of inquiry, continues to hold a central position in the field of L2 research. Despite general recognition of its role as a trait that significantly influences individual learners' success in L2 learning, a consensus regarding the magnitude of its role, specific nuances of its influence, and the dimensions of the construct remain elusive (Doughty & MacKey, 2021). This section seeks to present a comprehensive overview of the intricate relationship between the concept of aptitude and L2 learning. Specifically, it aims to delve into the nature of aptitude as it pertains to L2 learning, explore its role in explaining various aspects of learning outcomes, and shed light on the current consensus and ongoing debates surrounding its role in learning. Additionally, this section unpacks the implications of theoretical perspectives of language aptitude on the development of aptitude measurement.

2.1.1 Basic concepts of aptitude

Foreign language aptitude is a significant construct related to individual differences. It is defined as a special talent that reflects a general capacity for acquiring a second or foreign language (Carroll, 1981, 1990, 1993; Dörnyei, 2005; Skehan, 1998; Sparks et al., 2011). A prominent researcher in foreign language aptitude, John Carroll, characterised the concept as an individual's readiness and ability for learning a foreign language, along with the potential for achievement under favourable conditions. In this regard, aptitude is a synonym with ability, differing primarily in contextual use rather than meaning (Dörnyei, 2005). Accordingly, aptitude is considered an inherent trait that remains relatively stable over extended periods and is resistant to training. It may predict the rate or speed of learning and future learning progress through measures of achievement that serve as indicators of aptitude (Carroll, 1981).

While it is recognised as one of a number of key factors (among others such as motivation, learning experience, and context) that determines language learning success, foreign language aptitude is specifically defined as cognitive factors that are a composite construct encompassing various abilities of readiness to learn a second language (Carroll, 1981). It is also conceptualised as being distinct from other cognitive individual differences, such as general intelligence. The analysis of three measures, i.e., L2 proficiency, language aptitude, and two types of intelligence (verbal intelligence and reasoning) by Sasaki (1996) revealed that aptitude and intelligence were distinct from each other according to first-order factor analysis. However, the presence of a factor that is common between aptitude and intelligence, that is, analytic ability, was confirmed through Sasaki's second-order factor analysis. These findings suggest that aptitude and intelligence are interconnected, while components of aptitude in Carroll's model (1962, 1981) such as phonetic coding ability and rote memory are independent of intelligence. In addition, the meta-analytical findings from Li (2016) showing strong and positive correlations between aptitude and intelligence also suggests that the two constructs share a large overlap but remain non-identical.

2.1.2 Aptitude and L2 learning research

Decades of L2 learning research have shown that language aptitude, as the second strongest variable after the age of onset (that is, the age when learners are first meaningfully exposed to the L2), typically accounts for 10% to 20% of the variance in the prediction of the ultimate L2 attainment (Granena & Long, 2013). Despite the fact that it has experienced varying interests and attention in the field of second language acquisition (SLA) research, language aptitude research has (re)gained a prominent position in the research landscape over recent years (Vuong & Wong, 2019). An important shift of aptitude research involved the transition from a main focus on trying to predict the success of language learning using the results from aptitude tests towards a greater focus on explaining the outcomes of specific processes and aspects in L2 learning to gain a deeper understanding of the constituent elements of aptitude and their roles (Wen et al., 2019).

Comprehensive systematic reviews provide cumulative evidence that aptitude plays an important role in L2 learning. For example, Li (2016) compiles the results of 66 empirical studies that investigate the correlations between L2 learning achievement and aptitude components reflected in the batteries of the MLAT (Carroll & Sapon, 1959) and the PLAB (Pimsleur, 1966). The results showed an overall correlation of *r* = .49 (95% CI= [.45, .54]) between aptitude and general L2 proficiency, with the component of phonetic coding ability being the strongest predictor, and rote memory the weakest predictor. The examination of the effect sizes concerning aptitude, as measured through comprehensive test batteries, revealed moderate correlations (*r* > .30) with the criterion variables (that is, L2 knowledge and skills). However, aptitude had a weak association (*r* = .15) with vocabulary knowledge, and a non-significant correlation with L2 writing skill. The component of phonetic coding had a stronger correlation with vocabulary knowledge (*r* = .38) than with other outcome variables, and it had a weak correlation with listening skill (*r* = .12). Language analytic ability was reported as a stronger predictor for grammar knowledge (*r* = .39) and reading skill (*r* = .35) than for other L2 outcomes. Rote memory did not appear to be a strong predictor of L2 knowledge and skills. The key message conveyed from this meta-analysis is that the predictive power of aptitude outperforms other individual variables, such as working memory, motivation, and anxiety. Thus, aptitude can be considered as an effective predictor of L2 achievement, although its prediction of various elements within L2 learning outcomes differ. Li's earlier (2015) meta-analysis, based on 33 studies using either the MLAT or the adapted batteries modelled on the MLAT, synthesises the role of aptitude in explaining grammar acquisition. It reports an aggregate effect size of .31 with narrow confidence intervals, supporting the conclusion that aptitude plays a significant role in L2 morphosyntactic learning, and this effect has been confirmed in both naturalistic and instructed L2 learning contexts.

More recent studies continue to provide empirical evidence to support the important role of aptitude in different aspects of L2 learning. For example, in the aspects related to phonological attainment and comprehension of auditory input, Saito's (2019) investigation on the role of aptitude in L1-Japanese learners' English pronunciation attainment indicated that phonological analysis and memory in aptitude, measured by the LLAMA tests (Meara, 2005), may predict the occurrence of advanced L2 segmental proficiency attainment in an instructed context. Sok and Shin (2021) investigated the influence of individual differences on the listening comprehension of L1-Korean English-as-a-foreign-language (EFL) learners and found that aptitude (measured by the MLAT-elementary version) predicts the listening performance of L2 though it is mediated by metacognitive awareness. Furthermore, in terms of grammar and vocabulary learning and writing competence, Li et al. (2019) provides empirical evidence on the role of specific components of aptitude (i.e., language analytic ability measured by the PLAB and executive control working memory measured by an operation span test) in relation to L2 grammatical learning under different instructional conditions. The results suggest that language analytic ability can influence the outcomes of learning under conditions without external assistance (i.e., explicit instructions and interactional feedback), and WM is associated with performance when feedback was provided within the task. Mujtaba et al., (2021) examined the role of aptitude, WM and vocabulary size on L2 English learners' writing performance and found that subtests (i.e., LLAMA_B and LLAMA_E) of the aptitude measurement, receptive vocabulary size, and WM (measured by an operation span test) were evidenced as significant predictors of L2 writing performance.

A comprehensive synthesis that includes sixty-five studies by Li and Zhao (2021) provides an overview of the methods used to investigate the role of language aptitude in SLA. This synthesis reveals that, as a multi-facet construct, aptitude may display varied magnitudes of prediction on different aspects of L2 learning. The discrepancy of empirical evidence is due to (a) different measurements used to operationalise aptitude as a set of latent abilities, conceptualised in the underlying aptitude theoretical frameworks, (b) nuanced aspects of L2 learning and processing measured by a large variety of tests, and (c) moderating factors in relations to L2 learning settings (naturalistic and instructed learning) and other individual difference variables (e.g., learners' age and learning experience). In the following sections, aptitude theories and measurements will be reviewed, providing critiques of existing aptitude measures, and offering rationales for the current methodological endeavour.

2.1.3 Language aptitude theories

Driven by collaborative efforts from educational psychology, applied linguistics, and cognitive (neuro)science, the concept of foreign language aptitude has undergone significant modifications since its inception and continues to evolve (Wen et al., 2017). This section sets out to review the advances in language aptitude theoretical construction based on some comprehensive reviews in the field (e.g., Li, 2019; 2022; Wen et al., 2017), followed by the introduction of two theoretical models, i.e., the Stages Approach (Skehan, 2016) and the Phonological / Executive (P/E) WM Model (Wen, 2016), which are the theoretical foundations of the current research.

The Carrollian approach

The contributions of American educational psychologist John Carroll and his work nearly seven decades ago cannot be ignored in any review of foreign language aptitude research. Carroll's theory of aptitude lays the foundation for the Modern Language Aptitude Test (MLAT) (Carroll & Sapon, 1959), which is known to have the highest predictive validity among all language aptitude tests according to Li's meta-analytic review (2015, 2016). Carroll (1962, 1981, 1990, 1993) conceptualised aptitude construct as consisting of four measurable abilities: (1) phonetic coding ability to code and retain unfamiliar auditory input; (2) grammatical sensitivity to identify the linguistic functions of words in sentences; (3) inductive language learning ability to generalise patterns based on the examples from the input; and (4) associative memory to form and retain the link between verbal elements in L1 and L2 in memory.

The Carrollian approach was theoretically underpinned by behaviourism, a learning theory that also influenced the audiolingual approach to language teaching (Li, 2019). This influence is exemplified by methods of mechanical drills, rote memorising, and grammar instruction. As reviewed by Li (2019), the Carrollian approach exhibits the following characteristics that received criticism and sequentially informed the advancing of aptitude theories. First, aptitude serves as a determinant of learning success, focusing on learning outcomes rather than learning processes. Second, aptitude is considered a unified construct with validity based on a composite aptitude score measured by the MLAT. This implies that there is no room for the understanding that learners can be strong in one component while having weakness in another ability. Third, aptitude is conceptualised as a trait involved in the initial stages of learning. While this does not imply that aptitude becomes insignificant in advanced learning, it suggests the components measured by the MLAT may differ from the type(s) of aptitude more relevant to advanced learning. Furthermore, the componential abilities only pertain to the linguistic aspects of L2 learning and do not encompass the ability of using L2 appropriately in various communicative situations, as the MLAT solely focuses on measuring comprehension of linguistic materials without involving language production.

The Carrollian approach, according to Li's (2022) review, proposes that aptitude is primarily associated with instruction that involves learners' conscious, laborious processing of language materials, making it pertinent to explicit or instructed learning rather than implicit learning.

Since the foundation laid by Carroll's theory and the dominant applications of the MLAT battery anchored in the Carrollian theory, a large body of empirical evidence has been generated on the positive correlation between aptitude and L2 learning outcomes. This situation, however, did not seem to fuel further theoretical development of language aptitude of relevance to L2 learning. It has been claimed to fall far behind (before the start of the twenty-first century) when compared to the theoretical progress of L2 motivation, an equally important factor among all individual differences that has been purported to predict L2 learning outcomes (Wen et al., 2017). The relatively limited empirical research and theoretical development in the aptitude-related area were possibly due to a perception that language aptitude is a relatively fixed trait with relevance mainly to (outdated) language teaching approaches such as audiolingualism. The research landscape was predominantly driven by Carroll's theory and the comparison design of pre–post learning outcomes using the MLAT as the measurement (Skehan, 1998, 2002). This situation started to improve with several significant lines of inquiry in a new wave of theorising, leading to a variety of new conceptualisations of language aptitude that went beyond Carroll’s ideas.

Linguistic Coding Differences Hypothesis (LCDH) model

The LCDH model, proposed by Sparks & Ganschow (2001), is considered a variant of the Carrollian approach (Li, 2019). Based on the assumption that L1 and L2 learning share the same set of cognitive skills (Ganschow & Sparks, 1996), this model proposes that aptitude for L2 learning is simply a carryover of aptitude for L1 learning, and that L1 skills of learners are predictive of their additional language achievements. The model conceptualises four basic components of aptitude: (1) phonological / orthographical skills of L1 and L2, including phonetic coding and phonological processing ability; (2) language analytic ability of L1 and L2, including comprehension, grammar, vocabulary, and inductive language learning; (3) IQ (Intelligence Quotient) / memory skills, including intelligence and paired associate learning measures of L1 and L2; and (4) self-perceptions of motivation and anxiety of L2.

Two critical comments by Li (2019) question the theoretical conceptualisation of this model and its validation evidence: First, L1 skills needs to be clarified, particularly in terms of whether vocabulary and reading comprehension should be considered as language achievements rather than cognitive abilities. Second, the correlation between L1 skills and L2 aptitude could be the result of another unexamined variable such as motivation. Thus, the correlational or predictive relationships between L1 skills, L2 aptitude and L2 achievements need to be clarified. However, as an expansion upon Carroll’s four-factor model, LCDH model highlights the significant alignment of L1–L2 language analytic skills and purports to include a phonological measure of L1 and L2 in aptitude measurements. The implication of adding components of cross-linguistic phonology/orthography decoding skills in language aptitude research can be informative to the development of aptitude measurements.

Aptitude-Treatment-Interaction (ATI) approaches and Aptitude Complexes/Ability Differential (ACAD) frameworks

The ATI approaches, introduced by Snow (1991), conceptualises aptitude as any factors of individual differences, including not only cognitive factors but also affective factors that influence learning. It emphasises the interaction between aptitude and learning conditions, suggesting that learners with different attitude profiles may benefit differently from various types of instruction. Specifically, low-aptitude learners may benefit from structured materials, and high-aptitude learners may excel when they are encouraged to generalise patterns and rules through their own cognitive abilities.

Although the ATI approaches define aptitude differently from a narrow conceptualisation that specifies aptitude as cognitive variables in L2 research, these approaches have been instructive to the theoretical development of aptitude pertaining to L2 learning (Li, 2019). Robinson (2002a, 2005, 2007, 2012) proposed the Aptitude Complexes/Ability Differential frameworks, expanding on the ATI approaches, which provide a new perspective in conceptualising L2 field-specific aptitude as distinct clusters of cognitive abilities, known as aptitude complexes, involved in different learning conditions. The frameworks incorporate two hypotheses: the Aptitude Complexes Hypothesis contends that a collection of fundamental cognitive skills (e.g., processing speed, pattern recognition, phonological working memory capacity) combined to form higher-order aptitude complexes (e.g., noticing the gap, memory for contingent speed), which can be used in specific learning tasks; the Ability Differential Hypothesis states that L2 learners exhibit variations in cognitive abilities, resulting in their distinct profiles in the relevant aptitude complexes.

Robinson’s (2005) frameworks are presented in a hierarchical wheel-shape diagram (p. 52–53), as shown in Figure 2.1, in which ten first-order abilities situated in the innermost circle, the aptitude complexes in the second circle, the task aptitudes in the third circle, and the pragmatic/interactional abilities/trait in the most outward circle. The abilities and aptitude complexes in the two inner circles contribute largely to initial input-based learning, and the outer circles of task aptitudes and broader capacities are more related to task performance and transferable to real-world interactive settings. The frameworks also postulate that adult learning is predominantly a conscious process, implying very little role for implicit aptitude (Li, 2022).

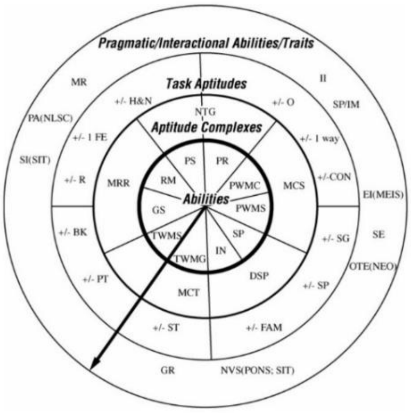
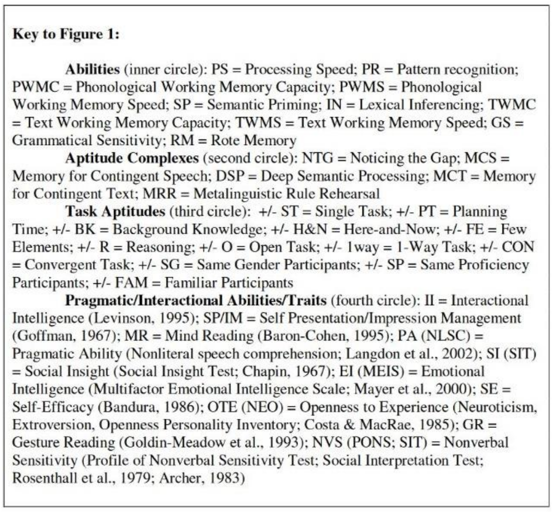


Figure 1: Aptitudes, development, and learning contexts: Changes in the relative contribution of aptitude factors to different aspects of L2 learning.
(Inner two circles: initial input-based learning; third circle: output practice and complex task performance; and outer circle: transfer of task performance to real-world interactive settings.)



Key to Figure 1:

Abilities (inner circle): PS = Processing Speed; PR = Pattern recognition; PWMC = Phonological Working Memory Capacity; PWMS = Phonological Working Memory Speed; SP = Semantic Priming; IN = Lexical Inferencing; TWMC = Text Working Memory Capacity; TWMS = Text Working Memory Speed; GS = Grammatical Sensitivity; RM = Rote Memory

Aptitude Complexes (second circle): NTG = Noticing the Gap; MCS = Memory for Contingent Speech; DSP = Deep Semantic Processing; MCT = Memory for Contingent Text; MRR = Metalinguistic Rule Rehearsal

Task Aptitudes (third circle): +/- ST = Single Task; +/- PT = Planning Time; +/- BK = Background Knowledge; +/- H&N = Here-and-Now; +/- FE = Few Elements; +/- R = Reasoning; +/- O = Open Task; +/- 1way = 1-Way Task; +/- CON = Convergent Task; +/- SG = Same Gender Participants; +/- SP = Same Proficiency Participants; +/- FAM = Familiar Participants

Pragmatic/Interactional Abilities/Traits (fourth circle): II = Interactional Intelligence (Levinson, 1995); SP/IM = Self Presentation/Impression Management (Goffman, 1967); MR = Mind Reading (Baron-Cohen, 1995); PA (NLSC) = Pragmatic Ability (Nonliteral speech comprehension; Langdon et al., 2002); SI (SIT) = Social Insight (Social Insight Test; Chapin, 1967); EI (MEIS) = Emotional Intelligence (Multifactor Emotional Intelligence Scale; Mayer et al., 2000); SE = Self-Efficacy (Bandura, 1986); OTE (NEO) = Openness to Experience (Neuroticism, Extroversion, Openness Personality Inventory; Costa & MacRae, 1985); GR = Gesture Reading (Goldin-Meadow et al., 1993); NVS (PONS; SIT) = Nonverbal Sensitivity (Profile of Nonverbal Sensitivity Test; Social Interpretation Test; Rosenthall et al., 1979; Archer, 1983)

Figure 2.1 The diagram of Aptitude Complexes/Ability Differential frameworks

The ACAD frameworks, as observed in the comprehensive reviews of aptitude theories and L2 learning research (e.g., Li, 2019; Wen et al., 2017), propose that the relevance of aptitude in L2 learning varies depending on the content of learning and instructional conditions. This proposal, in turn, suggests a detailed representation of dynamic interactions underpinning human cognitive abilities in specific educational contexts. The frameworks can be instructive to the development of aptitude theory as it identifies the dynamic interaction between aptitude profiles and task complexity and specific educational contexts. This characterises the ACAD frameworks as having a significant advantage by aligning with ATI approaches, which holds that different methods of L2 learning and teaching are differentially associated with components of language learning aptitude (Dörnyei, 2005). The concept of aptitude complexes, representing various combinations of aptitude components, implies their varying significance in diverse learning environments. Thus, the frameworks are in line with the current shifts to “micro” approaches that integrate aptitude measures into experimental or quasi-experimental research designs to examine instructional comparisons or intervention procedures (Skehan, 2019). However, the dynamic nature postulated in the aptitude complexes pose challenges for collecting empirical evidence to testify the plausibility of the theoretical hypotheses in the ACAD frameworks.

Cognitive Ability for Novelty in Language Acquisition-Foreign (CANAL-F) model

The CANAL-F model (Grigorenko, Sternberg & Ehrman, 2000) has been proposed, relying on Sternberg’s (1997, 2002) conception of successful intelligence that claims to capture the fundamental nature of human abilities. The model consists of three aspects: analytic, creative, and practical. Based on the hypothesis that intelligence and aptitude play different roles in foreign language learning, it emphasises the abilities to handle novel language occurrences when learning a new language. As such, the model proposes new parameters about how language aptitude should be tested. First, rather than simply testing memory and analytic skills, creative and practical language learning abilities should be tested. Second, the need for a more in-depth evaluation of learners' aptitude should be satisfied by providing sub-scores that can inform the appropriate types of training rather than just a general language-aptitude score. Third, the use of dynamic testing is considered, where testing and training should take place simultaneously, providing a simulation of learning in real time in which aptitude can be assessed (Sternberg, 2002).

To achieve the above aims, a new language aptitude test, the Cognitive Ability for Novelty in Acquisition of Language as applied to foreign language test (the CANAL-FT) (Grigorenko et al., 2000), was developed on the CANAL-F model to reflect the focus on the ability of coping with novelty in L2 learning. The test engages the participants in a simulated setting where they are gradually exposed to an artificial language and instructed to complete several learning activities by which their aptitude is measured. The authors propose five knowledge acquisition processes, i.e., selective encoding, accidental encoding, selective comparison, selective transfer, and selective combination, which, in turn, are operationalised at four linguistic levels, i.e., the lexical, morphological, semantic, and syntactic levels of processing. Specifically, the lexical level addresses how a person learns, comprehends, and uses vocabulary items. The morphological level tackles the structures and derivations of words. The semantic level, which is based on knowledge from higher order units of sentences and paragraphs, handles the understanding and use of the meaning of the words learners learn. The syntactic level manages the acquisition, comprehension and application of the grammatical rules that link the words to the higher order units. Unlike previously developed aptitude measurements, such as the MLAT, the CANAL-FT addresses the modes of input (visual and oral) in its design that are involved in different types of language tasks, that is, the visual mode in reading and writing, and the oral mode in listening and speaking. This test will be introduced in detail in Section 2.2.1.

The memory-incorporated models and the High-Level Language Aptitude Battery (the Hi-LAB)

The integration of WM in the aptitude model is clearly supported by a considerable body of evidence from studies that examine the relationships between WM, aptitude, and language learning, suggesting an important role for WM as components within foreign language aptitude. For example, Robinson (2002b) found that WM (measured by a reading span test) had a moderately strong correlation with language aptitude scores in Sasaki’s (1996) Language Aptitude Battery for the Japanese, which is a three-section measurement based on the MLAT and the PLAB. Similarly, Sáfár & Kormos (2008) found that phonological short-term memory (PSTM) of WM (measured by a nonword repetition task) had positive significant correlation with the inductive ability of aptitude though not with the total aptitude scores. The authors, therefore, suggested that the storage capacity, not the processing and executive control function, of WM could be a cognitive ability distinct from traditional aptitude constructs. Bolibaugh & Foster (2013) also discovered that the variations of adult learners’ individual differences in PSTM predicted both learning rate and ultimate attainment. The evidence from these studies suggest that various aspects of WM play important roles in L2 learning and processing and they should be represented in aptitude test (DeKeyser & Koeth, 2011).

The development of the High-Level Language Aptitude Battery (Hi-LAB) (Linck et al. 2013) has contributed significantly to the theoretical development of foreign language aptitude. The battery was designed to target gifted adult learners, and thus it could accurately predict and explain their high levels of L2 competence based on the componential structure of the battery built on advancement of the underlying theoretical model. Specifically, the Hi-LAB improves the measurement of aptitude in several ways, which include the theoretical advancement of (i) incorporating WM measures for executive control function and PSTM; and (ii) conceptualising components that underlie implicit learning, which is claimed as crucial during advanced stages of L2 learning by adults. In addition, methodological innovations exist for measuring cognitive components of aptitude through computer-based cognitive tasks (Doughty, 2019). Thus, the final componential battery measures multiple facets of cognitive abilities, i.e., WM of executive function and PSTM, Associative memory, Long-term memory retrieval, Implicit learning, Processing speed, and Auditory perceptual acuity.

The validation study of the Hi-LAB (Linck et al., 2013) evidenced that this battery could distinguish between a learner group of “successful” and a group of “very successful”. However, the effectiveness of its sub-tests in separating the two groups varied. The results revealed that PSTM, Associative memory, and Implicit learning were three subtests that had the highest discriminative power, whereasthe Executive function of WM, Long-term retrieval memory, Processing speed, and Auditory perceptual acuity were not as effective in distinguishing the two groups. Although this sophisticated battery underpinned by theoretical foundations in cognitive science was perceived as an exciting advancement in measuring aptitude, it has not been validated by further research conducted by researchers other than the developers and their associates, which is regretfully due to the restriction of its availability to the research community. Section 2.2.3.2 will provide an overview on this aspect.

The integration of WM in the theoretical framework of language aptitude is also reflected in the Stages Approach (Skehan, 2002, 2012, 2016) that is proposed to take both language and memory into account in conceptualising language aptitude. This theoretical framework will be introduced in the following section.

As reviewed above, aptitude research has generated fruitful knowledge over the past few decades, yet no peculiar hypothesis has emerged as the dominant theoretical framework to conceptualise aptitude construct, mirroring the dynamic and complex nature of L2 learning theories. The emerging theoretical perspectives in aptitude can provide frameworks for the development of instruments to measure the proposed aptitude construct, a critical need for advancing the evidence-based knowledge of aptitude for L2 learning. It is worth noting that the above review of aptitude theories does not include a synthesis of perspectives from cognitive neuroscience, being suggested as a vital part of aptitude conceptualisation in the field (see the reviews in Li, 2019, 2022; Wen et al., 2017), simply because this branch of work engages different research paradigms and techniques that the current study does not involve.

The following two sections will introduce two instructive theoretical frameworks, that is, Skehan's Stages Approach integrated with Wen's Phonological / Executive Model, which inspired the current research and provide its theoretical foundations.

2.1.4 Theoretical frameworks of the current research

2.1.4.1 The Stages Approach

The Stages Approach (Skehan, 2002, 2012, 2016) aims to incorporate the SLA-related insights that have emerged since the creation of the MLAT (Carroll & Sapon, 1959). These insights include, for example, the recognition of different learning processes involved in the development of interlanguage and the importance of specific types of instruction for learners of ages beyond the close of the critical period in various learning contexts. To achieve this goal, Skehan introduces the concept of exploring sequential stages of interlanguage development into the conceptualisation of aptitude, particularly in SLA. He proposes an alternative approach to comprehending the relationship between aptitude and the development of L2. This SLA-compatible aptitude approach posits that individual differences existing in the developmental stages of language learning can predict learning outcomes. The model outlines a series of developmental phases in SLA and estimates the effects of individual differences on learning outcomes in suggested phases. Based on these considerations, measurements of aptitude can be established (Wen & Skehan, 2021).

The Stages Approach explores the theoretical aspects of foreign language aptitude beyond Carroll's classic four-factor framework. It emphasises the importance of incorporating not only general cognitive skills but also language-specific abilities within the construct to achieve a balanced conceptualisation of aptitude related to L2 learning. The theoretical foundation of the model proposes that L2 learning is based on partial access to the innate capacity (i.e., Universal Grammar) for acquiring language. However, this innate capacity is supplemented by other structures and processes (Skehan, 2019). As a result, language is argued to involve "a hybrid system in which domain-general and domain-specific capacities co-exist" (p. 57). Consequently, language aptitude is purported to consist of components related to individual differences in both the language acquisition device—entailing domain-specific capacities for identifying language patterns and processing mechanisms—and the language-making capacity, involving domain-general cognitive operations like general implicit learning, pattern learning, and working memory. The construction of the model also reflects a shift in aptitude research from a 'macro' approach, where aptitude is explored in relation to L2 ultimate attainment (thus proposing its role in predicting learning achievement, as represented by the MLAT), to a 'micro' approach, which involves incorporating measures of aptitude into experimental or quasi-experimental research designs to investigate instructional comparisons or intervention procedures (Skehan, 2019).

The stages proposed in the model and the cognitive processing involved in each stage are listed in Table 2.1 (based on Wen et al., 2017; Wen & Skehan, 2021) together with the corresponding aptitude constructs that can inform the development of aptitude measurements.

The model originally outlines nine stages of language learning (Skehan, 2002, 2016), which are labelled as L2 cognitive processes. Skehan (2016) suggests a broad distinction between the first half of the stages (that is, the L2 cognitive processes of Input processing, Noticing, Pattern identification, Complexification, and Handling feedback) and the second half of stages (that is, the cognitive processes of Error avoidance, Automatisation, Creating a repertoire and achieving salience, and Lexicalisation). The first half of the stages relates to developing knowledge (which involves noticing linguistic input and applying analytic ability to extrapolate patterns from noticed input extensively), whereas the second half focuses on developing control over that knowledge developed in actual use (which involves accessing the language system and proceduralising learned knowledge for production). The Stages Approach essentially addresses the aptitude construct within the framework of cognitive abilities in SLA. It aligns with the notion that working memory and language aptitude (conceptualised by the Carrollian approach) play significant roles in aspects involving input processing, noticing, pattern recognition, complexification, and feedback (Wen & Skehan, 2021).

Table 2.1 The Stages Approach

SLA stages	L2 cognitive processes	Aptitude constructs
Input-oriented	Input processing (segmentation)	*Attentional control* *Working memory*
	Noticing	Phonetic coding ability *Working memory*
Development of interlanguage	Pattern identification	Phonetic coding ability *Working memory* Language analytic ability
	Complexification (e.g., generalising, extending, restructuring, integrating)	Language analytic ability *Working memory*
	Handling feedback	Language analytic ability *Working memory*
Performance-oriented	Error avoidance	*Working memory* *Retrieval memory*
	Automatisation	*Retrieval memory*
	Creating a repertoire, achieving salience	*Retrieval memory* *Chunking*
	Lexicalisation	*Chunking*

Note: Aptitude constructs in italics are components not included in Carroll's four-factor model.

The preliminary assumption, as elucidated by Skehan (2016), is that each stage in the list possesses sufficient distinctiveness to warrant research, thereby informing the creation of potential aptitude subtests. The first half of the set of knowledge–acquisition (*establishing* knowledge) phases generally aligns with the aptitude constructs proposed in theoretical frameworks. However, the latter half of the stages, which focuses on enhancing knowledge *control*, presents greater challenges in establishing connections with aptitude constructs. Regarding the comprehensiveness of the subtests in representing the componential constructs, existing aptitude measures—across the totality of measures available—perhaps encompass all the constructs involved in the first half of the stages, even though a single battery may not yet have included subtests for all constructs completely. This will be discussed in Section 2.2.2.1.

A second point asserted by Skehan (2016) is that auditory or memory processes are not explicitly included in the list of the stages. This omission is not intended to downplay the importance of these processes as verbal learning abilities, which remain central to any aptitude measurements. They should undoubtably be included and examined in all existing aptitude measurements (p. 19).

The construction of the Stages Approach provides valuable insights into the development of measures that require multiple subtests to represent various components of aptitude construct. This is elucidated by outlining a sequential list of phases in SLA, with the aim of enhancing the construct validity of aptitude measures (Skehan, 2016). Consequently, the model serves as one of the foundations for constructing aptitude subtests, enabling the consideration of the influence of cognitive individual differences on learning effectiveness at different stages. It also facilitates the application of distinct aptitude subtests to predict stage-specific learning outcomes and reflect the progress of SLA (Wen & Skehan, 2021).

2.1.4.2 WM and the P/E Model

Working memory and L2 learning

WM is a vital cognitive function enabling individuals to store and manipulate task-relevant information in their mind during various cognitive activities. It serves as a limited capacity memory system underlying critical cognitive processes, including language comprehension and production (Miyake & Shah, 1999). WM involves the temporary storage, manipulation, and maintenance of information essential for ongoing mental operations, whether linguistic or visual (Cowan, 2017; Oberauer et al., 2018; Schwieter, et al., 2022). Despite its limited capacity, WM is assumed to play a larger role than long-term memory in subserving human cognitive and action (e.g., Baddeley et al., 1988; Miyake & Shah, 1999; Lieder & Griffiths, 2020, cited in Wen & Jackon, 2022).

Although WM has been considered a general cognitive function since its original conceptualisation and has been argued as domain-general cognition in the processes of language acquisition (Roque-Gutierrez & Ibbotson, 2023), extensive research has been conducted to re-conceptualise the construct of WM, aiming to tailor it to specific domains of human cognition and behaviour (Logie et al., 2021). Notable early studies that discussed the role of WM in explaining individual differences in language learning and outcomes, such as those by Gathercole and Baddeley (1993) and Baddeley (2003), have significantly influenced how L2 researchers conceptualise the role of WM in L2 learning. This conceptualisation is closely tied to the seminal model of WM introduced by Baddeley and Hitch (1974), which has evolved over almost four decades of research (Williams, 2012).

This model initially posited two functional systems: one for processing and temporarily storing information (the short-term memory), consisting of two dissociable components, that is, the Visuo-spatial Sketchpad for visual and spatial information and the Phonological Loop for verbal information. The other system, termed the Central Executive, with limited capacity, facilitates interaction between the two components and the entire system, including long-term memory (LTM), to handle task requirements. However, this model faced a significant challenge due to extensive evidence for the predictive capacity of WM Span. Specifically, it struggled to explain good performance on the WM span test given that neither the phonological loop nor the sketchpad was modelled to be capable of holding *multiple* sentences, and when the central executive was primarily seen as an attentional system. Consequently, the model was expanded to include an additional component, the Episodic Buffer, added by Baddeley (2000). This component functions to bind multi-dimensional representations or episodes with storage capacity for conscious access (Baddeley, 2017). Hence, the model incorporates the episodic buffer functioning as a temporary storage system to absorb inputs from diverse perceptual sources and integrate them with information from different long-term memory components. This process is assumed to be accessed through conscious awareness to regulate the distribution of attentional resources across various storage subcomponents under the control of central executive, given the limited capacity of WM (Baddeley, 2022).

Within the scope of SLA research, various features of WM, including its limited capacity, attention control and allocation, executive functions in maintaining task-related information and inhibiting interference, as well as the phonological loop for storing novel linguistic occurrences, have gained significant attention. These features play vital roles in explaining individual differences involved in different aspects of L2 learning processes and outcomes, such as sentence processing, reading, speaking, lexical development, and overall proficiency (Juffs & Harrington, 2011; Williams, 2012). Meta-analysis results indicate a positive association between WM and both L2 processing and proficiency outcomes, with an estimated effect size of .255 (Linck, et al., 2014). Different WM components, such as PSTM and central executive function, have also been shown to influence various aspects of L2 acquisition and processing. For example, PSTM is particularly relevant to the learning of vocabulary items (Abu-Rabia, 2001; Speciale et al., 2004; Gathercole, 2006) and formulaic sequences and grammatical rules (French & O'Brien, 2008; Robinson, 1997; Williams & Lovatt, 2003), whereas the central executive plays a more crucial role in cognitive resource-demanding processing and real-time performance (Havik et al., 2009; Mackey et al., 2010; Sagarra, 2007).

Phonological / Executive (P/E) model

In alignment with the recent trend of modelling a domain-specific WM system, Wen (2016) introduces an integrated framework for WM. This framework aims to unveil emerging patterns in hypothetical relationships postulated between WM and the processes and outcomes of L2 learning, which it predicts and explains. Wen's (2016) framework is rooted in the understanding of WM's components and functions regarding potential effects on specific aspects of L2 learning and processing. The framework draws on theoretical assumptions underlying the conceptualisation of WM construct in various models, whether as a multi-componential system proposed by Baddeley and colleagues (see Baddeley, 2022 for reviews), or as a cognitive system with embedded executive processes by Cowan (1999), or as a cognitive capacity for attentional control by Engle (2002).

This SLA domain-specific WM is further elucidated by Wen and Jackson (2022) and Wen et al. (2021) as (a) a collection of cognitive resources of limited capacity, (b) a construct consisting of multiple components and embedded mechanisms, and (c) a set of micro-level subprocesses that can be operationalised and measured separately. The framework, known as P/E model, proposes two key components of WM in language learning and processing: First, phonological WM (PWM) consists of a short-term phonological store and an articulatory rehearsal mechanism, functioning as "a language learning device" playing the roles in the storage, chunking (grouping information into meaningful units), consolidation, and retrieval of novel phonological forms (Wen & Skehan, 2021, p. 10). Second, executive WM (EWM) refers to the attention control and executive function of WM that composes subprocesses for information updating, task-switching, and inhibitory control (Miyake & Friedman, 2012), serving to control and modulate processes during L2 comprehension and production, as well as L2 interactions of feedback or recasts (Wen & Jackson, 2022).

The P/E model provides a foundational theoretical framework that proposes criteria for empirical investigations to uncover the role of WM as a critical construct of individual differences in SLA research. The model's structure is also compatible with the two categories of WM span tasks informed by measurement construction in cognitive psychology (Conway et al., 2005). Simple memory span tasks, such as digit span, letter span, and nonword span tasks, can measure PWM's storage aspect, providing empirical evidence for understanding PWM's role underlying aspects of learning lexical items, phrases, or formulaic chunks, and morphosyntactic structures in L2 learning (Wen & Skehan, 2021). Conversely, complex span tasks, like reading span and operation span tasks, which involve processing-plus-storage in task design, can measure the inhibitory control functions of EWM (Wen et al., 2021). In summary, the P/E Model of is built on the presumption that WM constraints are integral to the language learning mechanisms, influencing and constraining L2 processing, and the evolution and long-term development of language (Wen & Skehan, 2021).

However, despite P/E model's availability to inform WM-related research in SLA, few methodological endeavours have been reported in SLA that aim to develop WM measurements tailored for L2 research since the proposal of P/E model. The heterogeneity of measurements used to investigate the role of WM as a multi-faceted construct in explaining variations in L2 learning may be resulting in an increasingly large body of inconsistent findings. Systematic review (e.g., Shin & Hu, 2020) of WM tasks used in L2 research underscore the need to develop standardised measurements that can accurately assess WM in ways relevant to L2 learning. It is important to note that while the P/E model offers guidance for WM-related research in SLA and types of span tasks have been extensively validated in the field of cognitive psychology, developing reliable and valid WM measures for specific use in SLA still needs to address several theoretical and methodological challenges, which will be discussed in the following sections.

2.2 Measuring language aptitude

There exists a consensus within the field that aptitude comprises a set of cognitive abilities (Doughty & MacKey, 2021). Given that this construct is reflected in the performance of learners in responding to the tests used to measure the construct, aptitude construct is defined by what is measured in the aptitude tests (Li & Zhao, 2021). This section offers an overview of major aptitude batteries that have been widely used or reviewed (e.g., by Li & Zhao, 2021; Skehan, 2023). Subsequently, associated theoretical inquiries and methodological considerations will be presented, providing a foundation for the current research, which aims to develop and validate a new aptitude battery.

2.2.1 An overview of aptitude batteries

The Modern Languages Aptitude Test (MLAT)

Based on Carroll's aptitude theory, the Modern Languages Aptitude Test (the MLAT) (Carroll and Sapon, 1957) stands as the most influential aptitude measurement (Li & Zhao, 2021). It provided compelling evidence in predicting L2 learning and influenced our understanding of the relationship between aptitude and L2 learning (Roehr-Brackin, 2022). For example, in systematic reviews, the MLAT demonstrates the highest predictive validity relative to other batteries (Li, 2015; 2016) and it constitutes contributes 58.8% of all aptitude measures used in L2 aptitude research (Chalmers et al, 2021).

The MLAT contains five subtests, namely Number Learning, Phonetic Script, Spelling Clues, Words in Sentences, and Paired Associates. Specifically, the subtest of **Number Learning** assesses learners' ability to acquire numbers in a new artificial language. Initially, learners listen to number names presented in one-, two-, and three-digit forms. During the testing phase, they are required to write down the numbers they hear in the testing phase. **Phonetic Script** is to evaluate learners' ability to associate sounds with a written symbol. Participants learn phonetic scripts for specific sounds and are subsequently tested by matching the scripts with the corresponding sounds presented. The subtest of **Spelling Clues** measures learners' aptitude for associating sounds with symbols based on their knowledge of English vocabulary. A typical question in this subtest involves presenting learners with a disguised word (e.g., *kloz*) spelled in an unconventional manner. Learners are then prompted to select one of five given options (e.g., *attire*, *nearby*, *stick*, *giant*, *relatives*) that best corresponds in meaning to the disguised word. The subtest of **Word in Sentences** requires learners to identify the linguistic functions of elements within sentences. It begins with practice questions. In each instance, a word in the first sentence is underlined and capitalised. Learners are instructed to identify a word in the second sentence that serves a similar role to the underline word in the initial sentence. Through this exercise, learners grasp the testing format and are subsequently required to complete the testing questions that measure their grammatical sensitivity to sentence elements. The subtest of **Paired Associates**, focusing on rote memory in language learning, assesses memorisation of words in a foreign language (Maya) alongside their corresponding English meaning. Learners are then tested on their ability to match each given test item with the appropriate meaning from five options.

Although the MLAT is rooted in Carroll's aptitude theory, it does not offer the one-to-one correspondence with all four components (i.e., phonetic coding ability, grammatical sensitivity, inductive language learning ability, and associative memory) proposed by the theoretical framework. For example, Number Learning assesses both phonetic coding ability and associative memory. Moreover, concerns have been raised about the MLAT's predictive power beyond the specific context of intensive language training using the audiolingual approach, which is conceived as the methodological assumption underpinning

Carroll’s model and the MLAT. However, these doubts regarding the applicability of the MLAT in diverse learning contexts have been resolved through empirical studies, indicating that aptitude as measured by this battery remains relevant as a predictor of learning in both naturalistic and instructed learning conditions (Roehr, 2012; Roehr-Brackin, 2022).

The Pimsleur Language Aptitude Battery (PLAB)

Sharing the similar theoretical foundation of Carrollian approach in conceptualising aptitude, the Pimsleur Language Aptitude Battery (PLAB) (Pimsleur, 1966) is directed at a specific learner demographic, encompassing grades seven through twelve (ages 12 to 18). This battery consists of six distinct parts, with the initial two parts centring on learners’ self-reported GPAs and a questionnaire regarding their interest in language learning. It is important to note that these components do not directly assess language knowledge or the aptitude construct. The remaining four subtests, integral to the measurement of aptitude, includes Vocabulary, Language Analysis, Sound Discrimination, and Sound–symbol Association. Specifically, the subtest of **Vocabulary** evaluates learners’ grasp of English vocabulary. Stimuli words are presented, and learners are required to select those that closely approximate the given stimuli in meaning. For example, the word *extended* should be chosen from three other words (*prompt, decreased, difficult*) because it closely matches the meaning of the stimulus word *prolonged*. **Language Analysis** is the subtest designed to measure language analytic ability (also known as inductive learning ability). It engages learners in extracting grammatical rules from an artificial language. Test takers then apply these rules to answer questions. The process commences with vocabulary items and a sentence containing those items. Test takers induce grammatical rules and subsequently apply them to select the correct sentence conveying a designated meaning from four available options. The subtest of **Sound Discrimination** assesses the ability to differentiate similar sounds within a new language. Isolated words or phrases containing similar sounds are presented. Learners are subsequently prompted to identify which word occurs in each given sentence during the testing phase. The subtest of **Sound–symbol Association** measures phonetic coding ability by exposing learners to auditorily presented pseudo words. Their task is to select the accurate spelling of the given words from four options based on their knowledge of English pronunciation and orthography.

Despite its inclusion of a subtest for inferring language structures from provided stimuli (a dimension not operationalised in the MLAT) and validation involving 6,000 language learners (Li & Zhao, 2021), the PLAB has not gained popularity as an aptitude measurement for several reasons. First, the battery is primarily perceived as suitable for adolescent learners, thereby constraining its applicability in aptitude-related research that targets diverse learner populations. Second, two of its components (i.e., GPA and motivation) are not cognitive abilities, potentially undermining the battery's exclusivity as a measurement for language aptitude. Third, although the PLAB assesses certain components (such as phonetic coding ability and grammatical sensitivity based on inductive language learning ability) proposed in Carroll's theoretical model, it notably omits associative memory, a crucial cognitive ability essential for learning a new language (Skehan, 2016).

The Cognitive Ability for Novelty in Acquisition of Language–Foreign Test (the CANAL-FT)

Developed by Grigorenko et al. (2000), the Cognitive Ability for Novelty in Acquisition of Language–Foreign Test (the CANAL-FT) is grounded in the CANAL-F theoretical model, which postulates that the ability for handling novelty plays a central role in language learning. This test is designed to simulate a naturalistic learning situation, progressively introducing an artificial language, Ursulu, across various cognitive processes on lexical, morphological, semantic, and syntactic levels. These cognitive processes are (a) selective encoding, which involves discriminating between relevant and less relevant information within a stream of input, aligned with learners' specific goals; (b) accidental encoding, which is to encode and comprehend background information within the context, enhancing both comprehension and knowledge for production; (c) selective comparison, an integral process that involves determining the applicability of previously acquired knowledge to current tasks. This process is linked to learners' capacity to retain ambiguous information in working memory; (d) selective transfer, which entails applying rules decoded or inferred from a previous situation to a new context; and (e) selective combination, combining the segmental data gathered from prior processes (selective and accidental encoding) with existing knowledge, to generate new knowledge. This new knowledge eventually modifies preexisting cognitive representations.

The CANAL-FT consists of nine sections, with five involving immediate recall and four (1–4 below) mirroring these same sections but also with delayed recall. These sections are (1) Learning Meanings of New Vocabulary from Context; (2) Understanding the Meaning of Passages; (3) Continuous Paired-Associate Learning; (4) Sentential Inference; and (5) Learning Language Rules (tested solely on immediate recall). In these sections, learners are exposed to the characteristics of Ursulu contextualised within the knowledge of English. Their cognitive abilities related to encoding, storage, and retrieval of information are evaluated through immediate recall (occurring immediately after learning) and delayed recall (taking place at substantial interval after learning). Notably, the CANAL-FT is distinguished by its incorporation of various input modalities, accounting for information processing in both visual and oral formats.

Grigorenko et al. (2000) validated the CANAL-FT on grounds of its convergent validity with the MLAT, discriminant validity compared to general intelligence measured by two tests, and predictive validity in relation to participants' performance in a language course, as assessed by instructors. However, due to its government sponsorship with a diplomatic focus (Skehan, 2023), this aptitude battery has not been widely used in aptitude–L2 learning research, largely due to its limited availability to other researchers. While sample items are disclosed in the appendix of Grigorenko et al. (2000), the testing items employed remain largely unexplored by subsequent studies. The challenges posed by restricted accessibility in providing reliability and validity evidence will be further discussed in Section 2.2.3.2.

The LLAMA Language Aptitude Tests

The battery of the LLAMA tests (Meara, 2005; Meara & Rogers, 2019) is another aptitude measure related to Carroll's model and loosely based on the components of the MLAT. This battery aims to provide a more user-friendly, freely available, and engaging computer-based interface, leveraging technological advancements (Rogers, et al., 2017). In this regard, the LLAMA tests, administered in a computer-based format, represent a methodological enhancement in measuring aptitude based on the Carrollian approach.

Since its initial release in 2005, the developers of the LLAMA tests have maintained a continuous process of refinement, guided by feedback from the research community (Rogers et al., 2023). This iterative approach reflects a responsible and commendable academic endeavour. The three versions released (the 2005 Windows-based version, the web-based LLAMA v.2.0 in 2018, and the recently launched LLAMA v.3.0) primarily involve cosmetic enhancements to the user interface, rectification of a small number of item errors, scoring revisions, and the inclusion of user identification recording. Notably, the fundamental structure of subtests and content design remains largely unchanged.

The battery consists of four subtests: (i) LLAMA_B is a vocabulary learning module that evaluates users' associative memory by linking unfamiliar vocabulary items (names) to non-existing (fictional) objects presented in images; (ii) LLAMA_D is a subtest designed to measure phonetic recognition ability, in which participants are required to discriminate repeated sounds from new sounds; (iii) LLAMA_E is adapted from Phonetic Script in the MLAT. This subtest measures the ability to form sound–symbol associations. Participants memorise visually presented symbols and their corresponding syllables, later applying these associations to complex two-syllabic new sounds; and (iv) LLAMA_F, a grammatical inferencing test to measure inductive or analytic ability in language learning. Participants inductively work out grammatical and morphological rules (i.e., word order, nominal affixes for gender difference and plurality, conjugating prepositions) from pictures and corresponding verbal forms. They are then assessed on the application of these rules in forming sentences. LLAMA_F has undergone iteration in version 3, transforming into the sole subtest for participants' language production (i.e., composing sentences based on the given lexical options) in the LLAMA battery.

The LLAMA tests have emerged as widely used measurements in recent aptitude-related research, largely attributed to their independence from a test taker's L1 (Rogers, et al., 2017; Roehr-Brackin, 2021) and their open accessibility to researchers in the field, especially when other batteries like the MLAT and the Hi-LAB (Linck et al., 2013) are not available to individual researchers (Chalmers et al., 2021; Li & Zhao, 2021; Roehr-Brackin, 2021). By 2021, this battery and its subtests had been referenced over 4,000 times in Google Scholar (Rogers et al., 2023) and featured in approximately 50 empirical studies published in international journals or book chapters (Bokander & Bylund, 2020).

However, while the LLAMA tests have gained prominence, few studies have rigorously examined their internal validity before employing them in substantive research. Notable researchers, including the battery's creator (Meara, 2005) and others (Singleton, 2017), have raised concerns about their uncritical utilisation. Recent empirical evidence has further raised questions about its internal validity as an aptitude battery (Bokander & Bylund, 2020) and the predictive validity of the original LLAMA (version 1) in relation to L2 learning outcomes (Bokander, 2023). In addition, a relationship between LLAMA_D and implicit learning has been suggested (Granena, 2013, 2019), followed by an increasing number of studies using LLAMA_D to measure aptitude in relation to implicit learning (e.g., Artieda & Muñoz, 2016; Saito, et al., 2019; Yalçın & Spada, 2016; Yi, 2018). However, construct validity of LLAMA_D as a measure of implicit learning aptitude has been questioned by Suzuki (2021a) and Iizuka and DeKeyser (2023).

The High-Level Language Aptitude Battery (the Hi-LAB)

The development of the High-Level Language Aptitude Battery (the Hi-LAB) (Linck et al. 2013) aligns with the theoretical advancements in modern cognitive research, providing a robust foundation for conceptualising language aptitude and incorporating cognitive factors like WM in the battery. The Hi-LAB, as its name suggests, is designed to differentiate exceptionally successful language learners from other individuals. Its conceptualisation of aptitude is informed by research such as DeKeyser (2000), which underscores the significance of aptitude for adult learners. Consequently, the components within this battery are operationally defined, focusing on cognitive and auditory perceptual abilities pertinent to adult learners who exhibit the potential for high-level achievement in L2 learning.

As outlined in Section 2.1.3, the Hi-LAB battery specifically targets gifted adult learners, enabling accurate predictions and explanations of their elevated levels of L2 competence based on its proposed aptitude constructs. These constructs encompass several key elements: Working Memory, consisting of two sub-constructs of Executive Functioning (with Updating in Inhibitory Control and Task Switching components) and Phonological Short-term Memory (PSTM), Associative Memory, Long-term Memory Retrieval, Implicit Learning, Processing Speed, and Auditory Perceptual Acuity. To measure these constructs, the Hi-LAB employs eleven tests, including: (1) **Running Memory Span** measuring sub-construct of Updating of Executive Functioning of WM, (2) **Antisaccade** and (3) **Stroop** assessing Inhibitory Control of Executive Functioning, (4) **Task Switching Numbers** evaluating Task Switching of Executive Functioning, (5) **Letter Span** and (6) **Non-Word Span** targeting the sub-construct of PSTM, (7) **Paired Associates** assessing Associative Memory, (8) **Available Long-term Memory Synonym** evaluating Long-term Memory Retrieval, (9) **Serial Reaction Time** measuring both Implicit Learning and Processing Speed, and (10) **Phonemic Discrimination** and (11) **Phonemic Categorization** both measuring Auditory Perceptual Acuity.

The Hi-LAB has distinctive features compared to the MLAT, which performs well at predicting learning at earlier stages in instructioned context. The Hi-LAB battery includes WM measurements for executive control functions and PSTM, incorporating the concept of implicit learning, and employs computer-based cognitive tasks (Doughty, 2019). As a result, the battery is anticipated to measure diverse aspects of both domain-general cognitive abilities and domain-specific language perceptual abilities, combined in order to define the aptitude construct for high-level attainment among adult learners (Linck, et al., 2013).

The validation study of Hi-LAB (Linck et al., 2013) yielded results supporting the robust predictive validity, particularly of PSTM, implicit learning, and associative memory in predicting high-level attainment. It is also indicated that the predictive capabilities of the Hi-LAB are more pronounced for listening than reading achievement, suggesting that potential benefit of incorporating measures of visual perceptual acuity in future versions of the battery (Doughty, 2019).

While the initial validation study demonstrated promising results of predictive validity, further research is imperative, specifically to establish internal validity (including item quality and subtest construction based on the theoretical framework) and convergent validity (in comparison to other aptitude measurements). However, due to limited availability of this battery to the research community, studies on the Hi-LAB conducted by other researchers rather than the authors and their affiliates have been scarce in the last decade since its inception. This limitation hampers comprehensive knowledge about this battery and the underlying theoretical framework, despite claims of it being a significant advancement in language aptitude research (Linck et al., 2013).

Summary of the section

This section provided an extensive overview of various batteries that have been used to collect data related to language aptitude, significantly contributing to our understanding of this multi-faceted construct. These batteries are of great significance, as they provide scores that represent aptitude and contribute to its definition based on the aspects measured (Li & Zhao, 2021). Nonetheless, these batteries face a few theoretical and methodological challenges that could potentially compromise their instrumental validity. Consequently, these challenges underscore the need for efforts to design, validate, and introduce new aptitude measures. Subsequent sections will delve into these pertinent questions and challenges, providing the foundation for the development of a new aptitude battery in the current research.

2.2.2 Theoretical inquiries on measuring aptitude

2.2.2.1 What components should be included?

Debates persist regarding the theoretical underpinnings of the multi-faceted construct of language aptitude and its role in explaining intricate aspects of L2 learning (Doughty & MacKey, 2021). The theoretical exploration of aptitude construct necessitates the creation of aptitude batteries, intrinsically linked to the development of aptitude measures. Given that an aptitude test needs to include various components measuring diverse aptitude constructs (DeKeyser & Koeth, 2011), this section aims to provide an overview of the componential constructs proposed in the theoretical models reviewed above, as illustrated in Table 2.2. These constructs are particularly relevant to the initial phases of language learning, focusing solely on language perception and excluding language production, as categorised by the Stages Approach (Skehan, 2016). Subsequently, this section presents a summary of whether these componential constructs are incorporated within existing aptitude batteries, as illustrated in Table 2.3.

Table 2.2 List of aptitude componential constructs in the theoretical frameworks

Construct	Theoretical frameworks					
	Carrollian	LCDH	ACAD	CANAL-F	Hi-LAB	Stages
Phonetic coding ability	√	√	×	√	√	√
Associative memory	√	√	√	√	√	×
Language analytic ability	√	√	√	√	×	√
Working memory	×	×	√	√	√	√

Note. Keys to column headings: Carrollian = Carroll’s classic model; LCDH = the Linguistic Coding Differences Hypothesis model; ACAD = the Aptitude Complexes/Ability Differential frameworks; CANAL-F = the Cognitive Ability for Novelty in Language Acquisition-Foreign model; Hi-LAB = the High-Level Language Aptitude Battery; Stages = the Stages Approach; MLAT = the Modern Languages Aptitude Test; PLAB = the Pimsleur Language Aptitude Battery; CANAL-FT = the Cognitive Ability for Novelty in Language Acquisition-Foreign Test; LLAMA = the LLAMA Tests; √ = included, × = not specified, * = related

Table 3.3 List of aptitude componential constructs in major existing batteries

Construct	Aptitude batteries				
	MLAT	PLAB	CANAL-FT	Hi-LAB	LLAMA
Phonetic coding ability	√	√	√	√	√
Associative memory	√	×	√	√	√
Language analytic ability	√	√	√	×	√
Working memory	×	×	*	√	×

Note. Keys to column headings: MLAT = the Modern Languages Aptitude Test; PLAB = the Pimsleur Language Aptitude Battery; CANAL-FT = the Cognitive Ability for Novelty in Language Acquisition-Foreign Test; LLAMA = the LLAMA Tests; √ = included, × = not specified, * = related

Phonetic coding ability

The construct of phonetic coding ability, that is, the ability to code and retain unfamiliar auditory input, has been conceptualised as an important component of aptitude and included in almost all aptitude theoretical frameworks (except ACAD), as reviewed in Section 2.1.3. To be exact, the classic Carrollian model (Carroll, 1962, 1981, 1990, 1993) includes phonetic coding ability as one of the four measurable abilities. The LCDH model (Sparks & Ganschow, 2001), similarly, includes phonetic coding ability as one of the four basic components of aptitude. Although the ACAD frameworks (Robinson, 2002, 2005, 2007, 2012) do not specify phonetic coding ability in the core cognitive abilities, the frameworks suggest the supplementation and expansion rather than the replacement of traditional aptitude model (the Carrollian model) and its components. The CANAL-F model (Grigorenko et al., 2000) proposes five acquisition processes when learners handle novel language occurrences while learning a new language. These processes are operationalised at lexical, morphological, semantic, and syntactic levels and involve selective encoding and accidental encoding abilities in both oral and visual input modes. The Hi-LAB model also includes auditory perceptual acuity as one of the cognitive abilities conceptualised, which is operationalised by phonemic discrimination and phonemic categorization. Among the stages outlined in the Stages Approach (Skehan, 2002, 2012, 2016) of L2 learning, phonetic coding ability is involved in the cognitive processes of Input processing, Noticing, and Pattern recognition of linguistic input.

Associative memory

The construct of associative memory, also known as rote memory, has been included in all theoretical frameworks. This construct generally refers to the ability to encode and store unfamiliar linguistic input in memory, with the stored forms being retrievable for future use. The Carrollian model employs associative memory to define the ability to establish and retain links between verbal elements in L1 and L2, offering a similar definition. In the LCDH model, memory skill for L2 paired-associate learning is incorporated. The ACAD frameworks go further, encompassing more memory-related constructs within the model, among which rote memory is positioned within the innermost cognitive abilities. Additionally, memory for contingent speech and memory for contingent text are proposed as aptitude complexes in the frameworks. Although a separate construct of associative memory is not explicitly outlined in the CANAL-F model, the relevant abilities involved encoding, storage, and retrieval of information are operationalised within the framework. In the Hi-LAB model, memory is notably emphasised as a significant cognitive ability in aptitude conceptualisation. Both associative memory and long-term memory retrieval are integral components of the aptitude construct. Finally, while associative memory is not not explicitly listed in the categories of L2 cognitive processes in the Stages Approach, Skehan (2016) acknowledges that auditory or memory processes are central to any aptitude measurement.

Language analytic ability

Language analytic ability refers to learners' ability to 'infer rules of language and make linguistic generalizations or extrapolations' (Skehan, 1998, p. 204). This construct has been integrated in Carroll's model and the LCDH model, particularly in relation to L2 learning within instructed context, except in the Hi-LAB model. The ACAD frameworks incorporate pattern recognition and grammatical sensitivity within the ten first-order abilities in the innermost circle, and metalinguistic rule rehearsal as a complex in the second circle. While the CANAL-F model does not explicitly label a construct as language analytic ability, analytic abilities for managing structures, deviations of words, and the acquisition and application of grammatical rules are involved in the five acquisition processes. In the Stages Approach, the development of interlanguage involves cognitive processes such as pattern identification, complexification (generalising, extending, restructuring, and integrating), and handling feedback—all of which engage the construct of language analytic ability.

An exception to including language analytic ability as a componential construct of aptitude is the Hi-LAB model. This model emphasises components grounded in implicit learning mechanism, which are believed to be predictive of highly advanced levels of L2 learning, thus excluding analytic ability as a primary explicit learning mechanism.

Working memory

The inclusion of WM as a component in the theoretical framework of language aptitude has sparked more controversy than any of the previously reviewed components. Notably, WM has not been proposed in Carroll's model or the LCDH model. Skehan (2016) contends that the phonetic coding ability in Carroll's model entails processing unfamiliar sounds in a way that involves some basic structural processing, akin to the conceptualisation of the phonological buffer of WM. This suggests that WM should be integrated into aptitude model to reflect the evolution of aptitude theory. In the ACAD model, Phonological Working Memory Capacity, Phonological Working Memory Speed, Text Working Memory Capacity, and Test Working Memory Speed are first-order abilities situated in the innermost circle, underpinning Aptitude Complexes in the second circle. The CANAL-F model relies on memory-related cognitive abilities to support its five knowledge acquisition processes, which involve comprehending linguistic information, encoding it into WM, transferring it, and storing it in long-term memory. While the CANAL-F Test does not specifically measure WM, its sections (Learning Meanings of Neologisms from Context, Understanding the Meaning of Passages, Continuous Paired-Associate Learning, Sentential Inference, and Learning Language Rules) likely engage WM to varying extents. The Hi-LAB model includes distinct Executive Functions in WM alongside Phonological Short-term Memory, offering a comprehensive conceptualisation of WM through various subtests. Wen and Skehan (2021) further elaborate on WM as a componential construct of aptitude, highlighting the interaction between an acquisition-oriented approach and memory-related assumptions, as well as the interdependent nature of memory and language aptitude.

In summary, Table 2.2 and Table 2.3 provide an overview of the reviewed components of aptitude in theoretical models and major existing aptitude batteries. Interestingly, these aptitude batteries do not appear to cover all four componential constructs proposed in the evolution of language aptitude theories. Since aptitude is defined by what is measured in aptitude tests (Li & Zhao, 2021), the development and refinement of aptitude batteries should align with theoretical advancements. However, this alignment often falls short in practice. The CANAL-FT is the battery that loosely incorporates all four constructs in its instrumentation. However, its restricted access to other researchers is a regrettable issue, which will be further discussed in Section 2.2.3.2.

2.2.2.2 Should aptitude measures be domain specific or general?

The question of domain generality versus specificity has been extensively discussed (Skehan, 2016, 2019; Wen et al., 2017) in the context of whether cognitive abilities for language learning differ fundamentally from those for other domains. This argument essentially centres on the unknown nature of language learning mechanisms, debating whether infants rely on mechanisms specifically developed for language acquisition or pre-existing mechanisms for general learning to understand language (Saffran & Thiessen, 2007).

When extending the discussion to L2 learning, the definition of language aptitude as cognitive abilities facilitating success in learning an additional language raises the question of whether it is distinct from general intelligence or learning abilities (Li, 2019), thus rendering language aptitude domain specific.

The domain generality–specificity debate for language aptitude has practical implications for the content design in aptitude batteries. Should aptitude measures be based on language materials, or can non-linguistic materials suffice? Two primary justifications support the former, as outlined by Skehan (2016, 2019) and Wen et al., (2017). First, language material processing might involve language-oriented capacities resulting from a critical period, accruing as individual differences for adult learners (Carroll, 1973). Second, Carroll's (1993) hierarchical theory proposes sub-abilities such as verbal, mathematical, musical, and mechanical aptitudes, indicating that the presence of these specialised abilities within the verbal domain may contribute to language aptitude. Existing aptitude batteries, except the Hi-LAB, align with the domain-specific approach.

Contrarily, counter-justifications for a domain-general approach consider that learning, whether L1, L2, or non-language domains, shares similar processes driven by usage or frequency (Ellis, 2002). Such perspectives suggest that aptitude measurements could rely on general learning tasks rather than specific language materials.

The domain-general perspective is embodied in the design of the Hi-LAB, which posits that proposed cognitive abilities, irrespective of their theoretical language-related associations, predict L2 learning (Li & Zhao, 2021). The Hi-LAB includes subtests (i.e., Task Switching Numbers Test for Executive Functioning and Serial Rection Time Test for Sequence Learning) completely eschew language specific materials, following the broad cognition approach in the test design. Two other subtests for Inhibitory Control in Executive Functioning (i.e., Antisaccade Test and Stroop Test) incorporate language materials (letters and words) in the instrumentation, though, the design of the subtests does not rely much on the linguistic characteristics of the materials (Skehan, 2016; Wen et al., 2017). The remaining seven subtests use language materials to measure memory-related cognitive abilities (see Linck et al., 2013) but again do not focus on or require attention to the language properties.

Recent research also incorporates domain-general cognitive abilities in language aptitude conceptualisation, aiming to elucidate the predictive power of other domain-general cognitive abilities beyond WM in explaining L2 learning. For example, Saito et al. (2021) measure different aspects of auditory processing abilities (i.e., audio-motor integration and auditory acuity for temporal and spectral information) and investigate the relationship between these domain-general abilities proposed as a perceptual-cognitive foundation of human language learning and L2 speech learning. However, it is important to consider whether their research is domain general in nature, as the materials employed in the measurement can arguably be perceived as linguistic materials, suggesting that the approach they proposed could in fact be domain specific (Wen & Skehan, 2021).

Although recently developed aptitude measure (the Hi-LAB) attempts to include subtests for the components of higher order cognitive abilities that are generally predictive for learning, research on the relationship between domain-specific components and other primary cognitive abilities is still notably lacking (Wen & Skehan, 2021). This is particularly pronounced in the measures of WM. Most researchers may agree that WM, as a multi-faceted construct in relation to language learning, comprises the storage system being domain specific and the executive component being domain general (Williams, 2012). In the recent achievements in modelling language aptitude, the P/E model (Wen, 2016) provides specific implications that the phonological 'P' part of WM is a language learning device (following Baddeley et al., 1998, cited in Wen, et al. 2023b), which is a domain-specific ability for L2 learning. The executive control 'E' part of WM, on the other hand, is the ability to control attention and suppress competing resources that may involve cognitive processes, such as recall of the temporarily held information, rather than memory *per se* (Juffs & Harrington, 2011). This ability is used to achieve cognitively demanding tasks related to L2 subskills (Wen, et al. 2023b), hence it could be domain general. This notion has been represented in a noticeable increase of using operation span tasks (solving mathematical operations while trying to remember words) to measure the executive functioning part of WM in L2 research. This leads to substantially more occurrences of using nonverbal WM tasks than verbal tasks (such as reading span tasks in which grammatical or semantic judgements of sentences are made while trying to remember words embedded in the sentences) in recent studies (see the systematic review in Shin & Hu, 2020).

However, the executive function of WM measured by complex span tasks can be operationalised in dual tasks reflecting both processing and storage and in which domain-general *and* domain-specific stimuli can both be used in instrumentation. Which type of stimuli should be used to operationalise executive function of WM has not been specified in L2 research. The convergent validity of verbal and non-verbal span tasks as WM measurements in L2 research has not been systematically investigated (cf. Draheim et al., 2018, which compares the validity of three nonverbal span tasks, that is, operation span, symmetry span and rotation span tasks, but does not include reading span tasks). Therefore, it has not been established whether using domain-specific stimuli would display the same construct validity as using domain-general stimuli in complex span tasks to measure WM in L2 learning research. This casts doubt on the synthesis of research findings in relation to the predictive validity of WM tests on L2 learning outcomes, as methodological disparities are apparent.

Given the availability of domain-specific complex span tasks, such as reading span tasks and listening span tasks, it seems prudent to include them in language aptitude batteries. Cai and Dong (2012) suggest that researchers need to attend to specific aspects of WM theory and make decisions in a hierarchical order of weighting. Specifically, the information types of the WM measure (verbal versus non-verbal) related to the domain generality–specificity are foremost decisions to make, followed by the encoding modalities (listening versus reading span tasks). The encoding languages (i.e., whether the test is conducted in participants' L1 or L2) are the final set of decisions to make.

Practical reasons may explain a decision to use more domain-general measurements. The potential confounding of language skills and experience can contribute to the variance of test results (see, e.g., Farmer et al., 2016). Domain-specific measures may have to deal with more challenges in terms of whether task stimuli should be in L1 or L2, as the language used may influence the outcomes (Linck et al., 2014). Therefore, a domain-specific measurement that has been validated with a cohort of participants in one study would require further scrutiny of cross-linguistic mapping, accurate translation, and revalidation when it is used in other research that has a different cohort of participants in another context, given that diverse language backgrounds may be involved (Wen et al., 2017).

In summary, most subtests of the existing aptitude batteries have been developed using a domain-specific approach. However, the tests for the operation-related aspect of WM (referred to as the Executive control part of the P/E model) does not commonly use linguistic materials. As a result, there is an incompatibility between the Stages Approach, which emphasises the domain specificity of language aptitude components, and the P/E model, which emphasises the domain generality of (part of) WM construct (Wen & Skehan, 2021). This calls for methodological efforts in developing measures to generate empirical evidence that can lead to inform a better understanding of these two theoretical models.

2.2.2.3 Does modality matter?

The theoretical frameworks of aptitude discussed in previous sections have conceptualised the componential constructs of cognitive abilities as the foundations for developing aptitude measures. Among these components, some involve the processing linguistic occurrences perceived in either aural or written (visual) forms. For instance, the ACAD frameworks include both oral content (memory for contingent speech and deep sematic processing) and written content (deep semantic processing and memory for contingent text) for incidental learning, suggesting that the construct of aptitude may not always be operationalised in one modality.

However, except the CANAL-FT, the input modality used in major aptitude batteries has been left unspecified. Components like associative memory, language analytic ability, and executive control capacity in working memory can be measured using either aural or written stimuli. The choice of modality for operationalising aptitude components is not purely methodological as it holds theoretical significance for understanding the interplay between components and their roles in explaining various aspects of L2 learning. Despite studies demonstrating the impact of input modalities on language learning (e.g., Kim & Godfroid, 2019; Plonsky et al., 2020; Webb & Chang, 2020; Zhao, et al. 2021) and the effects of task modality on cognitive individual differences and language learning (e.g., Zalbidea, 2017; Zalbidea & Sanz, 2020), the influence of modality on measuring language aptitude remains an unanswered question, warranting empirical investigation.

This section addresses the overlooked issue of modality concerning specific aptitude constructs and their representation in existing aptitude batteries, establishing one of the rationales for the current research.

Phonetic coding ability

Despite all aptitude batteries featuring subtests for phonetic coding ability (analysing and retaining unfamiliar sounds), their design involves mixed modalities. For example, in the MLAT, Phonetic Script tests sound-to-writing symbol associations, and Spelling Clues links sounds (presented in mis-spelled written forms) with vocabulary items. Both subtests measure participants' phonetic coding ability in the way that the form-meaning connection is retrieved orthographically. Similar approaches appear in the PLAB (Part 6 Sound-Symbol Association) and recent LLAMA_E. In this design, while phonetic coding involves handling sounds with primitive structures for better retention (Skehan, 2016), the ability to apply orthographical rules to recognise new sounds and link them to scripts is also tested. This ability, referred to as perceptibility, involves perceiving an auditory contrast that is systematically represented in writing (see the systematic review in Hayes-Harb & Barrios, 2021). Thus, these tests may introduce confounds to results and potentially compromise any (assumed) unidimensionality of the psychometric measurements for phonetic coding ability.

Another concern with using sound-text association for measuring phonetic coding ability is that this method may introduce confounds related to learners' knowledge of Roman alphabetic writing systems (which have specific principles for mapping between graphemes and spoken language units; Chang et al., 2016), especially for those with different L1s like Chinese. Such methodological challenges emerge in aptitude battery design.

Associative memory

Associative memory, often termed rote memory, is involved when forming links between representations of the L1 and a new language. This ability is typically evaluated through learning vocabulary items (some can be pseudo words to control the confound of the existing knowledge of vocabulary) in the written modality, with the corresponding meaning displayed in the form of textual equivalent in L1, or as a pictorial referent. Despite vocabulary research indicating the effects of modality in testing vocabulary knowledge (e.g., Masrai, 2019; Milton & Hopkins, 2006; Mizumoto & Shimamoto, 2008), the methodological convention of using vocabulary items solely in the written form to measure associative memory in aptitude batteries is widely accepted. A notable exception is the CANAL-FT (Grigorenko, et al., 2000). In Section 3 (Continuous Paired-Associate Learning) of the battery, the selective comparison and combination of lexical and morphological materials encoded into WM and retained in long-term memory are measured, using materials in both visual and oral forms.

Relying solely on written vocabulary items to measure associative memory assumes that orthographical forms can accurately operationalise the ability to store and retrieve phonological forms. This may compromise the construct validity of the subtest of associative memory in aptitude measures and the predictive power of the measurements in explaining the learning outcomes, particularly when aural input (or production) is substantially involved.

Language Analytic Ability

Language Analytic Ability (LAA) involves inferring linguistic rules and generalising or extrapolating linguistic concepts. It has been proposed to play a central role in L2 learning (Skehan, 1998) and has been included in nearly all aptitude models, except for the Hi-LAB.

Although LAA is not explicitly outlined in Carroll's aptitude model, it underlies two constituent abilities, i.e., grammatical sensitivity (the ability to recognize the grammatical function of words) and inductive learning ability (the ability to infer grammatical rules from language examples) (Carroll, 1990; Roehr, 2008). The Stages Approach situates LAA in the interlanguage development, in which LAA involves three consecutive processes, i.e., Pattern identification, Complexification (e.g., generalising, extending, restructuring, and integrating), and Handling feedback (Skehan, 2016).

The construct of LAA is measured, often though not always, by engaging learners attending to the target features in the learning phase, and then testing their ability to apply the rules to complete the questions or tasks in the testing phase. In the subtests for LAA in all aptitude batteries except the CANAL-FT (see Table 2.3), materials are presented in the written modality. The CANAL-FT addresses the modality in its design. In Section 2 (Understanding the Meaning of Passages), questions requiring inference and application of semantic information are used to measure the ability to apply selective and accidental encoding, comparison, and combination for both visually and orally presented materials that are beyond the lexical level. Section 3 (Continuous Paired-Associate Learning) measures the selective comparison and combination of lexical and morphological materials in both visual and oral forms. Tasks presented visually or orally in Section 4 (Sentential Inference), which measures selective encoding, comparison, transfer, and combination, firstly at the syntactic and morphological levels, then at the lexical and semantic levels. The section that has stimuli presented in written form only is Section 5 (Learning Language Rules), which is administered as the last section to measure participants' overall learning of the Ursulu language by using items (lexical, semantic, morphological, and syntactic) provided throughout the test.

Similar to the concerns about the construct validity of the subtest of associative memory in aptitude measurements, measuring LAA solely in the written modality has not been explicitly justified in research to date. Since LAA operates in a critical phase of input processing during the creation of rules, it handles the information received for processing, which stems from "the product of the phonemic coding stage" and is subsequently processed, inspected for consistent patterns, and serves as the foundation for rule formation (Skehan, 1998, p. 204). Therefore, the input involved in LAA does not rule out language examples being presented in the aural form. This raises concerns about the construct validity of the LAA subtests and their predictive validity in explaining learning outcomes, particularly when measures solely employ stimuli in the written modality in learning situations where novel auditory input is substantially involved.

Working memory

In Baddeley's model (1986, 1992), the WM system consists of two distinct subsystems—a phonological loop and a visuo-spatial sketchpad—that independently process information. Specifically, the phonological loop processes auditory verbal information (such as listening to words), while the visuo-spatial sketchpad handles visual verbal information (e.g., reading words). The theoretical model of WM in L2 learning, such as the P/E model (Wen, 2016), specifies two components: the Phonological Short-term Memory (PSTM) and the Executive Control capacity. When measuring these two components, using stimuli presented in the aural modality appears suitable for PSTM. However, the question of which modality should be employed to measure executive control and whether variations in modality influence test outcomes remain unexamined in existing measures.

Using the Hi-LAB as an example, which is the only aptitude battery featuring specific subtests for WM (see Table 2.2), the stimuli's modality has not undergone systematic control. This battery incorporates two domain-specific span tasks, namely the letter span tasks and the nonword span tasks, both of which present stimuli visually on the screen to assess PSTM. Additionally, it includes four subtests to measure executive functioning of WM, using domain-general materials that do not necessarily involve cognitive processing of linguistic attributes. Among these, three subtests (that is, Antisaccade Test, Stroop Test, and Task Switching Numbers Test) employ visual stimuli, while the Running Memory Span Test uses auditorily-presented letters.

Two types of complex span tasks using linguistic materials—the reading span tasks and the listening span tasks— are available to measure executive control functioning in WM. However, a systematic investigation into whether these two measurements yield consistent (highly correlated) results has not been conducted. Considering the potential impact of modality effects on measuring WM (e.g., Fougnie & Marois, 2011) would facilitate evaluation of potential pitfalls in instrument design. As discussed in the preceding section, the encoding modalities (listening versus reading span tasks) of WM tests could be important (Cai & Dong, 2012).

2.2.3 Methodological challenges in measuring aptitude

2.2.3.1 Reliability and validity of aptitude batteries

Reliability and validity are critical for the robustness of measurements, given the impact that the design of measurements can exert on results. Aptitude, being a construct composed of various cognitive individual differences, finds its definition somewhat shaped by the measures that assess it (Li & Zhao, 2021). Thus, it is crucial to ascertain the reliability and validity of aptitude measurements *prior* to conducting aptitude–learning research. This section reviews the reliability and validity evidence pertaining to existing aptitude batteries. Subsequently, it will introduce established frameworks that can offer insights into the validation of these aptitude measures.

Reliability and validity of research measures

Reliability and validity of educational measures are succinctly summarised by Cohen et al. (2011). Reliability pertains to the stability, equivalence, and internal consistency of an instrument, while validity refers to the extent to which an instrument measures its intended target. To elaborate more, reliability, often referred to as stability, denotes a measure’s ability to consistently yield comparable data from the same (or similar) respondents. Equivalence related to reliability is the capacity of different iterations of a data collection tool to yield comparable outcomes, as well as the level of agreement among various raters when human judgement is involved in the tool. Internal consistency, a facet of reliability, assesses how well a set of elements cohesively converge to produce coherent examination. This is often quantified through coefficients like Cronbach's (1951) alpha, reflecting inter-item correlations that represent the relationship of each item with the total of all other items within the measurement. In terms of the quality or acceptability of an educational measurement, reliability signifies a form of accuracy in obtaining consistent outcomes when the measurement is conducted on different occasions or by different samples of test takers from the same population.

Validity of a measurement is defined as "the extent that a test measures what it is supposed to measure" (Henning, 1987, p.89). In most empirical kinds of validity, a necessary but not a sufficient condition in the form of reliability is introduced: a test can exhibit reliability without necessarily possessing validity for a specific purpose, but validity cannot exist without prior reliability. Consequently, the emphasis often leans toward 'validating' a test rather than solely establishing its reliability (Henning, 1987).

Transparent reporting of reliability and validity evidence of measurements is fundamental in research, serving to ensure the credibility of results obtained. However, this crucial practice has often not been granted the necessary attention. Questionable measurement practices in the field of Psychology have drawn criticism for downplaying the importance of measurement, which obscures a remarkable range of choices available to researchers (i.e., the source of researcher degrees of freedom). Ultimately, these practices pose a significant risk to the integrity of accumulated knowledge within psychological science, characterised as a 'measurement schmeasurement' attitude (Flake & Fried, 2020).

In the field of SLA, research has revealed gaps in this aspect (e.g., Plonsky & Gass, 2011; Plonsky, 2013), raising concerns about the validity of methods and the scientific rigour of research findings, especially given the prevalence of developing or modifying instruments to measure a range of constructs in L2 research (Li & Prior, 2022). Measurements have often not been adequately described, leading readers to rely on them without knowing about their reliability and validity, initially discussed by Bachman and Cohen (1998), and subsequently by Cohen and Macaro (2013) and Norris and Ortega (2012). More recently, the practice of incomplete reporting of validity and reliability have been characterised as a 'sin' that undermines statistical quality in psychometrics within the field (see Al-Hoorie & Vitta, 2019).

Systematic reviews of reporting practices in L2 research have revealed areas that demand enhanced transparency. For instance, Plonsky (2013) found that a mere 21% of the 606 articles published between 1990 and 2010 in two key journals (*Language Learning* and *Studies in Second Language Acquisition*) in L2 research reported instrument reliability. Similarly, Derrick's (2016) investigation disclosed that reliability coefficients were reported for only 28% of the instruments employed across 385 empirical research articles published in three leading journals (*Modern Language Journal*, *Language Learning* and *Studies in Second Language Acquisition*) between 2009 to 2013. Moreover, Shin and Hu's (2022) meta-analytic review of working memory measures used in L2 research underscores the inadequate reporting of instrument reliability, with just 15% of the samples providing adequate reports, even lower than the rate of 21% reported by Plonsky (2013). These studies emphasise the need for greater reporting of reliability across various instruments in L2 research.

Returning to the topic of reliability and validity of aptitude measurements, it is worth noting that validation research has been conducted on both the MLAT and the PLAB, involving planned variations on substantial numbers of participants and extensively detailed in publications. This not only provides insights into how aptitude batteries should be validated but also underscores the importance of funding in facilitating validation research to refine aptitude measures (Skehan, 2023). However, when it comes to more recently developed batteries like the CANAL-FT and the Hi-LAB, there remains a gap in comprehensive validation evidence, partly caused by limited accessibility to these batteries.

In addition to the lack of reporting, low coefficients perhaps raise even more serious questions about the quality of aptitude-related research in the field. A comprehensive overview of methodology of research on language aptitude is offered by Li and Zhao (2021). In this review, reliability, as a foundation for validity, has been reported as high or at least acceptable for tests of explicit aptitude (like the MLAT), falling within the coefficient range of 0.7 and 0.9, which surpasses the recommended threshold of > .70 (Field, 2013).

However, one widely used aptitude battery, the LLAMA tests, has faced challenges in terms of its reliability and internal validity in empirical studies. For example, Bokander and Bylund (2020) revealed a complicated pattern of reliability coefficients (Cronbach's α) across all subtests in the LLAMA tests (version 2). These coefficients could not be easily interpreted as evidence supporting the notion that measures related to explicit learning yield higher reliability coefficients than those related to implicit learning. Although LLAMA_B (measuring explicit learning ability in associative memory) has the highest reliability coefficient (.81), LLAMA_F did not demonstrate a high reliability coefficient (.66), even with misfitting items being excluded, despite also aiming to measure another aspect of explicit learning ability (language analytic ability). LLAMA_D, reported the lowest reliability coefficient (.60), intended to measure implicit learning ability (Granena, 2013). This led the authors to conclude that internal validity of the LLAMA battery is questionable and express significant concerns about its widespread usage in SLA research. Similarly, the developers of the battery, Rogers and Meara (2019), reported low reliability coefficients for the LLAMA tests of the same version. Furthermore, recent investigations exploring LLAMA_D have raised doubts about its construct validity as a measure of implicit learning ability (see Iizuka & DeKeyser, 2023; Suzuki, 2021a).

There are various strategies to enhance instrument reliability, such as removing misfitting items or items with low discrimination, increasing the number of items in the test, or applying different scoring criteria (see Bokander & Bylund, 2020; Shin & Hu, 2022; Li & Zhao, 2021). However, although reliability is a necessary condition, it is not sufficient to validate measures (Henning, 1987). Thus, the focus should (also) be on rigorously scrutinising the validity of a measure, particularly when the measure is newly developed for a multi-faceted construct like language aptitude. In the welcome message from the editorial team of a new journal, *Research Methods in Applied Linguistics*, dedicated to advancing methods and approaches in language-related research, Li and Prior (2022) highlights two broad categories—internal validity and external validity—when it comes to evidence for construct validity. They also propose prioritising the examination of internal validity for newly developed measures and external validity for measures that already exist in the field.

In the following section, an established validation framework will be introduced, particularly focusing on the investigation of internal validity within the context of aptitude battery. To illustrate this framework, an example related to the examination of internal validity of an aptitude battery will be introduced.

A framework for validating aptitude measures

The concept of validity is central to measurement evaluation to determine what constitutes a valid test and how scores derived from such tests can be deemed valid. Messick (1989) introduces a comprehensive perspective on validity, which not only includes the construct validity of the measured trait but also considers the broader consequences associated with the influence of the measurement. He suggests that we can assess this unitary concept of validity by considering it as an integrated accumulation of evidence, rather than as isolated pieces of evidence.

The Interpretation/Use Argument, as proposed by Kane (2006, 2013), shifts the focus of validity from being an inherent property of a single test to being an argument or notion of how scores are appropriately employed. This perspective acknowledges that validity resides not in the test itself, but in the argument and rationale underlying the uses of scores. Within the framework of the Interpretation/Use Argument, the interpretive argument forms a chain of inferences and assumptions that underpin the interpretation and application of test scores. In essence, validation involves scrutinising the coherence of this argument and assessing the validity of both the theoretical underpinnings and empirical evidence through a series of analyses and empirical investigations. This process aims to establish the soundness of the proposed inferences by aligning them with the theoretical framework and the empirical evidence (Kane, 2006, 2013).

To addressing the need for specificity regarding the types of appropriate evidence and the required level of sufficiency within the framework, particularly in the context of applied linguistics, Purpura et al. (2015) engaged in an extensive exploration of how Kane’s (2006) principles could be effectively employed. They investigated the contextualisation of the underlying principles of validation within the domain of SLA, using examples from a study by Révész (2012) that examined the effects of recasts. In this endeavour, Purpura et al. adapted Kane’s validation framework to systematically identify, scrutinise, and substantiate claims related to the constructs being measured in the exemplary study. This involves the idea that specific language usage patterns can be elicited, manipulated, or measured through the execution of specific empirical tasks.

Specifically, Purpura et al. proposed an additional *Domain Description Inferences* to establish connections between the target domain and the performance sample obtained through tasks. These inferences are introduced sequentially as follows:

(1) *Scoring Inference* (also referred to as Evaluation Inference) connects performance samples to observed scores. This inference assumes that the performance produced can generate the observed scores, indicating the intended test construct. It involves statistical analysis to either support or refute claims regarding the functionality of the measurement.

(2) *Generalisation Inference* links the observed scores to expected scores, assumed to be consistently obtained regardless of variations in measurement conditions.

(3) *Explanation Inference* connects the expected scores to the underlying test construct. The underlying assumption is that observed scores generated from the tasks accurately reflect the constructs being measured or the associated network related to the construct. This inference provides support for the appropriate operationalisation of the theoretical construct, which needs to be substantiated through examinations of the internal structure of the test, comparisons of sample group differences, or assessments of task difficulty.

(4) *Extrapolation Inference* connects the construct-related score with a real-life or target score. It offers evidence that a score on a measure corresponds to performance on real-like tasks that engage the same knowledge, skills and abilities.

(5) *Utilization Inference* connects the target score to the use of the score for decision-making purposes, representing the practical implications of applied linguistics research.

Building upon the research of Kane (2006) and Purpura et al. (2015), Bokander & Bylund (2020) applied a schema for the validation of aptitude scores from the LLAMA tests. The schema includes inferences relating to scoring, generalization, explanation, extrapolation, and implication. Specifically, the first inferential level is the *scoring inference*, which results in an observed score and involves converting individual responses into a scale score for each subtest. This initial level of inference also encompasses factors such as the item format and variables related to the test context, such as the test takers' comprehension of the task. The *generalization inference*, which leads to a universe score, pertains to determining whether a scale score is a reliable measure of its underlying construct. In other words, it examines whether the scale score validly represents all theoretically possible items that assess the construct. The *explanation inference*, which leads to a construct interpretation, relies on construct validity evidence that specifically addresses the theoretical justification of each subtest. The authors explain that in the case of the LLAMA, this inference involves determining whether each subtest accurately measures its intended construct and does not capture any extraneous construct. The *extrapolation inference*, along with its corresponding target score, shares similarities with criterion validity (concurrent or predictive) in classical test theory. This involves assessing the correlations between the target score and other pertinent measures, such as L2 proficiency test. The *implication inference* relies on evidence that supports valid and well-grounded interpretations of the aptitude construct(s) being assessed. Additionally, it considers the impact of the test on all stakeholders, particularly when the test results are utilised in decision-making processes like recruitment or course admittance. In order to clarify where to search for the evidence, the authors have separated the test results into three levels: single item, subtest or scale scores, and test battery level with compound scores. This systematic approach enables a comprehensive assessment of the test performance and aids in understanding the reliability and validity related to the inferences drawn from the test results.

2.2.3.2 Accessibility of aptitude batteries

As reviewed above in Section 2.2.1, some aptitude batteries (e.g., the CANAL-FT and the Hi-LAB) involving cognitive constructs (e.g., WM) in the theoretical frameworks underpinning the subtests have provided considerable insights about the aptitude construct. However, the lack of accessibility to these aptitude measurements has significant implications for their utilisation in L2 research. Specifically, the MLAT has not been available to individual researchers and the PLAB is a commercially available battery. The CANAL-FT was a government sponsored battery that has not been used except by the authors of the battery, hence, little validity information is available other than what has been reported by Grigorenko et al. (2002). The Hi-LAB was also government sponsored measurement; hence the battery is selectively available and has not been used in research that involves non-government populations and across proficiency levels (Skehan, 2023). Despite the promising reliability and validity evidence provided by the authors of these measurements, these batteries have not been scrutinised independently. A more rigorous evaluation of the batteries necessitates a level of objectivity or scepticism that may frequently be beyond of the reach of developers themselves (Isbell & Kim, 2023, referring to Kane, 2013).

In contrast, the LLAMA tests are openly accessible research instruments that can be used for data collection through the internet. Being among the most widely used aptitude measurements in recent decade, their popularity in aptitude-related research is largely attributed to their open accessibility and continuous development based on feedback from the research community. This underlines the importance of accessibility of aptitude batteries for other researchers. As this battery facilitates internet-based research (IBR), it overcomes restrictions on the use of the battery, which, in turn, may enhance the progress of aptitude research.

This section primarily explores strengths of using IBR for data collection, along with potential concerns that could pose challenges to the reliability and validity of research findings derived from IBR.

Strengths of internet-based measures

The increased accessibility facilitated by IBR has been particularly highlighted during the COVID-19 pandemic. Furthermore, advancements in technology and the rapid development of IBR capabilities have presented new opportunities to address the limitations of traditional lab-based, in-person research. For example, IBR promotes better sampling diversity, as individuals can conveniently participate in research remotely (Casler et al., 2013; Newman et al., 2021), and can yield larger samples in a quicker and more cost-effective manner (Newman, et al., 2021).

Despite advantages that IBR instruments and platforms offer, certain methodological concerns have been raised and summarised below, as discussed by Newman et al. (2021) in their review of IBR platforms for collecting survey and experimental data.

Challenges to the internet-based measures

Newman et al. (2021) highlights three specific concerns—sampling, quality, and ethical—that could threaten the ethical principles and validity of IBR measures.

First, *sample bias* occurs when the participant population is skewed towards specific demographic groups, such as those with higher levels of education or computer literacy (Follmer, et al., 2017). Such biases can lead to non-representative data, reducing the generalisability of the results to the intended population. Another related concern is *self-selection bias,* where participants opt to join studies based on their interests, thereby violating the assumptions of randomness, and introducing potential biases (Stritch et al., 2017). Outcomes of self-selected participation in particular types of studies can be influenced by unobservable factors (Cheung et al., 2017). While random assignment can mitigate this bias in experimental design, its application becomes more complex in longitudinal studies where information about non-participants is absent. *Data non-independence* and *in-group bias* further raise concerns. Participants on the same IBR platform might form close communication networks and collaborate, introducing potential biases into the collected data (Gray, et al., 2016). The issue of *non-naivety* arises from participants' increasing familiarity with research instruments over time. As participants become more experienced, they might align their responses with researchers' expectations (Hauser, et al., 2019), thereby compromising the data quality (Chandler et al., 2015; Devoe & House, 2016).

Second, a range of quality concerns, such as inattentiveness and fraudulent/dishonest behaviour, can undermine the validity of measures. In particular, participants' *inattentiveness* may result in non-compliant or careless responses that threaten the internal validity of the measures and the statistical conclusions drawn from the obtained responses. This could potentially lead to the inflation or attenuation of observed relationships between variables (Aruguete et al., 2019; Buhrmester et al., 2018; Cheung et al., 2017). Furthermore, insufficient response effort on IBR platforms can introduce random measurement errors, further compromising the validity of the generated findings (Huang et al., 2015). Another pertinent quality concern is *fraudulent or dishonest behaviour*, involving actions where participants submit duplicated responses through multiple accounts or misrepresenting their eligibility for participatory compensation (Dennis et al., 2020; Kan & Drummey, 2018; MacInnis et al., 2020).

Lastly, ethical concerns relate to *participatory compensation* and *procedural transparency in participant recruitment*. The issue of financial incentives provided to participants revolves around whether researchers and platforms offer fair and ethical compensation, considering that participants recruited from these platforms often exhibit lower socioeconomic status, reduced well-being levels (Stone et al., 2019), and higher rates of clinical depression (Ophir et al., 2020) compared to the general population. To address these concerns, it is advisable for researchers to seek for approvals from their own Ethics Boards, aiming to establish compensation standards and addressing ethical dilemmas stemming from the potential exploitation of participants, who might be viewed as sources of inexpensive labour for scientific research (Palan & Schitter, 2018; Shank, 2016). Such unethical practices can also compromise data quality (Bohannon, 2016). Another ethical concern pertains to procedural transparency in the treatment of participants recruited from IBR platforms. This concern is related to the emerging criticisms about whether these participants are treated equitably in comparison to participants in traditional lab-based research. It has been suggested, in particular, that IBR participants may not have the same rights to withdraw from studies without adverse consequences or penalties (Gleibs, 2017), and that information about study risks and details may intentionally remain unclear (Pittman & Sheehan, 2016).

All of the aforementioned concerns highlight the necessity for researchers to thoroughly evaluate the limitations and potential biases inherent in the data collected through IBR platforms, which could threaten the reliability and validity of their findings. Newman et al. (2021) provide practical recommendations to address these concerns, as outlined in the table of recommendations for future research using online platforms (p. 1394–1396). Specifically, these recommendations include strategies to ensure appropriate sampling, such as striving for sample representativeness and minimising the inclusion of non-naïve participants. To tackle data quality issues, they suggest implementing attention checks, conducting per-screening of participants, and offering explicit instructions, questions, and warnings. Finally, the authors advocate for proper participant compensation and enhanced transparency in disclosing research details to participants and during the journal submission process.

While these concerns and recommendations primarily pertain to conducting research via IBR platforms in a general sense, they offer valuable guidance for the creation of internet-based aptitude measures to enhance accessibility. In the current research, several factors aligning with these recommendations have been carefully considered throughout the development of a novel IBR aptitude battery, which will be described in Chapter 3.

2.2.3.3 What language(s) should be used in measuring aptitude?

The language(s) used in aptitude batteries constitute another factor that necessitates consideration, although this matter has received relatively limited attention and hasn't been included as a major factor in systematic reviews of aptitude measures. Decisions regarding encoding language (the language of test items) and instructional language (the language in which test instructions are presented) taken by the creators of existing aptitude measures might appear self-evident. Nonetheless, a detailed examination of these measures, provide below, could offer insights into the conceptualisation and operationalisation of the aptitude constructs themselves.

Encoding language

The subtests within the MLAT, such as Phonetic Script and Spelling Cues, feature test items presented in semi-artificial languages, yet they assess participants' abilities based on their understanding of phonological and morphosyntactic rules in English. Similarly, the subtests (Number Learning and Paired Associates) for cognitive abilities like inductive learning ability and associative memory employ stimuli encoded in English. The subtest of Words in Sentences also uses English stimuli. However, the potential influence of knowledge of English on this battery has not been taken into account.

The instrumentation of the PLAB, much like the MLAT, primarily revolves around assessing linguistic knowledge related to English across its subtests. An exception is found in Part 5 Sound Discrimination. This particular subtest prompts participants to distinguish between pitch, orality, and nasality in words presented in a novel language. This could be difficult for L1 English speakers but less so for learners whose L1 has phonemes that vary in similar ways in terms of pitch, orality, and nasality. However, there is a lack of comprehensive research investigating the reliability and validity of this battery among learners with diverse linguistic backgrounds.

The design of the CANAL-FT incorporates linguistic features from an artificial language, Ursulu, in all nine sections. These features are embedded within an English-language context across the sections. Given that this battery has seen limited use in research beyond its initial validation study by its authors (Grigorenko et al., 2000), we do not know about the reliability and validity of the measure when employed with learners from diverse L1 backgrounds.

The Hi-LAB battery focuses more on domain-general cognitive abilities, resulting in the use of L1-neutral stimuli across most of its subtests, such as digits, coloured rectangles, and boxes. While certain subtests involve verbal stimuli like letters, words, and nonwords, the performance heavily relies on English-language knowledge. This design stems from the battery's intended purpose of distinguishing participants with aptitude for high-level attainment. Notably, the validation of the battery was conducted with U.S. government agencies and members of the U.S. military. However, due to the battery's unavailability, the opportunity for further investigation into its validity for learners from diverse L1 backgrounds is hindered.

In the design and iterations of the LLAMA tests, careful consideration has been given to the issue of encoding language issue. LLAMA_B uses unfamiliar vocabulary items, while LLAMA_F employs grammatical and morphological rules that differ from those of English. Consequently, the stimuli in these subtests are distinct from the English language. On the other hand, LLAMA_D and LLAMA_E have employed phonetic representations that closely resemble the phonological rules of Germanic languages. To investigate the potential impact of heterogeneous language backgrounds on test performance, empirical research has been undertaken by the authors. The findings of these studies indicated that there were no significant differences in test performance among participants from various L1s. As a result, the LLAMA tests were deemed to be language neutral in nature. However, it is noteworthy that participants' prior language learning experiences, such as bilingualism, monolingualism, or instructed L2 learning, may influence their test results. Specifically, participants with instructed learning experiences tend to outperform their monolingual and bilingual counterparts (Rogers et al., 2017).

Researchers have made efforts to adapt existing aptitude batteries to cater to the specific language backgrounds of L2 learners. For example, Li and Luo (2019) developed and validated an aptitude test tailored to L1-Chinese learners. Their test design was informed by the MLAT and the PLAB. However, they used Chinese and an artificial language to create stimuli, which encouraged participants to employ their L1 knowledge instead of relying on their L2-English. Specifically, for four subtests (i.e., Number Learning, Spelling Clues, Phonetic Script, and Paired Associates) that measure phonetic coding ability and associative memory, they used stimuli conforming to phonetic rules in participants' L1–Chinese. Furthermore, to measure grammatical sensitivity and language analytic ability, they employed stimuli featuring Chinese grammatical features and an artificial language in two subtests (Words in Sentences and Language Analysis). This strategy aimed to ensure that the stimuli were independent of L2-English proficiency. The reliability of this adapted battery was demonstrated through preliminary results, although some misfitting items were identified, underscoring the need for further refinement.

The potential influence of the language used in WM tasks on test results raises concerns. This is particularly relevant due to the fact that processing in L2 generally consumes more cognitive resources than processing in the L1. Nevertheless, the existing research evidence remains inconsistent in providing a definitive answer to the question of which language should be used to assess WM capacity. Insights emerge from Linck et al.'s (2014) meta-analysis, which delved into the connection between WM and L2 learning. In their study, WM span tasks' encoding language were categorised as either L1 (tasks requiring the processing or storage of numeric stimuli were coded as L1) or L2. The findings demonstrated stronger correlations between WM and L2 learning outcomes when WM measurements were conducted in L2. This outcome led the authors to suggest a potential confound of L2 proficiency with WM abilities when WM tasks are administered in L2. Distinguishing the true essence of what WM tasks measure becomes challenging where L2 proficiency potentially moderates WM abilities.

Researchers in SLA have taken steps to mitigate the potential confounding effect of using L2 stimuli in WM tasks. One approach is adapting reading span tasks into different language versions. For example, Gass et al. (2019) used Arabic and Chinese versions of the reading span tasks in Unsworth et al. (2009). However, further research is required, as the validity evidence for these translated versions of WM tasks relied on small sample sizes.

Instructional language

The choice of instructional language used in aptitude measures has received less attention than the encoding language issue. It is evident that instructions are more easily understood when provided in participants' native language. However, for openly accessible aptitude measures it can be challenging to ensure comprehensibility for participants with diverse linguistic backgrounds.

In the case of the LLAMA tests, efforts have been made to address this challenge. Rogers et al. (2023) detailed their exploration of instructional language in the LLAMA online v.2. In this version, they aimed to surpass the limitations of specific languages by using language neutral symbols along with a comprehensive set of references, moving away from relying solely on English instructions. However, this approach encountered issues, as the symbols displayed inconsistently across different operational systems, leading to confusion for test takers. Consequently, the LLAMA v.3 reverted to using English language instructions, acknowledging the need for clarity and consistency in instructions. However, the LLAMA team has acknowledged the necessity of adapting the battery into other language versions to accommodate the diverse linguistic backgrounds of participants and enhance the test's accessibility and comprehensibility (Rogers, et al., 2023).

2.3 Rationales for the current research

Language aptitude research has made significant advances in recent years, but there are still theoretical and methodological challenges that need to be addressed. One key limitation highlighted by Skehan (2023) is the fragmented nature of aptitude research, which is partly a result of the limited variety of available aptitude and working memory measures. This restricts researchers from obtaining a comprehensive understanding of the relationships between componential constructs of aptitude and aspects of L2 learning. Skehan emphasises that researchers have made fragmented contributions without fully developing complete test batteries. This has led to the dominance of two aptitude batteries in recent research, each with its own strengths and limitations. The Hi-LAB offers a wide range of aptitude subtests, but its accessibility is limited, and validation evidence is confined to a narrow population. On the other hand, the LLAMA tests have amassed substantial amount of empirical findings, but the battery itself lacks adequate validation. To address these limitations and to advance the aptitude research, Skehan proposes a combination of major aptitude batteries to leverage the strengths of different batteries and overcome their limitations. Moreover, Skehan suggests adopting a flexible approach to instrument construction. This involves creating a pool of validated tests that researchers can choose from based on the specific context of their study, a departure from the current 'one size fits all' approach (p. 233).

The current research responds to Skehan's (2023) call for urgent 're-evaluation' (p. 214) of existing aptitude instrumentation, by undertaking a comprehensive re-construction of an aptitude battery guided by aptitude–SLA theoretical frameworks. Instead of relying on unvalidated or inaccessible batteries, this effort aims to develop a new measure aligned with the latest theoretical advancements, effectively capturing the multi-faceted nature of aptitude across language specific domains, involving sound, working memory and processing (Skehan, 2023).

To be specific, the theoretical foundations of the new battery are rooted in the integration of some of the components of the first stage ('Input-oriented') and all of the components of the second stage ('Interlanguage development') of the Stages Approach, as well as certain components underlying major existing aptitude batteries such as the MLAT, the PLAB, and the LLAMA. Four primary components—associative memory, phonetic coding ability, language analytic ability, and WM—are operationalised. Additionally, the second-order components related to WM, i.e., PSTM and executive control, are constructed based on the P/E model. The initial validity check aims to establish this new battery's capacity to accurately measure aptitude and provide insights into its potential applications in aptitude-related research within the field of SLA.

The current research aims to address the gaps in aptitude measures, as outlined below.

2.3.1 Empirical gap 1: verifying theoretical frameworks

One significant research gap in aptitude-related research is the lack of empirical verification of certain aptitude theoretical frameworks. For example, the Stages Approach (Skehan, 2016) has been proposed with aptitude components unfolding in a sequential manner, corresponding to the developmental phases in SLA, which seems to be promising to understand the caveats of L2 learning processes in relation to the aptitude components. However, the model has not been used to guide the development of aptitude batteries, thus has not been rigorously verified by empirical studies. Similarly, although the P/E model (Wen, 2016) proposes componential constructs of working memory in relation to L2 learning, the need for standardised measurements to accurately measure the multi-facets of working memory in L2 research has not been fulfilled.

Wen and Skehan's (2021) theoretical endeavours seek to synthesise the Stages Approach and the P/E model. The primary aim is to address the challenge of fragmentation when conceptualising the multifaced constructs of aptitude and WM concerning the complexity of L2 learning development. The synthesised model introduces certain theoretical assumptions, which require empirical verification. The key prerequisite for verifying these theoretical assumptions is the availability of an aptitude battery developed on these frameworks. However, the current disconnect between the development of theoretical frameworks and aptitude measures constrains the verification of aptitude theoretical frameworks and impedes our understanding of interactions among aptitude dimensions and the roles of aptitude components on L2 learning. Therefore, there is a need to develop a new aptitude battery that incorporates the synthesis of the Stages Approach and the P/E model. Additionally, developing a new aptitude battery using domain specific measurements becomes appealing to address the tension between the concern for domain specificity (as related to the Stages Approach) and domain generality (as related to the P/E model), as pointed out by Wen and Skehan (2021).

2.3.2 Empirical gap 2: investigating effects of modality

The effects of modality in measuring aptitude have not been systematically investigated. Most existing aptitude measures (except for the CANAL-FT) do not consider the input modality as a potential factor and do not systematically control the possible confound of modality in instrumentation. However, exploring the impact of modality on measuring aptitude can provide valuable insights into optimal presentation formats and task design that operationalise aptitude components. Additionally, this exploration can shed light on the validity of aptitude measures in explaining specific aspects of L2 learning outcomes.

To bridge this empirical gap, a new aptitude battery must take modality into consideration. Specifically, for those constructs (e.g., associative memory, language analytic ability, and working memory) that can be operationalised in either aural or written modalities, it is necessary to design parallel versions of test items, allowing the comparison of the results obtained from tests administered in each modality. This can be better achieved through within-subject design, engaging the same participants tested in different modality conditions with different versions of the 'same' test in each modality.

2.3.3 Empirical gap 3: providing initial validation evidence

Although there are several aptitude batteries and some have gained popularity in SLA research, reliability and validity evidence has not been consistently established, leading to concerns about the quality of aptitude-related research. Transparent reporting of reliability and validity evidence is crucial to ensure the methodological rigour of aptitude research and the confidence in the findings.

To address this gap, a thorough evaluation of aptitude measures (or any other psychometric measurements) is required. By employing established validation schema, such as those discussed in Section 2.2.3.1, the current research aims to provide a transparent and replicable process to scrutinise the initial evidence of reliability and validity of a new battery.

2.3.4 Methodological gap: developing L1 sensitive measures

Most existing aptitude measures are written in English and incorporate stimuli that are more engaging for participants if their L1 is English or close to Germanic languages. This design may yield potential confounds if participants are non-native English speakers and have varying proficiency of the encoding language (i.e., the language in which the test items are written).

This gap may need the development of an aptitude battery that (initially, at least) targets a specific learner population rather than the creation of a uniform battery that could be suitable for all L2 learners regardless of their language experience. Therefore, the current research aims to develop a new aptitude battery that is sensitive to the L1 of participants who, in case of the current research, are native speakers of Chinese. This can initiate a long methodological journey and provide a prototype to invite further iterations of more language versions that can be used for different learner populations.

2.4 Research questions for the current research

These research gaps in aptitude measures necessitate the development of a new aptitude battery to facilitate empirical investigations through methodological advancements. In response to these gaps, the current research develops a new aptitude battery, *Tests of Aptitude for Language Learning (TALL)*, based on the synthesis of the Stages Approach and the P/E model, with the aim of measuring foreign language aptitude of L1-Chinese learners.

Three research questions are addressed concerning an initial validation of this battery:

Research question 1:

To what extent does TALL display satisfactory internal consistency and internal validity as a battery for language aptitude?

Research question 2:

To what extent does modality have effects on scores in the subtests that are administered in the aural and the written modality?

Research question 3:

To what extent does TALL predict the foreign language (English) proficiency of the participants?

CHAPTER 3: METHODS

3.1 Introduction

The first section of the chapter—Instrumentation—provides general considerations relating to (a) design decisions for the test suites of *Test of Aptitude for Language Learning* (TALL) and (b) technical issues involved in developing TALL into an IBR instrument. Following this, the details in the design of each subtest are elaborated, as well as the rationales and contributions of pilots to the final instrumentation. The second section—Main study—presents the study design, data collection procedure, and data analysis plan. This section also highlights the efforts made to address the issue of 'power', a problems that can affect the quality of quantitative L2 research (Isbell et al., 2022; Plonsky, 2013; Plonsky & Gass, 2011). These efforts include conducting a priori power analysis to compute required sample size, applying strategies to deal with outliers systematically, and using a multivariate data analysis plan.

3.2 Instrumentation

3.2.1 General considerations in developing TALL into an IBR instrument

3.2.1.1 Componential constructs

In Section 2.3, a theoretical framework for TALL has been proposed based on the Stages Approach (Skehan, 2016) and the Phonological/Executive (P/E) Model (Wen, 2016). Given that the aim of TALL is to measure the components of aptitude specifically involved in the early stages of L2 learning, which primarily focuses on handling novel input through central processing without involving language output, five componential constructs are postulated. These constructs are associative memory, phonetic coding ability, language analytic ability, and working memory (WM), with WM composed of phonological short-term memory and executive control capacity according to Wen (2016).

To measure the above constructs respectively, TALL was designed with five componential subtests: Vocabulary Learning (TALL_VL) for associative memory, Sound Discrimination (TALL_SD) for phonetic coding ability, Language Analysis (TALL_LA) for language analytic ability, Serial Nonwords Recall (TALL_SNWR) for phonological short-term memory, and Complex Span Tasks (TALL_CST) for executive control capacity.

3.2.1.2 Modalities and material versions

In Section 2.2.2.3, the effects of input modalities in L2 learning were reviewed. This leads to Research Question 2 about whether using stimuli in different modalities to measure aptitude would provide different results. Most existing aptitude batteries (except the CANAL-FT) have not taken into consideration the effects of modality in which stimuli are presented, suggesting a methodological gap in literature that needs to be addressed. To answer this RQ, three subtests, i.e., TALL_VL, TALL_LA, and TALL_CST, were designed to be administered in both aural and written modalities. However, aptitude, as a hybrid construct, should be measured by a complete test suite that contains all the subtests. Therefore, TALL was developed into two suites, each holding five subtests as shown in Table 3.1.

Table 4.1 Test suites and subtests of TALL

Test Suite	Subtest	Input modality	Targeted component in aptitude
Aural	TALL_VL	aural	associative memory
	TALL_SD	aural	phonetic coding ability
	TALL_LA	aural	language analytic ability
	TALL_SNWR	aural	phonological short-term memory
	TALL_CST	aural	executive control capacity
Written	TALL_VL	written	associative memory
	TALL_SD	aural	phonetic coding ability
	TALL_LA	written	language analytic ability
	TALL_SNWR	aural	phonological short-term memory
	TALL_CST	written	executive control capacity

Participants were tested in different modalities in a repeated within-subject design. Therefore, two versions of stimuli were developed, that is, each test suite had two material versions with stimuli counterbalanced so that a participant did not see the same item (or trial) in both modalities. For the three subtests that were administered in two modalities, stimuli in two versions of materials were consistent across the modalities.

3.2.1.3 Instructional and encoding languages

In Section 2.2.3.3, methodological issues related to the use of language in aptitude batteries are reviewed. To avoid potential confounds of participants' L2 knowledge, it is likely to be important to use participants' L1 for the instructions in psychometric measurements. In the current study, the target population was participants who had Mandarin, the dominant variety of Chinese in mainland China and Taiwan, as their L1 and English as their later learned foreign language, so TALL was developed using Mandarin for the instructions.

The target language in the subtests of TALL_VL, TALL_SD, and TALL_LA that involve language learning tasks was a miniature language adapted from Lithuanian, a Baltic language that was highly likely to be novel to the participants in the current study. The ideal of using artificially designed language stimuli adapted from a natural language was in line with the design of LLAMA subtests, which provided test stimuli that were neutral (arguably) to the L1s of participants, as evidenced by the insignificant differences in the performance of participants from different L1 backgrounds (Rogers, et al., 2017).

The encoding language in which the nonword stimuli were used in the TALL_SNWR followed the phonological rules of the participants' L1 Mandarin. Care also had to be taken to ensure that nonword could not be used as a cue to elicit associated sematic information. These considerations aimed to address two methodological restrictions (or confounds) summarised in Gathercole (2006). First, language knowledge can influence the accuracy of nonword repetition. Second, nonword stimuli containing syllables of lexical items or segments with high phonotactic frequencies can increase the accuracy of the task. Given that participants' L1 is Mandarin, nonword stimuli should be designed to conform to Mandarin phonology. Also, it was important to avoid real meaning associations. This methodological consideration can be challenging to ensure as individual syllables may correspond to one or more meanings in Mandarin, which may explain that nonword repetition task using stimuli in Mandarin is scarce in previous studies. In other words, the formation of a stimulus being legitimate phonologically yet illegitimate in meaning (i.e., a nonword) in Mandarin Chinese can be difficult relative to creating a nonword in English or other alphabetic language.

The design of nonword stimuli in TALL_SNWR was informed by the way of defining nonwords following phonological rules of Cantonese, a Chinese variety predominantly used in Hong Kong, Macau and Guangdong Province in China, introduced by Chan et al. (2011). Specifically, the principles for nonword formation in Mandarin Chinese were postulated as: (1) nonwords were in the form of consonant–vowel–consonant–vowel (CVCV) disyllabic words, which is the most frequent word formation in Mandarin; (2) all syllables in the nonwords were combination of a consonant and a vowel conforming to the phonotactic rules of Mandarin, being articulatable by a Mandarin native speaker. Meanwhile, the combined syllables should not exist in the Mandarin syllabary, hence they did not correspond to any Chinese characters. This, therefore, prevented the elicitation of associations of semantic meanings by the participants, as suggested by Gathercole (2006); (3) all nonwords and their componential syllables were presented in one consistent tone (i.e., the high-level tone) to minimise the demand for tonal encoding and the possibility of semantic association of any disyllabic words that had similar tones; and (4) two consonants and two vowels in each nonword were different so as to avoid possible prosodic effects, such as rhyme or alliteration. These principles were followed in the formation of all the nonwords in TALL_SNWR.

Verbal stimuli are commonly used in domain-specific complex span tasks to measure executive control capacity in WM (Conway, et al., 2005). The encoding language in which verbal stimuli are presented in the processing task has been proposed to be participants' L1 to eliminate the potential confound of other language knowledge. In other words, the design of verbal stimuli for meaning processing needs to be sensitive to the L1 of the participants. In the current study, participants' L1 Chinese was used to compose the sentence stimuli for meaning processing, and English letters were used as the stimuli for recalling. This task design was adopted from the reading span tasks in Gass et al. (2019), which is a Chinese version adapted from the reading span tasks in Unsworth et al. (2009), developed in the Attention & Working Memory Lab in Georgia Institute of Technology (https://englelab.gatech.edu/).

3.2.1.4 Techniques of the Internet-based research methods

TALL was intended to be conducted in person for data collection before the COVID-19 pandemic. However, logistical challenges during the pandemic imposed the necessity of developing an instrument that can be used on the Internet to collect data remotely. The applications of IBR methods in measuring aptitude have been reviewed in Section 2.2.3.2, in which the strengths, concerns and recommendations have been highlighted. In this section, specific technical considerations are introduced. These considerations served the purpose of minimising the potential issues that may threaten the internal validity of TALL as an IBR instrument.

Archival techniques

Archival techniques were applied to record response times and to identify invariant responding, the measures recommended to ensure data quality in IBR (Newman et al., 2021). Specifically, response times in all testing items were recorded and downloadable as part of the raw datasets. This allowed attention checks to identify inattentive participants and manually exclude the relevant data points in the data preparation protocol (to be introduced in Section 3.3.5.1). In addition, the design of TALL did not allow participants to skip an item without any response.

Explicit instructions, warnings, and time features

Explicit instructions and warnings were used throughout TALL to ensure the efficacy of the study and validity of research findings based on the responses at the level of reduced dishonesty (Buhrmester et al., 2018; Hunt & Scheetz, 2019). For example, explicit instructions forbidding note taking were communicated at the beginning of all subtests except in TALL_LA, in which participants were instructed that they could take notes in the learning phase. The instruction of “Note-taking Are Forbidden” was displayed on the screen throughout the testing phases in two WM subtests, i.e., TALL_SNWR and TALL_CST.

It is highly recommended to increase procedural transparency in IBR, as reviewed in Section 2.2.3.2. This purpose can be achieved not only in the process of recruiting participants and collecting their consents but also in the procedure of the experiment. Specifically, in all the subtests, countdown timers were displayed on the screen to indicate the available time for the learning tasks, as well as the progress of testing.

Techniques to limit non-naivety

As reviewed in Section 2.2.3.2, non-naivety of participants may compromise data quality. Given that TALL is a novel battery that has not been publicised or used by participants other than those in the current study, it was unlikely that participants had become familiar with TALL when they took the test. Despite the unlikeness of participants’ non-naivety, techniques were still applied to avoid participants taking the same test more than once by providing a one-time test code for each invited participant. This ensured that a participant could only use an assigned test code once and was not able to reattempt the access to the test. The test codes were within the control of the researcher, hence eliminating the potential fraudulent or dishonest behaviour of submitting multiple responses to achieve higher scores or to obtain more financial compensation.

Service and backend data protection

TALL was developed by outsourced developers using Java scripts. The test platform was powered by a commercial service Tencent Cloud (http://www.tencentcloud.com), with the server IP:139.186.128.135 based in China. The study complied with the General Data Protection Regulation (GDPR) and received ethical approval. Additional measures were taken to protect data collected through the test platform, as follows.

First, the website server has been connected to the Tencent Cloud T-Sec host security service for privileged clients. The service is based on the massive threat data accumulated by Tencent Security and uses machine learning to provide users with security protection services such as asset management, Trojan file detection, hacker intrusion detection, and vulnerability risk warning.

Second, a dual strategy of carrying out regular back-up on both the server and local backend has been implemented to maintain the database. This measure ensures that data can be recovered against the situations of the backend database being damaged due to network attacks or force majeure. The backup data has been archived in the School of Foreign Languages and Cultures at Chongqing University, the institution that financially sponsored the technical development of TALL.

Third, a database security audit mechanism has been adopted, whereby the server can constantly record the operations performed by users and save them in the relevant log files. These audit logs allow the administrator to monitor the details of accesses and operations to detect security vulnerability of the system through log analysis, and to fix the vulnerability in time. This admin role has been undertaken by one of the developers.

3.2.2 Development of subtests

The following sections describe the considerations in the design of each subtest of TALL. They provide details about the selections of stimuli and the comparisons of the subtests to similar tests in other aptitude measures. All materials used in TALL subtests, including written, audio, and pictorial forms of stimuli, have been uploaded to the project portfolio in the OSF repository (https://osf.io/bhca3/), and all the subtest manuals can be found in Appendix A.

3.2.2.1 TALL_VL: Vocabulary Learning

TALL_VL, informed by Paired Associates in the MLAT (Carroll & Sapon, 1959) and LLAMA_B in the LLAMA tests (Meara, 2005; Meara & Rogers, 2019), was developed to measure associative memory of the participants, demonstrated by the participants matching the vocabulary items in a novel language they learnt to the corresponding meanings displayed by pictures. This subtest was designed in both aural and written modalities, with each modality having two counterbalanced versions of vocabulary items.

Phases of TALL_VL: learning and testing

Figure 3.1 showed the experimental paradigm of this subtest having two sequential phases. In the learning phase in two minutes, participants were exposed to 20 pictures of objects that were arranged in a fixed layout on the screen. Participants listened to a name that matched with an object by clicking a picture of an object with the mouse if they were assigned to take the aural test suite, or they read the name of an object on the screen by hovering over the corresponding picture with the mouse cursor if they took the written suite. Participants were allowed to click on or hover over any picture as many times as they wanted, with a two-minute countdown bar displayed on the screen helping them to manage the learning pace.

In the testing phase, participants were presented with the same 20 pictures of objects on the screen as in the learning phase, but the layout of the pictures was randomly arranged. They were then presented with acoustic forms (in the aural suite) or written forms (in the written suite) of the 20 vocabulary items, one at a time in random order (random for different participants). Participants were tested on the ability to identify the correct object that matched the form they listened to or read by clicking on the corresponding picture (thus, it is a receptive, meaning recognition test, as all pictures were available for every answer). Participants were allowed to take the test at their own pace to complete the testing phase. Performance was scored based on the number of vocabulary items that a participant correctly matched with the corresponding pictures, with a total score of 20. A participant's final score was displayed immediately upon the completion of this subtest.

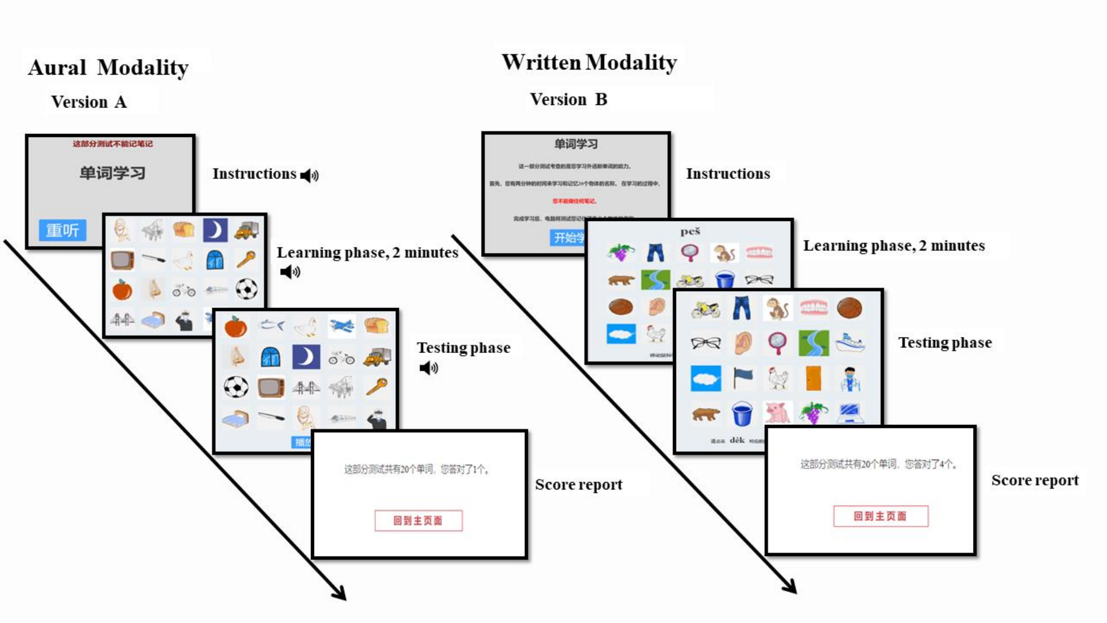


Figure 3.1 Experimental paradigm of TALL_VL

Characteristics

In existing major batteries, associative memory is operationalised in vocabulary learning tasks that measure participants' ability to establish connections between verbal stimuli (e.g., words in L1) and responses (e.g., equivalents in a foreign language) and retain such associations in memory. The relevant subtests in the MLAT, CANAL-FT, LLAMA, and Hi-LAB all employ this paradigm, though they differ in the creation of test items in terms of modality, contextualization, and time of recall. As discussed in Section 2.2.1, the CANAL-FT is the only battery using test items in both aural and written forms, embedding target items in learning context, and assessing both immediate and delayed recall. In contrast, the other batteries solely use written items and assess immediate recall in non-contextualised learning tasks.

In the design of TALL_VL, the issue of modality has been addressed by presenting items in either aural or written forms in different modality suites. This approach allows for the investigation of the effects of modality on measuring associative memory (part of RQ2). Moreover, the decision to only employ non-contextualised leaning task and immediate recall was influenced by the practical considerations of the experimental process and length, as well as the alignment of two sets of test items. The creation of test items drew extensive inspiration from the MLAT and the LLAMA_B. However, the items in TALL_VL are distinct from those tests in several ways.

First, TALL_VL used vocabulary items, adapted from Lithuanian words that were novel to L1 Chinese participants, as target forms to learn, with the meaning of the target items presented in pictures. This design was different from that in the MLAT, in which the target vocabulary items in Turkish were paired with words written in English. The MLAT may be suitable for participants of L1 English learning other languages. However, this learning design can add a potential confounding factor when participants have different levels of knowledge of the English vocabulary if their L1s are not English. The moderating effect of proficiency in the language through which learning took place has been evidenced in recent vocabulary learning research (e,g., Degani & Goldberg, 2019). Therefore, the design of TALL_VL broke away from the language paired association format and allowed participants' learning of novel vocabulary items not depending on their knowledge of an earlier acquired language.

Second, although the design of TALL_VL was similar to the vocabulary learning subtest (LLAMA_B) in the LLAMA tests, it is different from LLAMA_B in terms of the nature of the objects that represent the meanings. LLAMA_B uses non-existing objects (new entities with unfamiliar shapes and unknown functions) to match with the target vocabulary items in an artificial language, while TALL_VL uses objects that have corresponding lexical forms of concrete nouns in participants' L1 Chinese. The design of LLAMA_B aims to remove the confound of various L1s in other aptitude measurements, such as the MLAT, and it is claimed that by presenting non-existing but still describable objects "breaks away from the paired-associate format" and "allows test-takers a lot of flexibility" (Rogers, et al., 2017, p. 50). However, it is not clear the extent to which these non-existing objects could be challenging for the participants to decode, form the associations between the concepts of unfamiliar objects and their auditory or verbal formats, and store the associations for memory retrieval if the participants are from different L1 related cultural backgrounds or different age groups. Considering these potential additional variables such design of LLAMA_B could bring to the test, TALL_VL uses objects that have corresponding lexical forms of concrete nouns in participants' L1 Chinese. This design is to ensure that all the target vocabulary items are perceived by participants in a similar way that is unlikely to involve the visual recognition and processing of novel objects, which could be confounded by visual and perceptual capacities. One rationale for this design decision was that very often (and arguably, *most* often) learning a new language, requires learning new words for broadly known entities, that are broadly comparable to concepts and constructs already learnt, especially in the case of concrete nouns. Of course, sometimes language learning can involve learning new objects, categories, or concepts themselves, such as nuances of time, colour, emotions, or weather, as well as new constructs denoted by nouns and verbs. However, as it is possible that learning such new concepts and constructs draws on different mechanisms, relative to learning new forms for *known* entities, TALL_VL focused on just the learning of new forms for familiar objects.

In addition to these design steps, the target vocabulary items (adapted from Lithuanian words) and the corresponding objects in two versions in TALL_VL were also controlled in terms of the number of syllables, letters, and diacritics, and the frequencies of the lexical equivalents in *Chinese*. Specifically, in each version of vocabulary items, ten monosyllabic and ten disyllabic target words were used. In each version, ten words were in three letters and ten in four letters, and ten words were with diacritics. All the lexical equivalents were chosen from *A Frequency Dictionary of Mandarin Chinese* (Xiao et al., 2009) based on the index of 'usage rate' of a word and the occurrences per million tokens in the corpora. To maintain the consistency of the target words in the two versions, the mean and standard deviation of the usage rates of their Chinese equivalents were calculated and compared, which indicated non-significant differences between the two versions. In this way, the target vocabulary items in TALL_VL were controlled in terms of the likely degree of familiarity of the L1 lexis to the participants.

3.2.2.2 TALL_SD: Sound Discrimination

TALL_SD, informed by Part 5 in the PLAB (Pimsleur, 1966), was developed to measure participants' phonetic coding ability—specifically, their ability to encode and remember unfamiliar sounds in a new language. The test evaluates participants' phonological awareness, encompassing their capacity to encode incoming sounds, identify the sounds within different sound sequences, distinguish them from similar vowels, segment phrases into individual sounds, and associate the sounds with their given corresponding meanings. All of these skills are driven by their phonological awareness, as outlined by Anthony and Francis (2005). Following this design, participants were required to discriminate three similar sounds that were presented as embedded sounds in short phrases or sentences in the new language.

Phases of TALL_SD: learning and testing

Figure 3.2 shows the paradigm of this subtest, with two sequential phases. In the learning phase, participants listened to three isolated sounds while they saw three corresponding pictures of objects on the screen–these same three pictures were constant throughout. They then listened to four sets of phrases, each set having three phrases with one of the three sounds embedded in each phrase. While a phrase was played, the corresponding picture of the sound embedded in this phrase was highlighted to display the match of the sound and its meaning. For example, when participants were presented with an auditory phrase *mūsų sija pilna*, they saw three pictures on the screen, among which the picture of a corn highlighted in a red frame, corresponding to the meaning of the sound *sija*.

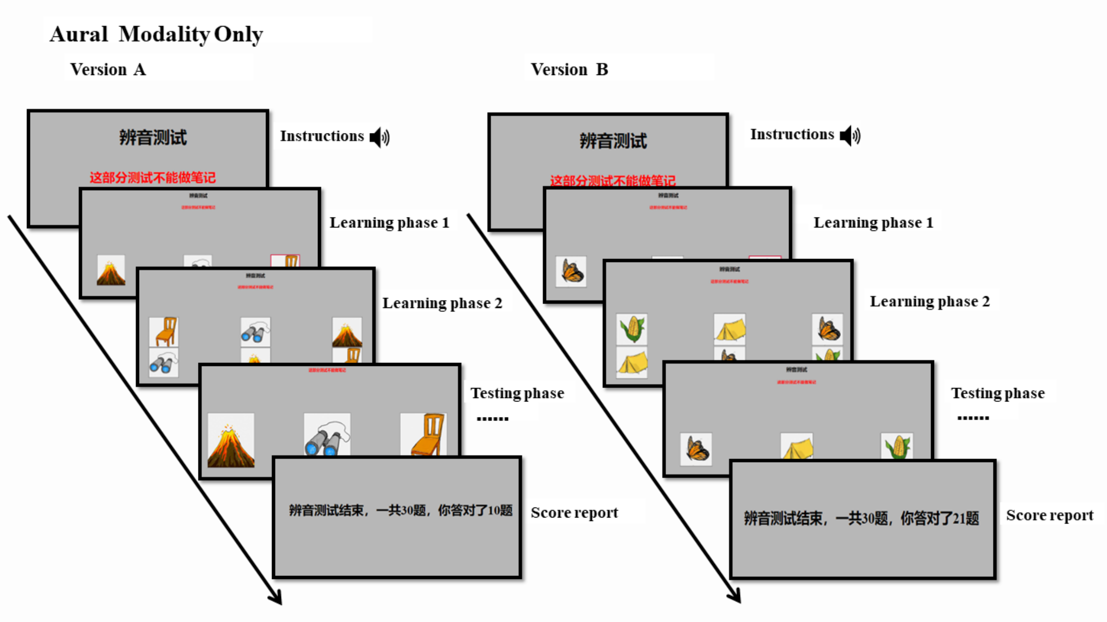


Figure 3.2 Experimental paradigm of TALL_SD

In the testing phase, participants were presented with the same three pictures that corresponded to the sounds on the screen as in the learning phase. They then listened to 30 test stimuli one at a time and were required to discriminate which of the three sounds was embedded in the stimuli by clicking on the corresponding picture associated with the sound. Participants were given a maximum of 15 seconds to make the choice for each stimulus. 10 sets of stimuli were designed, each having 30 stimuli in total in the testing phase, and each sound was presented 10 times. The playing order of all stimuli was random, and the progress through the test was displayed by a bar on the screen, showing the proportion of total testing items participants had completed. Although TALL_SD was administered only in the aural modality, it had two versions of materials that were counter-balanced across test suites. The performance of the participants in this subtest was scored by the number of the correct choices they made, and the total score was 30. A participant's final score was displayed immediately upon the completion of this subtest.

Characteristics

In existing major batteries, phonetic coding ability is operationalised in diverse subtests, characterised by features that inspired the design of TALL_SD. Firstly, subtests in the CANAL-FT and Hi-LAB measure phonetic coding ability without requiring participants to associate novel sounds with written forms. This approach avoids introducing confounds related to differences in participants' abilities to perceive phonological and orthographical forms, as well as their existing knowledge of Romanic writing systems. Secondly, in the PLAB and CANAL-FT, the subtests measuring perceiving and retaining novel sounds require participants to create form-meaning connections, extending beyond merely measuring participants' perceptual acuity of novel sounds. These subtests present test items contextually, embedded within short phrases or passages. This operationalisation engages participants' phonological awareness not only in encoding and identifying the novel sounds but also segmenting the novel sounds from the phonological input and associating them with their corresponding meanings. This paradigm simulates language learning processes involving phonetic coding ability, such as Input processing, Noticing, and Pattern recognition of linguistic input, as outlined in the Stages Approach.

In the design of TALL_SD, the modality has been decided to be exclusively aural. Additionally, practical considerations regarding the alignment of two sets of test items were taken into account. The design was informed by the test format of Part 5 in the PLAB but differs from this test in several ways.

First, being different from the target stimuli in the PLAB that are adopted from a tone language, TALL_SD used two versions of sounds in Lithuanian, and each version included three target sounds that were similar. These sounds were in the consonant-vowel-consonant-vowel ($C_1V_1C_2V_2$) formation, with only the first vowel (V_1) being different across the three sounds. The combinations of the three sounds were consistent across two versions, that is, one diphthong (i.e., two vowels in a single syllable in V_1) and two single vowels were used in each version.

Version A: *v**ie**ta* (火山, volcano) / *v**a**ta* (望远镜, telescope) / *v**y**ta* (椅子, chair)

Version B: *s**au**ja* (蝴蝶, butterfly) / *s**ė**ja* (帐篷, tent) / *s**i**ja* (玉米, corn)

The length of the stimuli in the learning and testing phases and the auditory elements other than the sounds in the stimuli were identical in the same set. For example, in Version B, the stimuli in Set 4 of the learning phase were: ***sauja*** *mūsų /* ***sėja*** *mūsų /* ***sija*** *mūsų*. In addition, the sounds were evenly embedded at the beginning, in the middle, or at the end of the stimuli in the two phases and two versions. In this way, the differences between the two versions of stimuli were controlled as close as possible.

In addition, unlike the design in the PLAB that presents the meaning of the target sounds in English words, TALL_SD used pictures of objects to present the meanings of the sounds. The objects were selected from *A Frequency Dictionary of Mandarin Chinese* (Xiao et al., 2009), with the mean and standard deviation of the usage rates of the corresponding lexical forms being consistent and statistically non-significantly different between the two versions. In this way, the L1 meaning of the sounds in TALL_SD were controlled in terms of the degree of familiarity to the participants.

3.2.2.3 TALL_LA: Language Analysis

TALL_LA subtest, informed by the test format of the Language Analysis in the PLAB and the LLAMA_F, was designed to measure language analytic ability in learning grammatical features in a miniature language. The target features and vocabulary items in this subtest were adapted from Lithuanian. This subtest was administered in both aural and written modalities.

Phases of TALL_LA: learning and testing

Figure 3.3 shows the paradigm of this subtest, in which participants were required to go through two sequential phases. In the learning phase, participants clicked one of the blue buttons arranged in a grid on screen and were presented with a picture displaying the semantic meaning of verbal phrases or sentences. In the aural modality, participants listened to verbal phrases or sentences that described the meaning of the displayed picture, while in the written modality, they read the verbal phrases or sentences on the screen. The design of the learning task was adapted from the cross-situational learning paradigm (see, for example, in Bovolenta & Marsden, 2020; Walker et al., 2020), in which learners are presented with vocabulary and grammar across multiple learning situations without any feedback. Specifically, in TALL_LA, when participants were presented with a picture about ‘a horse is sleeping’, they listened to or read a sentence description *plaktukas miega*; they were presented with another picture of ‘two horses are sleeping’ and its sentence description *plaktukai miega*. Inferences could be made that, for example, *plaktukas* was the singular form of *plaktukai*, and both words referred to the animal, the horse. Participants were given five minutes to examine 20 pictures corresponding to 20 phrases and sentences consisting of vocabulary items: two nouns, three verbs, and two adjectives (in Version Ａ) or adverbs (in Version B). Morphological and syntactic properties were presented in the verbal descriptions, including three morphosyntactic rules (i.e., nominal endings, verbal inflections, and word order) in the target language. In the learning phase, participants were allowed to click the pictures in any order and as many times as they wanted, with a five-minute countdown bar displayed on the screen helping them manage the learning pace.

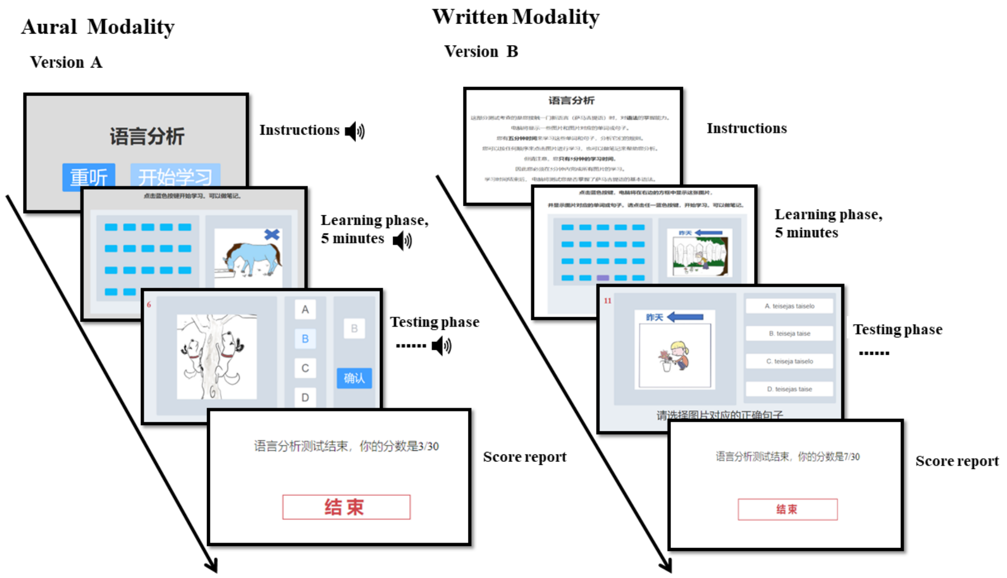


Figure 3.3 Experimental paradigm of TALL_LA

In the testing phase, participants were presented with pictures one at a time and were required to choose the correct verbal description of the picture from four given options given. In the aural modality, participants were required to click (as many times as they wanted) each button of the four options to listen to the optional descriptions and then make the choices, while in the written modality, the four options were displayed on the screen for them to choose by clicking the correct optional sentence describing the picture (as shown in Figure 3.3).

TALL_LA had no time limit for participants to complete all 30 testing items that were presented in random order, and the testing progress was displayed by a bar on the screen showing the proportion of total testing items a participant had completed. This subtest had two versions of materials that were counter-balanced across test suites. The performance of the participants in this subtest was scored by the number of the correct choices they made, and the total score was 30. A participant's final score was displayed immediately upon the completion of this subtest.

Characteristics

TALL_LA was informed by the test format of the Language Analysis in the PLAB and the LLAMA_F, both using miniature language forms that are different from the prior acquired languages of the participants. The potential confound of grammar knowledge among participants could be addressed by using target structures in a miniature language, which helped to achieve a baseline where the target forms are unfamiliar to all participants. However, TALL_LA had substantially different test design from the existing measures in several ways.

First, TALL_LA can be administered in both aural and written modalities, while both the Language Analysis in the PLAB and the LLAMA_F present language stimuli in the written modality. Therefore, TALL_LA allows investigation on the effect of modality in measuring language analytic ability when the target features are perceived in different modalities.

Second, the design of the stimuli in TALL_LA, as presented in Table 3.2, employed three morphosyntactic rules—nominal endings, verbal inflections, and word orders, each rule having pair of features. Each rule pair differed between two versions of stimuli. These features either did not exist in the participants' L1 Chinese (nominal endings and verbal inflections) or existed in the L1 but in a different form (word orders). The selection of these specific rules in TALL_LA was based on the consideration of interpretability and inferability to achieve form–meaning mapping. These features were either interpretable, contributing essentially to the meaning (e.g., verbal inflections related to tenses, and affirmative and negative forms), or inferable from the pictorial stimuli (e.g., agreement rules involving nominal endings and word orders involving adjectives and adverbs).

It should be noted that although TALL_LA drew inspiration from the subtests in the PLAB and the LLAMA, the features were distinct from those present in these batteries in the following aspects:

(1) The morphosyntactic rules used in TALL_LA were different from those in the PLAB and the LLAMA_F. Specifically, the Language Analysis in the PLAB focuses on word orders, an object marker, and verbal inflections. On the other hand, the features in the LLAMA_F pertain to word order (e.g., prepositions in sentence initial positions, adjectives preceding nouns, singular markers as sentence final elements, and numerals preceding nouns) and semantical agreement (e.g., colour adjectives agreeing with shapes, numerals agreeing with prepositions). A practical consideration factored into the selection of rules in TALL_LA, as two parallel sets of each target feature were necessary to generate stimuli for the counterbalanced design.

(2) Whereas LLAMA_F used pictures of non-existed objects (i.e., coloured shapes with an eye and legs), TALL_LA used pictures to simulate the scenarios that demonstrated two classes of nouns (i.e., animals in Version A and people in Version B) and their behaviours. The creation of the pictorial stimuli was challenging with a few considerations. For example, the appearances of people that were ‘typically’ feminine (with longer hair, ponytails, or a bun) and masculine (with shorter hair or a moustache) that could inform the biological genders related to the use of the nominal endings. The use of left or right arrows with Chinese characters of *昨天* (yesterday) and *明天* (tomorrow) conceptualised the past and future tenses. The background of the garden fence and the household furniture differentiated the uses of the adverbs (inside and outside) in the stimuli.

(3) In the testing phase, TALL_LA included a generalised vocabulary item (either a noun, a verb, or an adjective/adverb) that tested generalisation of a rule as it was not introduced in the learning phase. This is different to the previous measures, in which all the vocabulary items are either provided in the instructions (in PLAB) or presented in the learning phase (in LLAMA_F).

(4) TALL_LA provided explicit test instructions that allowed notetaking in the learning phase, which was the same as the subtest in the PLAB, while it was different from the notetaking instructions in the LLAMA_F. Whether notetaking is allowed during any learning (or testing) phase can be relevant to the validity of the test. For the PLAB, notetaking is not

Table 5.2 Target features in TALL_LA

Rule	Version	Paired features	Example of target feature	English meaning
Nominal ending	A	singular	*grotuvas bega*	a dog is running
		plural	*grotuvai bega*	dogs are running
	B	feminine	*vireja taise*	a grandma planted
		masculine	*virejas taise*	a grandpa planted
Verbal inflection	A	affirmation	*grotuvas verda*	a dog is eating
		negation	*grotuvas verdane*	a dog is not eating
	B	past tense	*kirpeja valge*	a woman watered
		future tense	*kirpeja valgelo*	a woman will water
Word order	A	adjective position in affirmation	*grotuvas melyn miega*	a blue dog is sleeping
		adjective position in negation	*plaktukas gelton begane*	a yellow horse is not running
	B	adverb position in past tense	*kirpeja viduje taise*	a woman planted inside
		adverb position in future tense	*kirpejas lauke valgelo*	a man will water outsdie

a potential threat to the validity of the test, given that there is no prior learning phase in the test (cues for learning are provided as a list of vocabulary items in the test directions). However, notetaking instruction needs more considerations in TALL_LA. It may threaten the validity of the tests because both are internet-based in which participants are not invigilated. Evidence on the effect of notetaking on test results is reported in the LLAMA validation study by Rogers et al.(2017). Although they found that the performance of note-taking participants was not significantly different from that of their counterparts who were not allowed to take notes, participants were observed writing out the complete sentences and drawing the corresponding pictures in the learning phase rather than trying to figure out the grammatical rules before the testing phase. Given that note-taking was therefore likely to happen, it was decided that it would better to provide clear instructions that explicitly allow notetaking in TALL_LA. However, the inclusion of the generalisation vocabulary test items (that were not in the learning phase) could mitigate the potential effects of note-taking behaviour.

Third, TALL_LA had a different testing format and scoring strategy from precious measures. Participants were required to choose the correct answer from four alternative forced choices (4AFC), similar to the format in the Language Analysis in the PLAB. Among the four alternative choices, target morphosyntactic features were presented correctly in one choice and incorrectly in other three alternatives. The lexical item in these four choices was the same, which ease the difficulty as participants were not tested on whether they remembered the lexical meanings of the words. The final score in TALL_LA was obtained by the number of correct choices the participants made, and the correct choices represent the target features and lexical items. As such, TALL_LA aimed to gauge participants' ability to inferentially analyse morphosyntactic features and correctly recognise the features in the new sentences.

This testing format was different from that in the LLAMA_F, in which participants need to construct a target sentence by choosing lexical items from the list of 16 items that they have been exposed to in the learning phase to create a sentence. In this respect, LLAMA_F involves production of sentences. However, LLAMA_F does not score participants' answers for complete accuracy, that is, each of the ten sentences is scored for including two items corresponding to two syntactic features, and any other items appearing in the answer are ignored regardless of being correct or not (Rogers et al., 2017; Rogers et al., 2023). For example, a participant may be scored with full credit if he or she correctly chooses two items of the two features (e.g., verbs/preposition being sentence initial, and sentences with a singular subject ending in a singular marker). Any response that correctly includes a verb/preposition and a singular noun in a sentence ending with a singular marker will be awarded two points, despite that the response includes an incorrect singular noun and/or an incorrect singular maker. This scoring strategy aims to ease the difficulty of scoring production by ignoring test-takers' learning of vocabulary items. However, it is not clear whether each test item contains *only* two grammatical features, as the full list of test items are not introduced in the literature. This point is important because partial scoring of the target features but ignoring other features included in the same response may threaten the validity of the test, as learning of the exposed features is not precisely reflected in the test scores.

In addition to the above considerations in test design, TALL_LA also controlled the number of target features appearing in each test item. Specifically, each stimulus in a material version included two grammatical features. Therefore, the 30 test stimuli had a consistent number of occurrences of target features across two versions. The stimuli in each version include 15 pairs of features of nominal endings, 15 pairs of verbal inflections, and 14 pairs of word order. The number of syllables in the same type of items was consistent in the learning and testing phases, and the number of lexical items was also consistent between the two versions of the materials.

3.2.2.4 TALL_SNWR: Serial Nonwords Recall

TALL_SNWR was developed to measure the participants' phonological short-term memory (PSTM), a component of working memory, by requiring participants to repeat a series of nonwords in the order they were presented. As discussed in Section 2.2.2.3, employing auditory stimuli seems appropriate for measuring the PSTM component because the phonological loop, proposed as a distinct subsystem of WM, processes auditory verbal information (Baddeley, 1986, 1992). Nonword repetition tasks, extensively used to measure PSTM (Gathercole, 1995, 2006; Gathercole & Baddeley, 1989), inherently involve auditory stimuli. However, in the Hi-LAB, the sole existing aptitude battery featuring specific subtests for WM, the nonword span tasks present stimuli visually on screen to assess PSTM, lacking a specified justification of this methodological deviation from the norm. The design of TALL_SNWR aligns with the nonword repetition tasks paradigm by Gathercole and colleagues, using auditory stimuli only. Despite using nonword stimuli exclusively in the aural modality, two versions of materials were developed and counterbalanced across both the aural and written test suites of TALL.

Phases of TALL_SNWR: practice and testing

Figure 3.4 shows the paradigm of this subtest. The participants were first provided with three trials to practise in the learning phase. These trials had two, three, and five nonwords, respectively. The purpose of the practice phase was to help participants become familiar with the experimental format. The participants listened to a trial containing a series of nonwords presented sequentially with constant speed (1000 ms) and intervals (1500 ms), then they were required to repeat the nonwords, in the same order in which they were heard, by clicking the corresponding '开始录音假词 1' (*start recording nonword No.1*) buttons on the screen. After they completed the recording of the nonwords in this trial, they were required to submit their recalling of the nonwords by clicking the '提交' (*submit*) button on the screen. After three practice trials, the testing phase began and followed the same procedure of the practice phase: participants first listened to a trial with a series of nonwords, then repeated the nonwords one by one in the presenting order by clicking the corresponding recording buttons, submitted their recalling to complete the test of the trial and continued to next trial till the end of this subtest.

17 trials with 74 nonwords in total were randomly presented in the test, each trial having between two to seven nonwords. Participants were allowed to record and submit their recalling of each trial in 30 seconds. If they did not submit the recalling of a trial in this time limit, the test program would automatically move on to the next trial. The progress of the testing phase was displayed by a bar on the screen that shows the proportion of the total testing trials a participant had completed.

TALL_SNWR was the only subtest in TALL that collected participants' production data, that is, their articulations of nonwords, hence the data of participants' performance were stored in TALL's backend and downloadable for the manual scoring after the completion of the experiment. The final score of TALL_SNWR was the number of nonwords that were assessed manually as being correctly articulated in the correct order. The total score was 74.

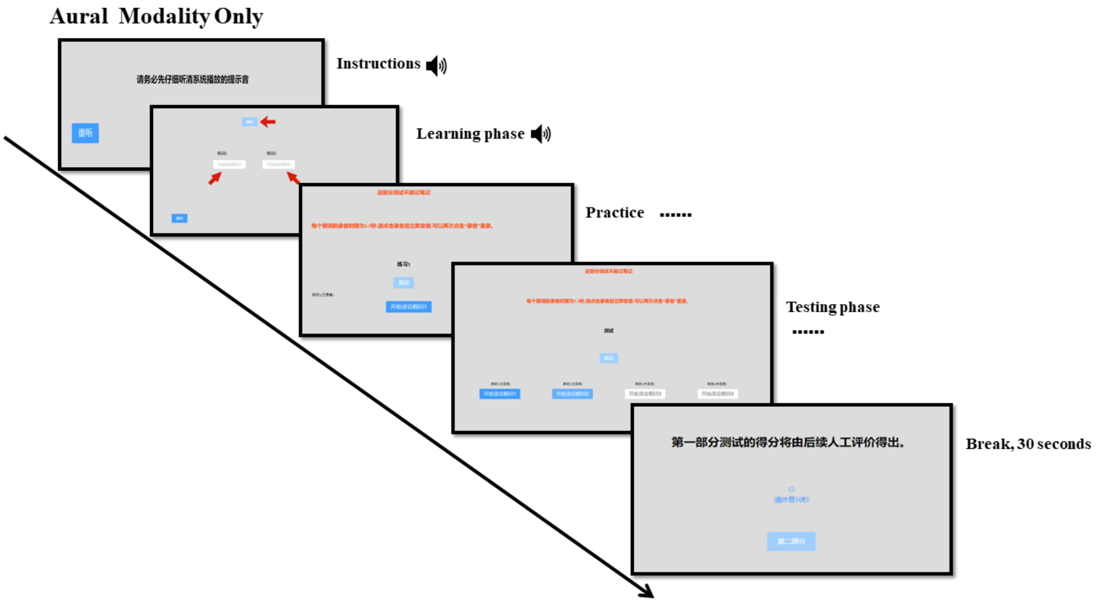


Figure 3.4 Experimental paradigm of TALL_SNWR

Characteristics

TALL_SNWR is the only subtest in TALL that elicit production data. Participants' ability to perceive and discriminate novel sounds could differ from their ability to articulate novel sounds, especially when the sounds are domain-specific and composed following phonological rules in different languages. Therefore, the creation of the nonwords followed participants' L1 phonological rules in Mandarin to ensure that the nonwords are articulatable. Additionally, as introduced in Section 3.2.1.3, to address the methodological limitations in the nonword repetition task described by Gathercole (2006), syllables of the nonwords do not exist in the Mandarin syllabary to avoid real meaning associations.

Given that the nonwords in TALL_SNWR adhere to the phonological rules in Mandarin, these nonwords were consistently created with two syllables, the most common word formation in Mandarin. Thus, the trials in TALL_SNWR consist of a serial of two-syllabic nonwords, ranging from 2 to 7. This design differs from the nonwords developed by Gathercole and colleagues, as they created English nonwords with varied number of syllables, ranging from one to five. The reason for not creating multi-syllable nonwords, as in the tasks designed by Gathercole, was that multi-syllable words were less observed in Mandarin than two-syllable words, and words with four or five syllables might be associated with prosodic effects, a potential confound that was undesirable.

Consequently, the formation of nonwords in TALL_SNWR was extensively informed by the creation of nonword stimuli for Cantonese speakers by Chan et al. (2011), as introduced in Section 3.2.1.3, and the nonword repetition task used in Suzuki (2021b) for Japanese speakers, in which three-mora nonwords—the most common word formation in Japanese—were created by combining three Japanese morae randomly. Taking into account the above-mentioned considerations, seven consonants (*b, c, d, f, m, p, r*) and seven vowels (*a, ai, e, ei, ia, ou, ua*) were chosen from the Mandarin syllabary to form 11 non-existing syllables (*be, bou, cei, dua, fai, fe, mia, pe, pia, ra, rei*).

As shown in Table 3.3, these 11 non-existing syllables were used to form 14 nonwords. Given that two versions of stimuli were needed for the counterbalancing design, the componential syllables in each nonword were combined in a different order to form a paired version of nonwords. For example, if *be-dua* was used in Version A, then its reversed form *dua-be* was used in Version B. In this way, each of the nonword stimuli could only be used in one version. The 14 disyllabic nonwords appeared 6 times throughout the practice and testing phases, and this generated 84 stimuli in each version. All nonwords were presented with a constant speed of 3 seconds per item. In the practice phase, the trials and the nonwords in each trial were presented in a fixed order, that is, the 3-nonword trial was the first trial presented, followed by the 4-nonword trial, and the 5-nonword trial was played last. In the testing phase, trials were presented in a random order, but nonwords in each trial were presented in the constant order. This design followed that of Gathercole et al. (1994).

Table 6.3 Two versions of nonwords in TALL_SNWR

Version	Nonwords (combined in different orders in each trial)
A	*cei-ra, pia-fe, dua-cei, rei-pia, fai-bou, fe-rei, pia-fai, cei-mia, be-dua, mia-fe, fe-ra*
B	*ra-cei, fe-pia, cei-dua, pia-rei, bou-fai, rei-fe, fai-pia, mia-cei, dua-be, fe-mia, ra-fe*

3.2.2.5 TALL_CST: Complex Span Tasks

Complex span tasks (CST) are commonly used to examine active and controlled mechanisms of primary memory more than the passive characteristics of short-term memory, as in the nonword recall test (Draheim et al., 2018). Essentially, complex span tasks comprise two intertwined components of storage and processing. The processing component is inserted between the stimuli to be retained by participants as a distractor to prevent rehearsal of the stimuli to be retained. The primary objective of the participant is to retain the stimuli presented despite the interference caused by the distractor, which gradually increases from one trial to another. CST using verbal stimuli in both processing and recalling tasks are used as domain-specific tasks to measure executive control capacity in working memory (Conway, et al., 2005). To investigate the modality effect in measuring aptitude, TALL_CST was adapted from the Reading Span Tasks in Gass et al. (2019), which is a Chinese version of the reading span tasks in Unsworth et al. (2009). The stimuli for sentence meaning processing were developed in both aural and written modalities.

Three methodological considerations were reflected in the design of the verbal stimuli in TALL_CST. First, the length of the sentence stimuli was controlled constantly. Second, the location of the lexical cues was fixed to be at the end of the sentence, on which the decision of implausibility of the sentence meaning can be made, following the methodological consideration in Gass & Lee (2011). Third, the lexical cues were controlled in terms of the mean and standard deviation of usage rates from *A Frequency Dictionary of Mandarin Chinese* (Xiao et al., 2009).

Phases of TALL_CST: practice and testing

Figure 3.5 and 3.6 shows the paradigms of TALL_CST that had dual tasks design to engage participants' trying to understand the meaning of the sentence stimuli and recalling the letters in the correct displaying order. A practice phase was designed to help participants become familiar with the testing procedure. Three steps were designed in the practice phase. First, the participants practised four recall trials, in which they listened to or read a string of letters (2 to 4 letters) in English, then clicked the corresponding letters on the screen based on the order in which the letters were displayed. Second, the participants practised the sentence processing task of listening to a sentence (in the aural modality) or reading a sentence on the screen (in the written modality) in Chinese, then making a semantic judgement about whether the sentence was sensible in terms of meaning by clicking the button '正确' (*correct*) or '错误' (*incorrect*). There were 15 sentences for practice in this step. Third, participants were provided with the opportunity to practise both the sentence processing task and letter recalling task in two further practice trials. That is, they made a judgement of the plausibility of the meaning of a sentence first, then they were presented with a letter, followed by another sentence judgement, then a letter to recall, until at the end of the trial when they were required to recall all the letters in the correct presenting order. The two practice trails were displayed in a fixed order, with the first trial having 3 sentences and 3 letters, followed by the trial having 4 sentences and 4 letters. All trials and the sentences and letters included in the trials were displayed in the fixed order in the practice phase.

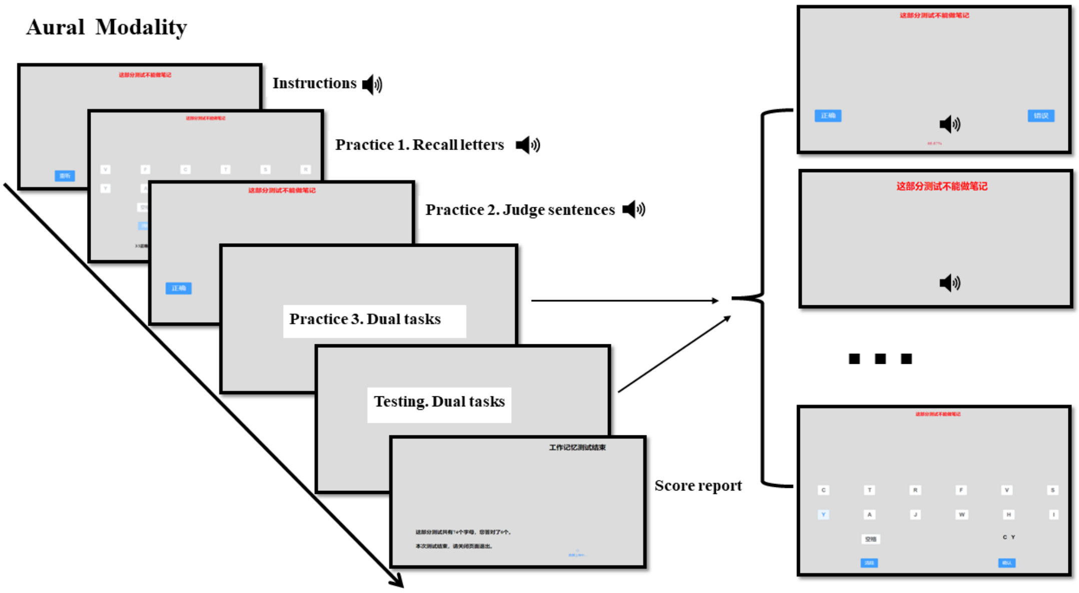


Figure 3.5 Experimental paradigm of TALL_CST (aural modality)

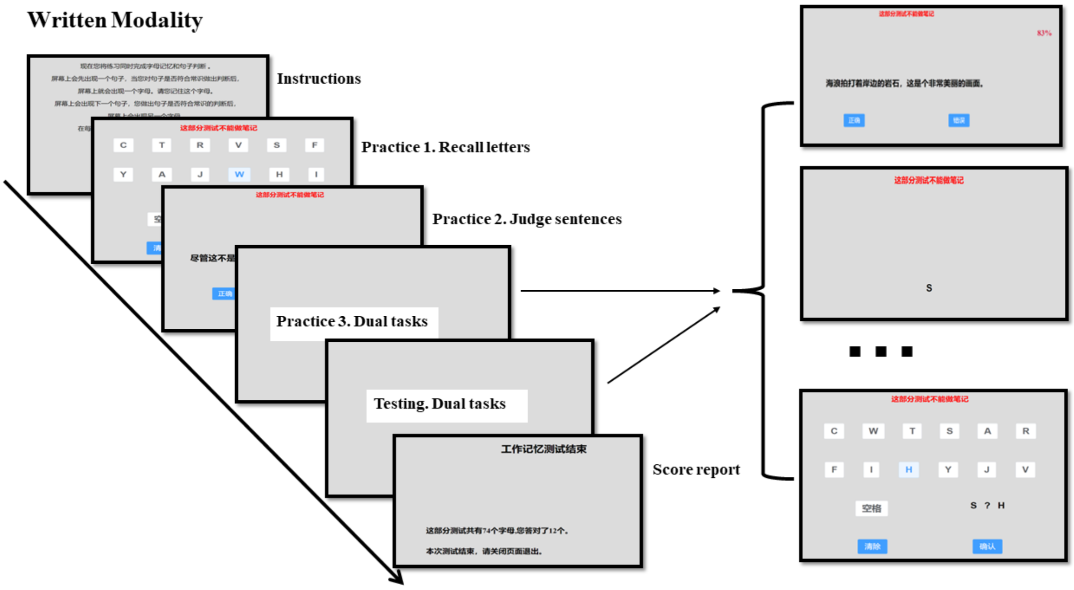


Figure 3.6 Experimental paradigm of TALL_CST (written modality)

The design of the tasks in the practice phase followed the design of the pre-test tasks of reading span tasks in Unsworth et al.(2009). These tasks also served as attentional checks for the data quality. Specifically, recalling a string of 2 to 3 English letters was unlikely to be challenging for the participants in the current study considering their level of literacy in L2 English as college students. Failure to recall the letters correctly would indicate inattentiveness of the participants, leading to the termination of the test. The sentence meaning judgement task required participants to process 15 sentences in order to make a semantic judgement as quickly as they can. The reaction time in the practice trials was recorded and used to compute a subject-adaptive response time limit for each individual participant to make a sentence meaning judgement in the testing phase. This procedure follows the processing time limit design (i.e., time limit is computed by the mean of the reaction time for responding to 15 sentences plus twice the standard deviation) in the reading span tasks in Unsworth et al. (2009).

There were 15 trials in the testing phase, each containing 3 to 7 sentences. The total number of letters to be recalled, 74, was the same as the total number of sentences. Participants followed the same procedure of the third step in the learning phase, that is, the combination of processing and recall tasks. They were randomly presented with a trial containing sentence stimuli for meaning processing and letter stimuli for recall. They were required to recall the letters in the correct order at the end of the trial. The test programme would proceed to the next sentence stimulus if a participant did not make a semantic judgment within the time limit based on the individual performance in the sentence processing tasks in the practice phase, and the stimulus without a response to the judgment would be recorded as incorrect. Participants were also required to complete the recall of the letter string at the end of each trial in 30 seconds. English letters were randomly assigned in all trials and presented (aurally in the aural modality and on the screen in the written modality) with constant speed (800 ms) and interval (200 ms). The sentence stimuli within each trial were presented in a constant order and the sequence of the trials was randomised. 50% of the sentence stimuli were semantically plausible, while the other half of the sentences did not make sense (they were grammatically correct, but the meaning was very strange). The participants' judgements of the semantic plausibility of the sentence stimuli and their recall of the string of letters were stored for data analysis.

For self-monitoring purposes, participants were also provided with the percentage of accuracy in sentence processing throughout the testing phase, and a bar on the screen showed the proportion of the total testing trials a participant had completed. The performance of the participants in this subtest was scored by the total number of correct letters that were recalled in the correct order, and the total score was 74. A participant's final score was displayed immediately upon the completion of this subtest.

Characteristics

The reading span tasks in Unsworth et al.(2009) has been translated into different language versions, and the Chinese version was used in Gass et al. (2019). TALL_CST used sentence stimuli that were different from the Chinese version of the stimuli in Gass et al. (2019). First, the composition of the sentences was rigidly controlled, which produced compound sentences, each containing 20 characters and a comma. Second, the cue word on which the implausibility of a sentence meaning can be judged was located at the end of the sentence. This design ensured that semantic judgement would not be apparent to the participants until the final word, following the methodological consideration in Gass & Lee (2011). This is an example:

(a)天气 暖和 起来，草地 上 到处 是 野餐的 学生 和 真理。

weather warm up, grassland on everywhere be picnicking student and truth

Literal translation: ‘The weather has been getting warm and the grass was full of students and truth for picnics.’

This sentence stimulus, as shown in (a), was for participants to process meaning and make a semantic judgement. The final word “真理” (*truth*) was the cue word to ascertain that this sentence did not make sense, as shown in the literal translation. Third, the final word, that is, the cue word, in all sentences was a two-character word, the most commonly used word form in Mandarin Chinese. Given that two versions of stimuli were needed for the counterbalanced design, all the final words were chosen from *A Frequency Dictionary of Mandarin Chinese* (Xiao et al., 2009) with consistent means and standard deviations of the usage rates across the two versions.

3.2.3 Stages of pilots

Two stages of pilots were conducted before the data collection of the main study, with a larger sample size than one would perhaps normally have in a pilot study. As a completely new internet-based test platform that was developed by outsourced technicians, TALL needed to be validated with a large number of test takers to testify to the stability of the system for data collection when multiple test takers could access the test simultaneously. This also helped to capture the nuances in the process of collecting data on an internet-based platform, which was different from conducting lab-based experiments in person.

3.2.3.1 Preliminary pilot

Stage 1 involved a preliminary pilot study conducted in two months, with the objective of a) establishing the logistic procedures of recruiting participants from different universities, distributing test codes and instructions to the participants online, and providing asynchronous technical support via a commonly used social media platform QQ (im.qq.com) powered by Tencent; b) identifying the time spent and the degree of fatigue of participants in completing one suite of five subtests; and c) assessing the stabilities of the test platform and the back-end data storage and data retrieval, especially in the circumstance when multiple users took the test online in an overlapping time window. Given that the preliminary pilot was conducted before the finalising of the two material versions for the counterbalanced design in the main study, only one version of test items was used in the preliminary pilot, that is, participants took the *same* version of test items in both aural and written modalities. Hence, the data elicited in the preliminary pilot study were not used for analyses, as taking the same test twice would be a confound in the results.

In the preliminary pilot, participants (*N* = 205) who were undergraduates in the first or the second year from two universities completed two rounds of tests, one in the aural modality and one in the written modalities. The back-end data provided the information about the time it took, in general, to complete one test suite. Among these participants, 22 people agreed to conduct an online think-aloud task while taking TALL in the first round. They were required to verbally report whatever came into their mind in the process of the test. Even though they were not required to think-aloud while they were engaging in the WM related subtests, their behaviours in those subtests, such as repeating the stimuli aloud to consolidate the memory, were also observed. These participants were also asked to describe their degree of fatigue in the testing process, as well as their feeling about the clarity and comprehensibility of the experimental instructions.

This pre-pilot study provided initial information on whether test items would be suitable (at all) for learners at their proficiency level (that is, college students who had passed the English proficiency test for the higher education admission, explained in detail in Section 3.3.3). It also informed the practicality of the test procedures, verified the interpretability of the experimental design, and ensured that completion of a TALL suite in an average of 45 minutes would not cause serious fatigue for participants. More importantly, the pre-pilot study with a sample size close to the main study provided nuanced information in terms of whether the test could be conducted with elicited data captured and stored under varied conditions/qualities conditions of internet connection. This information led to further technical refinements of using the resumable transfer in the backup system. This improved the capability of the test system to resume a data transfer from the point it was interrupted or stopped due to the unstable internet connection, hence it ensured the stabilities of the test system and data transfer.

3.2.3.2 The main pilot

Stage 2 involved a pilot study after the completion of the development of two versions of materials. The main objectives of the pilot study were to assess a) the accuracy of the random distributions of material versions in two sequential sessions; b) the stability of the test platform and the back-end data storage and data retrieval after the technical refinements informed by the preliminary pilot; c) the clarity and intelligibility of test stimuli in TALL, especially the pictorial stimuli used in TALL_LA which aimed to demonstrate linguistic features, such as the biological genders of human beings, affirmation and negation, and the past and future tenses; and d) the functionality of generating a complete TALL score report for participants individually.

Twenty-two volunteers showed interest in participating the pilot test, but only ten of them took the test at least once. Seven participants completed two rounds (aural and written suites) of tests within one week. Participants were provided with a report of their scores after the completions of the pilot study. Five of these participants agreed to give an online post-test debriefing. This 5-minute debrief collected the participants' retrospective comments about the degree of difficulties and fatigue they felt during the test process, as well as their comments on the clarity and comprehensibility of the experimental instructions and the acoustic and pictorial stimuli used in TALL. Considering that these volunteers were first year undergraduates majoring in foreign languages, in addition to providing them with a report of their scores, I organised an online workshop introducing TALL and its design to the volunteers who had enrolled in the pilot study. A Questions & Answers session was also included in this workshop to nurture participants' interests in language learning research, cognitive individual differences, and psychometric test design.

The pilot study provided reassurance that the completion time for the entire TALL suite would be around 45 minutes, and it would not cause serous fatigue to participants at the college level. It also informed us about the stability of the test platform and the back-end data storage after the technical refinements. A comparatively low attendance rate of participants who had shown their interest in taking the test also provided helpful information. This led to a strategy applied in the main study: online recruitment continued until the number of participants reached a threshold on the last planned day of the first test session. This threshold was set at least 25% higher than the minimum expected sample size (as detailed in Section 3.3.1.1 below). If the number of the attending participants did not reach the threshold, the data collection time would be extended.

3.3 The main study

3.3.1 Samples and sampling

3.3.1.1 Participants recruitment

The recruitment of participants in the current study was advertised with an online registration form on a Tencent-powered QQ social platform in the professional network of the researcher in Chongqing, China. The purpose of using a register form was to pre-screen participants if they were a particular cohort of population, for example, that they could provide L2 English proficiency scores from a specific set of the National Matriculation English Test (NMET), and if they were not from foreign language related undergraduate programmes. The prescreening considerations were to keep the proficiency scores, that is, the scores of NMET, comparable in the current study as college students took different sets of NMET for the college admission. The scores of different sets of NMET are eligible for higher education admission in China, and the different sets of NMET reflect regional differences in the foreign language proficiency tests in the country. To increase the representativeness of the target population, participants in this study were recruited from 11 colleges or universities at different academic tiers, ranging from community colleges to national key universities. To sum up, the participants who were invited to participate in this study were (a) Year One students from disciplines other than the major of Foreign Languages and Literature, and (b) test takers who took National Set 1 of NMET in the last six months before the recruitment date.

892 eligible participants were filtered from the 990 initial registrations, and they were randomly assigned into two conditions (i.e., taking aural suite in the first round of test then written suite in the second round, and vice versa). Each of them received a test invitation email including an information package of the introduction of the study, the instructions on how to set up the personal computer and how to access the test website, and an identical test code. Participants were also invited to join a social media chat group the researcher set up on the QQ platform, on which they could ask for help if they met difficulties in accessing the test and receive asynchronous technical support. The researcher was aware of the potential issues associated with data non-independence and in-group bias (as reviewed in Section 2.2.3.2) if all the participants were engaged in the same network. To mitigate these issues, the chat group was set up by disabling functions of friend requests and private chat, and all the members were reminded that test content related discussion was not allowed. Considering that it would be difficult to anticipant how many participants would take the test because the test was not scheduled individually, the recruitment of participants continued till the final day of the first test session according to the research

timeline. The halt of the recruitment was called with more confidence when the sample size had reached the minimum threshold. The next section will introduce the details about the prior power analysis for the sample size.

3.3.1.2 Prior power analysis

The power of a statistical test is defined as the likelihood that a test would provide statistically significant findings (Cohen, 1988). Given that quantitative research intently seeks for statistical significance to achieve the goal of generalising findings from samples in studies to wider population under investigation, a priori probability of statistical significance would be expected to be regularly calculated and well understood, and hence researchers are advised to determine the desired sample size so that they can ensure an adequate degree of statistical power of the findings (Nicklin & Vitta, 2021). However, as reminded by experts, researchers in SLA rarely conduct a power analysis. For example, only 6 studies out of 606 conducted a power analysis as reported in Plonsky's (2013) synthesis study on the quality of quantitative research in the field of second language research. The current study aimed to avoid this questionable research practice.

In the current study, the investigation of the effects of modalities in measuring aptitude (Research Question 2) relied on the comparison of the performance of the participants in taking the TALL subtests that were administrated in both modalities. Therefore, it was necessary to perform the prior power analysis to inform the sample size with which the study could have the ability to detect an effect of a particular magnitude. To achieve this purpose, statistical power analysis was conducted to predetermine the expected sample size using the software G* Power 3 (Faul et al., 2016). The software computed the number of participants based on the setting of the required power level (1 – β) (usually the convention is 0.80) and the estimates of effect size suggested by the meta-analytical results in previous studies. Given that each subtest was informed by the effect size relevant to the componential construct that it was designed to measure, the current study computed the expected sample size for each subtest, respectively. In cases that the information about the synthetical effect size in the investigation of the same construct was not available, the effect sizes from the pairwise comparisons between groups in previous studies would be used. If such effect sizes were not available, the benchmark effect sizes suggested in Plonsky and Oswald’s (2014) meta-analysis of effect sizes in the L2 research were consulted.

Table 3.4 shows the results of the planned sample sizes using G * Power 3 with the setting power level and the referenced effect sizes. The results suggested that 67 participants were needed for the repeated design of TALL_VL to obtain the effect of the modality reported in a study on the testing of vocabulary size in aural and written formats (Mizumoto & Shimamoto, 2008). Since there was no available information on the effect of modality in measuring grammar learning or executive control in WM, a benchmark effect size of a small magnitude in L2 research was adopted. This allowed for the determination that a sample size of 52 was needed in taking TALL_LA and TALL_CST in two modalities. Therefore, the predetermined sample size of the current study suggested by G*Power 3 was 67.

Table 7.4 Prior power analysis results

Subtest	**Effect size**	**Description**	**Power level**	**Sample size required**
TALL_VL	partial $\eta^2 = .42$	the main effect of test format (Mizumoto & Shimamoto, 2008)	$\alpha = .05$ Power = .80	n = 67 for one group in within-subject design
TALL_LA & TALL_CST	*d* = .4	the meta-analytical small effect of group comparisons in L2 research (Plonsky & Oswald, 2014)	$\alpha = .05$ Power = .80	n = 52 for one group in within-subject design
All subtests	*d* = .4	the meta-analytical small effect of group comparisons in L2 research (Plonsky & Oswald, 2014)	$\alpha = .05$ Power = .80	n = 61 for one group in within-subject design; n = 198 for two groups in between-subject design

Using G * Power 3 software to conduct power analysis has been applied in recent publications (e.g., Bovolenta & Williams, 2022; Walker et al., 2020). But it was not the only method available to achieve the purpose of prior power analysis. For the exploratory purposes, R code of a power analysis (see below) introduced by Norouzian (2020) was also applied in the current study for the mixed repeated measures designs, which means the same research groups were tested several times and the performance across the testing sessions was measured by an interaction factor, namely, the modality factor in the current study.

```
plan.mrm(d = .4, n.rep = 2, n.group = 1, factor.type = "within")
```

The results produced by this code suggested that in order to achieve a conventional statistical power of .80 when comparing a group tested using one test suite of one modality at the first session to the same group tested by the other suite in the second session, 61 participants were needed to detect an effect size of *d* = .40. This effect size was based on the benchmark of a small magnitude in L2 research as synthesised by Plonsky and Oswald (2014). Similarly, to achieve a conventional statistical power of .80 of a comparison of two groups tested by using different suites in the first session, the code was adapted as:

```
plan.mrm(d=.4, n.rep=1, n.group=2, factor.type="between")
```

The result suggested that 198 participants for two groups are required to detect the same effect size of *d* = .40.

The sample sizes obtained from the prior power analysis using different statistical methods did not have considerable differences. The results suggested that it would be necessary to have 198 participants in the first test to allow the between-subject comparison of the effect of modality, and 67 participants to complete two rounds of test for the within-subject design.

3.3.1.3 Final sample size

As mentioned earlier, the recruitment strategy involved continuously inviting participants to take the test until the scheduled final day of the first test session. This approach was adopted to account for the unpredictable drop-off rate and the attrition commonly associated with Internet-based research. Additionally, the predetermined sample size of 198, as explained in the previous section, ensured that the actual sample size met the minimum requirements for acceptable statistical power. To be specific, recruitment would end if more than 198 participants had taken the test by the planned final day of the first session. Otherwise, recruitment would continue until the number of participants reached a total of 198.

The data collection period spanned a total of 80 days. At the beginning of the procedure, 892 test codes were issued to invite participants to take the test. Of these, 276 participants accessed the test website using their assigned test code. Among these participants, 221 individuals (80%) completed all subtests during the first session, and their behaviour was verified by checking the recorded data of TALL_SNWR to ensure there were no instances of misbehaviour, such as talking to others, or chewing food during the process, or playing the audio stimuli on a separate device to help with the recall. These 221 participants became eligible to participate in the second session after a minimum 30-day interval. Out of these eligible participants, 194 returned to take the second session, and 181 of them (93.3%) completed the second session, which constituted the final sample size for the within-subject design employed in the main study. Figure 3.7 shows the overview of the number of participants in the experiment process.

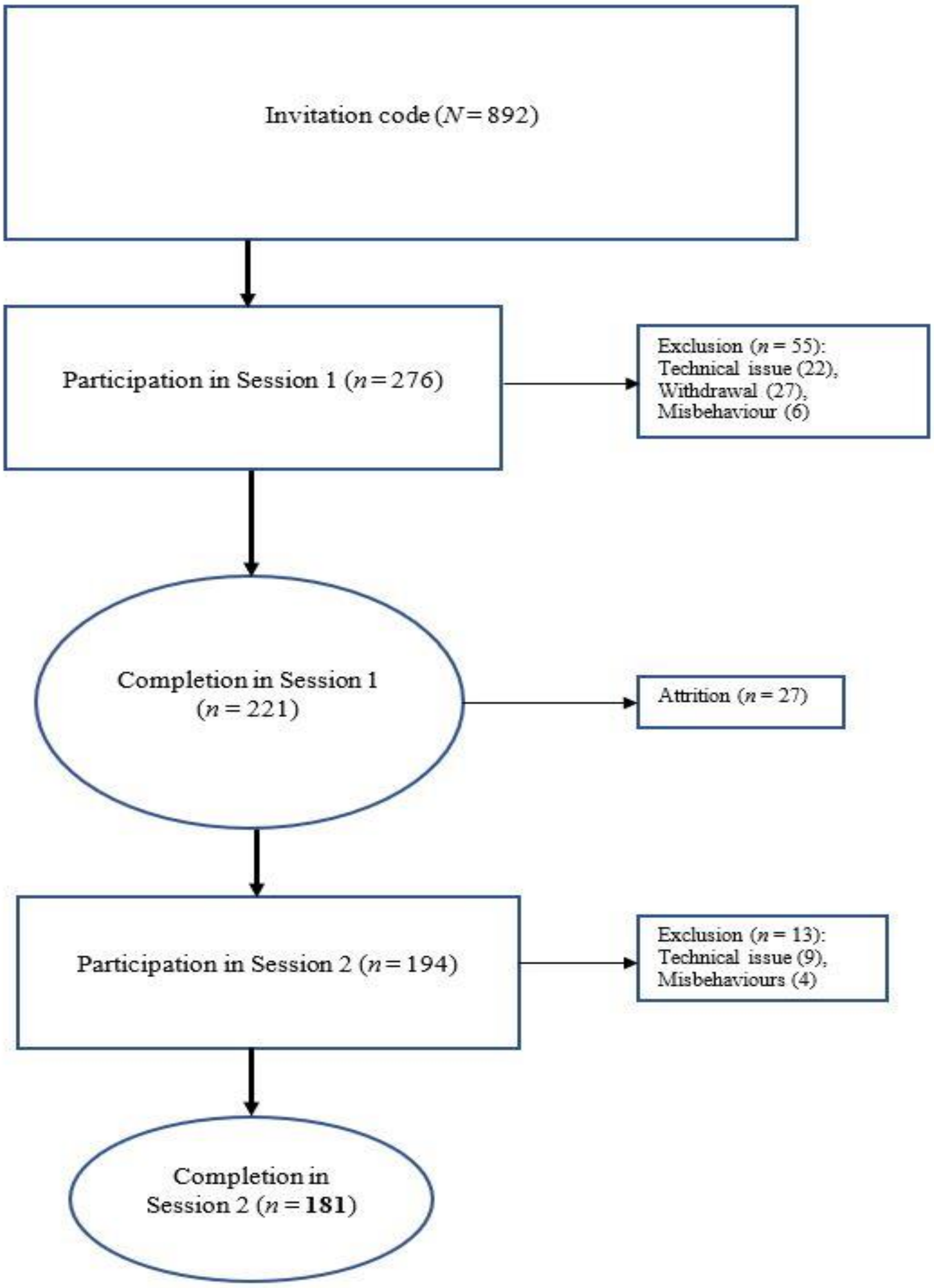

Invitation code (N = 892)
Participation in Session 1 (n = 276)
Exclusion (n = 55):
Technical issue (22),
Withdrawal (27),
Misbehaviour (6)
Completion in Session 1 (n = 221)
Attrition (n = 27)
Participation in Session 2 (n = 194)
Exclusion (n = 13):
Technical issue (9),
Misbehaviours (4)
Completion in Session 2 (n = 181)

Figure 3.7 An overview of participants' involvement

3.3.2 Ethical Considerations

Before commencing data collection, ethical approval for this study was obtained from the Ethics Committee of the Department of Education at the University of York. The ethical approval includes an explanation of the use of the internet-based instrument developed for data collection with participants recruited from China. This study was unlikely to present significant ethical risks since all participants were over 18 years old and were recruited through the researcher's professional network, without any involvement in the courses or modules for which the researcher was the instructor.

Each participant received an email with a set of test invitations. This package included an identical test code and detailed participation requirements, such as the necessary digital equipment, the expected time for completing the test, and the possibility of retaking the test after a minimum 30-day interval. Furthermore, the invitation package provided information about the lack of connection between test performance and participants' academic achievements, along with what they could expect in return: a 50-yuan cash payment made online and a report of their scores in all subtests.

All participants who used the provided test codes to access the test platform read, signed, and downloaded the consent forms written in their L1 Chinese, before proceeding with the test. Examples of equivalent consent forms in both English and Chinese are available in Appendix B. Participants were informed that they had the option to withdraw from participation at any point during the test without the need to provide a reason. They could also request the withdrawal of their data by emailing the research within two weeks after the data had been collected.

Although the overall ethical risk was low, a primary concern revolved around the potential anxiety experienced by test-takers. This anxiety could manifest as feelings of inadequacy in dealing with the test or in setting up their computers for the test's purpose. In order to mitigate this anxiety, the researcher established a social media chat group on the QQ platform to provide asynchronous technical support, as explained in Section 3.3.1.1, and to address any inquiries.

All participants had the flexibility to complete the tests in the comfort of their homes or dormitories at their own pace. The researcher expressed gratitude to the group continuously but ensured the anonymity of all participants. This social media chat group was maintained for six months after the completion of data collection, in case further consultation or support was required.

3.3.3 Measures

The TALL suites

As introduced in Section 3.2.1.2, TALL had two suites for use in the current research. One was the aural suite, with all instructions and materials presented in the aural modality. The other was the written suite, comprising all instructions and materials in TALL_VL, TALL_LA and TALL_CST presented in the written modality, while TALL_SD and TALL_SNWR had materials (necessarily) presented in the aural modality. The audio instructions were generated in a female voice using an online application (https://app.xunjiepdf.com/text2voice/). The audio stimuli in the target language for TALL_VL, TALL_SD, and TALL_LA were presented by a female native speaker of Lithuanian, while the audio stimuli in Chinese for TALL_SNWR and TALL_CST were presented by a female native speaker of Mandarin Chinese.

English Proficiency Test: the NMET

The proficiency of participants' L2 English was determined by referring to the NMET scores, which were self-reported by the participants in the pre-test background questionnaire (see Appendix C). The NMET is the college entrance exam for English undertaken in all provinces in mainland China and serves as the official reference for assessing test takers' English language ability for college and university admission. Although the NMET is a national wide high-stake examination regulated by the Ministry of Education, different test sets are developed and administered in various provinces.

In the current study, the National Set 1 was selected because this was the test set taken in Chongqing, the city where the universities and colleges from which participants were enrolled are located. Thus, a larger number of students would have taken this test set compared to other test sets. The National Set 1 consists of four sections with a total score of 150. Section 1 assesses listening skills with 30 points, comprising 20 multiple choice questions based on 10 recorded dialogues or monologues. Section 2 evaluates reading skills with 50 points, including 15 multiple-choice and 5 filling-in-gap questions related to four passages, each consisting of 250 to 300 words. Section 3 measures grammar and vocabulary knowledge with 30 points, involving cloze and word-filling questions. Section 4 assesses writing skills with 40 points, featuring two writing tasks. Test takers receive only their total NMET score and are not provided with a breakdown of scores for individual sections. As explained in Section 3.3.1.1, participants were selected based whether they reported their scores from the National Set 1, which they had obtained six months prior to the start of the current study.

In addition to collecting information about the participants' L2 English learning background, such as the years of instruction and scores of various proficiency tests, the background questionnaire also included questions about whether the participants had instructed learning experiences in other foreign languages. While 19 participants reported having experience learning six foreign languages (Japanese, French, Russian, German, and Korean) apart from English, none of them had studied any of these languages for more than one year or reported taking any proficiency tests. It is evident that, in general, the participants had a homogenous language background.

3.3.4 Procedure

Within-subject design

To address the research question regarding the reliability and validity of TALL as a battery for language aptitude (RQ1) and to investigate the effects of modality on measuring aptitude (RQ2), a within-subject design was employed to enhance statistical power during data analysis. In this design, the participants completed five TALL subtests in one suite (either aural or written) during the first session. They then took the test in the other suite during a subsequent session, separated by a minimum 30-day interval.

To mitigate any potential carry-on effect that might arise from repeated testing, the main study was designed as a longitudinal study, enabling participants to complete two rounds of tests with the specified interval. Additionally, the modality of the test suites and the material versions were counterbalanced between the two sessions. Furthermore, the order of items in the testing phases of all subtests was randomised, ensuring that every possible item sequence was presented to the participants with no control over how often each sequence was used.

Considering the possibility of a high attrition rate in the Internet-based longitudinal study, the within-subject design with a minimum 30-day interval between two test sessions allowed for an alternative comparison between participants based on the data collected in the first session. In other words, the performance of participants who completed the aural suite in the first session could be compared to those who took the written suite in the first session, in case many participants did not take the second round of testing[1]. The study design is illustrated in Figure 3.8.

The slight numerical imbalance between groups in different modality and material version conditions was expected. Participants were randomly assigned to two groups,

[1] The final sample size was sufficient for within-subject comparison. Therefore, the optional between-subject comparison was not exercised.

determining the order in which they took the test suites prior to their participation in the first session. Additionally, the material version that participants took in the first session were randomly assigned after they logged in using the assigned test code. Furthermore, the variation in sample sizes across modality and version conditions resulted from attrition, with some participants who completed the first session not participating in the second session.

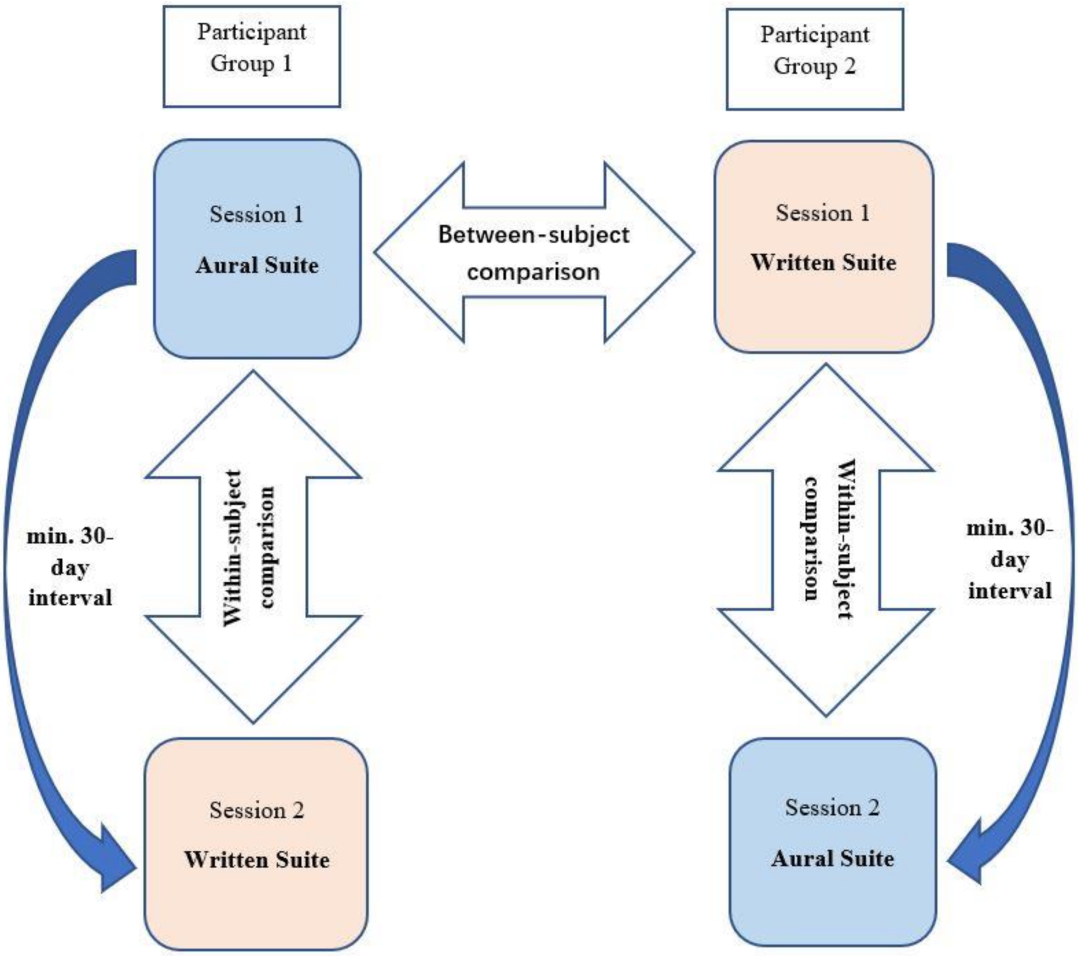


Figure 3.8 The study design

General experimental procedure

The main data collection phase occurred between November 2021 to February 2022. To complete all subtests of TALL in a single session, participants required approximately 45 minutes. They were instructed to complete these subtests on the test platform in a quiet environment, at their convenience. Participants were trusted to work independently, without seeking assistance from others. The data collection procedure is outlined in Figure 3.9.

Participants received the invitation package containing the URL of the test website. They were instructed to set up their computer to enable the recording function, following the provided instructions in the invitation package. Afterward, they initiated the audio testing procedure[2] by accessing the website, using the test code provided to login. They were then directed to the consent form, which they were required to review, sign off on, and download. Additionally, they were asked to provide demographic and foreign language learning background information.

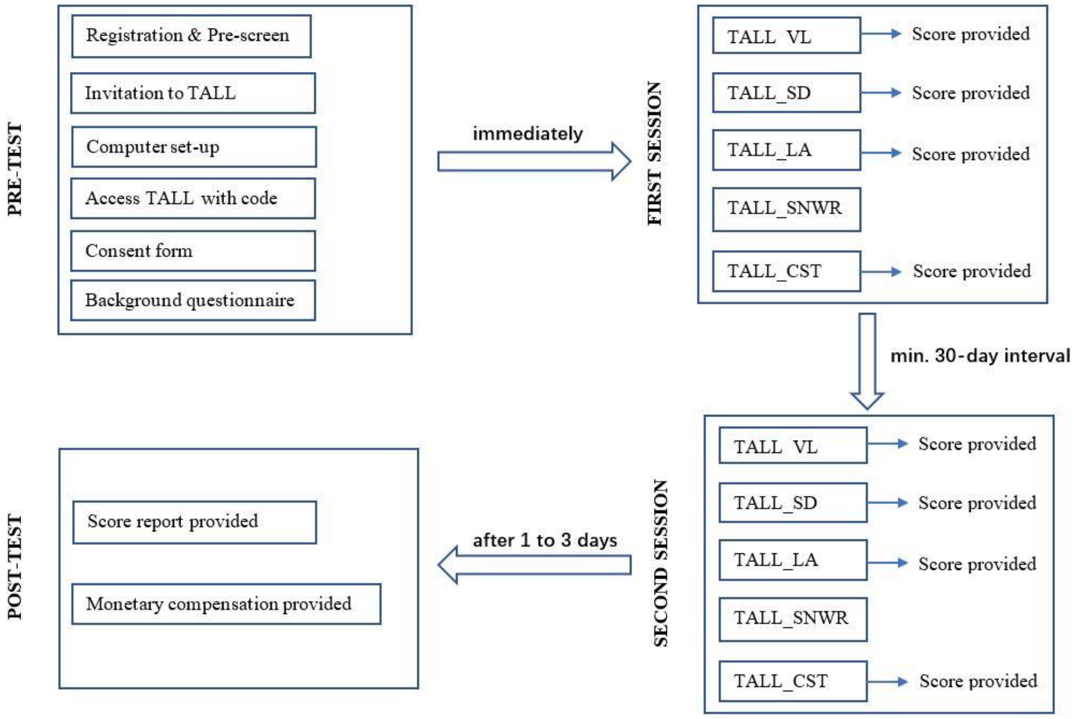


Figure 3.9 Procedure of data collection

Following the pre-test procedure described above, participants had the flexibility to take the subtests in a fixed order (TALL_VL, TALL_SD, TALL_LA, TALL_SNWR, and TALL_CST) at their own pace. The computer provided a mandatory 30-second break between TALL_SNWR and TALL_CST to alleviate potential fatigue. This break was

[2] Participants were instructed not to start the test if they were unable to hear the recorded voice in the testing trial. However, some participants, despite encountering setup issues, proceeded with the test, resulting in blank recordings in TALL_SNWR and an incomplete test.

implemented based on participants' feedback during pilot testing, where they reported that TALL_SNWR was cognitively demanding.

Participants who completed all the subtests in the first session were informed to take the second session after at least 30 days. A reminder email was sent to participants when the 30-day period elapsed. Following the completion of the second session, participants received a report containing their complete TALL scores for both modality suites and the monetary compensation of 50 yuan.

3.3.5 Data analysis plan

This section begins by outlining the steps taken to prepare data sets for analysis, focusing on general scoring principles and dealing with missing values and outliers. It then presents the general data analysis plan for each research question. To enhance readability and minimise the need for readers to refer back to this section, specific decisions regarding statistical procedures are presented separately in Chapters 4, 5, and 6.

Data collected in the current research were analysed using R (R Core Team, 2022), with R markdown files (see Appendix D) documenting data analysis procedures and present analysis results. This practice aligns with the recommendation of developing and disseminating data analysis tools to promote reproducibility and data sharing within the field of applied linguistics (Mizumoto, 2023).

3.3.5.1 Data preparation

The raw data were collected on the TALL test website and stored in the back-end database, with one data file generated for each subtest on a daily basis. Given that all raw data associated with test codes were collected and stored, it posed a significant workload to exclude data from participants who withdrew during the process or due to erroneous recordings (which will be discussed later in this section).

Data preparation was conducted using the `tidyverse` package (Wickham et al., 2019) in `R` to organise and rename variables for analysis. This process was carried out separately for each subtest.

Erroneous data removal

TALL generated raw data that included two types of erroneous data points. The first type comprised raw data sets with missing values, meaning that participants' responses were not recorded. This type of missing values differed from participants' failure to respond because the testing programme recorded non-responses as 0. In other words, regardless of whether a participant could respond to an item or not, the data set should have been complete with all test items recorded. However, in the data pool of the current study, four participants had incomplete data sets. This issue might have resulted from backend operational error related to the broken point transfer and data storage stability (see the technical improvement after the preliminary pilot in 3.2.3.1). Consequently, the data from these four participants were excluded from the final datasets for analysis.

The second type of erroneous datasets consisted of repetitive data points found within the raw data sets. There were two possible reasons for these problematic data points. First, it might have resulted from repetitive attempts by the researcher to download data files, leading to the duplication of responses for certain items in the data sets. Second, although unlikely, it could still be possible that participants took the same test more than once. Although the design of TALL was intended to prevent participants from taking the same test repeatedly by controlling access with one-time test codes, a small number of observations still appeared to be caused by repetitive testing. Consequently, a manual check was conducted to address this issue. Specifically, (1) only the repetitive portion of data pointes was removed if the responding times for repetitive items were the same (for the first possible reason), and (2) the entire data sets of participants were removed if they had taken the test more than once (for the second possible reason).

Outliers

An outlier in a dataset is defined as an observation that significantly deviates from the other observations in that dataset (Barnett & Lewis, 1994). Two types of outliers are recognised: *extreme observations*, which are values that are either extremely low or high but still belong to the same distribution as the other values in the data set, and *contaminants*, which are values from a different distribution and may not necessarily be extreme values. In psychological experiments, which typically involve small sample sizes, distinguishing between these two types of outliers can be challenging, as contaminants can also be extreme, and some extreme values can occur in heavy-tailed distributions that resemble a normal distribution (Wilcox, 1998).

The removal of outliers from a dataset is a common practice in quantitative research, as it can reduce the standard error of the estimates. However, the handling of outliers raises concerns in psychology, as inappropriate methods can lead to increased error variance, reduced sample size, diminished statistical power, and violations of test assumptions, such as the assumption of a normal distribution (Aguinis et al., 2013; Osborne & Overbay, 2004, cited in Nicklin & Plonsky, 2020). While various strategies have been recommended for handling outliers (e.g., Bakker & Wicherts, 2014), and diverse practices within specific subdomains of applied linguistics have been synthesised (see Nicklin & Plonsky, 2020), determining the appropriate approach can be challenging when the goal is to conduct exploratory analyses of data to uncover phenomena rather than initially testing hypotheses (Isbell et al., 2022).

Considering the research aim of providing initial evidence of reliability and validity for a new psychometric measurement, and recognising the exploratory nature of the analyses, a cautious approach to handling outliers was adopted in the current study. In general, automatic exclusion of outliers from the analysis was avoided, unless they could be attributed to errors, to prevent the risk of inflating the Type I error rate, as recommended by Bakker and Wicherts (2014).

Outliers in this study were identified based on observed behavioural responses that indicated participants were not engaging in the test in a typical manner. For example, in TALL_VL in the aural modality, participants who did not click all the buttons of the pictures of the stimuli during the learning phase had their extremely high scores identified as outliers. Similarly, in TALL_CST, participants who either did not provide any response or indiscriminately clicked the same button were identified, and their data were treated as outliers.

The determination of outliers in the data of TALL_CST was distinct. It was based on the performance of semantic judgments on sentence stimuli in the participants' L1, which may not have posed significant cognitive challenges for the college students in this study. Therefore, statistical criteria, specifically the z-score (calculated as the difference between the score and the mean score divided by the standard deviation), with a threshold of 3.29 (Field et al., 2012), to identify outliers in the processing data in TALL_CST.

It is important to note that this approach for removing *attentional* outliers in TALL_CST differs from the practice suggested by Conway et al. (2005). They recommended that data sets are removed if accuracy in the processing task falls below 85%. While this threshold aims to ensure "near perfect" precision in the processing task (p. 775), it can be somewhat arbitrary, as processing stimuli may impose varying demands on heterogenous populations with diverse L1 literacy backgrounds. Therefore, in the current study, data in TALL_CST were retained for analysis if two conditions were met: (1) the recorded behavioural responses in the sentence processing task did not suggest that participants were disengaged from the experiment, and (2) the processing data had a z-score of less than 3.29.

Data preparation protocol

A stepwise protocol (see Appendix E) was developed and followed to prepare datasets for analysis. The aim was to manage inconsistent sample sizes resulting from cases where participants either completed the test once or withdrew during the testing process.

Scores in the subtests

In the TALL_VL, TALL_SD, and TALL_LA subtests, the datasets used for analysis were derived from scores reflecting the number of accurate responses to test items in the raw data files.

The scoring approach applied in TALL_SNWR and TALL_CST followed the all-or-nothing scoring approach suggested in Conway et al. (2005), where methodological considerations for scoring procedures were discussed, and recommendations for decisions based on experimental design within the framework of WM theories were made. Specifically, an all-or-nothing scoring approach credits performance only when stimuli are accurately recalled in the correct serial position. For example, if three stimuli to be recalled in the correct sequence are as “ABC” and the participant produces “BCA”, 0 will be scored as none of the three stimuli is recalled accurately in the correct position. This scoring approach deviates from partial credit scoring, which gives credit for partly correct stimuli accurately recalled but not in the correct sequential position. The rationale behind choosing all-or-nothing scoring was to equally emphasise accuracy in both the form of items recalled and the sequential position of target items.

In TALL_SNWR, the raw data consisted of audio files containing productive data that required manual scoring. Participants’ raw scores in this subtest reflected the number of stimuli articulated correctly in the correct sequential position. Scoring was conducted by the researcher and an invited marker who was a native speaker of Mandarin. They achieved 98.6% agreement rate, resolving any discrepancies through further further crosschecks and discussions.

For TALL_CST, the raw data files contained both processing data (i.e., semantic judgements of sentence plausibility) and storing data (i.e., letters recalled in sequential order). However, only the storing data were used for analysis in this subtest, consistent with the practice of analysing the storing data only in previous research involving complex span tasks (e.g., Gass et al., 2019; Unsworth et al., 2009). The processing data were employed to identify potential outliers, a process described in detail in the preceding section titled “Outlier”.

Types of data

Two types of data, *split* data at the item or trial level, and *aggregated* data of the subtests, were generated for analysis.

First, in TALL_VL, TALL_SD, and TALL_LA, the item-level data were binary (or dichotomous) data, with ‘0’ indicating an incorrect response and ‘1’ representing a correct response.

The data files for TALL_SNWR and TALL_CST were transformed into polytomous data (i.e., data with more than two distinct response categories or levels) for analysis. The proportion of correct responses in each trial constitutes the item-level data. Specifically, unit-weighted scoring option was applied in data transformation. In this option, each trial was scored as a proportion of correctly recalled stimuli in this trial, with all stimuli being treated equally, regardless of the size of the trial. For example, in this research, recalling one stimulus in a trial of the size of two stimuli was scored the same as recalling two stimuli in a four-stimulus trial, both yielding a score of 0.50. This scoring option contrasted with load-weighted scoring, which assigns higher weight to stimuli in trials with a higher load (larger trial size). The decision of choosing unit-weighted scoring option was based on the empirical results that did not reveal a significant difference between unit-weighted and load-weighted options in Conway et al. (2005). Furthermore, unit-weighted scoring aligned with established procedures from psychometric measures in complex span tasks, as reviewed by Conway et al.

It is noteworthy that the approach used for transforming data (i.e., the proportion of correct responses in each trial as the item-level data) for analysis in TALL_SNWR and TALL_CST differed from the scoring approaches used in other studies involving aptitude or WM measurements. For instance, in the Hi-LAB (Linck, et al., 2013), the total number of correctly recalled stimuli functioned as the scores in WM-related subtests. Similarly, Kormos and Sáfár (2008) used weighted average scores (i.e., the weighted average of the number of repeated syllables) in a non-word repetition test. The lack of standardised scoring approach in studies about WM–L2 relationship has been highlighted in synthesis studies (e.g., Leeser & Sunderman, 2016; Shin & Hu, 2022), as the varied scoring approaches in WM tasks may yield discrepant and incomparable findings across different research endeavours.

Second, the aggregated data sets of TALL_VL, TALL_SD, and TALL_LA were generated using transformed percentage total scores, with the maximum score set at 100%. This scoring approach was consistent with the method reported by Bokander and Bylund (2020). In contrast, the aggregated datasets for TALL_SNWR and TALL_CST were derived from the average scores of the test trials, following the approach suggested by Conway et al. (2005).

3.3.5.2 Data analysis for RQ1: Validation plan

In order to address RQ1, which explores the extent to which TALL demonstrates satisfactory reliability and internal validity as an aptitude measure, data were analysed at the *item*, *subtest*, and *battery* levels. This validation plan was adapted from the schema used by Bokander and Bylund (2020) to investigate the internal validity of the LLAMA tests, which draws upon the multilevel inferences outlined by Kane (2006) and Purpura et al. (2015).

Data analysis for RQ1, as illustrated in Figure 3.10, starts at the *subtest* level, aiming to assess the internal consistency of the test items within each subtest – essentially, the reliability of each subtest as a measure of a specific component of aptitude. Prior to the reliability checks, unidimensionality of each subtest, which indicates the degree of commonality among the test items, was evaluated. The assumption of unidimensionality at the subtest level serves as the foundation for subsequent examinations at the item level.

At the *subtest* level, data were split according to different material versions and modalities, followed by analysis to estimate the reliability and factor structure of variables underpinning the data. This analysis was conducted based on the principles of the theory of generalizability (G theory) outlined by Cronbach et al. (1963) and Gleser et al. (1965). The outcomes of this analysis provided evidence for making generalisation inferences, that is, the extent to which scores can be generalised to a broader array of potential items targeting the same construct intended by the subtest.

At the *item* level, analyses aimed to assess the suitability of items within each subtest using Item Response Theory (IRT) models (see Appendix F for a brief introduction). These analyses provided insights into the scoring quality of each subtest, evaluating whether it consisted of well-functioning items with appropriate levels of difficulty for participants and the ability to discriminate participants' latent abilities. Separate analyses were conducted for different conditions of modalities and material versions, within each subtest at the item level.

In three subtests (TALL_VL, TALL_SD, and TALL_LA), the one-parameter logistic (1PL), two-parameter logistic (2PL), and three-parameter logistic (3PL) models were applied, respectively, to dichotomous datasets. These models were compared to determine which best fit the data. For polytomous datasets in two WM subtests (TALL_SNWR and TALL_CST), the Generalised Partial Credit Model (GPCM) was applied.

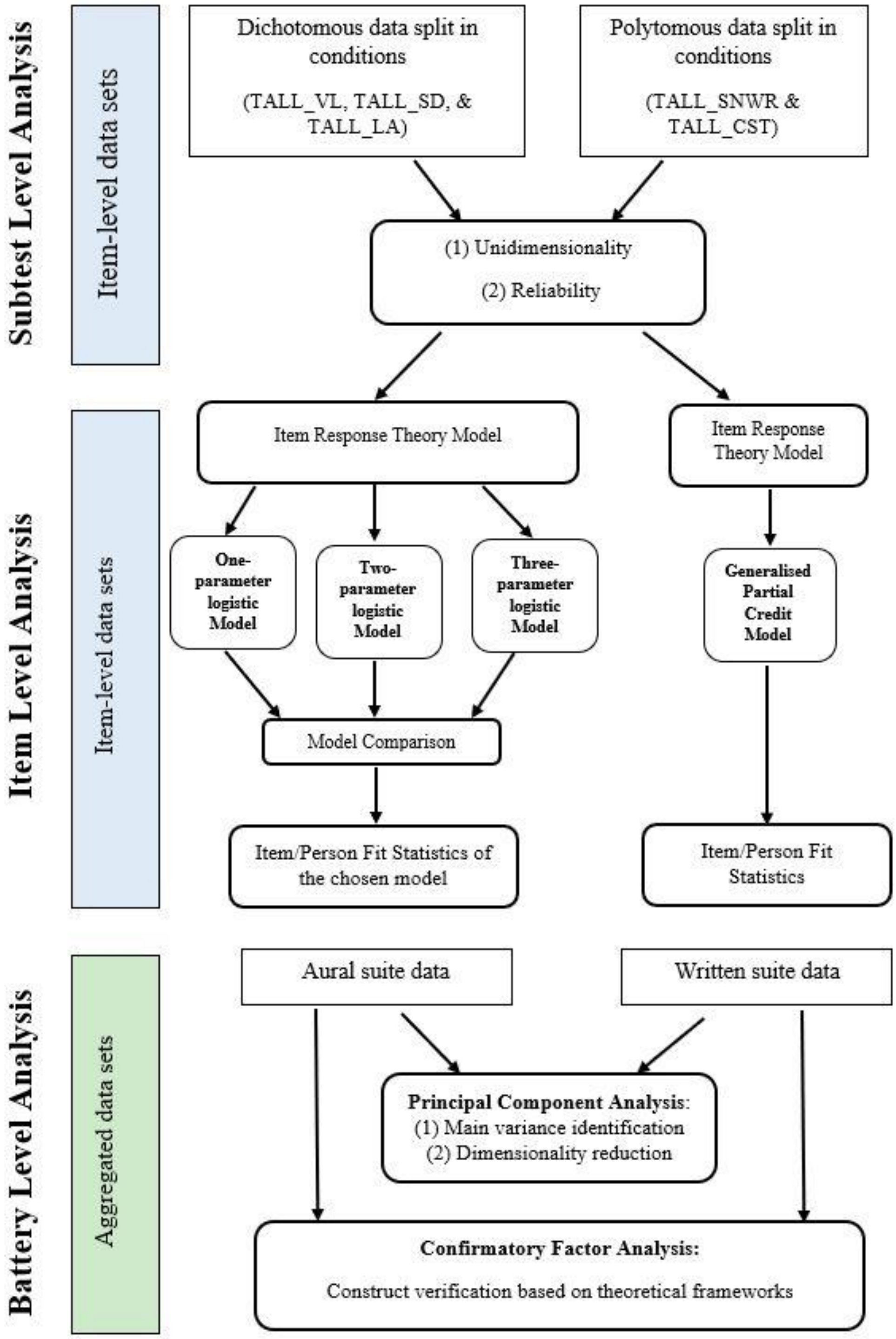

Subtest Level Analysis
Item-level data sets
Dichotomous data split in conditions
(TALL_VL, TALL_SD, & TALL_LA)
Polytomous data split in conditions
(TALL_SNWR & TALL_CST)
(1) Unidimensionality
(2) Reliability
Item Level Analysis
Item-level data sets
Item Response Theory Model
Item Response Theory Model
One-parameter logistic Model
Two-parameter logistic Model
Three-parameter logistic Model
Generalised Partial Credit Model
Model Comparison
Item/Person Fit Statistics of the chosen model
Item/Person Fit Statistics
Battery Level Analysis
Aggregated data sets
Aural suite data
Written suite data
Principal Component Analysis:
(1) Main variance identification
(2) Dimensionality reduction
Confirmatory Factor Analysis:
Construct verification based on theoretical frameworks

Figure 3.10 Data analysis flowchart for RQ1

At the *battery* level, analyses focused on the explanation inference within the validation plan, which assesses the internal validity of TALL as a componential measure for aptitude. Two statistical methods were employed on aggregated data sets, which included the scores of each subtest (rather than individual item scores), to provide evidence for the explanation inference. This evaluation aimed to determine the extent to which TALL subtests reflected a structure consistent with theoretical frameworks of language aptitude, specifically the Stages Approach (Skehan, 2016) and the P/E Model (Wen, 2016)), upon which TALL was conceptualised.

Firstly, Principal Component Analysis (PCA) was used to investigate whether TALL subtests could succinctly measure the dimensions of the aptitude construct without much redundancy. Confirmatory Factor Analysis (CFA) was used to verify whether TALL, as an aptitude battery, tapped into the components supported by the alignment of the data with the structure of four primary factors based on the theoretical models it was built upon.

3.3.5.3 Data analysis for RQ2: Mixed-effects Modelling

To explore the extent to which input modality differentially influenced scored in in TALL_VL, TALL_LA, and TALL_CST (RQ2), data were collected using a within-subject design to investigate modality effects in these three subtests.

Mixed-effects modelling (MEM) was used for this research question. MEM offers statistical advantages when dealing with related data points in experimental designs where participants provide repeated responses to multiple measurements for each stimulus or item, and when testing materials involve counterbalanced stimuli or items that share multiple characteristics (Gries, 2021).

In the current investigation, besides considering modality as the predictive variable, test sessions and material versions when analysing the data generated in the experimental design. Additionally, due to the use of convenience samples and the inability of the developed materials to encompass all possible linguistic options, crossed random effects for subjects and items were included in the data analysis (Baayen et al., 2008).

In summary, MEM allowed us to simultaneously consider multiple effects when dealing with repeated measurements, a task that *t* tests or ANOVAs could not efficiently handle.

3.3.5.4 Data analysis for RQ3: Correlations and multiple regressions

To address RQ3, which examines the extent to which TALL predicts participants' L2 English proficiency as measured by the NMET, scores were aggregated from the TALL subtests in two ways (as introduced in Section 3.3.5.1).

Correlation analyses were conducted to examine the relationships between the scores of TALL subtests and participants' self-reported NMET scores. The results of the correlation analyses informed the application of regression analysis, taking into consideration the potential collinearity of the data.

Before performing multiple regression analyses, the aggregated data were transformed and checked against the assumptions of the regression model. Multiple regressions were than employed to estimate the proficiency of participants' L2 based on the values of other predictive variables—namely, the scores obtained from the TALL subtests.

3.4 Chapter summary

In this chapter, the rationale and methodological considerations for developing TALL are presented. The focus is on the development of the subtests within the battery, including the selection and development of test stimuli, formats, and scoring options. Additionally, detailed methods for data collection and data analysis are reported, covering considerations in sampling, data preparation, and the data analysis plan. Subsequent chapters will present results related to the research questions.

CHAPTER 4: RESULTS & DISCUSSION FOR RQ1 – RELIABILITY AND INTERNAL VALIDITY

4.1 Introduction

This chapter reports the analyses that aimed to answer RQ1 about the extent to which TALL, as a measure for language aptitude, displayed satisfactory reliability and internal validity. The first section of this chapter presents the results from the data wrangling following the protocol and practices introduced in Section 3.3.5.1, which re-composes datasets into a usable format with cleaned data prior to further analysis. Different types of data files of all subtests and the aggregated data files of the two test suites (aural and written) were generated in this section. The second section reports descriptive statistics of the data of each subtest split by the conditions of material version and/or modality, as well as the self-reported L2 English proficiency test (NMET). These results were about normality checks, the means, the standard deviations, and the plots displaying the density of distribution and the error bars of the confidence intervals. The third section presents the results from the data analyses at the subtest, item, and battery level respectively, based on the validation plan introduced in Section 3.3.5.2. In this section, unidimensionality and reliability checks were performed at the subtest level to check the dimensional assumption prior to the analysis at the item level. IRT models were applied for the item level analysis to diagnose the item quality of each subtest. At the battery level, PCA was performed on the aggregated datasets to investigate whether TALL could succinctly measure the dimensions of aptitude construct without much redundancy. CFA was used to verify whether TALL, as an aptitude battery, measured the components (i.e., associative memory, Phonetic coding ability, language analytic ability, and WM) informed by the theoretical models of language learning aptitude. All results and analysis code in this chapter were rendered in R markdown files (see Appendix D).

4.2 Final data files

Following the stepwise data preparation protocol (Appendix E), the paired datasets were obtained from participants who had scores of five subtests from two sessions available for analysis. Further removing erroneous data and outliers resulted in the final data files for analysis from 165 participants (118 females and 47 males, age$_{mean}$ = 19.01, ranging from 17 to 20). This final sample size was larger than the intended sample size based on the results of the prior power analysis (see Section 3.3.1.2).

In this chapter, to provide descriptive statistics, datasets were formed for each subtest, respectively. The dataset of each subtest was further split by conditions of material version and/or modality and the unaggregated data (item scores) were analysed at the subtest and item level. Table 4.1 shows the split datasets in five subtests. The aggregated datasets (subtest scores rather than item scores) were also generated at the battery level for each test suite (see Table 4.2).

Table 4.1 Split datasets

Subtest	**Split dataset**
TALL_VL (Vocabulary Learning)	Version A in aural modality
	Version B in aural modality
	Version A in written modality
	Version B in written modality
TALL_SD (Sound Discrimination)	Version A
	Version B
TALL_LA (Language Analysis)	Version A in aural modality
	Version B in aural modality
	Version A in written modality
	Version B in written modality
TALL_SNWR (Serial Nonwords Recall)	Version A
	Version B
TALL_CST (Complex Span Tasks)	Version A in aural modality
	Version B in aural modality
	Version A in written modality
	Version B in written modality

Table 4.2 Aggregated datasets

Test suite	Subtest	Aggregated dataset
Aural	TALL_VL	Combined (Version A and B) in aural modality
	TALL_SD	Combined (Version A and B) in aural suite
	TALL_LA	Combined (Version A and B) in aural modality
	TALL_SNWR	Combined (Version A and B) in aural suite
	TALL_CST	Combined (Version A and B) in aural modality
Written	TALL_VL	Combined (Version A and B) in written modality
	TALL_SD	Combined (Version A and B) in written suite
	TALL_LA	Combined (Version A and B) in written modality
	TALL_SNWR	Combined (Version A and B) in written suite
	TALL_CST	Combined (Version A and B) in written modality

Note. The combined datasets were created by aggregating scores from different participants who took each version.

4.3 Subtest level analysis

Subtest level analysis was performed using data of each subtest separately. Datasets were split by the conditions of materials version and/or modality respectively. Descriptive statistics will be reported first, followed by results from the unidimensionality and reliability checks, providing the evidence to demonstrate the assumptions were met for the analysis at the item level (to be reported in Section 4.4).

4.3.1 Descriptive statistics

Descriptive statistical analyses were performed on each subtest using datasets split by the conditions of material version and/or modality, as well as on participants' self-reported NMET scores.

TALL_Vocabulary Learning

The total score of TALL_VL is 20. The results (see Table 4.3 for descriptive statistics) did not indicate substantially skewed distribution of the scores in any condition according to the rules of thumb for skewness (between -2 to +2) and kurtosis (between -7 to +7) (Hair et al., 2010). However, the results from Shapiro-Wilk normality test showed that only the scores of Version B in the written modality were normally distributed (p = 0.12), while the datasets of scores in other conditions did not have normal distributions, with p values less than .01.

Table 4.3 Descriptive statistics of TALL_VL

modality	version	*n*	min	max	mean	std.dev	skew	kurt.	std.err	Shapiro-Wilk	
										W	*p*
aural	A	81	0	19	5.46	3.32	1.46	2.85	0.37	0.88	< .001
aural	B	84	0	19	7.14	3.98	0.82	0.62	0.43	0.95	0.002
written	A	84	0	20	8.23	4.47	0.58	-0.32	0.49	0.96	0.009
written	B	81	1	20	9.95	4.68	0.03	-0.80	0.52	0.98	0.12

The violin box plots of the scores (Figure 4.1) displayed the information on density of the distribution (i.e., the shape of the violin), the range between the upper and lower quartiles (i.e., the box showing 50 % of the scores), the mean (i.e., the point in the box), and the error bar for confidence intervals (i.e., the error bar in the box) of the split datasets of materials versions in two modalities (aural and written) from two test sessions (first and second).

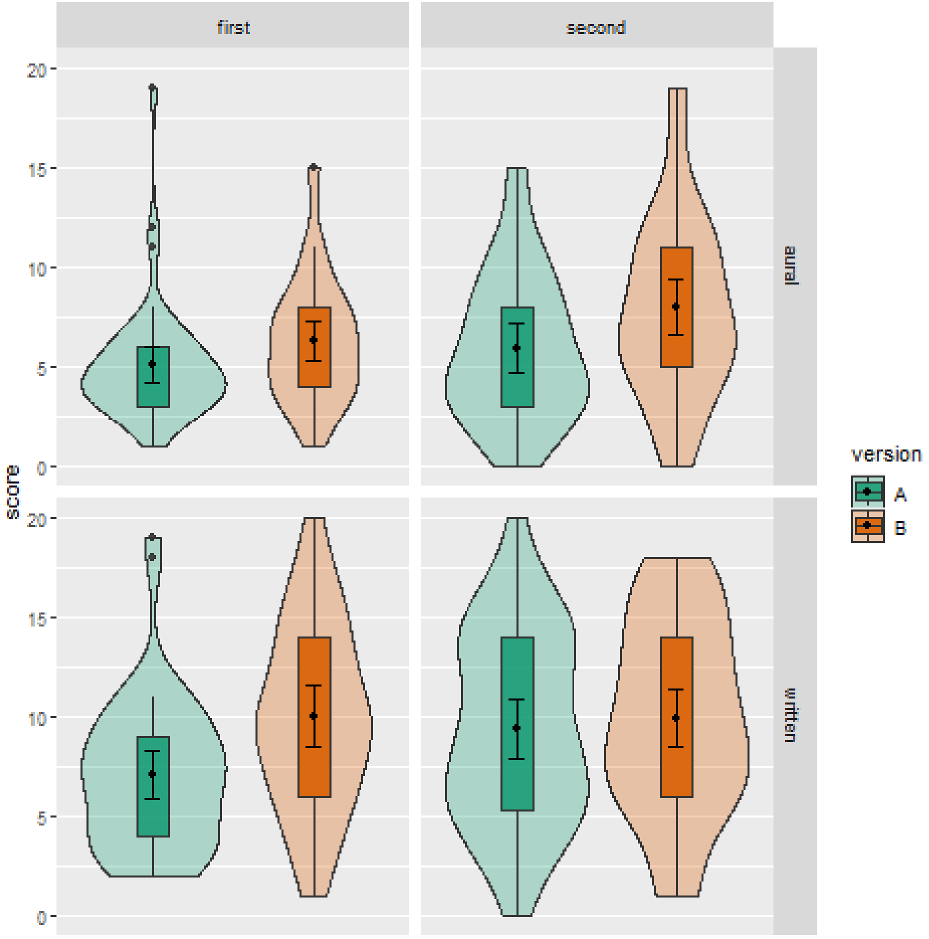


Figure 4.1 Violin box plot of TALL_VL

TALL_Sound Discrimination

The total score of TALL_SD is 30. The results (see Table 4.4 for descriptive statistics) did not indicate substantially skewed distribution of the scores in any condition according to the rules of thumb for skewness. However, the results from Shapiro-Wilk normality test showed that scores of both versions did not have normal distributions, with *p* values less than .001.

Table 4.4 Descriptive statistics of TALL_SD

version	*n*	min	max	mean	std.dev	skew	kurt.	std.err	Shapiro-Wilk	
									W	*p*
A	165	6	30	21.80	6.50	-0.50	-0.85	0.51	0.93	< .001
B	165	4	30	23.61	4.09	-1.84	6.03	0.32	0.86	< .001

The violin box plots of the scores are shown in Figure 4.2.

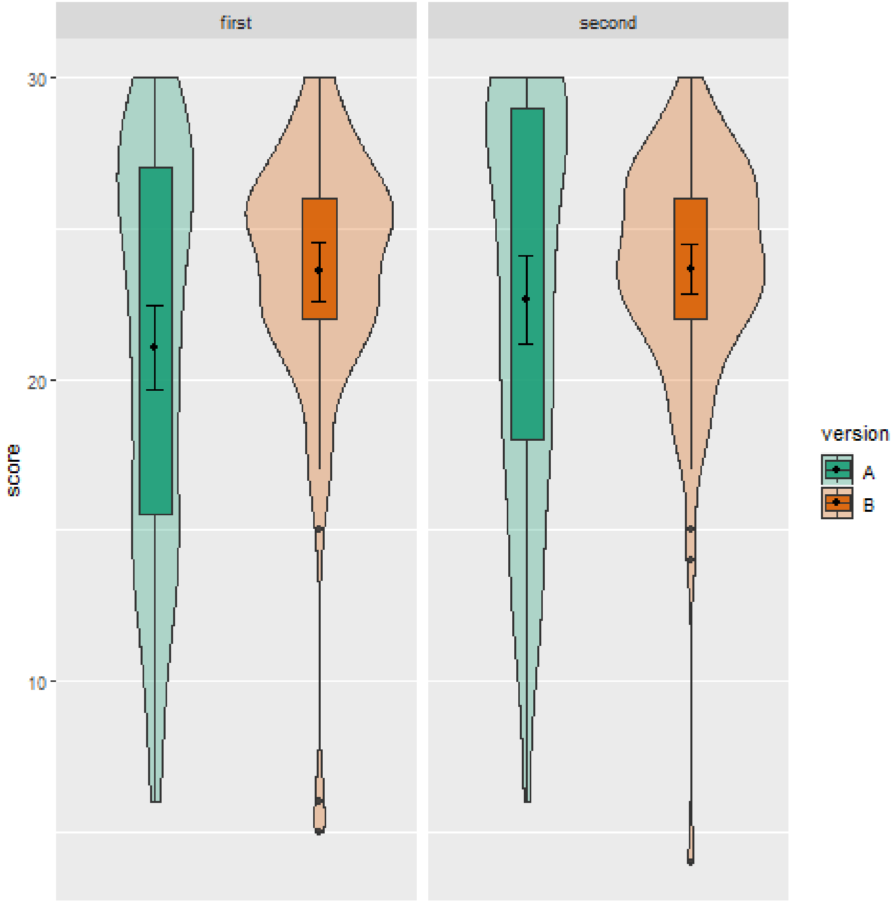


Figure 4.2 Violin box plot of TALL_SD

TALL_Language Analysis

The total score of TALL_LA is 30. The results (see Table 4.5 for descriptive statistics) did not indicate substantially skewed distribution of the scores in any condition according to the rules of thumb for skewness. However, the results from Shapiro-Wilk normality test showed that all scores did not have normal distributions, with *p* values less than .001.

Table 4.5 Descriptive statistics of TALL_LA

modality	version	*n*	min	max	mean	std.dev	skew	kurt.	std.err	Shapiro-Wilk	
										W	*p*
aural	A	81	3	30	19.73	9.34	-0.41	-1.42	1.04	0.87	< .001
aural	B	84	4	30	20.24	7.36	-0.58	-0.88	0.80	0.92	< .001
written	A	84	4	30	24.73	7.55	-1.46	0.73	0.82	0.71	< .001
written	B	81	0	30	23.88	7.43	-1.44	1.02	0.83	0.76	< .001

The violin box plots of the scores are shown in Figure 4.3.

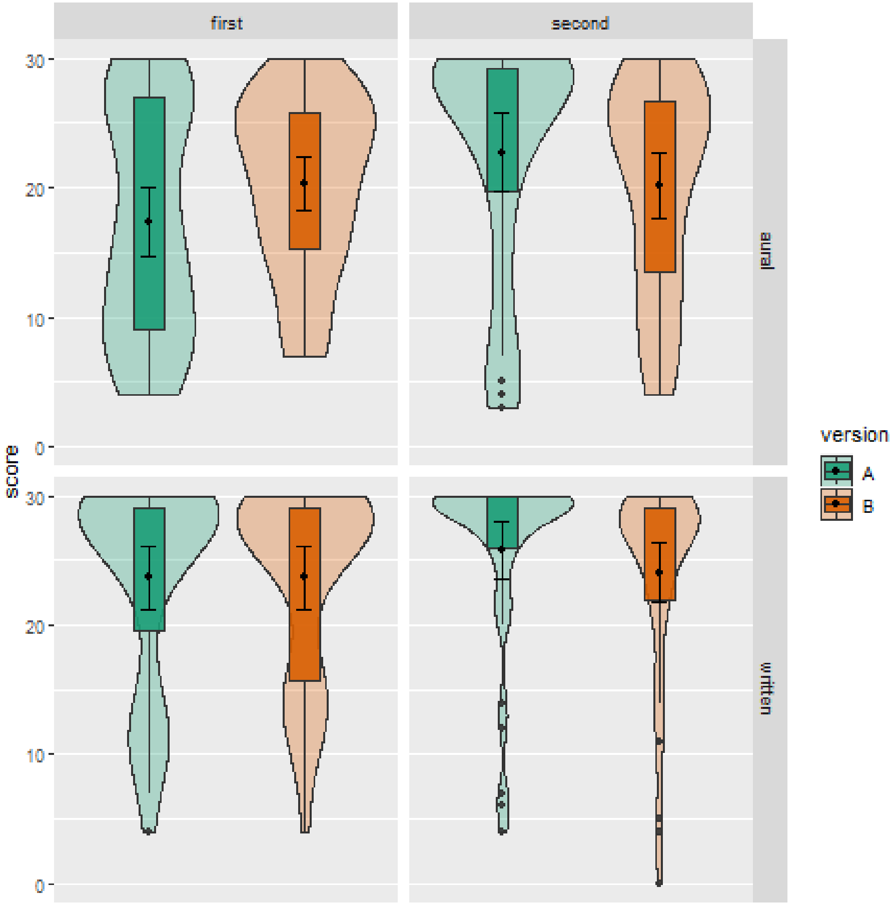


Figure 4.3 Violin box plot of TALL_LA

TALL_Serial Nonwords Recall

The total score of TALL_SNWR is 17. The results (see Table 4.6 for descriptive statistics) did not indicate substantially skewed distribution of the scores in any condition according to the rules of thumb for skewness. However, the results from Shapiro-Wilk normality test showed that scores of both versions did not have normal distributions, with *p* values less than .001.

Table 4.6 Descriptive statistics of TALL_SNWR

version	*n*	min	max	mean	std.dev	skew	kurt.	std.err	Shapiro-Wilk	
									W	*p*
A	169	0.25	14.76	5.70	2.56	0.74	0.75	0.20	0.97	< .001
B	169	0.84	15.20	5.71	2.69	0.68	0.27	0.21	0.97	< .001

The violin box plots of the scores are shown in Figure 4.4.

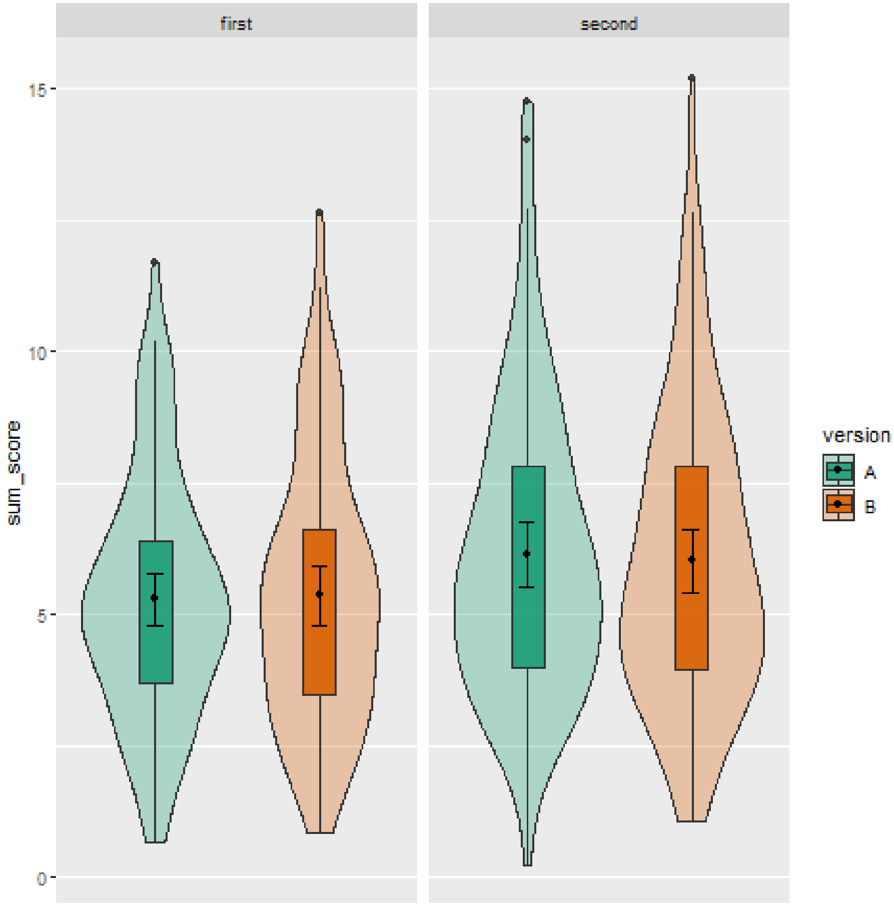


Figure 4.4 Violin box plot of TALL_SNWR

TALL_Complex Span Task

The total score of TALL_CST is 15. The results (see Table 4.7 for descriptive statistics) did not indicate substantially skewed distribution of the scores in any condition according to the rules of thumb for skewness. However, the results from Shapiro-Wilk normality test showed that all scores did not have normal distributions, with *p* values less than .001.

Table 4.7 Descriptive statistics of TALL_CST

modality	version	*n*	min	max	mean	std.dev	skew	kurt.	std.err	Shapiro-Wilk	
										W	*p*
aural	A	81	7.16	15	12.25	1.90	-0.87	0.11	0.21	0.93	< .001
aural	B	84	5.44	15	12.25	2.00	-0.95	0.72	0.22	0.93	< .001
written	A	84	7.41	15	13.32	1.20	-1.61	5.51	0.13	0.89	< .001
written	B	81	6.08	15	13.00	1.48	-1.63	4.61	0.16	0.89	< .001

The violin box plots of the scores are shown in Figure 4.5.

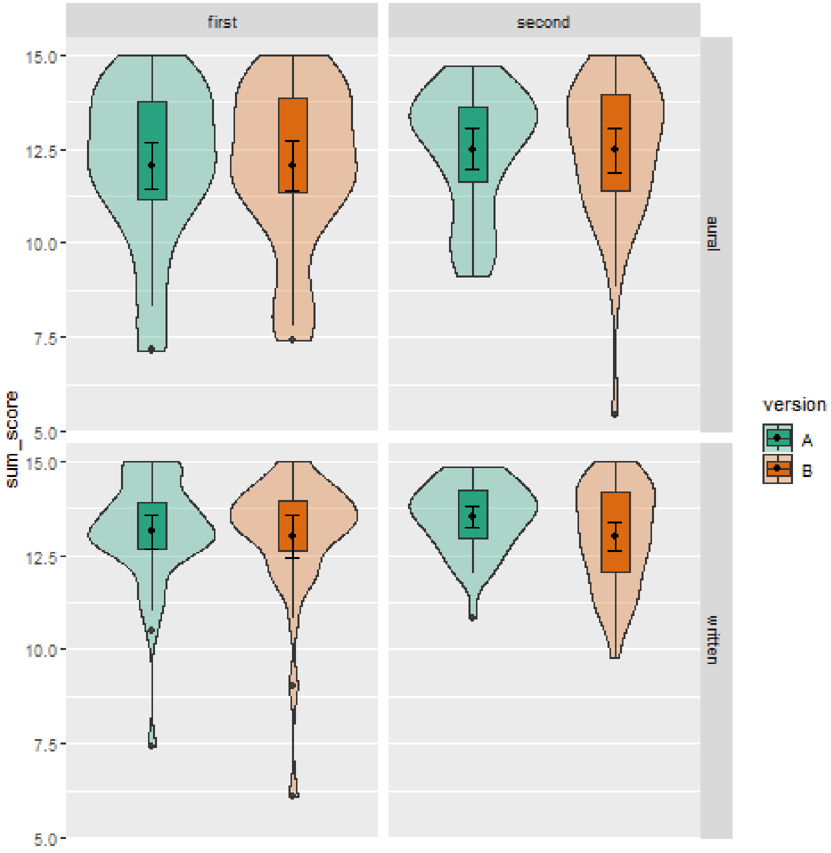


Figure 4.5 Violin box plot of TALL_CST

NMET Scores

The total score of NMET is 150. The results (see Table 4.8 for descriptive statistics) did not indicate substantially skewed distribution of the scores according to the rules of thumb for skewness. However, the normality of the NMET scores was evaluated using the Shapiro-Wilk normality test and the result showed that the scores did not have normal distributions, with *p* values less than .001.

Table 4.8 Descriptive statistics of NMET scores

n	min	max	mean	std.dev	skew	kurt.	std.err	Shapiro-Wilk	
								W	*p*
165	53	146	126.11	15.81	-1.9	4.5	1.23	0.83	< .001

The histogram and box plots of the scores are shown in Figure 4.6.

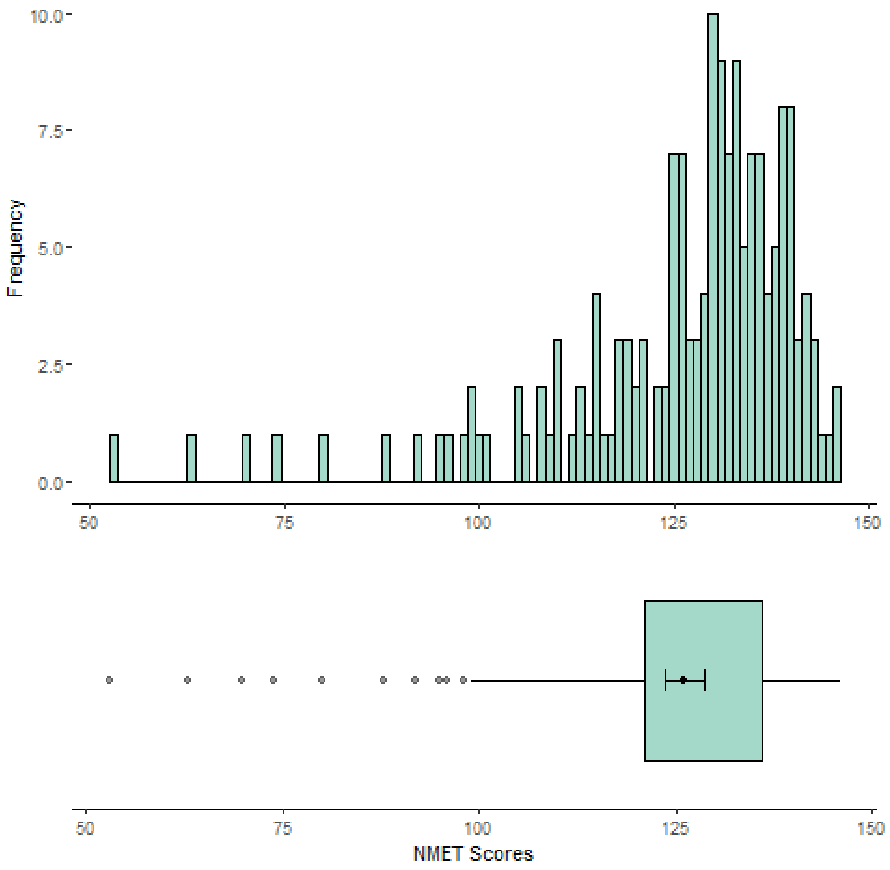


Figure 4.6 Histogram and box plots of NMET scores

4.3.2 Reliability and unidimensionality

This section is mainly about the internal consistency of the items in each subtest, that is, the reliability of each subtest as a measure for a component of aptitude. Reliability checks can also inform the unidimensionality of the subtests, that is, the degree of commonality of the latent abilities that can be reliably captured by each subtest. The reliability and unidimensionality analysis aimed to (1) ascertain that each subtest measures a latent construct with the internally consistent set of items, and (2) retain as many items in the measurement as possible whilst adhering to the internal consistency of the items.

To achieve the first aim, the unidimensionality indices and the reliability coefficients of each dataset collected by the same instrument (i.e., the instrument using the same version of materials in the same modality) were examined. To achieve the second aim, two solutions were needed. First, the changes of the reliability coefficient of each dataset were checked if each item was dropped. Additionally, items were checked on the loading to the general factor in the statistical outputs. If the reliability coefficient was incrementally increased via the deletion of a certain item, or if an item had factor loading less than 0.2 to the general factor (Revelle, 2022), this indicated that the item would receive further scrutiny in further IRT analysis at the item level (see section 4.4).

4.3.2.1 Reliability coefficients: from α to ω

Reliability is crucial to determine whether psychometric instruments can accurately measure latent variables. The fundamental issue of reliability is the extent to which scores obtained in one time and place using one instrument can predict scores obtained in a different time and/or place, perhaps using a different instrument (Revelle & Condon, 2019). Instrument reliability can also take various forms, such as the estimates of *internal consistency* provided by items within a test, the estimates of *alternate form reliability* provided by the observed correlation between alternate forms or parallel tests administered simultaneously, the *dependability* measured by the same tests given at similar times, and the *stability* assessed by the same measures taken over a longer period.

In the current research, two forms of reliability were examined. The main concern was the internal consistency of the subtests as reliable instruments that could obtain scores by participants taking a particular version of materials (i.e., the form) in one or two modalities (i.e., the occasion). Another type of reliability was alternate form reliability that only applied to two subtests (i.e., TALL_SD and TALL_SNWR), which were administered in the aural modality only. In these two subtests, alternate form reliability was estimated by the correlations between the scores obtained by two different versions of materials that were conceptually similar but with different items. It is worth noting that test session was not considered as a condition for splitting data for reliability assessment because the measurements in the same material version and the same modality were treated as equivalent for reliability analysis, regardless of the session in which they were administered. This is also reasonable given that the carry-over effects have been controlled by counterbalancing design of the test items in different sessions.

In addition to poor L2 instrument reliability reporting and the low reliability of aptitude measures (Bokander & Bylund, 2020; Rogers & Meara, 2019; see the review in Section 2.2.3.1), the question about which reliability coefficient should be used has also been raised by researchers in the field (e.g., O'Reilly & Marsden, 2020). Derrick's (2016) systematic review reports that reliability coefficients were reported for 28% of the instruments, with Cronbach's alpha (α) being the most frequently used index in 22% of the instruments. Although Cronbach's alpha has been a widely used reliability coefficient index in the literature in L2 research broadly, and of aptitude research specifically, it has been argued that its assumptions are often not met or corrected. For example, McNeish (2018) argues that Cronbach's alpha may impose limitations that result in underestimating reliability of psychometric measures as its assumptions may often be violated. These assumptions include that Cronbach's alpha assumes items having *tau* equivalence (that is, all test items contribute equally to the total score), a normal continuous distribution of data, unidimensionality (that is, all test items measure the same construct), and uncorrelated errors (that is, items are not correlated via any other sources). To overcome these limitations, McNeish proposed other alternatives (omega coefficients, coefficient *H*, and greatest lower bound) that use factor analysis of item loadings on a single latent dimension to determine instrument reliability (see, however, Raykov and Marcoulides (2019) who argue that under appropriate conditions, Cronbach's alpha should continue to be used, challenging McNeish's interpretation of the original assumptions of Cronbach's alpha). Revelle and Condon (2019) also discuss two problems with *α* as a poor estimate of reliability despite that it can be easy to calculate from just the item statistics and the total score. These problems are that, first, *α* underestimates reliability of a test in cases where there is a lack of *tau* equivalence and overestimates of the proportion of test variance if that is associated with the general variance in the test, and second, α does not measure internal consistency because it is a function of the number of items and the average correlation between the items, rather than a function of the unidimensionality of the test. Therefore, they provide practical guides to use a variety of coefficients, including McDonald's (1999) omega estimates, as powerful alternatives to Cronbach's alpha according to the appropriateness for certain purposes.

The choice of reliability coefficient is critical in the field of applied linguistics, though scarce empirical studies provide thought-provoking evidence. For example, O'Reilly & Marsden's (2020) use of ordinal omega as an alternative to Cronbach's alpha provides a clear illustration for omega being a superior reliability check than alpha when item responses are of the non-continuous nature and when *tau* equivalence is violated due to the variation in terms of items' relation to the construct. Considering the important psychometric property of the measures in the current research, such as the proportion of variance in scores associated with the general factor that the instrument deemed to measure, and the assumption of *tau* equivalence of items being violated, it would be reasonable to use omega estimators, based on factor analytic approaches, as alternative to Cronbach's alpha.

Specifically, the current research used the `omega` function in `psych` package (Revelle, 2022) in `R` to examine two omega coefficients: omega hierarchical (ω_h) estimated the reliability of the general factor of a test after controlling the variance of other factors (Green & Yang, 2015), and it was the coefficient index of the general factor loadings on the items with the exploratory Schmid-Leiman procedure (Zinbarg, et al., 2007). Omega total (ω_t) was the estimate of the total reliable variance of the instrument, i.e., the estimate of the total reliability of the test. The difference between these two omega coefficients is that the former is based upon the sum of squared loadings on the general factor, while the latter is based upon the sum of the squared loadings of all factors. In addition, Cronbach's alpha estimate with its 95% confidence intervals was also included in the report of the results to facilitate comprehensible comparison to the α coefficients reported in previous literature about the reliability of aptitude measurements.

4.3.2.2 Unidimensionality indices

Dimensions are known to be the constructed variables of psychometric measures. Most psycho-educational measures are not essentially unidimensional, i.e., the measures may not merely gauge the latent variable of interest, and secondary minor latent variables may be included in the measures (Slocum-Gori et al., 2009). As explained in the above section, omega hierarchical (ω_h) can be used as the estimate of the general factor that accounts for the data to a full extent. It is also recommended that researchers should always examine the pattern of the estimated general factor loadings prior to estimating ω_h, an informal assumption test to avoid possible misinterpretations of the estimate (Zinbarg et al., 2006). The assumption holds that if the factor loadings were salient for only a relatively small subset of the 'indicators' (i.e., the items of the subtest in the current research), this would suggest no true general factor underlying the data. Accordingly, unidimensionality checks should be an indispensable part of the results that inform our understanding of the reliability of instruments.

Although the unidimensionality check is important before examining reliability, there is no uniform or well-accepted objective index to represent the unidimensionality of a test (Slocum-Gori & Zumbo, 2010). The current research used `omega` and `unidim` functions in `psych` package to conduct exploratory analysis of item-response data *prior* to the check of reliability. Specifically, in the diagnostic statistical outputs of the `omega` function, the relative size of the eigen value of a general factor (*g*) compared to the other eigne values and the Explained Common Variance (*ECV*) were reported to indicate unidimensionality of the amount of test variance accounted for by a general factor (Revelle, 2022). To be more precise, *g* greater than 1 provided evidence that the measurement captured one factor with reliable commonality, i.e., *the eigenvalue-greater-than-one* rule (Mulaik, 1972). This rule was applied to interpret the unidimensionality evidence of the test. It suggests that any factor that accounts for more variability than a single observed variable should be considered as a dimension measured by the instrument. However, this rule has been criticised in the literature as performing poorly, especially for small sample sizes (see the discussion in Slocum-Gori & Zumbo, 2011). The eigenvalue for the first factor (*F1*) was also reported in the current research as it was suggested that this eigenvalue can be particularly important for external evaluation when the eigenvalue greater than 1 is used as the only retention criterion (Henson & Roberts, 2006). The `unidim` function also calculates several indices to assess whether the items of a subtest measured one latent trait. In the output of this function, *u* provides high values when the data are unidimensional. The product of *fa.fit* (that is, the fit of the one factor model to the correlations) is the measure of unidimensionality based on a simple logic: a one factor model of the data fits the covariances of the data if the data are unidimensional. Thus, the closer *fa.fit* to 1, the higher degree of unidimensionality is evidenced, and *fa.fit* would be 1 when the factor model is perfect (Revelle, 2022). It is worth noting that other approaches can be used to check unidimensionality of the data, e.g., using confirmatory one-factor model to check the goodness of fit of the data to the model (Flora, 2020). However, considering the exploratory purpose of the current research to diagnose the pattern of factor loadings of test items in each subtest as a measure for a latent construct, the current research did not apply confirmatory approaches to examine unidimensionality of the item level data.

4.3.2.3 The results

The results of reliability and unidimensionality of each subtest were obtained using `psych` package in `R`.

TALL_Vocabulary Learning

As shown in Table 4.9, although the eigen value of general factor (*g*) was greater than one in each dataset, the relative value of *g* compared to F1 in the dataset of Version A in the aural modality was the lowest of the four datasets, suggesting that it was possible the first factor could be accounting for the observed scores more than the general factor could. Other unidimensionality indices also showed that this dataset had the lowest unidimensionality of the four datasets in this subtest though its fa.fit value (= .60) indicated that the 60% covariances of the data fitted the one factor model, which could still be considered as the evidence of unidimensionality of this instrument version measuring a general latent variable.

In Table 4.9, the results supported, in general, the satisfactory reliability of the split datasets in TALL_Vocabulary Learning evidenced by omega_total (ω_t) and Cronbach’s alpha (α) coefficients. Omega hierarchical (ω_h) estimate of Version A in the aural modality was the lowest (.31) among all datasets of this subtest. Similarly, this dataset also had the lowest reliability coefficient in α (.70), which was lower than the acceptable threshold of .74 proposed by Plonsky and Derrick (2016). But its 95% confidence intervals of alpha estimate (CI = [.60, .79]) included the threshold value of .74.

Changes of the reliability coefficient of each dataset were checked if each item was dropped. The results showed that the incremental deletions of two items (that is, Item A1 of Version A in the aural modality and Item A1 of Version A in the written modality) would slightly increase the *α* coefficient by .01. These items were retained in further analysis because the goal was to keep as many items as possible, as long as their inclusion would maintain acceptable internal consistency. However, several items had factor loadings less than 0.2 to the general factors. Specifically, the following items needed further checks in IRT analysis: Items A1, A16, A17, A18, and A19 of Version A in the aural modality, Item B17 of Version B in the aural modality, Item A1 of Version A in the written modality, and Item B18 of Version B in the written modality.

Table 4.9 Reliability and unidimensionality of TALL_VL

Subtest	Dataset	n	k	Unidimensionality					Reliability				
												95% CI	
				g	F1	ECV	u	fa.fit	ω_h	ω_t	α	lower	upper
Vocabulary Learning	Version A in aural modality	81	20	1.70	2.00	.23	.29	.60	.31	.86	.70	.60	.79
	Version B in aural modality	84	20	3.70	1.60	.42	.49	.77	.56	.90	.78	.70	.84
	Version A in written modality	84	20	3.30	2.80	.36	.58	.82	.49	.91	.81	.75	.87
	Version B in written modality	81	20	4.40	1.80	.45	.69	.88	.58	.93	.84	.78	.88

Note. Key to column headings: n = number of participants (each did two of the tests, e.g., version A in aural modality and version B in written modality); k = number of test items; g = eigen value of the general factor; F1 = eigen value of the first factor other than the general factor; ECV = explained common variance; u = unidimensionality value; fa.fit = unidimensionality measure; ω_h= omega hierarchical; ω_t = omega total; α = Cronbach’s alpha; 95 % CI = 95% confidence intervals of Cronbach’s alpha with lower and upper bound

TALL_Sound Discrimination

As shown in Table 4.10, although *g* values of both datasets were greater than one, the dataset of Version B had the eigen value of the first factor (*F1* = 6.74) greater than the *g* value of the general factor, suggesting that it was possible the first factor could be accounting for the observed scores more than the general factor could. The results of other unidimensionality indices (i.e., *ECV*, *u* and *fa.fit*) provided similar evidence that the dataset of Version B had a slightly lower degree of unidimensionality compared to the dataset of Version A.

The results provided strong evidence of high reliability of the two instrument versions in TALL_Sound Discrimination, evidenced by ω_t and α coefficients. Both versions had α coefficients higher than the acceptable threshold of .74. ω_h estimates of the datasets also indicated satisfactory reliability of the general factor the subtest measured.

Changes of the reliability coefficients were checked if each item was dropped. The results showed that the deletion of any item would not increase the coefficient of the instrument of Version A, while the incremental deletions of two items (that is, Item B2B and B8B) of Version B would slightly increase the α coefficient by .01. In addition, no items of Version A had factor loadings less than 0.2 to the general factor, while seven items (i.e., B1B, B3A, B3C, B5C, B6C, B9B and B10B) had factor loadings less than 0.2 to the general factor, and hence these items needed further check in IRT analysis.

Table 4.10 Reliability and unidimensionality of TALL_SD

Subtest	Dataset	n	k	Unidimensionality					Reliability				
				g	F1	ECV	u	fa.fit	ω_h	ω_t	α	95% CI	
												lower	upper
Sound	Version A	165	30	7.40	2.20	.50	.79	.91	.62	.96	.89	.87	.92
Discrimination	Version B	165	30	6.18	6.74	.41	.48	.83	.72	.92	.78	.73	.83

Note. Key to column headings: n = number of participants; k = number of test items; g = eigen value of the general factor; F1 = eigen value of the first factor other than the general factor; ECV = explained common variance; u = unidimensionality value; fa.fit = unidimensionality measure; ω_h= omega hierarchical; ω_t = omega total; α = Cronbach's alpha; 95 % CI = 95% confidence intervals of Cronbach's alpha with lower and upper bounds

TALL_Language Analysis

As shown in Table 4.11, all indices provided clear evidence to support that all versions of this subtest had high degree of unidimensionality in terms of measuring a general latent variable.

In Table 4.11, ω_t and α coefficients provided strong evidence of high reliability of all instrument versions in TALL_Language Analysis. The α coefficients were much higher than the acceptable threshold of .74. ω_h estimates also indicated strong reliability of the general factor the subtest measured.

Changes of the reliability coefficient of each dataset were checked if each item was dropped. The results showed that the incremental dropping of only one item (TB26 of Version B in the written modality) would slightly increase the α coefficient of alpha by .01. This item was retained in further analysis because its inclusion would maintain acceptable internal consistency. Additionally, no items having factor loadings less than 0.2 to the general factors that needed further scrutiny in IRT analysis.

TALL_Serial Nonwords Recall

As shown in Table 4.12, all indices provided clear evidence to support that all versions of this subtest had high degree of unidimensionality in terms of measuring a general latent variable.

In Table 4.12, ω_t and α coefficients provided strong evidence of high reliability of all instrument versions in TALL_Serial Nonwords Recall. α coefficients were much higher than the acceptable threshold of .74. ω_h estimates also indicated satisfactory reliability of the general factor the subtest measured.

Changes of the reliability coefficients were checked if each item was dropped. The results showed that the incremental deletion of only one item (that is, AT1) would slightly increase the α coefficient of Version A by .01. This item was retained in further analysis because its inclusion would maintain acceptable internal consistency. In addition, only Item AT1 of Version A had factor loadings less than 0.2 to the general factor, and so only this item needed further checking in IRT analysis.

Table 4.11 Reliability and unidimensionality of TALL_LA

Subtest	Dataset	n	k	Unidimensionality					Reliability				
				g	F1	ECV	u	fa.fit	ω_h	ω_t	α	95% CI	
												lower	upper
Language Analysis	Version A in aural modality	81	30	14.70	1.70	.72	.94	.98	.80	.98	.95	.94	.97
	Version B in aural modality	84	30	12.00	.00	.71	.75	.93	.84	.97	.91	.88	.94
	Version A in written modality	84	30	14.30	5.20	.63	.91	.97	.74	.99	.96	.94	.97
	Version B in written modality	81	30	14.81	.04	.73	.92	.97	.82	.98	.94	.93	.96

Note. Key to column headings: n = number of participants (each did two of the tests, e.g., version A in aural modality and version B in written modality); k = number of test items; g = eigen value of the general factor; F1 = eigen value of the first factor other than the general factor; ECV = explained common variance; u = unidimensionality value; fa.fit = unidimensionality measure; ω_h= omega hierarchical; ω_t = omega total; α = Cronbach's alpha; 95 % CI = 95% confidence intervals of Cronbach's alpha with lower and upper bounds

Table 4.12 Reliability and unidimensionality of TALL_SNWR

Subtest	Dataset	n	k	Unidimensionality					Reliability				
				g	F1	ECV	u	fa.fit	ω_h	ω_t	α	95% CI	
												lower	upper
Serial Nonwords Recall	Version A	165	17	4.43	1.60	.66	.80	.95	.72	.89	.86	.82	.89
	Version B	165	17	5.04	1.10	.69	.84	.96	.76	.91	.88	.85	.90

Note. Key to column headings: n = number of participants; l = number of test items; g = eigen value of the general factor; F1 = eigen value of the first factor other than the general factor; ECV = explained common variance; u = unidimensionality value; fa.fit = unidimensionality measure; ω_h= omega hierarchical; ω_t = omega total; α = Cronbach's alpha; 95 % CI = 95% confidence intervals of Cronbach's alpha with lower and upper bounds

TALL_Complex Span Tasks

As shown in Table 4.13, *g* values of all datasets were greater than one. However, the dataset of Version B in the written modality had the *g* value of the first factor (*F1* = 1.34) greater than the *g* value, suggesting that the first factor could be accounting for the observed scores more than the general factor could. In addition, the relative sizes of *g* values compared to the *F1* values in two datasets of the written modality were lower than those of the aural modality, suggesting lower degree of unidimensionality of the instruments when they were administered in the written modality. Given that *fa.fit* values of the datasets in the written modality (.69 of Version A and .80 of Version B) indicated that at least 69% covariances of the datasets fitted the one factor model, unidimensionality of these instrument versions measuring a general latent variable was still sufficiently convincing.

Table 4.13 provides evidence, in ω_t and α being higher than the acceptable threshold of .74*,* of acceptable reliability of the datasets in the aural modality of this subtest. ω_h estimates of the datasets in the aural modality were higher than those in the written modality. Although α coefficients of the two datasets in the written modality (.64 of Version A and .72 of Version B) were lower than the acceptable threshold of instrument reliability coefficient (α = .74), the 95% confidence intervals of the coefficients still included the threshold value.

Changes of the reliability coefficient of each dataset were checked if each item was dropped. The results showed that the incremental deletion of only one item (Item B_T3 of Version B) would slightly increase the α coefficients of alpha by .01. This item was retained in further analysis because its inclusion would maintain acceptable internal consistency. However, several items had factor loadings less than 0.2 to the general factors. Specifically, the following items needed further checks in IRT analysis: Item A_T4 of Version A in the aural modality, Items A_T6, A_T7, A_T11, A_T13, and A_T15 of Version A in the written modality, and Items B_T3 and B_T6 of Version B in the written modality.

Table 4.13 Reliability and unidimensionality of TALL_CST

Subtest	Dataset	n	k	Unidimensionality					Reliability				
												95% CI	
				g	F1	ECV	u	fa.fit	ω_h	ω_t	α	lower	upper
Complex Span Tasks	Version A in aural modality	81	15	3.10	1.20	.53	.69	.88	.64	.87	.84	.78	.88
	Version B in aural modality	84	15	2.59	1.21	.46	.73	.89	.56	.86	.82	.76	.87
	Version A in written modality	84	15	1.09	0.64	.30	.36	.69	.35	.68	.64	.51	.74
	Version B in written modality	81	15	1.17	1.34	.26	.60	.80	.33	.79	.72	.62	.81

Note. Key to column headings: n = number of participants (each did two of the tests, e.g., version A in aural modality and version B in written modality); k = number of test items; g = eigen value of the general factor; F1 = eigen value of the first factor other than the general factor; ECV = explained common variance; u = unidimensionality value; fa.fit = unidimensionality measure; ω_h= omega hierarchical; ω_t = omega total; α = Cronbach’s alpha; 95 % CI = 95% confidence intervals of Cronbach’s alpha with lower and upper bounds

4.4 Item level analysis

This section mainly addresses the check of item quality of the data split by the conditions of material versions and/or modalities in each subtest using IRT models. The following sections introduces IRT models that were suitable to the characteristics of the data of the subtests. The results and summary plots will be presented for each subtest.

4.4.1 Item Response Theory (IRT) models

In the current research, two types of data at the item level were generated in five subtests: dichotomous data from TALL_VL, TALL_SD, and TALL_LA, and the polytomous data from TALL_SNWR and TALL_CST. The following section introduces different IRT models that were applied to these two types of data.

4.4.1.1 IRT models for dichotomous data (TALL_VL, TALL_SD, and TALL_LA)

Rasch modelling was used in this analysis, although concerns about the appropriateness of using Rasch analyses in language testing and assessment research have been raised (see Appendix F for a summary of the debate about the use of Rasch).

1PL, 2PL, and 3PL models

Basic Rasch model (also known as One-Parameter Logistic model, 1PL), Two-Parameter Logistic (2PL), and Three-Parameter Logistic (3PL) model were applied on the dichotomous data. The purpose of using different IRT models was to examine if adding the parameter of discrimination in the 2PL model and two parameters, i.e., discrimination and guessing, in the 3PL model, the data would have a better model fit than only having a parameter of item difficulty in the 1PL model. To achieve this purpose, `ltm` package (Rizopoulos, 2006) in `R` was used. The current research also used `eRm` package (Mair & Hatzinger, 2007) to provide supplementary analysis if 1PL model was evidenced as the model with the best fit to the data from the analyses through `ltm` package (see Appendix F for explaining reasons for using two packages).

Stepwise analyses

As introduced above, analyses were performed in the following steps, using IRT models on the dichotomous data in TALL_VL, TALL_SD, and TALL_LA, respectively. More detailed explanations on the stepwise analyses can be found in Appendix F.

Step 1. Model comparisons

In the first step of the analysis, Rasch models were created using `rasch` function in the `ltm` package. Initially, a basic Rasch model with equal discrimination parameters was generated and compared to an unconstrained Rasch model. The comparison involved a likelihood ratio test (through the `anova` function), including the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC) and the *p*-value of the log-likelihood ratio to assess model fitness. Lower AIC and BIC values indicated better fit.

Subsequently, more complex 2PL and 3PL models were explored, and their performance was compared to the initially identified Rasch model. Model comparisons were used to determine if adding discrimination and guessing parameters were needed. If the 2PL or 3PL models were found to have a better fit, further analyses were conducted in the `ltm` package. However, if the Rasch model was superior, supplementary estimations were made using the `RM` function in the `eRm` package to assess item and person fitness. It's worth noting that no comparisons between the 2PL and 3PL models across different `R` packages were made because the `eRm` package does not support the estimations of the 2PL and 3PL models.

Step 2. Model Fitness

In this step, model fitness was assessed to ensure the quality of the chosen model. Fit statistics were obtained to determine if the Rasch model met the assumption of unidimensionality, identifying items that deviated from the expected pattern. A non-significant *p*-value (>.05) from the `GoF.rach` function in the `ltm` package and the `LRtest` function in the `eRm` package suggested acceptable model fit for the Rasch model (Mair & Hatzinger, 2007; Rizopoulos, 2006). For the 2PL or 3PL models, the `margins` function examined fitness through two-way χ^2 residuals, with a rule of thumb of 3.5 as an indicator of goodness of fit.

Step 3. Model estimations and item/person fitness

In this step, the chosen model with the best fit was analysed using various functions in the two `R` packages. Descriptive statistics were obtained using `summary` function in both packages, including coefficients of difficulty (and discrimination for 2PL and 3PL models) and standard errors for all items.

In the ltm package, `item.fit` and `person.fit` functions can be used to compute item and person fit statistics for 1PL, 2PL and 3PL models. The χ^2 (displayed as X^2) statistic tested the null hypothesis that the item responses follow the chosen model against the alternative hypothesis that they did not. A large χ^2 value indicated poor fit of the item to the model, while a small χ^2 value indicated good fit.

For the Rasch model, the `itemfit` and `personfit` functions in the `eRm` package were used, which produced outputs easier to interpret. Several cut-off values of the item and person model fit were applied as the criteria to ascertain the adherence of the items and participants to the model's expectations. The value of mean square (MNSQ) fit statistics is expected to be close to 1.0 if the item or person fits well to the model. The current research took the reference of the acceptable ranges of MNSQ values [0.50, 1.50] (Wright & Linacre, 1994), and [0.70, 1.30] with the infit *t* statistics ranging of [-2, 2] (Bond & Fox, 2015). If the MNSQ values of any items and persons were greater than 1.50, further examinations were conducted to investigate the reasons of the misfitting.

Step 4. Plots of model estimations

Plots of IRT model estimations provides visualisations of the items and persons characteristics and the test information.

Items or Persons Pathway Maps visually identify misfitting items or persons outside the range of satisfactory infit t-statistic between -2 and +2. Person-item maps display item and person parameters along the latent dimension, ideally showing a spread across the entire range with adequate test takers.

The slopes of Item Characteristic Curve (*ICC*) indicate discrimination ability of items, with steep slopes reflecting effective discrimination. A positive slope signifies the item's effectiveness in measuring the construct and distinguishing levels of ability between test takers.

Item Information Curve (*IIC*) reveals how much information about the latent ability an item provided at each level of the latent ability. Similarly, Test Information Curve (*TIC*) shows how much information about the latent ability the instrument provided at each level of ability, evaluating the test's appropriateness and difficulty level for participants.

4.4.1.2 IRT models for polytomous data (TALL_SNWR and TALL_CST)

In this part of analysis, the term "item" was used interchangeably to "trial", both referring to the sets of stimuli to be retained. The scoring method and datasets composition have been introduced in Section 3.3.5.1. The `ltm` package was used for the analyses in this section, as it applies Marginal Maximum Likelihood Estimation (MMLE) to estimate data based on the assumption that individual person parameters conform to a specific distribution (Nicklin & Vitta, 2022). Generalised Partial Credit model (GPCM) (Muraki, 1997) in the this package was used to examine the model fit of the polytomous data from TALL_SNWR and TALL_CST.

Stepwise analyses

Analyses were performed in the following steps, applying the GPCM on the polytomous data sets split by material versions in TALL_SNWR and by versions and modalities in TALL_CST. More detailed explanations on the stepwise analyses can be found in Appendix F.

Step 1. Model comparisons

To build the GPCM, the `gpcm` function was used in the `ltm` package that allows three constraint options, i.e., gpcm option assuming each trial having an estimated discrimination parameter, 1PL option assuming the discrimination parameter being equal for all trials, and Rasch option assuming the equal discrimination parameter being fixed at one. All three options were applied to the split data sets separately, and the models were then compared with each other to ensure that the chosen model had the best fit to each dataset.

Step 2. Model Fitness

Fitness statistics of the chosen model to the data were obtained in this step. For the Rasch model, the `GoF.gpcm` function was used to perform a parametric bootstrap goodness-of-fit test using χ^2 statistic. Based on 50 iterated datasets, the non-significant *p*-value > .05 would suggest an acceptable fit of the model (Rizopoulos, 2006).

Step 3. Model estimations

The descriptive statistics of the chosen model with better fitness were analysed by the `summary` function, providing coefficients of the category threshold parameters and the discrimination parameter. The category threshold parameters represented the points on the latent trait scale that determine when the test takers were equally likely to endorse one answer option versus the next. The lower values of the category threshold parameters indicated that the item was more difficult to endorse, and the higher values indicated that the item was easier to endorse. The discrimination parameter provided information about the how well the item distinguished between individuals with different level of the latent ability. In the outputs, the z-value for each coefficient was also displayed, which was obtained by the coefficient divided by the standard error and indicated whether the coefficient was statistically significant. As a rule of thumb, a z-value with the absolute value greater than 1.96, indicating the statistical significance of the coefficient at the 5% level, suggested that the parameter was unlikely to have arisen by chance.

Step 4. Plots of model estimations

Plots of IRT model estimations provides visualisations (the ICC, the IIC, and the TIC) of the items' characteristics and the test information, as introduced in the previous section about the stepwise analysis of dichotomous data.

4.4.2 The results

The results of IRT analyses on the data of the subtests are presented separately. The analysis code and output are documented in `R` markdown files (see Appendix D) separately for dichotomous data from TALL_VL, TALL_SD, and TALL_LA, and for polytomous data from TALL_SNWR and TALL_CST.

4.4.2.1 Item quality in TALL_VL

Model comparisons

The results of the AIC, BIC and *p*-value of likelihood ratio tests for the 1PL, 2PL and 3PL models suggested that the inclusions of different discrimination parameters per item in the 2PL model and a guessing parameter in the 3PL model did not increase the model fitness to all datasets in this subtest. Therefore, the 1PL (Rasch) model was chosen to provide model estimations using the `ltm` package and the `eRm` package.

Model fitness

The results of the parametric bootstrap goodness-of-fit test using χ^2 statistic suggested that all datasets had acceptable fit of the chosen models, with *p*-values > .05. Similarly, the Andersen likelihood ratio tests also returned *p*-values >.05, indicating that all datasets fitted well to the Rasch models.

Model estimations and item/person fitness

The results of item fitness statistics of all datasets showed that no items had associated χ^2 *p*-value < .05, suggesting no items had poor fit to the model.

Additional evidence of item fitness was obtained through MNSQ fit statistics and infit *t* statistics. Although three items (see Figure 4.7) fell outside the range of [-2, +2] for infit *t* statistics, none of them displayed a poor model fit, as all had Infit MNSQ values lower than 1.5. This suggested that none of the items in this subtest required removal due to poor model fit.

The person fit statistics revealed that two participant (refer to Figure 4.8) had infit *t* values fell outside the range of [-2, +2]. However, their Infit MNSQ values were below 1.5, indicating acceptable person fit statistics. Thus, there was no evidence from the person fit statistics to suggest that any individual estimations poorly fit the Rasch models.

Plots of model estimations

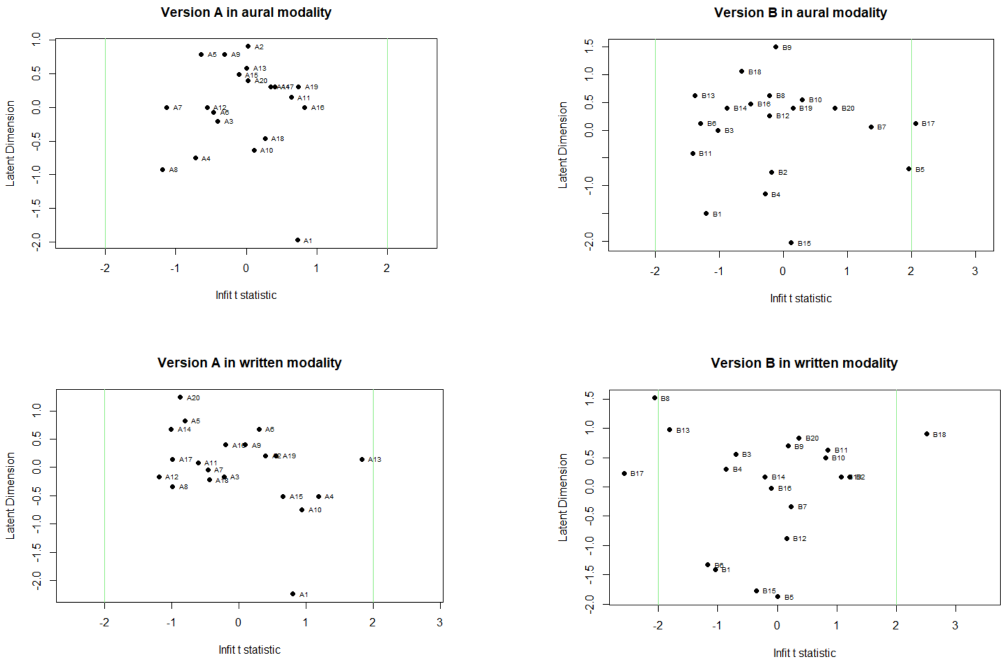


Figure 4.7 Items Pathway Maps for TALL_VL

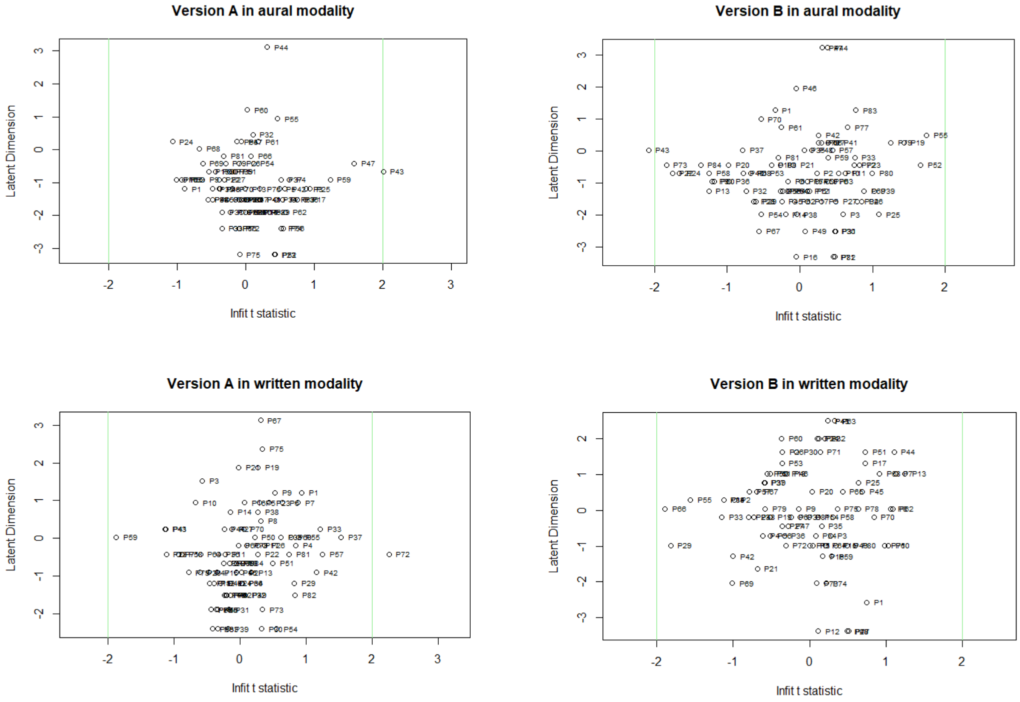


Figure 4.8 Persons Pathway Maps for TALL_VL

The person-item maps (Figure 4.9) showed that, in general, the items were well-targeted to the ability levels of the participants without any item that located at the extreme ends of the map. Item A1(alk, *lion*) in both aural and written modalities displayed abrupt transitions to the items next to it, which suggested that this item was much less challenging to the participants. The maps also showed that the participants' ability levels were not evenly distributed (see the upper bar for Person Parameter Distribution). Most of the participants who took Version B in the aural modality, or Version A or B in the written modality roughly located in the middle of the map. The exception was the participants who took Version A in the aural modality, with the majority locating on the lower dimension of ability between -1 and -2.

The ICCs plotted (Figure 4.10) showed that all items in both modalities had strong positive slopes, indicating that the items were measuring the unidimensional construct. The curves also showed that, in general, all items had a range of difficulty levels to discriminate individuals' levels of the vocabulary learning ability, and they were similarly effective to measure the construct.

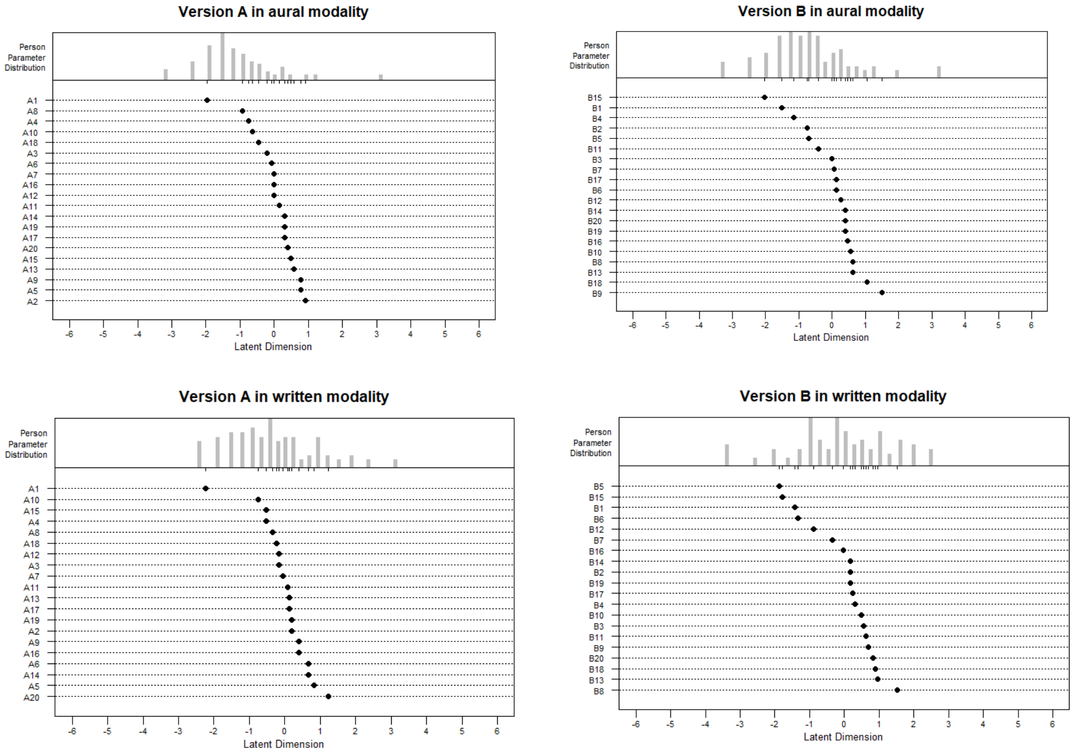


Figure 4.9 Persons-Item Maps for TALL_VL

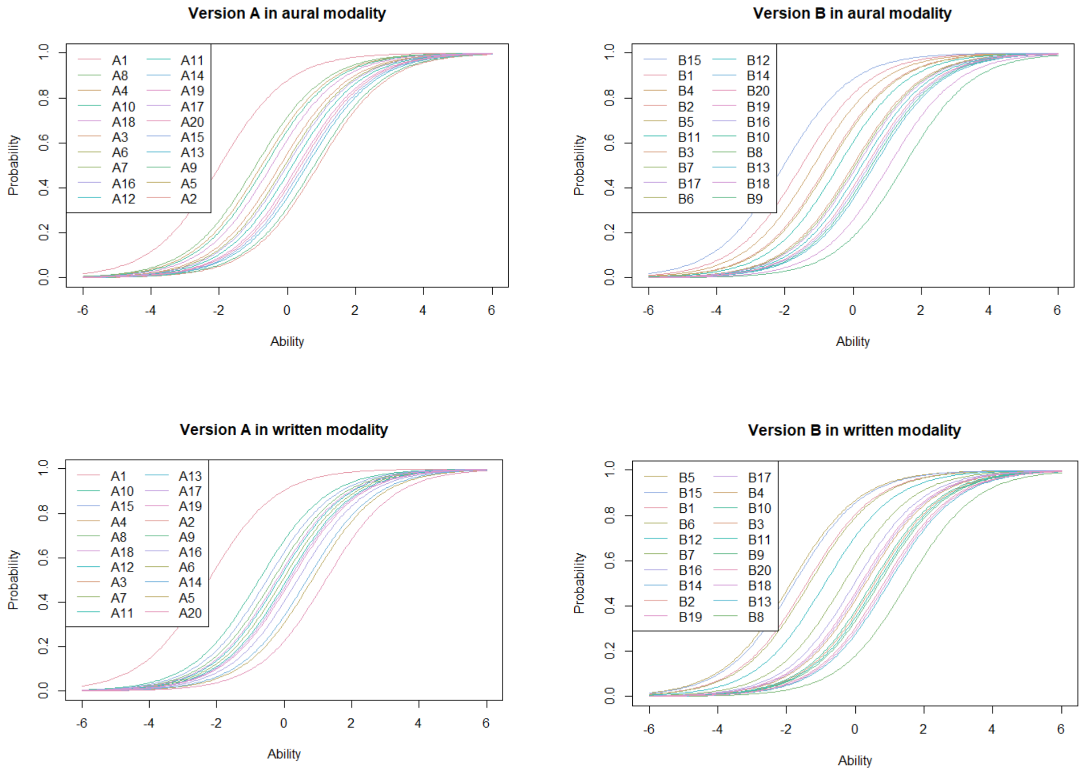


Figure 4.10 Item Characteristic Curves for TALL_VL

The information provided by the IICs (Figure 4.11) showed that, compared to the items of Version A, items of Version B had better quality in terms of providing information evenly along the scales of ability. However, one item (Item A1) of Version A, however, was too easy compared to other items of the same version.

Finally, the TICs and the statistics in Table 4.14 indicated that all instruments were appropriate to evaluate participants' ability, with consistent information provided about the latent ability scales. To be specific, instruments administered in the written modality had more evenly distributed information about the latent ability, especially with Version B providing nearly equal information (48.96% and 50.31%) about participants who had negative (poor) and positive (strong) ability along the scale (see the nearly symmetric curve spread across the x axis). Instruments administered in the aural modality provided much more information about the higher-level ability, suggesting that the instruments were more challenging comparing to the same instruments but administered in the written modality.

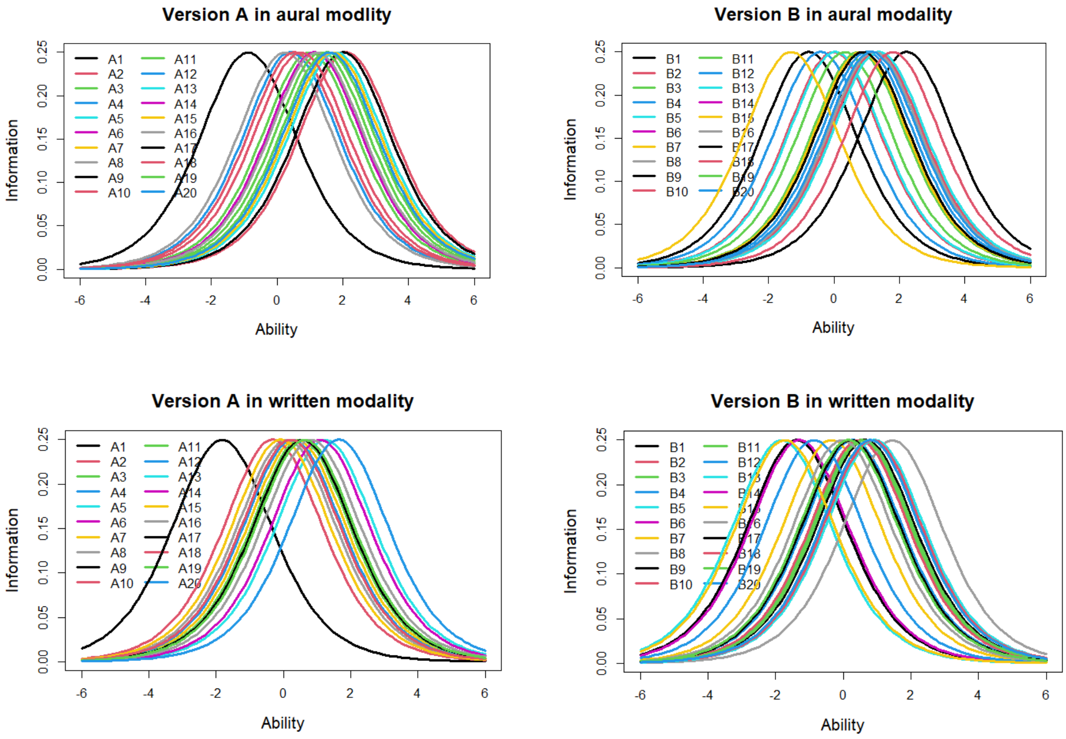
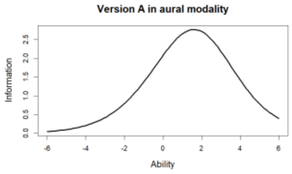
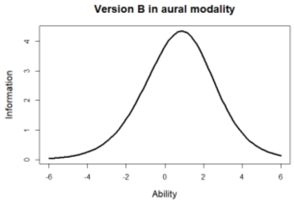
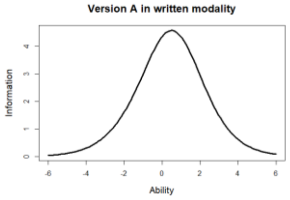
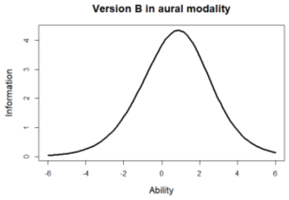


Figure 4.11 Item Information Curves for TALL_VL

Table 4.14 Test Information for TALL_VL

Modality	**Version**	**Total Information**	**Information (-6, 0)**	**Information (0, 6)**	**Test Information Curve**
aural	A	15.54	3.9 (25.09%)	11.07 (71.24%)	
	B	20	6.74 (33.68%)	13.09 (65.46%)	
written	A	20	7.9 (39.49%)	11.96 (59.83%)	
	B	20	9.79 (48.96%)	10.06 (50.31)	

4.4.2.2 Item quality in TALL_SD

Model comparisons

1PL, 2PL models were generated to the data of the subtest of Sound Discrimination obtained by using Version A and B separately, but 3PL models with a guessing parameter added could not converge to the datasets. The results of the AIC, BIC and *p*-value of likelihood ratio tests for the 1PL and 2PL models suggested that the inclusions of different discrimination parameters per item in the 2PL models did not increase the model fitness to the datasets from this subtest. Therefore, the 1PL (Rasch) model was chosen to provide model estimations using the `ltm` package and the eRm package.

Model fitness

The results of the parametric bootstrap goodness-of-fit test using χ^2 statistic suggested that datasets obtained by using Version A had acceptable fit to the Rasch model, with the *p*-value (.09) > .05. Similarly, the Andersen likelihood ratio tests also returned non statistically significant *p*-value (.76) >.05 to indicate that the data fitted well to the Rasch model. However, the results of model fitness check did not suggest that the data obtained by using Version B had an acceptable fit to the Rasch model, with the *p*-value (.01) < .05 of the parametric bootstrap goodness-of-fit test and the *p*-value (.024) < .05 of the Andersen's likelihood ratio test.

Model estimations and item/person fitness

The results of item fitness statistics of two datasets of both versions showed that none of the items in this subtest exhibited a poor model fit (having associated χ^2 *p*-value <.05) that would necessitate their removal. As displayed in Figure 4.12, items in Version A fitted well to the the Rasch model, evidenced by the infit *t* statistics. These items also had values of MNSQ fit statistics below 1.5. In Version B, only one item, Item B2B (**sėja** su vėju), had a large value of infit *t* statistics (4.039) out of the acceptable range [-2, +2], suggesting *overfit* of the item to the model. However, this item had acceptable Outfit and Infit MNSQ fit statistics less than the cut off value of 1.5. Given that MNSQ statistics are more relevant to the impact of *underfit* on the test (Bond et al., 2020; Nicklin & Vitta, 2022), this item should not be diagnosed as problematic item that could potentially threat the validity of the measurement.

As shown in Figure 4.13, when participants took Version A, two persons had infit *t* statistics beyond the acceptable range, but their MNSQ fit statistics remained below 1.5, indicating no misfit. In contrast, when participants took Version B, several individuals had infit *t* statistics falling outside [-2, +2]. Among these, four participants exhibited MNSQ statistics beyond the acceptable threshold of 1.5, suggesting misfit to the Rasch model. Given that the current analysis is exploratory, focusing on checking the internal validity of a new instrument, the data of these four participants were not excluded from further analyses. The results that more participants had exhibited mist fit in Version B than in Version A provided insights into the comparison of the two versions, which will be discussed in Section 4.7.2.1.

Plots of model estimations

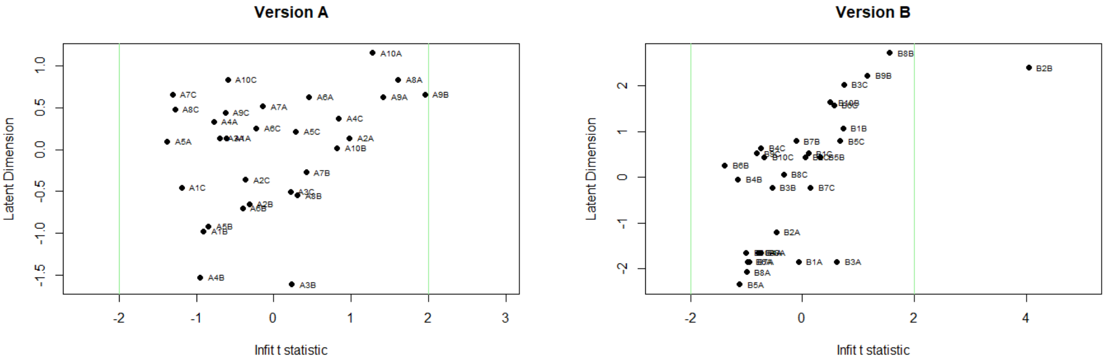


Figure 4.12 Items Pathway Maps for TALL_SD

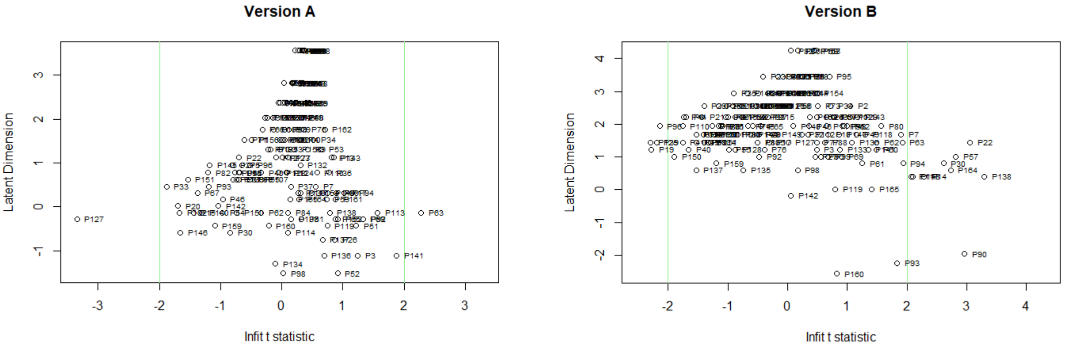


Figure 4.13 Persons Pathway Maps for TALL_SD

The person-item maps (Figure 4.14) showed that, in general, the items were well-targeted to the ability levels of the participants without any item that located at the extreme ends of the map. However, some differences between two material versions were evidenced. Version A yielded more evenly distributed items of difficulty parameters, with all items having smooth transitions in-between. Items of Version B, however, had abrupt gaps between the items containing the sound **sauja** and the items containing the sounds of **sėja** and **sija**, indicating that the sound **sauja** was systematically less challenging than the other two sounds.

The maps also revealed uneven distribution of participants' ability levels in two versions (see the upper bar for Person Parameter Distribution). In version A, most participants clustered in the positive half of the latent dimension, indicating better performance and stronger sound discrimination ability as measured by Version A compared to Version B. Conversely, participants who took Version B showed a wider spread along the latent dimension than Version A. The majority still resided in the positive half, signifying stronger abilities in Version B than in Version A. However, a minority of participants positioned around the lower end of ability scales (around -2), reflecting a preference for the less challenging items of Version B compared to Version A.

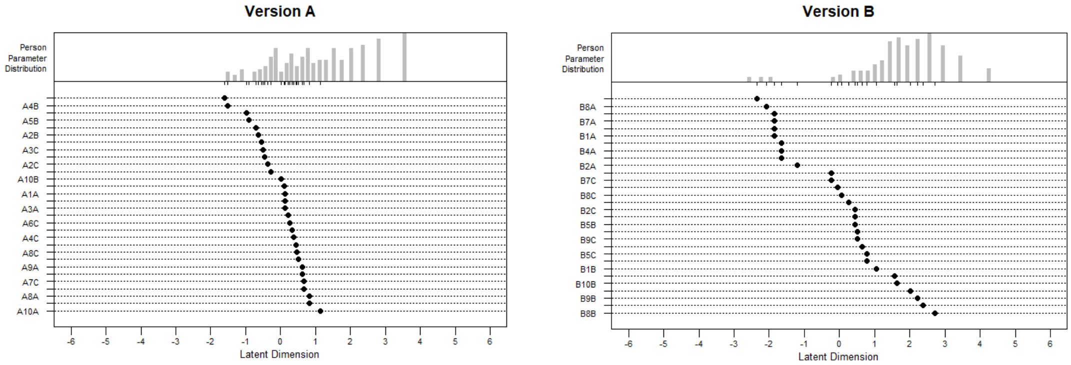


Figure 4.14 Persons-Item Maps for TALL_SD

The ICCs plotted (see Figure 4.15) showed that all items in both modalities had strong positive slopes, indicating that the items were measuring the unidimensional construct. The curves also showed that, compared to the items of Version A, items of Version B had a wider range of difficulty levels to discriminate individuals' levels of the sound discrimination ability, and these two versions of materials were not equally effective to measure the same construct.

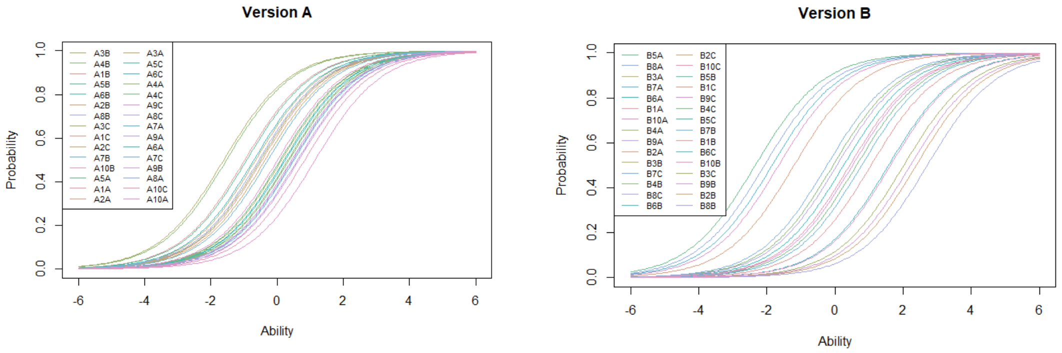


Figure 4.15 Item Characteristic Curves for TALL_SD

The information provided by the IICs (see Figure 4.16) revealed that, compared to items of Version B, items of Version A provided more information about the latent ability and demonstrated superior quality in terms of evenly distributing information along ability scales. In contrast, items of Version B, particularly those containing the sound ‘sauja’, were less challenging than other items within the same version. Consequently, these items of Version B provided limited information about higher levels of ability due to their low difficulty levels.

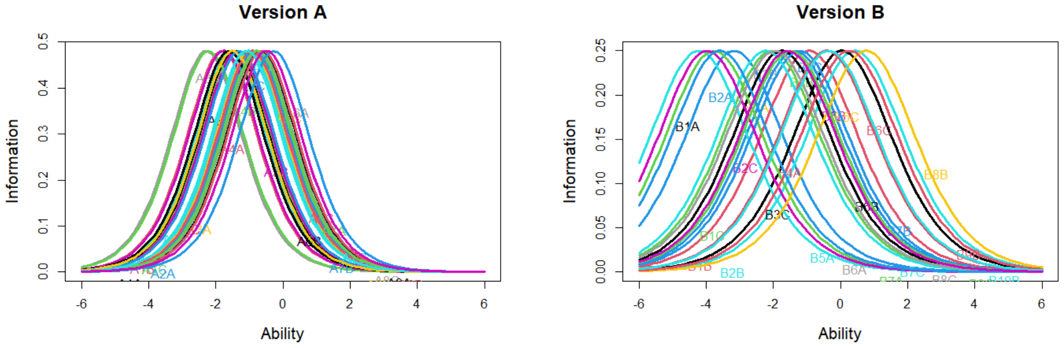
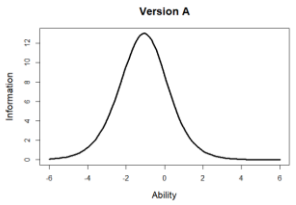
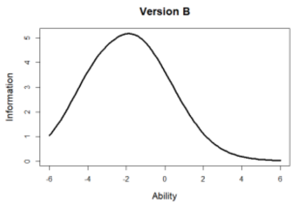


Figure 4.16 Item Information Curves for TALL_SD

Finally, the TICs and the statistics in Table 4.15 indicated that both versions were appropriate, in general, for evaluating participants’ ability, as they provided inconsistent amounts of information about the latent ability scales of [-6, +6]. Specifically, Version A provided more information about the latent scales than Version B did, although both versions provided more information for participants with low ability, that is, the amounts of test information for ability levels in the interval (-6, 0) were about 80% and 77%. This suggested that, for the participants in the current research, both versions were not challenging to measure their ability to discriminate the sounds.

Table 4.15 Test Information for TALL_SD

Version	**Total Information**	**Information (-6, 0)**	**Information (0, 6)**	**Test Information Curve**
A	41.6	33.25 (79.92%)	8.29 (19.93%)	
B	29.98	23.02 (76.78%)	5.81 (19.39%)	

4.4.2.3 Item quality in TALL_LA

Model comparisons

1PL, 2PL and 3PL models were applied to the data of Language Analysis from Version A and B, separately for aural and written modalities. For Version A in the aural modality and Version B in the written ability, AIC, BIC, and likelihood ratio test *p*-values favoured the 1PL model over the 2PL and 3PL models. However, for Version B in the aural modality and Version A in the written ability, the data did not support that adding extra parameters in 2PL or 3PL models improved model fit significantly compared to the 1PL model. The choice of the 1PL model was based on its parsimony.

Model fitness

The results of the parametric bootstrap goodness-of-fit test using χ^2 statistic and the Andersen likelihood ratio tests suggested that, in general, most datasets had acceptable fit of the chosen models, with *p*-values > .05. The only exception was the results from the parametric bootstrap goodness-of-fit test on the data of Version B in the written modality, which had a *p*-value = .01. However, the Andersen likelihood ratio test on this dataset had a *p*-value =.802, supporting an acceptable model fitness to the data.

Model estimations and item/person fitness

The results of the parametric bootstrap goodness-of-fit test using χ^2 statistic suggested that most items did not have associated χ^2 p-value < .05, indicating the poor fitness of the items to the model. The only exception was Item B26 (virejas viduje ruke, *a grandpa swam inside*) in the written modality, which might be an item having a poor model fit (*p* value =.04).

Additional evidence of item fitness was provided by MNSQ fit statistics and the infit *t* statistics. The results showed that although there were a few items (see Figure 4.17) having infit *t* statistics out of the range of [-2, +2], none of the items exhibited a poor model fit with Infit MNSQ values lower than 1.5. This suggested that they should not be removed as being threaten to the validity of the instruments.

Furthermore, the results of person fit statistics (see Figure 4.18) showed that six participants (four from Version A and two from Version B) in the written modality had infit *t* values beyond [-2, +2]. But they all had Infit MNSQ values lower than 1.5. Consequently, there was no evidence from the person fit statistics to suggest that any person estimations poorly fit the Rasch models.

Plots of model estimations

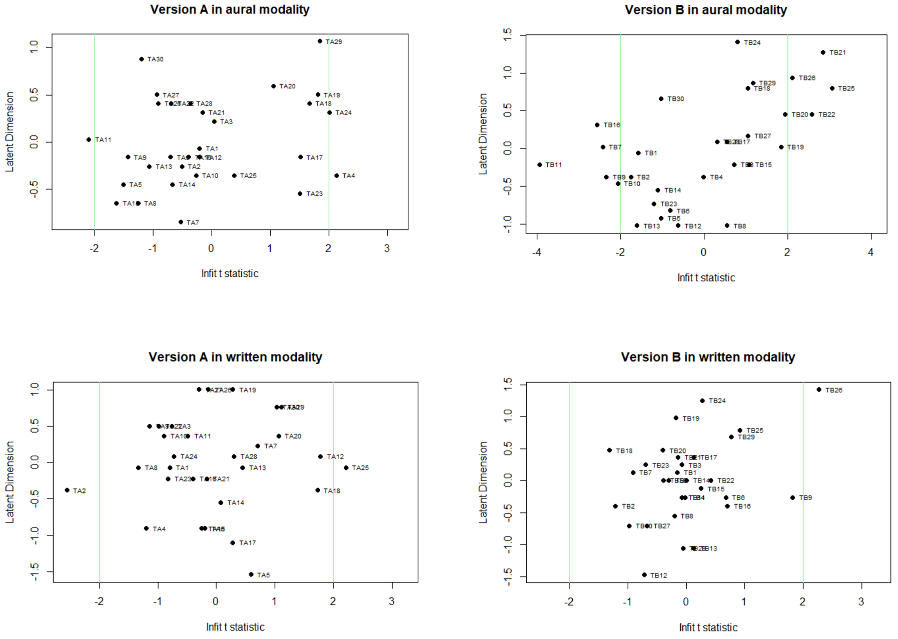


Figure 4.17 Items Pathway Maps for TALL_LA

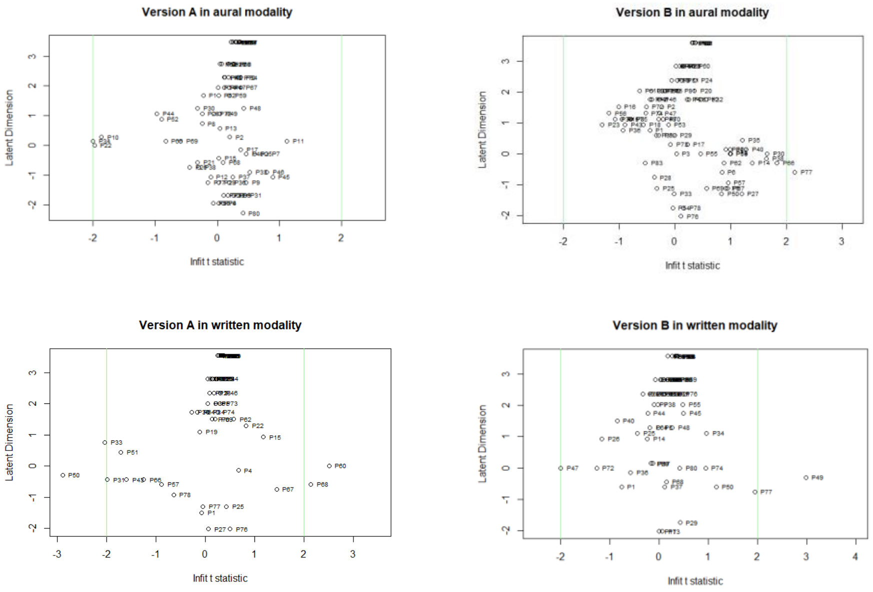


Figure 4.18 Persons Pathway Maps for TALL_LA

The person-item maps (Figure 4.19) showed that, in general, the items were well-targeted to the ability levels of the participants without any item that located at the extreme ends of the map. All items of both versions in two modalities yielded evenly distributed difficulty parameters as they had smooth transitions in-between. The maps also showed that the participants' ability levels were not evenly distributed (see the upper bar for Person Parameter Distribution), and most of them located on the positive half on the scales of ability dimension. The results suggested that generally the participants had strong ability of language analysis measured by the instruments. This was particularly reflected when the participants took the test in the written modality. Most of them located on the right end of the scales (3 to 4) on the maps, indicating clear ceiling effects when they were tested in the written modality.

The ICCs plotted (Figure 4.20) showed that all items in both modalities had strong positive slopes, indicating that the items were measuring the unidimensional construct of language analytic ability. The symmetry of the curves also showed that, in general, all items had a range of difficulty levels to discriminate individuals' levels of the latent ability, and they were similarly effective to measure the construct.

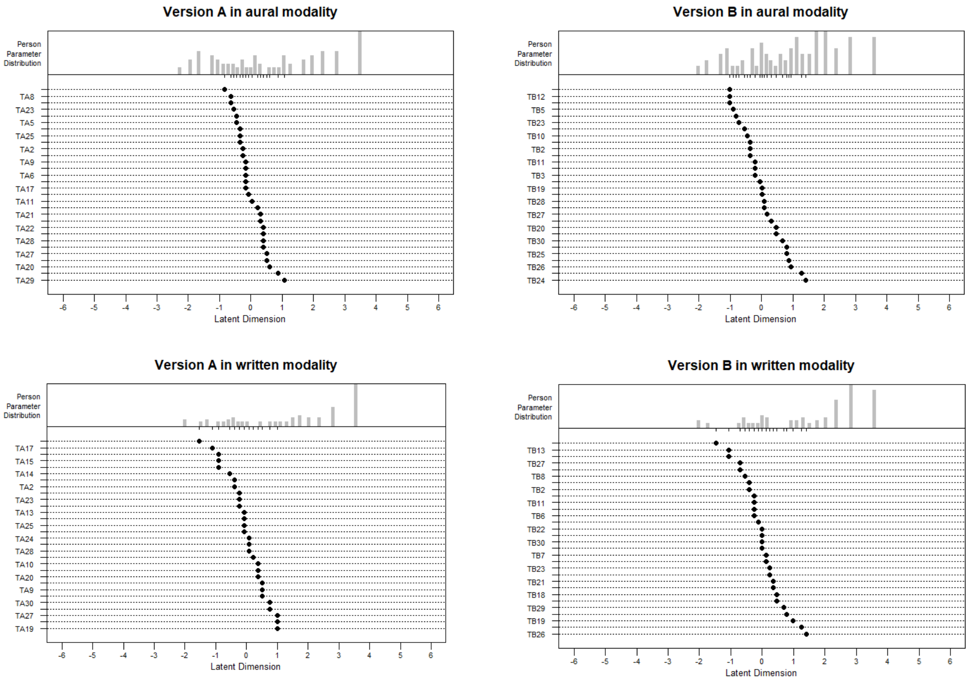


Figure 4.19 Persons-Item Maps Maps for TALL_LA

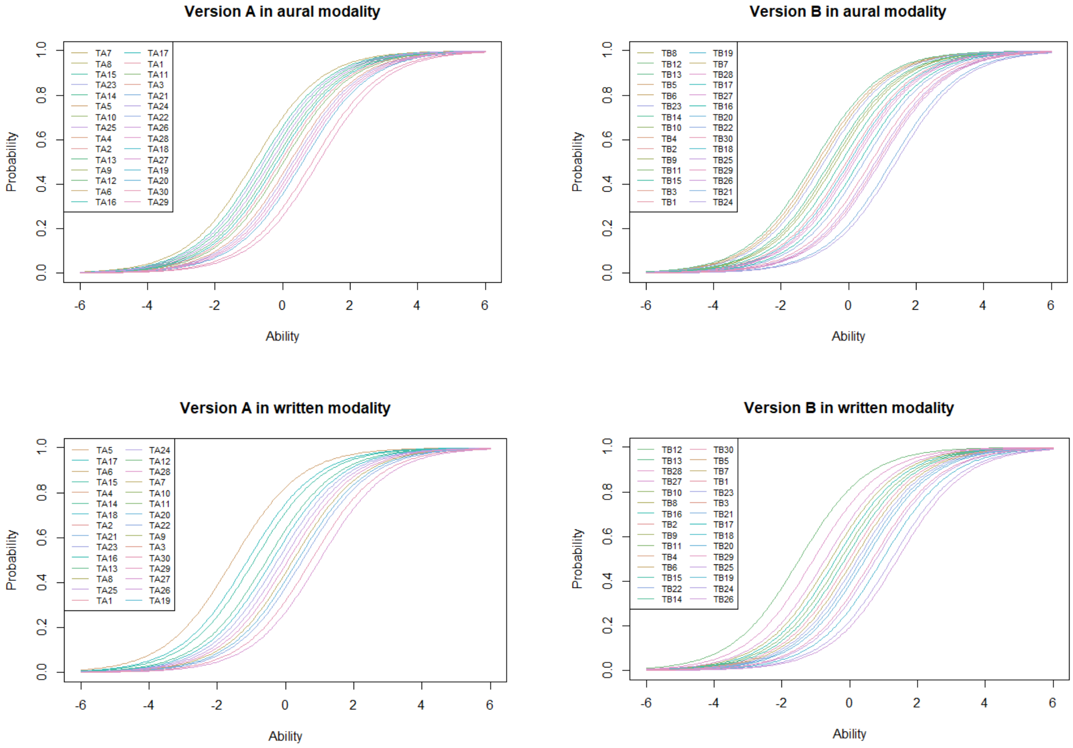


Figure 4.20 Item Characteristic Curves for TALL_LA

The information provided by the IICs (Figure 4.21) indicated that, compared to the items of Version B, items of Version A provided more information along the scales of ability. Most items provided more information about lower-level ability than higher-level ability.

Finally, the TICs and the statistics in Table 4.16 indicated that, in general, all instruments were appropriate to evaluate participants' ability. However, Version A in both modalities provided more information about the ability compared to Version B. All instruments did not provide evenly distributed information about the latent ability, with much more, from 72% to 90%, test information provided for lower ability levels in the interval (-6, 0). This pattern was particularly reflected in the written modality, which suggested that all instruments were not very challenging, especially when they were administered in the written modality, to measure the language analytic ability.

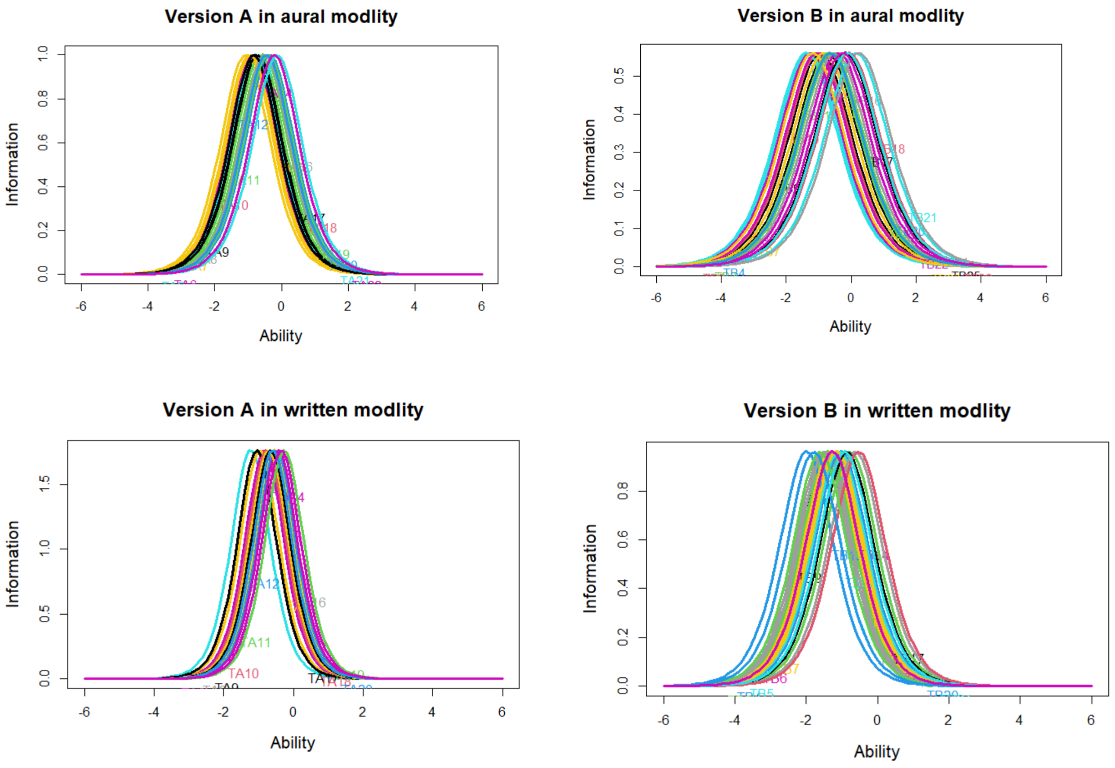
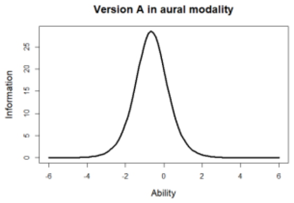
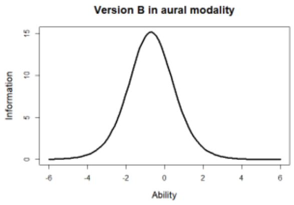
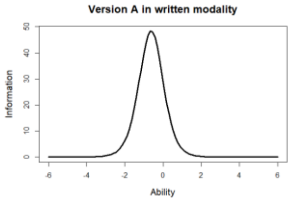
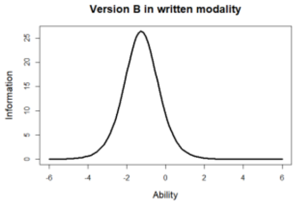


Figure 4.21 Item Information Curves for TALL_LA

Table 4.16 Test Information for TALL_LA

Modality	Version	Total Information	Information (-6, 0)	Information (0, 6)	Test Information Curve
aural	A	60	46.62 (77.69%)	13.39 (22.31%)	
	B	44.95	32.39 (72.07%)	12.53 (27.88%)	
written	A	79.72	65.8 (82.54%)	13.92 (17.46%)	
	B	58.82	53.48 (90.91%)	5.34 (9.08%)	

4.4.2.4 Item quality in TALL_SNWR

Model comparisons

GPCM was applied to data from Version A and Version B of the subtest, using three constraint options. The “gpcm” constraint option failed to converge with the data from both versions. AIC, BIC and likelihood ratio test *p*-values indicated that the GPCM with the the “1PL” constraint option provided a better fit for both versions. Consequently, the GPCM with the “1PL” constraint, assuming equal discrimination parameters for all test items (trials), was selected for model estimations using the `ltm` package.

Model fitness

The results of the parametric bootstrap goodness-of-fit test using the χ^2 statistic showed that *p*-values for both versions of the data were .02, which is smaller than the predefined criterion of .05. These results suggested that, assuming the null hypothesis (i.e., the hypothesis that the observed data follows the expected data distribution), there is a probability of 2% or less of obtaining data as extreme as what was observed. In other words, the data from this subtest did not have a good overall model fit to the chosen model.

Model estimations and item parameters

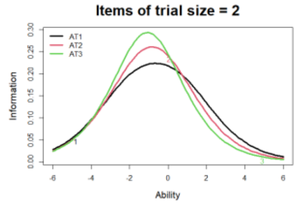
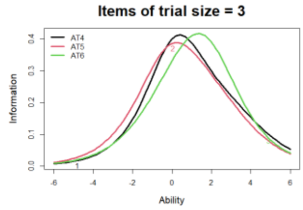
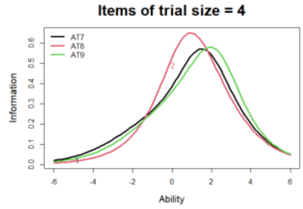
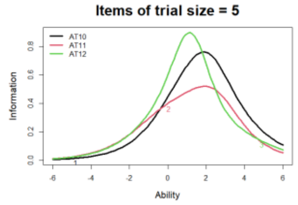
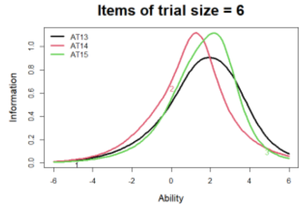
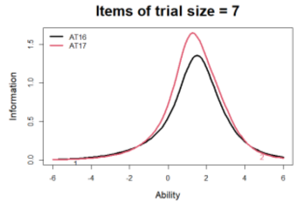
The chosen GPCM constrained the discrimination parameters to be equal for all items, and the coefficients of discrimination parameters were similar for data from both versions. The z-values of the coefficients for both versions were greater than the 1.96, corresponding to a two-tailed *p*-value of .05, indicating non-statistical significance at the 5% level. These results suggested that the discrimination parameters estimated by the GPCM for data from both versions were similarly reliable.

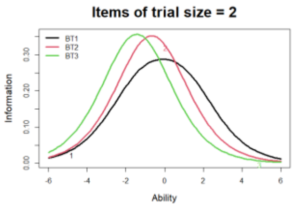
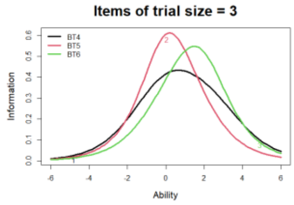
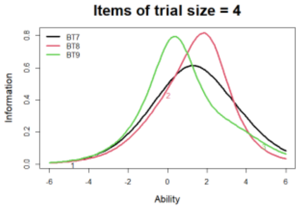
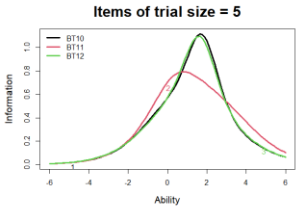
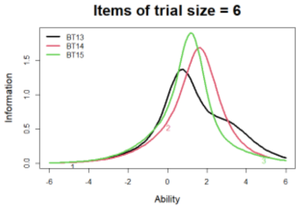
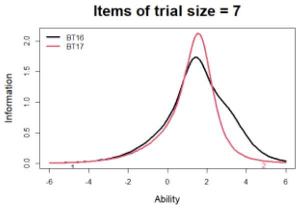
The coefficients of the category threshold parameters and the corresponding z-values generally indicated reliable estimations of difficulty parameters for data from both versions. Fewer than 25% of the total coefficients were below 0, indicating difficulty in endorsing higher category thresholds. Approximately half of these coefficients had z-values greater than 1.96, signifying reliable estimations. However, fewer than 27% of all coefficients had statistically significant z-values (that is, absolute values < 1.96), raising doubts about the reliability of model estimations for difficulty parameters. Items with the smallest set sizes (i.e., 2 stimuli to retain) were the easiest to remember, while items with larger set sizes appeared more challenging. Exceptions occurred in items with larger set sizes (e.g., 6 or 7 items in one trial), where the last category threshold coefficients were lower than the earlier ones, which may be due to testing errors.

Plots of model estimations

As shown in Table 4.17, IICs were plotted for trials with the same set size (number of stimuli in one trial) in both versions. Additionally, information for trials with the same set size was calculated. The plots revealed that trials with larger set size offered greater information than those with smaller set size. Furthermore, most IICs were positioned on the positive side of the interval (0, 6) on the ability axis, indicating that trials were more informative about the higher ability levels than the lower ability ones. An exception was observed for trials with a set size of 2, which provided more information about lower ability levels within the interval (-6, 0) on the ability scale.

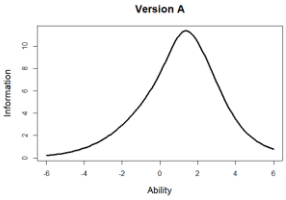
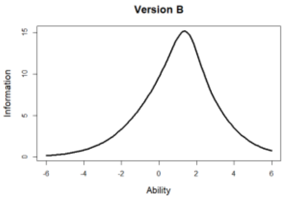
Table 4.17 Trial Information for TALL_SNWR

Version	Trial size	Total Information	Information (-6, 0)	Information (0, 6)	Item Information Curve
A	2	4.441	2.77 (62.37%)	1.539 (34.66%)	
	3	6.661	2.258 (33.9%)	4.192 (62.93%)	
	4	8.884	2.48 (27.92%)	6.161 (69.35%)	
	5	10.359	2.49 (24.04%)	7.509 (72.49%)	
	6	13.333	2.823 (21.17%)	10.256 (76.92%)	
	7	9.634	1.582 (16.42%)	7.958 (82.6%)	

Version	Trial size	Total Information	Information (-6, 0)	Information (0, 6)	Item Information Curve
B	2	5.125	3.211 (62.66%)	1.822 (35.55%)	
	3	7.687	2.617 (34.04%)	4.937 (64.22%)	
	4	10.248	2.9 (28.3%)	7.111 (69.4%)	
	5	12.809	3.067 (23.94%)	9.437 (73.67%)	
	6	15.377	3.022 (19.65%)	12.146 (78.99%)	
	7	11.109	1.841 (16.57%)	9.201 (82.83%)	

Finally, TICs and statistics in Table 4.18 indicated how much information about the nonword recall ability associated with the phonological short-term memory each version provided at the different levels of ability. In general, although both versions were appropriate to evaluate participants' ability, Version B provided more information about ability compared to Version A. Both versions, however, did not provide evenly distributed information about the latent ability, with much more amounts (above 70%) of test information for ability levels in the interval (0, 6), suggesting that both versions were challenging to the participants.

Table 4.18 Test Information for TALL_SNWR

Version	Total Information	Information (-6, 0)	Information (0, 6)	Test Information Curve
A	53.313	14.404 (27.02%)	37.615 (70.56%)	
B	62.354	16.658 (26.71%)	44.654 (71.61%)	

4.4.2.5 Item quality in TALL_CST

Model comparisons

GPCM was applied the split data of both versions in two modalities, using three constraint options. All three constraint options were successfully applied to all data sets except the data of Version A in written modality, on which the model with the “gpcm” constraint option could not be successfully converge. The results of model comparisons showed that the models with the “1PL” constraint option had better fit on the data, evidenced by the AIC, BIC and p-values of likelihood ratio tests. Therefore, the GPCM with the “1PL” constraint, assuming equal discrimination parameter for all trials, was selected for model estimations using the `ltm` package.

Model fitness

The results of the parametric bootstrap goodness-of-fit test using χ^2 indicated that the *p*-value of Version B in the aural modality was greater than the .05 criterion, indicating a better model fit. The *p*-value from the analysis of the data from Version A in written modality was marginally smaller than the .05 criterion. However, the data from Version A in the aural modality and Version B in the written modality had a poor fit to the model, with *p*-values smaller than .05, indicating that the observed data had a probability of less than 5% being generated from the expected data distribution under the assumption of the null hypothesis (i.e., no significant difference between the observed data and the expected data according to the model) being true. Hence, the results suggested that the models did not adequately fit the datasets.

Model estimations and item parameters

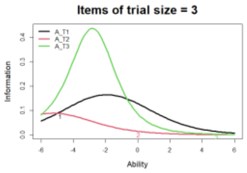
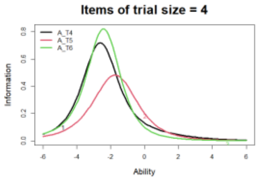
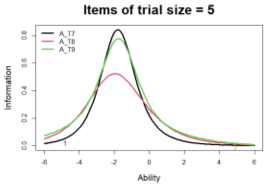
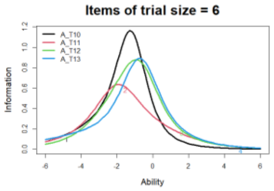
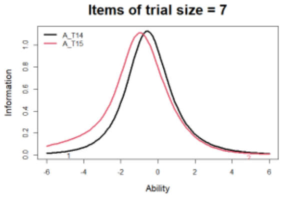
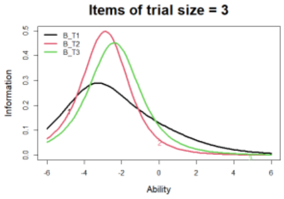
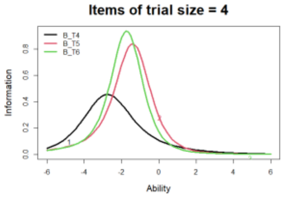
The chosen GPCM constrained the discrimination parameters to be equal to all items, and the coefficients of discrimination parameters of the data from two versions in the same modality were similar. The results indicated non-statistical significance of the coefficients at the 5% level, hence suggested that the discrimination parameters estimated by the GPCM for all data were reliable. The coefficients of discrimination parameters in the aural modality were higher than those in written modality.

The coefficients of the category threshold parameters and the corresponding z-values indicated complicated patterns of difficulty parameters of data in this subtest. Except a small number of coefficients, most coefficients of all datasets were smaller than 0, showing that most category thresholds were easy to endorse. Among the coefficients greater than 0, only two (Category 5 of Trial 10 in Version A and Category 6 of Trial 14 in Version B) in the written modality had z-values greater than 1.96, indicating the reliable estimations. Second, about one third of the coefficients had statistically significant z-values (that is, absolute values < 1.96), raising doubts about reliability of model estimations for difficulty parameters. In addition, the coefficients of the last category threshold parameter tended to be consistently lower than those of the former category threshold parameters on the data obtained in the written modality, which was reasonable, with a few exceptions: Trials 11, 12 and 14 in Version A, and Trials 11, 13 and 14 in Version B.

Plots of model estimations

As shown in Table 4.19, IICs were plotted for trials that had the same set size in both versions of aural modality separately. Additionally, information for trials with the same set size were calculated. The plots revealed that trials with larger set sizes provided more information than those with smaller set size. Furthermore, all IICs were positioned on the negative side of the interval (-6, 0) on the ability axis, indicating that all trials were more informative about the lower ability levels than the higher ability ones.

Table 4.19 Trial Information for TALL_CST in aural modality

Version	Trial size	Total Information	Information (-6, 0)	Information (0, 6)	Item Information Curve
A	3	3.522	2.664 (75.62%)	0.431 (12.23%)	
	4	6.543	5.759 (88.02%)	0.584 (8.93%)	
	5	7.731	6.365 (82.33%)	1.126 (14.56%)	
	6	13.662	10.216 (74.78%)	2.909 (21.29%)	
	7	7.734	5.299 (68.52%)	2.246 (29.04%)	
B	3	5.235	4.388 (83.82%)	0.512 (9.78%)	
	4	6.422	5.707 (88.87%)	0.56 (8.73%)	

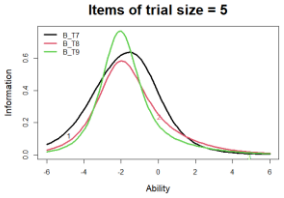
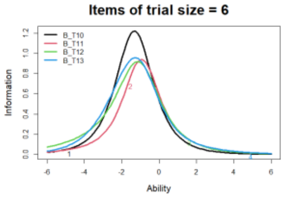
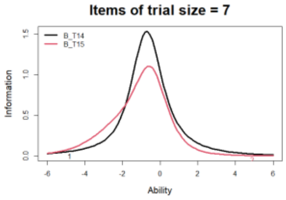
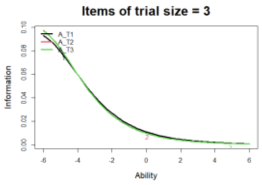
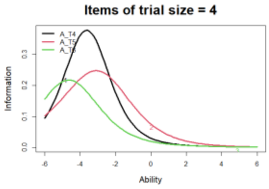
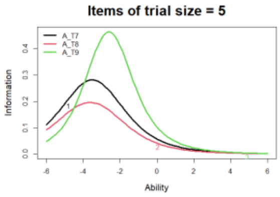
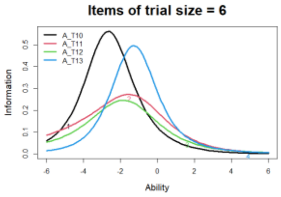
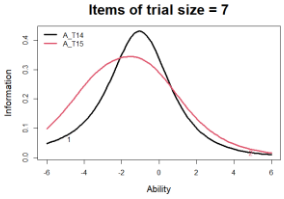
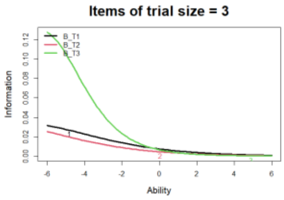
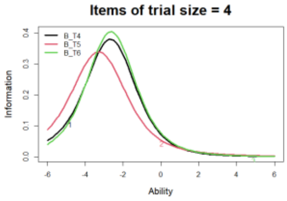
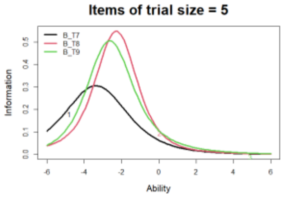
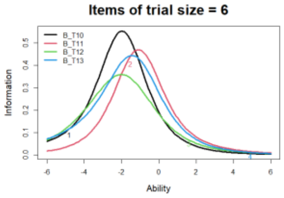
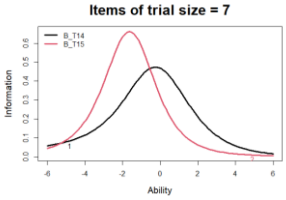
Version	Trial size	Total Information	Information (-6, 0)	Information (0, 6)	Item Information Curve
	5	7.592	6.203 (81.7%)	1.215 (16.01%)	
	6	13.43	10.521 (78.34%)	2.634 (19.61%)	
	7	7.597	5.591 (73.6%)	1.895 (24.95%)	

Table 4.20 provided information of trials with the same set size in both versions of written modality. The patterns were similar to those from aural modality, indicating that trials were more informative about the lower ability levels than the higher ability ones. Importantly, all trials in written modality were less informative when compared to their counterparts with the same set sizes in aural modality.

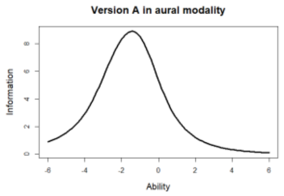
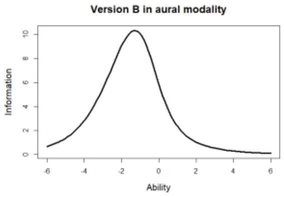
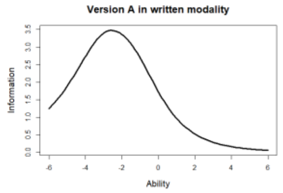
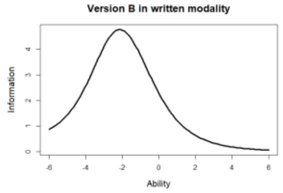
Table 4.20 Trial Information for TALL_CST in written modality

Version	Trial size	Total Information	Information (-6, 0)	Information (0, 6)	Item Information Curve
A	3	1.825	0.825 (45.18%)	0.065 (3.56%)	
	4	3.783	3.051 (80.65%)	0.194 (5.12%)	
	5	4.149	3.461 (83.42%)	0.298 (7.18%)	
	6	6.888	5.333 (77.42%)	1.107 (16.06%)	
	7	4.497	2.999 (66.68%)	1.197 (26.62%)	
B	3	1.166	0.514 (44.06%)	0.041 (3.54%)	
	4	4.375	3.828 (87.5%)	0.268 (6.14%)	

Version	Trial size	Total Information	Information (-6, 0)	Information (0, 6)	Item Information Curve
	5	5.118	4.419 (86.35%)	0.403 (7.88%)	
	6	7.994	6.1 (76.31%)	1.426 (17.84%)	
	7	5.095	3.443 (67.58%)	1.408 (27.63%)	

Finally, TICs and the statistics in Table 4.21 indicated that Version B was slightly more informative than Version A in both modalities about the ability of executive control in WM. Both versions in aural modality provided more information about the ability than those in written modality. All instruments, however, did not provide evenly distributed information about the latent ability, with much more amounts (above 74%) of test information for lower ability levels in the interval (-6, 0), suggesting that they were not challenging to the participants.

Table 4.21 Test Information for TALL_CST

Modality	Version	Total Information	Information (-6, 0)	Information (0, 6)	Test Information Curve
aural	A	39.192	30.303 (77.32%)	7.296 (18.62%)	
	B	40.276	32.411 (80.47%)	6.817 (16.93%)	
written	A	21.143	15.669 (74.11%)	2.861 (13.53%)	
	B	23.747	18.304 (77.08%)	3.547 (14.94%)	

4.5 Battery level analysis

This section addresses the explanation inference (as part of the validation plan) which concerns the construct validity of TALL as a componential measure of aptitude.

As introduced in Section 3.3.5.2, two statistical methods were applied on the aggregated datasets (that is, the score of each subtest rather than the scores at the item level of each subtest). The aggregated data for each subtest were split by the condition of modality only, i.e., the data from the two versions of materials in the same modality were not split. Prior to taking this decision, the mean differences of the data of Version A and Version B in the same modality, in each subtest, were compared using pairwise *t*-tests. This step was to ensure that the differences between the data obtained using Version A and Version B in the same modality were not statistically significant from each other.

Descriptive statistics (see Table 4.22 and Table 4.23) and Shapiro-Wilk Test revealed that most datasets were not normally distributed, except for Vocabulary Learning in the written suite. Consequently, non-parametric *t* tests were deemed appropriate. To assess homogeneity of variance between data from both versions in the same suite, Levene's Test, less sensitive to deviations from normality, was employed. If the homogeneity of variance was statistically significant, the Yuen–Welch (Y–W) Test (a robust method for non-normality distributed data with unequal variances) was applied using `yuen` function in the `WRS2` package (Mair & Wilcox, 2020). Alternatively, when Levene's Test indicated no significant variance difference, the Mann-Whitney-Wilcoxon (MWW) Test was used to compare non-normally distributed data with assumed equal variance, conducted through the `wilcox.test` function.

The results of the non-parametric *t*-tests indicated that the mean differences between the data collected using Version A and Version B, in general, were not statistically significant. Therefore, it should not be problematic to combine the data from the two material versions for the analyses at the battery level, as these datasets from different versions in the same modality were independent, taken by different participants (and at different times). Specifically, Principal component analysis (PCA) was used to investigate whether TALL subtests could succinctly measure the dimensions of aptitude construct without much redundancy. Confirmatory factor analysis (CFA) was used to verify whether TALL, as an aptitude battery, tapped into the components evidenced by the fitness of the data to the structure of factors informed by the theoretical models. The applications and the conventions of output interpretation of PCA and CFA will be introduced in the following sections, together with the results.

Table 4.22 Summary of descriptive analysis for TALL aural suite

subtest	version	n	mean	sd	min	max	skew	kurtosis	se	Shapiro-Wilk	
										W	*p*
VL	A	81	0.27	0.17	0.00	0.95	1.46	2.85	0.02	0.88	< .001
	B	84	0.36	0.20	0.00	0.95	0.82	0.62	0.02	0.95	< .01
SD	A	81	0.75	0.22	0.27	1.00	-0.61	-0.91	0.02	0.9	< .001
	B	84	0.79	0.14	0.13	1.00	-1.48	4.55	0.02	0.9	< .001
LA	A	81	0.66	0.31	0.10	1.00	-0.40	-1.43	0.03	0.87	< .001
	B	84	0.67	0.25	0.13	1.00	-0.57	-0.88	0.03	0.92	< .001
SNWR	A	81	0.33	0.14	0.04	0.87	0.72	1.04	0.02	0.96	< .05
	B	84	0.35	0.15	0.07	0.74	0.56	-0.35	0.02	0.96	< .01
CST	A	81	0.82	0.13	0.48	1.00	-0.85	0.08	0.01	0.93	< .001
	B	84	0.82	0.13	0.36	1.00	-0.95	0.71	0.01	0.93	< .001

Note. Key to subtest: VL = Vocabulary Learning; SD = Sound Discrimination; LA = Language Analysis; SNWR = Serial Nonwords Recall; CST = Complex Span Tasks

Table 4.23 Summary of descriptive analysis for TALL written suite

Subtest	version	n	mean	sd	min	max	skew	kurtosis	se	Shapiro-Wilk	
										W	*p*
VL	A	84	0.41	0.22	0.00	1.00	0.58	-0.32	0.02	0.96	<.01
	B	81	0.50	0.23	0.05	1.00	0.03	-0.80	0.03	0.98	0.12
SD	A	84	0.70	0.22	0.20	1.00	-0.41	-0.80	0.02	0.94	< .01
	B	81	0.79	0.14	0.17	1.00	-2.18	7.48	0.02	0.81	< .001
LA	A	84	0.82	0.25	0.13	1.00	-1.47	0.75	0.03	0.71	< .001
	B	81	0.80	0.25	0.00	1.00	-1.43	1.02	0.03	0.76	< .001
SNWR	A	84	0.34	0.16	0.06	0.82	0.80	0.34	0.02	0.95	< .01
	B	81	0.33	0.17	0.05	0.89	0.78	0.55	0.02	0.96	< .01
CST	A	84	0.89	0.08	0.49	1.00	-1.63	5.60	0.01	0.89	< .001
	B	81	0.87	0.10	0.41	1.00	-1.61	4.43	0.01	0.88	< .001

Note. Key to subtest: VL = Vocabulary Learning; SD = Sound Discrimination; LA = Language Analysis; SNWR = Serial Nonwords Recall; CST = Complex Span Tasks

4.5.1 Principal Component Analysis (PCA): Exploring factor structure

PCA is a descriptive method which is included under the broad term of Exploratory Factory Analysis (EFA) to obtain the factor structure based on the data clustering pattern (Plonsky & Gonulal, 2015). As an umbrella term of statistical analysing method, EFA covers both PCA and EFA, which can be used when no particular expectations regarding the number and nature of the underlying factors (i.e. latent variables) that exist in the data (Loewen & Gonulal, 2015). Conceptually, the difference between EFA and PCA lies in how the variance in the data are treated, that is, PCA analyses variance whereas EFA analyses covariance, while the importance of the distinction between these two methods can be controversial (Field et al., 2012). The goal of PCA is to identify principal components, which are directions that maximize the variation in the data. This method is particularly useful when the variables in the dataset are highly correlated, indicating that there may be redundancy in the data. Therefore, PCA can be applied to reduce the dimensionality of multivariate data to two or three principal components that can be visualized graphically, while preserving most of the original information. When the goal of a research is to reduce the number of variables, PCA is a high-quality choice for statistical analysis (Conway & Huffcutt, 2003).

The PCA procedures, as outlined in Field et al. (2012), involves using a correlation matrix to calculate the variates of the variables, which are represented by eigenvectors. The largest eigenvalue associated with each of the eigenvectors serves as a measure of the significance of the corresponding component. The eigenvalue for a factor is calculated by summing the square of the loadings for that factor. This allows for an assessment of how much of the variance in the variables could be explained by that factor, based on the loadings of the variables. Consequently, a higher loading indicates a greater extent to which a factor accounted for the variance in the variables. Through PCA, the factors with large eigenvalues are retained or extracted.

The main purpose of PCA was to investigate the extent to which the five subtests measure the dimensions of the componential construct of aptitude without much redundancy, rather than to reduce or consolidate variables. Ideally, the results of PCA should demonstrate that the subtests were correlated and made distinct contributions to the componential construct represented by principal factor dimensions, and hence no subtests should be considered as being redundant which may lead to their removal.

PCA in the current research was conducted through functions in the `psych` package (Revelle, 2022), the `FactoMineR` package (Husson, et al., 2017), and the `factoextra` package (Kassambara & Mundt, 2017) in R, following a stepwise protocol (see Appendix G).

4.5.1.1 PCA of the aural suite

First, preliminary analysis was conducted prior to conducting PCA. The results of Shapiro-Wilk test for normality suggested that all scores of the subtests were not normally distributed, informing that Kendall's *tau* should be the appropriate correlation coefficient for the non-parametric data. The correlation matrix with statistically significant coefficients displayed with an asterisk (see Figure 4.22) suggested a moderate level of association between most variables (here the subtests), indicating a fundamental assumption of PCA, that is, there is some degree of linear relationship between variables. Bartlett's test of sphericity, χ^2 (10) = 84.76, *p* < .001, provided additional information to indicate that correlations between subtests were sufficiently large for PCA. The Kaiser-Meyer-Olkin measure verified the sampling adequacy for the analysis KMO = .65 ('mediocre' according to Kaiser, 1974), and all KMO values for individual subtests were above the acceptable limit of .5. The value of the correlation matrix determinate was greater than 0.00001, indicating that the matrix did not have a heuristic problem.

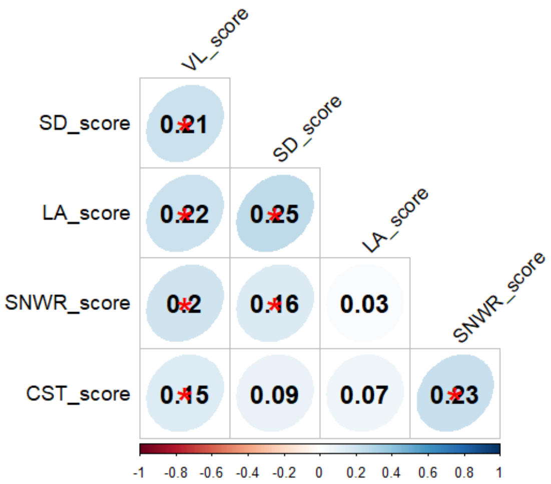


Figure 4.22 Correlation matrix of subtests in aural suite

PCA was conducted on the data from the aural suite using oblique rotation (oblimin). The initial analysis aimed to obtain eigenvalues for each component in the data. Two components had eigenvalues over Kaiser's criterion of 1, jointly explaining 60.5% of the variance. However, when applying Jolliffe's (1972) criterion, three components exhibited eigenvalues greater than 0.70, with all five components exceeding Stevens' (2002) criterion of 0.512 for a sample size of 100. These results implied that a single principal factor explanation was less than ideal due to the very low correlations between certain subtests (see Figure 4.22). Furthermore, it appeared that none of the five underlying factor dimensions should be omitted, given that even the least important component could account for 11.2% of the variance (eigen value = 0.56). The scree plot (Figure 4.23) illustrated the percentage of variance explained by the five dimensions.

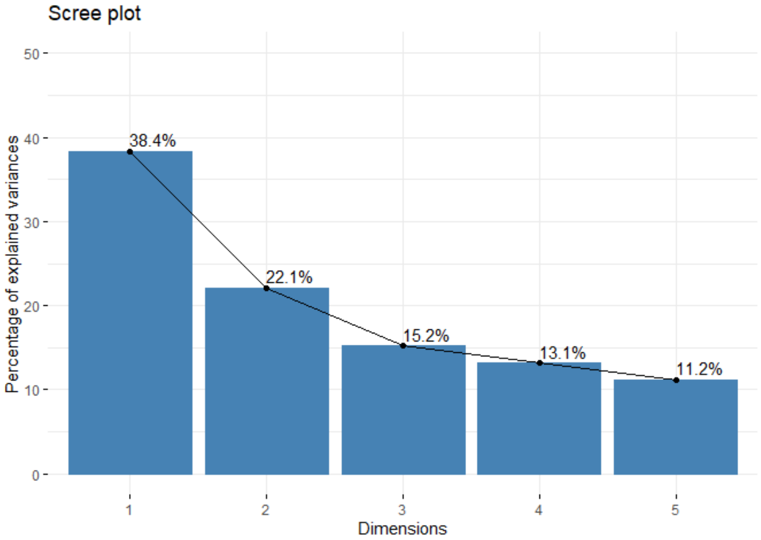


Figure 4.23 Scree plot of aural suite

Table 4.24 reported the factor loadings after rotation, along with the percentage of variance explained by each dimension and the cumulative percentage of variance explained by the extracted factors. The clustering of the subtests on the components indicated that all subtests contributed nearly evenly to the first dimension (component 1). It might be reasonable to consider extracting the first two dimensions as the principal components, explaining 60.5% of the total variance, which aligns with the field-specific criterion of 60% (Plonsky & Gonulal, 2015). Alternatively, one could consider extracting the first three dimensions as the principal components, explaining 75.68% of the total variance with higher confidence. Regardless of the number of factors chosen as principal components, the PCA results clearly indicated minimal redundancy, making it unadvisable to reduce the variables (in this case, the subtests), as all variables made distinct contributions to the principal components. This interpretation was visualised in the factor map of contributions (Figure 4.24), describing the relationships between the subtests and the underlying factor dimensions.

Table 4.24 Summary of factor loadings of subtests in aural suite

	Oblimin rotated factor loadings				
Subtest	Dimension 1	Dimension 2	Dimension 3	Dimension 4	Dimension 5
TALL_VL	**.75**	.02	–.05	**–.53**	**–.39**
TALL_SD	**.60**	**.44**	**–.45**	**.46**	–.14
TALL_LA	**.49**	**.65**	**.47**	–.07	**.33**
TALL_SNWR	**.65**	**–.48**	**–.30**	–.10	**.49**
TALL_CST	**.57**	**–.51**	**.49**	**.39**	–.18
Eigenvalues	1.92	1.11	.76	.66	.56
% of variance	38.38	22.12	15.19	13.14	11.18
Cum. % of var.	38.38	60.50	75.68	88.82	100.00

Note. Factor loadings over .30 appear in bold. Key: VL = Vocabulary Learning; SD = Sound Discrimination; LA = Language Analysis; SNWR = Serial Nonwords Recall; CST = Complex Span Tasks; Cum. % of var. = cumulative percentage of variance

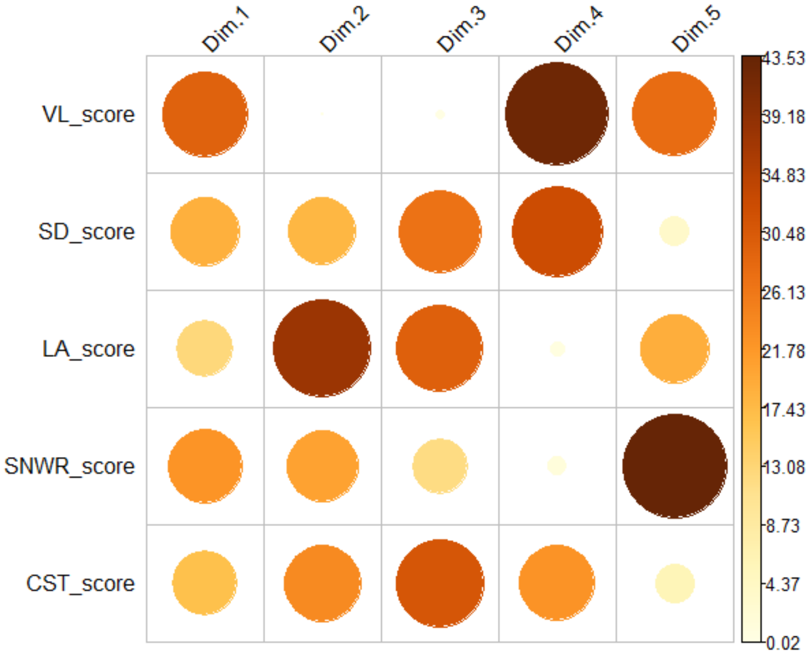


Figure 4.24 Factor map of contributions in aural suite

4.5.1.2 PCA of the written suite

The results of Shapiro-Wilk test for normality indicated that all data from the written suite were not normally distributed, informing the use of Kendall's tau as the correlation. The correlation matrix with statistically significant coefficients displayed with an asterisk (see Figure 4.25) suggested a moderate level of association between most variables (here the subtests), indicating that a fundamental assumption of PCA was met. Bartlett's test of sphericity, χ^2 (10) = 52.51, p < .001, provided additional information to indicate that correlations between subtests were sufficiently large for PCA. The Kaiser-Meyer-Olkin measure verified the sampling adequacy for the analysis KMO = .69 ('mediocre' according to Kaiser, 1974), and all KMO values for individual subtests were above the acceptable limit of .5. The value of the correlation matrix determinate was greater than 0.00001, indicating that the matrix did not have a heuristic problem.

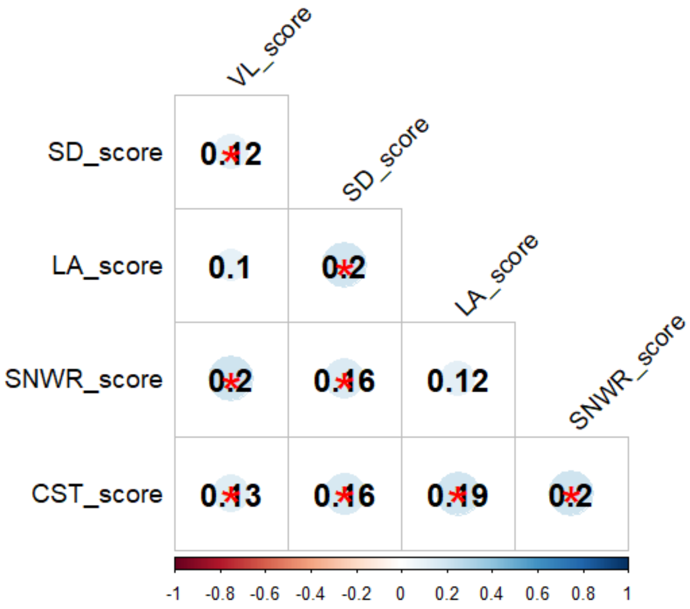


Figure 4.25 Correlation matrix of subtests in written suite

PCA was conducted on the data from the written suite using oblique rotation (oblimin). The initial analysis was run to obtain eigenvalues for each component in the data. One component had eigenvalue over Kaiser's criterion of 1 and explained 36.1% of the variance. However, when applying Jolliffe's (1972) criterion, four components exhibited eigenvalues greater than 0.70, with all five components had eigenvalues greater than the

Stevens' (2002) criterion of 0.512 for a sample size of 100. These results implied that a single principal factor explanation was less than ideal due to the very low correlations between certain subtests (see Figure 4.25). Furthermore, it appeared that none of the five underlying factor dimensions should be omitted, given that the least important component could account for 14% (eigen value = 0.7) of the variance. The scree plot (Figure 4.26) illustrated the percentage of variance explained by the five dimensions.

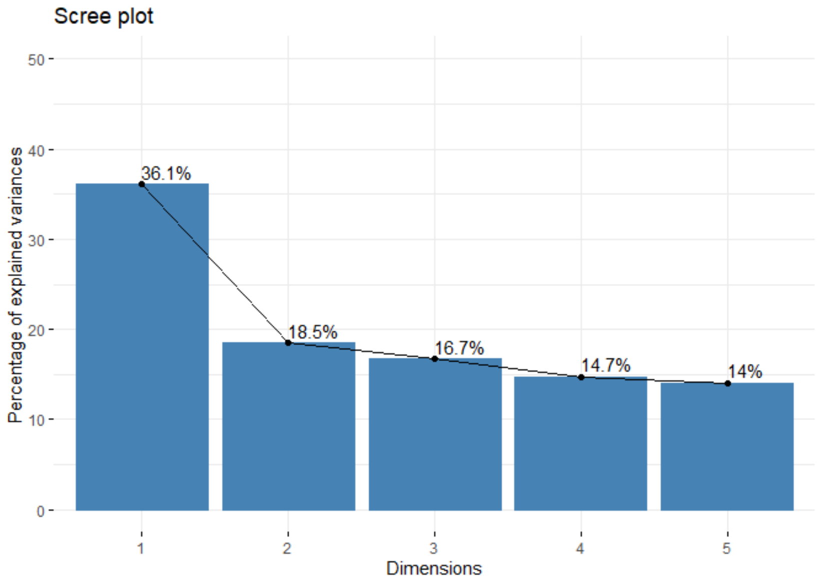


Figure 4.26 Scree plot of written suite

Table 4.25 reported the factor loadings after rotation, along with the percentage of variance explained by each dimension and the cumulative percentage of variance explained by the extracted factors. The clustering of the subtests on the components indicated that all subtests contributed nearly evenly to the first dimension (component 1). It might be reasonable to consider extracting the first two factor dimensions as the principal components, explaining 54.56% of the total variance, although the eigenvalue of the second dimension was slightly lower than Kaiser's criterion of 1. This cumulative percentage of variance (54.56%), however, was slightly lower than the field-specific criterion of 60% (Plonsky & Gonulal, 2015), suggesting that it would be reasonable to consider extracting the first three dimensions as the principal components, explaining 71.26% of the total variance. Regardless of the number of factors chosen as principal components, the PCA results clearly indicated minimal redundancy, making it unadvisable to reduce the variables (subtests), as all variables made distinct contributions to the principal components. This interpretation was visualised in the factor map of contributions (Figure 4.27), describing the relationships between the subtests and the underlying factor dimensions.

Table 4.25 Summary of factor loadings of subtests in written suite

	Oblimin rotated factor loadings				
Subtest	Dimension 1	Dimension 2	Dimension 3	Dimension 4	Dimension 5
TALL_VL	**.55**	**–.58**	**.46**	–.03	**.39**
TALL_SD	**.64**	**.33**	–.16	**–.66**	.13
TALL_LA	**.53**	**.57**	**.54**	.28	**–**.15
TALL_SNWR	**.65**	**–.38**	–.15	.00	**–.64**
TALL_CST	**.63**	.09	**–.53**	**.46**	**.32**
Eigenvalues	1.8	.92	.84	.74	.7
% of variance	36.09	18.47	16.7	14.74	14.01
Cum. % of var.	36.09	54.56	71.26	85.99	100.00

Note. Factor loadings over .30 appear in bold. Key: VL = Vocabulary Learning; SD = Sound Discrimination; LA = Language Analysis; SNWR = Serial Nonwords Recall; CST = Complex Span Tasks; Cum. % of var. = cumulative percentage of variance

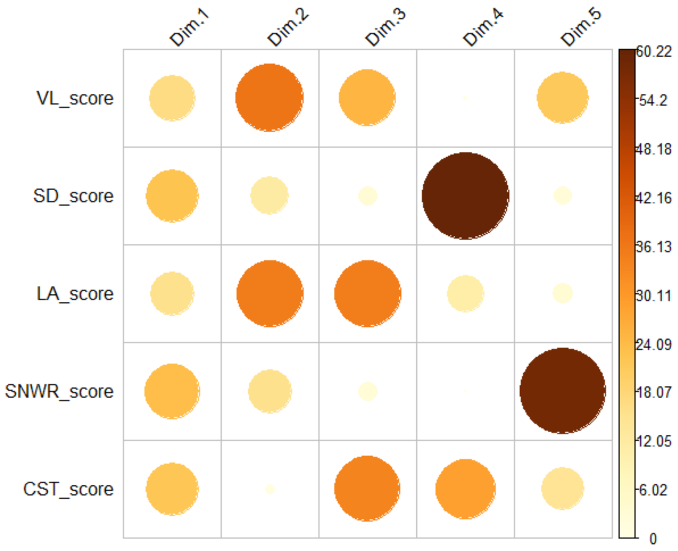


Figure 4.27 Factor map of contributions in written suite

In summary, the results from PCA showed that both aural and written suites of TALL did not have much redundancy. All five subtests loaded nearly evenly to the first factor dimension (or the principal component 1) that explained about 36% (written suite) or 38% (aural suite) of the total variance. It seemed to be more reasonable to extract the first two or three dimensions that could explain up to about 71% (written suite) or 75% (aural suite) of the total variance in cumulation, which fitted in the 70-90% range of the score variance (Jolliffe, 2002) that the extracted factors should explain. The results also suggested that all subtests, in general, contributed similarly to the principal components underlying the construct of aptitude.

4.5.2 Confirmatory Factor Analysis (CFA): Verifying a four-factor model

CFA, as a part of Structural Equation Modelling (SEM), is one of the two main analytical methods based on the common factor model, the other being the Exploratory Factor Analysis (EFA). While both methods aim to identify the underlying structure of observed relationships among the measures with a smaller set of latent variables, CFA differs from EFA in terms of the number and nature of prior specifications and restrictions made on the factor model. Unlike EFA, which is a data-driven approach, CFA requires a strong empirical or conceptual foundation to guide the specification and evaluation of the factor model. The number of factors and the patten of measure-factor loadings need to be specified in advance, and the prespecified factor solution is evaluated based on its ability to reproduce the sample correlations (covariance) matrix of the measured variables. As such, CFA is typically used in later phase of scale development and construct validation, after the underlying structure has been established on prior empirical (EFA) and theoretical grounds (Brown, 2015).

In the preceding section, PCA results provided empirical evidence supporting that TALL subtests can effectively measure a componential construct with minimal redundancy. To further validate whether TALL, as a battery for aptitude, aligns with the proposed components outlined in theoretical models—specifically, the Stages Approach and the P/E Model—CFA was employed.

The theoretical framework of TALL, upon which the five subtests were developed, posits that these subtests should be structured around four *primary* factors: associative memory, phonetic coding ability, language analytic ability, and WM. Therefore, CFA was applied separately on the data from both suites, with a pre-specified factor structure. In this structure, the subtests of TALL_VL, TALL_SD, TALL_LA were expected to load onto the three factors (associative memory, phonetic coding ability, and language analytic ability), while the subtests of TALL_SNWR and TALL_CST were anticipated to load onto a single factor representing WM.

The objective of using CFA was to evaluate how well the hypothesised model fitted the empirical data. CFA was conducted using the `lavaan` package (Rosseel, 2012) in R, following a stepwise protocol (see Appendix H) of data preparation, model constructing and model fit checking.

4.5.2.1 CFA of the aural suite

The transformed data of all subtests in the aural suite were substantially improved and had non-significant skew and kurtosis, i.e., the absolute values of skew.2SE and kurt.2SE were less than 1. However, the results of Shapiro-Wilk test showed that, except for the data of TALL_SNWR, the data were still not normally distributed even after transforming the outliers, with *p* values less than .05 (TALL_CST), less than .01 (TALL_VL and TALL_SD), and less than .001 (TALL_LA).

To deal with the non-normal data, the estimator option of "MLM" was specified in the `cfa` function through the `lavaan` package, which fit the model using standard maximum likelihood to estimate the model parameters but with robust standard errors and Satorra-Bentler correction for scaled test statistics. The robust model fit indices showed that the model fit the data well, as indicated by the non-significant chi-square test for the User Model (χ^2 (2) = 2.47, *p* = .29), and that the specified model provided a significantly better fit to the data compared to the baseline model (χ^2 (10) = 102.50, *p* < .001). The approximate fit indices, i.e., CFI = .995, TLI = .975, RMSEA = .021, and SRMR = .019, also indicated that the data of the aural suite fit well to the four-factor model hypothesised based on the theoretical framework of aptitude. The estimates for the factor loadings, intercepts and variances were all significant, showing that the model explained a substantial amount of variance in the observed variables (i.e., the subtests).

4.5.2.2 CFA of the written suite

The transformed data of all subtests in the written suite were substantially improved. Most of the transformed data had non-significant skew and kurtosis with the absolute values of skew.2SE and kurt.2SE below than 1. The only exception was the data of TALL_LA, which had a significant skew with skew.2SE = 1.96. The results of Shapiro-Wilk test showed that, except for the data of TALL_VL, the data were still not normally distributed even after transforming the outliers, with *p* values less than .05 (TALL_SNWR), less than .01 (TALL_SD and TALL_CST), and less than .001 (TALL_LA).

To deal with the non-normal data, the estimator option of “MLM” was specified in the cfa() function through the lavaan package, which fit the model using standard maximum likelihood to estimate the model parameters but with robust standard errors and Satorra-Bentler correction for scaled test statistics. The robust model fit indices showed that the model fit the data well, as indicated by the non-significant χ^2 test for the User Model (χ^2 (2) = 1.99 with *p* = .40), and that the specified model provided a significantly better fit to the data compared to the baseline model (χ^2 (10) = 66.48, *p* < .001). The approximate fit indices, i.e., CFI = 1.000, TLI = 1.001, RMSEA = .000, and SRMR = .018, also indicated that the data of the written suite fit almost perfect to the four-factor model hypothesised based on the theoretical framework of aptitude. The estimates for the factor loadings, intercepts and variances were all significant, showing that the model explained a substantial amount of variance in the observed variables (i.e., the subtests).

4.6 Summary of the results

The focus of this chapter was to investigate the RQ1:

- To what extent does TALL display satisfactory internal consistency and validity as a measure for aptitude?

A data cleaning protocol was followed to prepare the final data for analysis from 165 participants, and the data files were formed for each subtest and split by conditions of material version and/or modality as unaggregated data sets for analyses at the subtest and item levels. Aggregated datasets (joining together different versions of the tests) were also generated for each test suite at the battery level, though modality was always kept separate.

The results of **reliability and unidimensionality checks** revealed that, in general, all subtests had satisfactory reliability evidenced by the coefficients of Omega and Cronbach’s alpha according to the field-specific acceptable threshold of coefficient alpha > .74 proposed by Plonsky & Derrick (2016). Two data sets, i.e., the one of Subtest TALL_VL using Version A in the aural modality and the other of Subtest TALL_CST using Version A and B in the written modality had the coefficients lower than .74. However, their 95% CI of alpha estimates still included the threshold value. The unidimensionality of all subtests was evidenced, as the fa.fit statistics indicated satisfactory proportions (from the lowest of 60% to the highest 98%) of covariance being accounted for by the general factor.

The **item analysis** using IRT models did not provide evidence for the deletion of any items in any of the subtests that could have been considered as poor discriminators of the latent abilities the subtests intended to measure. The results also showed test information provided by the subtests varied. TALL_SD, TALL_LA, and TALL_CST were not challenging to the participants, while TALL_VL and TALL_SNWR were difficult. TALL_SNWR was the most challenging subtest to the participants.

Finally, **internal validity** at the battery level was investigated on the aggregated data sets of the aural suite and the written suite, using PCA and CFA respectively. The results of PCA revealed that both suites of TALL did not have much redundancy (that could have led to the reduction of any specific subtests). All subtests, in general, contributed similarly to the first principal component underlying the construct of aptitude. The CFA results provided strong evidence of model fit statistics, indicating that the data of both suites fitted well to the hypothesised four-factor model based on the theoretical frameworks on which TALL was constructed.

4.7 Discussion for RQ1

RQ1 sought to scrutinise the internal validity of TALL as a theory-based aptitude battery by analysing the data on the subtest level, item level, and battery level sequentially, following the validation plan. The results suggested that the internal validity of TALL can be established, based on evidence to make (a) the generalisation inference that scores may be generalised to the componential construct each subtest intended to measure, (b) the scoring inference that all subtests were composed of well-functioning items with appropriate levels of difficulty and discrimination, and (c) the explanation inference that TALL can be the measure for the construct of aptitude that is conceptualised by the theoretical frameworks. The discussion of RQ1 below is under four general themes: reliability and unidimensionality, item quality, and verification of theoretical frameworks.

4.7.1 Reliability and unidimensionality

4.7.1.1 Unidimensionality

The results of unidimensionality checks for all datasets provided a general indication that each subtest effectively measures a componential construct of aptitude. Following the *eigenvalue-greater-than-one* rule (Mulaik, 1972), all datasets had extracted a general factor with its eigenvalue (g) greater than one, which indicated that a common factor should be retained and the common factor in each dataset accounts for a substantial portion of the variance in the dataset. However, the eigenvalues of F1 in most of the datasets (except Version B in the aural and written suites of TALL_LA, and Version A in the written suite of TALL_CST) were greater than one as well, indicating that other factors, not only the general factors, also contributed significantly to the variance associated with other variables. This finding was not surprising, as most of the psycho-educational measures are not purely unidimensional (Slocum-Gori et al., 2009). Considering that the factors in the datasets were highly likely to be correlated, the ratio of explained common variance (ECV) and the *fa.fit* were used to inform unidimensionality of each measurement. The indices of ECV and the unidimensionality measure represented in the *fa.fit* suggested that the datasets of TALL_VL in the aural suite and TALL_CST in the written suite need cautious interpretation. Specifically, these datasets had the lowest values of the explained common variance (ECV), which were .23 and .26, respectively, and they had lowest values of .60 and .69 of *fa.fit*.

The unidimensionality issue related to measuring associative memory using vocabulary learning task in the aural modality may suggest that retaining new words in the acoustic forms involves more factorial dimensions than remembering words in the written forms. This can be explained by Skehan's (2006) notion that the ability to process sound and retain new language occurrences in long-term memory being related to vocabulary learning and, therefore, central to measuring aptitude. Thus, measuring associated memory by vocabulary learning task in the aural form taps into both sound processing and form retaining.

A methodological explanation of the results could be related to the testing equipment used. Since participants were tested in the online condition, it was unavoidable that they used their own headphones or speakers. Hence, perceiving the input could be more susceptible to the external factors involved in the aural modality than in the written modality. However, this heterogeneity of audio quality did not seem to lead to higher variability (reflected in the standard deviations in Table 4.3) in the results from TALL_VL in the aural suite compared to the results in the written suite.

Regarding the subtest of TALL_CST in the written modality, the indices of unidimensionality were also relatively lower than those in the aural modality. This issue may be related to the reliability of this subtest in the written modality, which will be discussed in the following section.

4.7.1.2 Reliability

The results indicated satisfactory reliability for most subtests, as evidenced by the coefficients of Omega hierarchical (ω_h), Omega hierarchical (ω_t), and Cronbach's alpha (α) estimators. However, scores in the datasets of Version A in the aural suite of TALL_VL and Version A and Version B in the written suite of TALL_CST displayed lower coefficients, albeit still close to the acceptable threshold of .70 (Field et al., 2012). The results suggested variability attributable to subfactors in these datasets, as they had the lowest values of ω_h (.31, .35, and .33, respectively), indicating that after controlling the variance of other factors, the estimated reliability of the general factor was lower than .35.

The suitability of using Cronbach's alpha (α) as the reliability coefficients for the current study was considered. Given that the assumption of normal distribution was violated in all datasets and the assumption of unidimensionality was only approximately met in several datasets as mentioned above, Cronbach's alpha could underestimate the actual value of reliability (McNeish, 2018). This was reflected in the current study as the coefficients in α were systematically lower than those in ω_t in all datasets, and the discrepancies were particularly apparent for the above-mentioned datasets that had concerns about unidimensionality. Specifically, Version A in the aural suite of TALL_VL had .70 of coefficient in α but the coefficient in in ω_t was .86; Version A and Version B in the written suite of TALL_CST had coefficients of .64 and .72 in α, whereas their coefficients in ω_t were .68 and .79, respectively. Importantly, reporting the reliability coefficient solely in α, which is almost always the case in the research field, may possibly miss an important psychometric property of a measurement, the proportion of variance in the scores associated with a general factor, as reflected in ω_h (Zinbarg et al., 2005). This suggested that the choice of reliability coefficient in the field of applied linguistics merits more considerations, and the practice of using other suitable estimators rather than relying on Cronbach's alpha should be encouraged (see the similar discussion in O'Reilly & Marsden, 2020).

The purpose of using Cronbach's alpha in the current study was to provide comparable coefficients to the reliability coefficients of other aptitude measurements reported in literature. The results showed that all subtests had coefficients greater than .70, a threshold of the moderate benchmark suggested in the L2 domain by Brown (2014), except one dataset of Version A in the written suite of TALL_CST which had the coefficient of .64.

When compared to the reliability coefficients of specific subtests of the LLAMA tests from which TALL were informed, the two versions of TALL_VL in the written suite had coefficients (.81 and .84) that were similar to the coefficient of LLAMA_B (.81) reported by Bokander and Bylund (2020). The two versions of TALL_LA in the written suite had coefficients (.96 and .94) that were much higher than the coefficient of LLAMA_F (.60).

Compared to the reliability coefficients of the subtests in the Hi-LAB that measure the construct of phonetic coding ability, the two versions of TALL_SD had coefficients (.89 and .78), which were in line with the coefficients of Phonemic Discrimination in the Hi-LAB (.81 and .80) reported by Tseng et al. (2015), and were higher than the coefficients of Phonemic Categorization (.66 and .77) reported by Mislevy et al. (2010) (cited in Hughes, et al., 2023). The two versions of TALL_SNWR displayed the coefficients (.86 and .88), which were very much in line with the coefficient of Non-Word Span (.86) measuring phonological short-term memory in the Hi-LAB (Mislevy et al., 2010) and slightly lower than the coefficient of Non-Word Span (.93) (Tseng et al., 2015). TALL_SNWR also displayed coefficients at the higher bound of the mean reliability values (ranging from .72 to .89) of simple span tasks, as calculated by Shin and Hu (2021) in their meta-analysis.

Two versions of TALL_CST in the written suite had lower coefficients (.64 and .72) compared to the coefficients of the subtests of Operation Span (.81) and Running Memory Span (.77) in the Hi-LAB, both measuring executive control capacity of WM, reported by Mislevy et al. (2010). Furthermore, the coefficients of TALL_CST in the written suite were lower than the coefficient (.86) of the translated Reading Span Tasks (Gass et al., 2019), though their coefficient of .86 was computed jointly on the Chinese and Arabic versions with a small sample size of 19 and so may not be an appropriate comparator.

It is worth noting that the scoring approach used in the WM subtests in this study (i.e., the proportion of correct responses in each trial as the item-level data) differed from the scoring methods employed in the Hi-LAB, which relied on the total number of correctly recalled stimuli. Had the total number of correctly recalled stimuli been applied in the current study, the numbers of items for analysis would have increased from 17 to 74 (in TALL_SNWR) and from 15 to 76 (in TALL_CST). Mathematically, the computation of Cronbach's alpha would have been inflated due to a larger number of homogenous items, despite the instrument remaining unchanged. Consequently, the reliability results obtained through the current scoring approach might be comparatively more conservative than those observed in the Hi-LAB, but the coefficients generally surpassed the acceptable threshold. Therefore, the overall reliability of the WM subtests in TALL was considered satisfactory.

To sum up, except TALL_CST in the written suite, subtests of TALL had displayed similar or higher reliability compared to the subtests in other aptitude batteries that measure the constructs in the same input modality.

The lower coefficients of TALL_CST in the written modality compared to the coefficients in the aural modality might be related to the stimuli for processing. The sentences presented in participants' L1 for semantic judgement probably lacked challenge for the participants in the current study, as they were college undergraduates. This was evidenced by the ceiling effect in the scores of this subtest and participants' responses lacking variability (See Table 4.7 in Section 4.3.1).

A technical arrangement in the experiment design also contributes to the discrepancy between the levels of challenge in the two modalities. Participants were not allowed to make a sentence judgement by clicking on the ‘correct’ or ‘incorrect’ button until the complete sentence had been played in the aural modality, while their responses to select the answer of the judgement were not controlled in the written modality. This means that they could make the choice at any time once a sentence was displayed on the screen. The data recorded at the backend revealed that the reaction time for semantic judgements of sentence stimuli was systematically shorter in the written suite compared to the aural suite. This shorter time interval between sentence processing and letter recall suggests that cognitive load in retaining letters in WM was reduced when the sentence stimuli were presented in the written suite.

4.7.2 Item quality

4.7.2.1 Dichotomous datasets from TALL_VL, TALL_SD, and TALL_LA

In general, Rasch models applied to the dichotomous datasets of TALL_VL, TALL_SD, and TALL_LA obtained good model fitness statistics. The results of item parameters and item fit for the Rasch models provided no clear evidence that any items were of poor quality and could threaten the internal validity of the instruments, and so deletion of items was not necessary. Some nuances of item quality, particularly the alignment of two equivalent versions used in some measurements, still merit discussion.

Subtest of TALL_VL

As revealed in Section 4.4.2.1, four datasets of this subtest fitted well to the Rasch models applied, and no items displayed particular concerns that led to their deletion. However, Version A in the aural suite had the lowest reliability coefficient in ω_h (.31), albeit its coefficients of ω_t (.86) and α (.70) were satisfactory. Version A seemed to have yielded less reliable data compared to Version B in both test suites, and the discrepancy of reliability between the two versions was more apparent in the aural suite than in the written suite.

The possible reason may relate to the design of items. In version A, 8 out of 10 three-letter items end with the same consonant *k*, whereas in version B, 5 out of 10 three-letter items end with the consonant *k*. To discriminate more words that share a same consonant and to retain them in memory could pose extra challenges, especially when the words are presented in the aural modality. This observation is evidenced by the descriptive statistics, in which the mean score of Version A in the aural modality was the lowest, and the discrepancy between the mean scores of two versions was greater in the aural modality than in the written modality. This point should be considered in future refinement of the instrument, important when two versions of stimuli are needed for a counterbalanced design.

In addition, the coefficients of the difficulty parameters suggested that two versions in the aural suite were more challenging to the participants than the versions in the written suite. The total information of test also evidenced that this subtest provided more information about participants whose ability was above the average, especially in the aural suite. Considering that the participants in the current study were expected to have high ability of associative memory, this subtest may need to be revised to decrease the level of difficulty. It could be refined with a simple methodological revision, that is, to increase the time allowed for learning from the current two minutes to three minutes.

Subtest of TALL_SD

In this subtest, Version A may have better item quality than Version B, as the dataset of Version B of TALL_SD had poor model fit to the Rasch models applied. The reason may be due to the larger differences in the parameters of difficulty and discrimination between items in Version B compared to those in Version A. It was revealed that the items of Version B had a wider variability in difficulty, which was particularly reflected in that items having the first diphthong *s**au**ja* were consistently less challenging to the participants compared to the other sounds (*s**i**ja* and *s**ė**ja*) that had single vowels. This variability was not reflected in the dataset of Version A, in which the first diphthong *v**ie**ta* was not perceived as the easiest sound compared to other single-vowel sounds (*v**a**ta* and *v**y**ta*). In addition, four participants had misfitted person fit statistics in Version B, but none had misfitted person fit statistics in Version A.

The results also suggested that the two test versions were not aligned in terms of item difficulty and discrimination though they both displayed satisfactory reliability coefficients and could be used as reliable measurements for the same construct. Therefore, a refinement may be necessary if two versions are needed for a within-subject design.

Subtest of TALL_LA

Most of the datasets from TALL_LA showed good fit to the applied Rasch models, except the dataset of Version B in the written suite. In this case, the results of the χ^2 statistic yielded inconsistent outcomes, with a significant *p* value of .01 in the parametric bootstrap goodness-of-fit test and a non-significant *p* value of .80 in the Andersen likelihood ratio test. This inconsistency can be attributed to the different assumptions underlying these two statistical tests. The parametric bootstrap goodness-of-fit test assumes that the Rasch model is correctly specified and assesses how well the model fits the data, while the Andersen likelihood ratio test compares the Rasch model to a null model that assumes no relationship between the observed responses and the latent ability being measured. It is worth noting that while overall model fit statistics are valuable, especially for model comparison, they may not provide detailed insights when the primary objective is to assess individual items in isolation. In this case, a thorough examination of item quality, as discussed in Section 4.4.2.3, did not reveal any concerning evidence related to any specific item within this subtest.

While the total information curves suggested that the subtest performed better in distinguishing participants with below-average abilities, indicating that it was not challenging for the college undergraduates in this study, this observation is not unexpected or concerning. It is reasonable to anticipate that college undergraduates, typically possessing high analytic abilities, might find certain tasks not challenging. Furthermore, it is crucial to note that the total information provided by this subtest was sufficient, and the reliability and internal validity indices for this subtest were deemed satisfactory in the current study. As a result, there does not appear to be a pressing need for revisions to the test items or experimental design unless additional evidence emerges that warrants refinement.

4.7.2.2 Polytomous datasets from TALL_SNWR and TALL_CST

The results of analyses on the polytomous datasets of the subtests of TALL_SNWR and TALL_CST using the Generalised Partial Credit Models (GPCM) suggested complicated findings on the item quality.

Options of constraints on discrimination parameters

In the analysis of the TALL_SNWR datasets, it was observed that the GPCM failed to converge when discrimination parameters were not constrained equally across trials. Models assuming equal discrimination parameters for all trials exhibited better fit than those assuming unequal discrimination parameters for individual trials. In the case of the four datasets from TALL_CST, the GPCM converged successfully without constraints on equal discrimination parameters, except for the dataset of Version A in the written suite. However, model comparison indicated that models assuming equal discrimination parameters for all trials yielded the best-fit statistics.

These results align with findings reported by Draheim et al. (2018) in their application of the GPCM to other complex span tasks (Operation span, Symmetry span, and Rotation span). The finding in the current study contradicts the assumption that item/trial discrimination would vary across different set sizes. One potential explanation could be that the datasets in this study comprised a relatively small number of data points, limiting their ability to successfully converge to models with more intricate estimations, such as those without constraints on equal discrimination parameters.

Model fit statistics

The results of the parametric bootstrap goodness-of-fit tests showed that, except the dataset of Version B of TALL_CST in the aural suite, *p* values of χ^2 statistics were smaller than .05 (though greater than .01), indicating that in most datasets the observed data significantly deviated from what was expected according to the models. This suggested that the model estimations may not fit the data adequately. This finding echoes the poor model fit (with *p* values lower than .01) reported by Draheim et al. (2018). Given that the overall model fit statistics were useful especially for model comparison, it was not particularly informative when the main purpose of the current study was to check individual items in isolation (Draheim et al, 2018, referring to Kang et al., 2005).

Difficulty parameters

TALL_SNWR

In TALL_SNWR, most difficulty parameter coefficients were greater than 0, indicating that the subtest presented a high level of overall difficulty for the participants in the current study. Total information curves, as shown in Figure 4.26 in Section 4.4.2.4, indicated that the subtest was most discriminative for participants with above-average abilities. Given that the participants were college undergraduates with high abilities, it may be beneficial to consider revising the subtest. This revision could involve removing two trials with a set size of 7, which provided over 80% of information about above-average ability and adding trials with set sizes of 2 to 4. This adjustment would provide more information about below-average ability while maintaining an acceptable level of participant fatigue during the experiment.

The results also revealed that the coefficients of the first category threshold parameters in all trials were consistently lower than those of the latter category threshold parameters in all trial. This pattern indicated that it was easier for participants to get all nonwords recalled incorrectly than to get some nonwords recalled correctly, suggesting that the task was challenging. However, some inconsistent patterns were observed in the trials involving large set sizes (e.g., 6 or 7 stimuli in a single trial), making it unclear whether recalling fewer nonwords correctly was consistently easier than recalling more nonwords correctly in these trials. This variability might be related to a testing strategy possibly employed by the participants: when faced with trials having large set sizes, they may have recorded the nonwords they could recall without adhering to the requirement of recalling the sequence of nonwords. In other words, participants could click the ‘Recording’ button for any nonword they remember, disregarding the order of the nonwords, rather than recalling them in sequence. This testing approach could have been used by participants in the uninvigilated online testing environment, potentially introducing errors in the testing results. Addressing this technical issue merits refinement in future iterations.

TALL_CST

The coefficients of the category threshold parameters of difficulty in the subtest of TALL_CST showed that most category threshold parameters were below 0, suggesting the overall difficulty of the subtest was low in the current study. The total information curves (shown in Figure 4.29 in Section 4.4.2.5) indicated that the discrimination of this subtest was maximised for below average participants. TALL_CST in the written suite provided even lower total information than in the aural suite, and it performed very poorly at discriminating participants whose ability were above the average.

The poor performance of TALL_CST in providing information on participants who performed above the average in this subtest suggested concerns about whether complex span tasks would be appropriate to measure executive control capacity of working memory of the populations with high WM ability. Similar concerns are discussed by Draheim et al. (2018), particularly considering that complex span tasks are widely used to measure participants at the college level who may have high cognitive ability. Complex span tasks used in the current study are essentially the same as operation span tasks discussed by Draheim et al., that is, using English letters as the stimuli for recall and applying the partial span score (i.e., the total number of letters recalled in proper serial position). Letters for recall can be the simplest stimuli, allowing participants to engage in more articulatory rehearsal of the stimuli. Although TALL_CST was designed to include domain-specific tasks that used verbal stimuli (i.e., sentences) for processing, the stimuli were written in participants' L1, which was not challenging for the populations at high L1 literacy level, like the participants in the current study. The processing stimuli of sentences in L1 used in TALL_CST could be even less challenging than the simple arithmetic used in the operation span tasks and are likely to be automatically processed by participants at high literal level of L1. Therefore, the processing part of the complex span tasks is not effective as a distractor to prevent articulatory rehearsal.

This invites reconsiderations about the verbal stimuli used in the paradigm of the listening and reading span tasks. TALL_CST used verbal stimuli in participants' L1 to address the confound of L2 English proficiency in the processing tasks, as most listening and reading span tasks are written in English. This methodological implementation should be kept as the core element in this subtest. What can be refined may be the verbal stimuli for recall, similar to the Klingon characters used in Hicks et al. (2016) and Ruiz et al. (2021).

For example, in the reading span tasks, using verbal stimuli presented in the unfamiliar orthographical forms would be able to suppress articulation and help remove the influence of articulatory rehearsal (Baddeley et al., 1984). Participants may need to retain the orthographical forms in short-term memory and select the stimuli from given multiple options. In this scenario, the storage component of executive control would entail the retrieval of visual representations of novel stimuli, rather than the recall of auditory stimuli. In listening span tasks, using stimuli presented in unfamiliar but still articulatable phonological forms might also be effective. However, it may require technical refinement to record participants' recall of the stimuli, as opposed to having them choose answers from a provided list of visually presented letters.

One straightforward solution could involve technical refinements that include trials with increased set sizes of 8 and 9, as suggested by Draheim, et al. (2018). It is worth noting that Draheim and colleagues reported that the difficulty of trials with 8 and 9 stimuli remained below average ability level, resulting in an overall low level of difficulty. Given the technical arrangement applied in TALL_CST, which necessitates a longer interval between sentence processing and stimuli recall, adding an extra load in retaining stimuli for an extended duration in the aural suite compared to the written suite (as discussed in Section 4.7.1.2), the technical solution of increasing set size merits consideration in future refinement of this subtest, particularly in the aural suite. Furthermore, empirical evidence demonstrating digit span and monosyllabic word span among Mandarin speakers (e.g., Mattys et al., 2018) may provide additional support for the option of increasing set size in span tasks, especially if the subtest targets L1-Chinese participants.

To gain a deeper understanding of the internal validity of CST, it is highly recommended that both the processing component, involving processing accuracy and processing time, and the storage component, be considered and reported (Conway et al., 2005; Shin & Hu, 2020). This comprehensive approach is crucial as both components may contribute substantial unique variance to the results. For instance, Unsworth et al. (2009) used CFA and the Structure Equation Model to examine the relationships among the components of CST and their effectiveness in predicting higher-order cognition (represented by a number of general fluid abilities tests covering different content domains, including spatial, numerical, and verbal). Their findings indicated that CST is multi-faceted task that relies on components offering unique information and are not redundant. This underscores a limitation in the current study's data analysis and the need for future research to employ sophisticated analyses to provide a more detailed breakdown of the variance explained by each component of CST.

4.7.3 Verification of aptitude theoretical frameworks

The validation of TALL as a battery for language aptitude includes the investigation into the explanation inference in the validation plan that relates to whether the data collected by this battery displayed satisfactory fitness to the theoretical frameworks, i.e., the combination of the Stages Approach (Skehan, 2016) and the P/E model (Wen, 2016). The results from PCA confirmed that all subtests representing their corresponding componential constructs of aptitude in both suites of TALL contributed similarly to the first principal component underlying the construct of aptitude, and they did not have much redundancy in factor dimensions that might led to the reduction of any subtest. The results from CFA results provided strong evidence of model fit statistics, indicating that data of both suites fitted well to the hypothesised four-factor model based on the theoretical framework on which TALL was developed.

Using PCA to investigate the interrelationship between the different subtests has been reported in a limited number of studies, which are almost all on the data elicited by the LLAMA tests. Bokander and Bylund (2020) reported the results of PCA that shared a similar component structure to that revealed in literature (e.g., Artieda & Muñoz, 2016; Granena, 2013), that is, the LLAMA subtests loaded on two separate components, with LLAMA D loading on one component and other subtests loading on the other. Although the results from PCA in the current study are not comparable to the results reported by Bokander and Bylund, some points related to their findings suggested by PCA are intriguing.

First, what information can be provided by PCA needs careful consideration. The primary aim of PCA is to reveal the interrelationship between subtests included in aptitude batteries, which underlines the nature of this analysis being exploratory for the factors under the existing data, rather than being confirmatory for testing hypothesis predetermined according to theories. When the two-component or the three-component solution is selected in PCA, it has changed the exploratory nature of PCA. Instead of this approach to using PCA, how many principal components should be extracted from the data comes as the result, not the other way around.

Second, even if the two-component factor structure is validly extracted from PCA, could this lead to the conclusion that these two components represent the two dimensions of implicit and explicit nature of aptitude operationalised in the LLAMA tests (Bokander & Bylund, 2020; Granena, 2013)? The answer is NO, with certainty. An alternative interpretation of the two-component factor structure of the LLAMA tests could be that LLAMA_D is the only subtest that relies on the ability to process phonological forms, whereas the other subtests engage the ability to process stimuli presented visually. In fact, the two-component factor structure can only provide the evidence that LLAMA_D is a subtest essentially different from the other subtests, and it has loaded on a separate dimension from the others. It cannot lead to the conclusion that LLAMA_D is implicit in nature, which has been criticised as a premature conclusion in recent studies (see Iizuka & DeKeyser, 2023; Suzuki, 2021a), nor that the other subtests are explicit.

Third, the current study used the Oblimin method of Oblique factor rotation rather than the Varimax solution used by Bokander and Bylund, as the former allows factors to be correlated, which seems to be plausible for the interrelationships between the subtests and the components of aptitude they represent.

The current study also used CFA to verify the predetermined four-factor structure that is proposed in the Stages Approach. The model fit statistics provide strong support to indicate that the data of both suites fit well to the models. However, running CFA with most latent variables represented by single indicators (i.e., the subtests in this study) may not be a typical practice. Nevertheless, this concern was addressed by choosing the four-factor model after the unifactor baseline model was conducted. Future research may be needed to try out other statistical methods for this verification purpose.

CHAPTER 5: RESULTS & DISCUSSION FOR RQ2 – EFFECTS OF MODALITY

5.1 Introduction

This chapter presents the results to answer RQ2: *To what extent does input modality have effects on participants' scores in the subtests of TALL_VL, TALL_LA, and TALL_CST?* The prediction relevant to this RQ was that participants could perform better in these subtests if the stimuli were presented in the written form because the stimuli in the written form would be easily coded by the visual processing mechanism in comparison with the stimuli input in the aural form. The first section of this chapter introduces the rationale of using Mixed-effects Modelling (MEM) as the statistical method for data analysis. It also includes the summary of model parameters in applying MEM. The second section presents the results of MEM analyses on the unaggregated data of three subtests (i.e., TALL_Vocabulary Learning, TALL_Language Analysis, and TALL_Complex Span Task) that were administered in two input modalities. The third section summarises the results in this chapter, which leads to the final section of discussion about the results and findings in relation to the effect of modality in measuring aptitude. All results and analysis code in this chapter were rendered in an R markdown file (see Appendix D).

5.2 Mixed-effects Modelling (MEM)

MEM was used to answer this research question. MEM has its significant advantage in modifying or enriching generalised linear models in terms of the assumption that data points need to be statistically independent of one another. The statistical merit of MEM is to deal with data points that might be related when participants in the experimental design produce repeated responses (i.e., the same participant responds to many different items), and when specific versions of testing materials with counterbalanced stimuli or items that might share multiple characteristics (e.g., many participants respond to the same stimuli) (Gries, 2021). Although repeated measures ANOVA, which may take into account person or item-level variability, have been used to analyse the data obtained from participants responding to many test items, the analytic techniques have several drawbacks. Since they cannot simultaneously account for both sources of variability, observations within a condition must be compressed over either items or persons. As such, the statistical power of the study, or the possibility of identifying an existing effect, is reduced when essential information about variability among individuals or items may get lost during data aggregation (Barr, 2008). This limitation of data-being-independent assumption can be addressed by LMM that takes the dependencies in the data into account. Furthermore, other than indicating whether an effect is significant or not by ANOVA, MEM provides information about the magnitude or direction of the effect by offering individual coefficient estimates for each predictor that show growth or trajectory (Brown, 2021).

RQ2 aimed to investigate the effect of input modality as the predictor variable on the scores of three subtests (i.e., TALL_VL, TALL_LA, and TALL_CST) that were administered in both aural and written modalities. As introduced in Section 3.3.4, participants took two rounds of TALL in a repeated design, with order of the test suites (aural and written) and material versions (A and B) counterbalanced. The accuracy of their response to the test items/trials in three subtests administered in both modalities were the dependant variable in MEM analysis. Input modality was the primary fixed effect in MEM on the test scores (as reviewed in Section 2.2.2.3). Additionally, previous research has shown that repeated exposure to the experimental paradigm may potentially improve participants performance due to the familiarity participants gained with the test procedure and paradigm (Suga & Loewen, 2023; Suzuki & Koizumi, 2020). Therefore, test session could be a potential source of variability that should also be taken into account when analysing the data collected in the repeated experimental design. This hypothesised fixed effect could also be postulated from the eyeballing of the violin box plots presented in Section 4.3.1.

Given the fact that convenience samples were used in the current research and the materials developed would not exhaust all possible options, test scores might vary across levels of the grouping factors of subjects and items. More importantly, data were collected in the current research when all participants were tested more than once and when each item/trial were measured more than once. This suggested a crossed random-effects structure of the current research design, that is, the multiple repeated-measurements structure coexisted in the single experimental design. Therefore, by-subject and by-item are both sources of variation that need to be counted in as random effects in a single model, which can be achieved in MEM (Baayen et al., 2008; Gries, 2021). Furthermore, individual items/trials were related in a specific version of materials, hence they shared extra organised characteristics, known as nested random effects (Gries, 2021).

To summarise, multiple factors (i.e., input modality, test session, participants and test items nested in the material version) in the current research need to be taken into account simultaneously when investigating their effects on the test performance, and MEM helps to consider these multiple factors when repeated measurements were used in a way that ANOVA could not.

MEM analyses were performed in the following steps, using data from TALL_VL, TALL_LA, and TALL_CST, respectively. These steps followed the practice guidance for linear MEM in psychological science (e.g., Meteyard & Davies, 2020) and the general overview of a MEM process with specific example in linguistics (e.g., Gries, 2021, Winter, 2013). The software used was the `lme4` package (Bates, et al., 2015) in R.

5.2.1 Data preparation and assumption checks

The first step of MEM analyses involved exploring and preparing data for analyses. Data used in the MEM analysis were wrangled into unaggregated long format through the tidyverse package, in which each row represented an individual observation that did not aggregate across either participants or test items/trials.

Prior power analysis had been conducted (see Section 3.3.1.2) to show that 67 participants were needed to facilitate a within-subject design with two rounds of test. In the current step, the total sample size should generate 900 to 2500 data points for the MEM analysis, as suggested by Meteyard and Davies (2020) for psychological research.

In addition, correlation between the results from two modalities was analysed after the normality check of distribution on each dataset. Non-normality of data distribution of the results from all three subtests indicated that correlation coefficients (i.e., Kendall's *tau*) should be used on the non-parametric ranks of the data (Field et al., 2012).

Assumption checks are essential to ensure that the necessary conditions are met when conducting a meaningful analysis using MEM. First, **Linearity** of data was checked through the residual plot of a *simple regression formula* of the fixed effects. For categorical binary data from TALL_VL and TALL_LA, the residual plot of linearity indicated the pattern of two lines, while for continuous data from TALL_CST, the residual plot did not indicate a nonlinear or curvy pattern. Second, **absence of collinearity** of the fixed effects were checked. It was to ensure that the multiple predictors were not highly correlated, in other words, they were not very similar to each other. Otherwise, the interpretation of the model became unstable as it would be very difficult to decide which predictor played a big role (Winter, 2013). This assumption was checked by calculating variance inflation factors (VIFs), with value less than 5 for each predictor indicating that multicollinearity was not a significant issue (Fox, 2020). Third, **homoskedasticity**, which refers to the approximately equal variability of the data across the range of predicted values, was assessed using the residuals of the model. Ideally, the residual plot should exhibit a blob-like pattern of the data points. Forth, **the normality of residuals** assumption was checked for the continuous data from TALL_CST. It is worth noting that this assumption is considered the least important, as linear models are relatively robust against violations of normality assumption (Winter, 2013). Finally, **the absence of influential data points** was checked to detect any that could drastically change the results. This assessment was performed using the dfbeta function in R on the regression models, which yielded 'leave-one-out diagnostics' (p. 19). In this approach, coefficients needed to be adjusted if a specific data point was excluded. If any value caused the slope to change sign, that data point was identified as influential and required special attention. Another approach involved visually inspecting values to identify any differed by at least half of the absolute value of the slope (Winter, 2013).

5.2.2 Model selections

This step first involved the selection of random effects according to the experimental design. As mentioned above, the current research had a crossed random-effects structure from the multiple repeated-measurements, therefore, subjects and test items should be considered as the random effects. Given that individual items/trials were related in a specific version of materials, material version could be accounted in the model as a nested random effect. However, the complex random effect structures may inhibit model convergence, therefore the selection of subject and material version as the random effects was used as a way of simplifying model structure to deal with any convergence problems.

To enhance transparency of the practice in this step, a table was used, adapted from a supplementary example in Meteyard and Davies (2020), to document the model comparison and the model building/selection process. This documentation included the specifications and statistics of all models, as well as the approach taken to address convergence issues. The rationale for employing model simplification methods was also included in this step, with the results of comparison methods (such as Likelihood-Ratio-Test, AIC, BIC) being reported.

The model building process started by establishing the model with all random effects and subsequently simplifying it by removing random effects to check whether the model fit improved. In essence, this involved a process from a maximal model to a minimal model that converges, with the aim of selecting the most appropriate random effects for an improved model fit. Once the model with selected random effects was established, fixed effects were added to construct the main effects models. These models incorporated all fixed effects, both with and without interactions, along with random effects featuring random intercepts and random slopes corresponding to each fixed effect in the models. It is important to note that random slopes should be included in the models, as it is reasonable to assume that the responses from participants or items to the fixed effects (modality and session) did not follow precisely the same pattern.

5.2.3 Model interpretation

This step presented the output of the final selected model, i.e., the model that did not have convergence issues and had the best fit. The results are provided in a table with parameters estimates for the fixed effects (coefficients, standard errors with confidence intervals, and associated test statistics and *p*-values) and random effects (intercepts and slopes). The model's performance was summarised using the marginal R^2 and the conditional R^2, necessary to identify and estimate sources of variance in regression models (Plonsky & Oswald, 2017). The marginal R^2 showed the proportion of variance explained by the fixed effects alone in the model, while the conditional R^2 indicated the proportion of variance explained by both the fixed and random effects. In addition, to obtain an intuitive interpretation of magnitude of effect sizes, both marginal R^2 and conditional R^2 were converted to correlation coefficient of *r*. This allowed comparisons to the benchmarks (i.e., *r* close to .25 is considered small, .40 medium, and .60 large) suggested by Plonsky and Oswald (2014) for interpreting effect sizes in L2 research.

5.3 Results of MEM analyses

Generalised Linear Mixed-effects Model (GLMM) was applied to the binary data from TALL_VL and TALL_LA through the `glmer` function in the `lme4` package, while Linear Mixed-effects Model (LMM) was applied to the continuous data from TALL_CST through the `lmer` function.

Assumption checks were performed on three subtests separately. All subtests had the total sample sizes that generated data points significantly surpassing the range of 900 to 2500 data points recommended by Meteyard and Davies (2020). This suggested that the sample size in the current study did not indicate any statistical power issue in applying GLMM or LMM.

Correlation coefficients of Kendall's *tau* (.16 for TALL_VL, .32 for TALL_LA, and .20 for TALL_CST) indicated positive correlations between the scores obtained in two modalities across all three subtests, reaching a significant level. However, based on Cohen's (1988) general benchmarks for effect sizes—where *r* = .1 signifies a small effect, .3 a medium effect, and .5 a large effect—the observed effect sizes of correlations were within the small to medium magnitude range. Alternatively, considering the benchmarks suggested by Plonsky and Oswald (2014) for interpreting findings in L2 research, these effect sizes might be interpreted at a lower magnitude.

The results from assumption checks indicated that the linearity of data was evidenced. The residual plot of linearity showed the pattern of two lines for TALL_VL and

TALL_LA, and the pattern of four lines for TALL_CST. All three subtests had the VIFs values for modality and session (both were 1.00) smaller than 5, indicating that these two predictor variables were not highly correlated with each other. This indicated that the assumption of absence of collinearity was met. The residual plot did not exhibit a clear pattern indictive of homoskedasticity for TALL_VL and TALL_LA. While the datasets met statistical power requirements for GLMM, their sample sizes might be limited and affected this assessment. The evidence of homoskedasticity for TALL_CST was clear as the variability of the data were approximately equal across the range of predicted values. No influential data points were detected in the datasets from all three subtests. In general, the assumption checks on a model of the simple regression formula did not raise issues that need special attention prior to GLMM or LMM analysis for the datasets from all three subtests.

5.3.1 TALL_VL

Model selection

Model selection process (see Table 5.1) started from the comparisons of null models with random effects only. The results from AIC, BIC, and χ^2 with *p* values of the Likelihood-Ratio-Test suggested the Reduced Null 1 model having the best fit among null models. Two fixed effects (modality and session) were added into the chosen null model to build two main-effects models. The Main Effects 1 model included random slopes for both fixed effects, while the Main Effects 2 model introduced an interaction between these two fixed effects. Both main-effects models converged. Initially, the Main Effects 1 model was compared to the Reduced Null 1 model using a likelihood-ratio test, indicating a better model fit to the data. Subsequently, the two main-effects models were compared to each other, with the likelihood-ratio test revealing that the Main Effects 1 model exhibited lower AIC and BIC values. However, the difference between the two models was statistically insignificant. Considering the lack of statistical evidence for the interaction between the fixed effects ($\hat{\beta}$ = 0.001, SE = 0.29, z = 0.004, *p* = .997) in the Main Effects 2 model, the final model selected was the Main Effects 1 model. This choice enables the interpretation of individual fixed effects as isolated predictors (Brown, 2021).

The final selected GLMM was:

```
accuracy ~ 1 + modality + session + (1 + modality | subject) + (1 + session
| subject) + (1 + modality | item) + (1 + session | item)
```

The output of analysis on the final chosen model (see Table 5.2) showed the parameters estimates for the fixed effects and random effects, as well as the model performance.

Fixed effects

Examination of the output for the fixed effects in the model showed that the regression coefficient for the intercept was -1.08, indicating the average expected value when all predictor variables (modality and session) were zero. The *p*-value was below the alpha level of .001, indicating that the intercept was statistically different from zero. Furthermore, the analysis indicated that scores, on average, were estimated to be 0.71 points higher in the written modality compared to the aural modality ($\widehat{\beta}$ = 0.71, SE = 0.13, z = 5.60), with a *p*-value less than .001. Additionally, scores were estimated to be 0.29 points higher when participants took the subtest in the second session compared to their performance in the first session ($\widehat{\beta}$ = 0.29, SE = 0.10, z = 2.81), with a *p* value of 0.005.

Additionally, Table 5.2 includes odds ratios (ORs) derived from exponentiating the coefficients of the chosen GLMM. These ORs can be interpreted as the ratio of the odds of an event occurring in one group compared to the odds of the same event in another group. In this study, ORs represented the odds of correctly answering test items with a one-unit increase in the fixed effects, such as switching from the aural to the written modality or from the first test session to the second session. The results showed an OR of 2.04 for the effect of modality on accuracy, suggesting that switching from the aural to the written modality increased the odds of a correct response by a factor of 2.04 (equivalent to a 104% increase), while holding the session constant. Similarly, the OR for the effect of session on accuracy was 1.34, indicating that taking the test in the second session increased the odds of a correct response by a factor of 1.34 (equivalent to a 34% increase), while keeping the modality constant. Both predictor variables had statistically significant impacts on the test outcomes.

Table 5.1 The model selection process (TALL_VL)

Sampling Units		N total observations = 6600 N Subjects = 165; N items = 40										
Model specification	**Model name**	**Nested / simpler Model**	**Fixed Effects added**		**Random Effects**		**Model fit**				**LRT Test against nested**	
					Subjects	**Items**	**AIC**	**BIC**	**LL**	**df**	**df**	**χ^2**
RE only	Null 1: subject + item + version/item	-	-		intercepts	intercepts	unidentification warning – simplify the model by removing version/item				-	-
	Null 2: subject x item	-	-		“	“	Unidentification issue – simplify the model by removing the subject/item interaction				-	-
	Reduced null 1: subject + item	Reduced null 2: subject + version	-		“	“	7967.8	7988.2	–3980.9	6597	0	424.57
	Reduced null 1: subject + item	Reduced null 3: subject	-		“	“	7967.8	7988.2	–3980.9	6597	1	470.12***
	Reduced null 1: subject + item	Reduced null 4: item	-		“	“	7969.8	7988.2	–3980.9	6597	1	421.79***
FE main effects	Main effects 1	Reduced null 1: subject + item	modality + session		Slopes for 2 FEs	Slopes for 2 FEs	7613.3	7715.2	–3791.6	6585	12	378.49***
FE two-way interactions	Main effects 2	Main effects 1	modality x session		“	“	7615.3	7724.0	–3791.6	6584	1	0

Notes: **RE** – Random effect; **FE** – Fixed effect; **AIC** – Aikake Information Criterion; **BIC** – Bayesian Information Criterion; **LL** – Log Likelihood; **df** – degrees of freedom; **LRT** – Likelihood Ratio Test; **χ^2**– Chi-square. A “ in the table cell indicates no changes from the previous model.

Table 5.2 Results from final selected model (TALL_VL)

	Fixed Effects				
	Log-Odds	**SE**	**95% CI**	**z**	***p***
Intercept	–1.08	0.14	[–1.37, –0.89]	–7.53	< .001
Modality	0.71	0.13	[0.46, 0.96]	5.60	< .001
Session	0.29	0.10	[0.09, 0.49]	2.81	0.005

	Odds Ratios	**95% CI**	***p***
Intercept	0.34	[0.26, 0.45]	< .001
Modality	2.04	[1.59, 2.61]	< .001
Session	1.34	[1.09, 1.64]	0.005

	Random Effects		
	Variance	**SD**	**Correlation**
Modality | Subject (intercept)	0.03	0.19	
Modality | Subject (slope)	0.60	0.77	-0.11
Session | Subject (intercept)	0.38	0.61	
Session | Subject (slope)	0.54	0.73	0.11
Modality | Item (intercept)	0.30	0.55	
Modality | Item (slope)	0.25	0.50	-0.28
Session | Item (intercept)	0.28	0.53	
Session | Item (slope)	0.02	0.16	-0.44

	Model performance	
	Marginal	Conditional
R^2	0.037	0.2

Notes: *p*-values for fixed effects calculated using Satterthwaites approximations. Confidence Intervals have been calculated using the Wald method.

Random effects

The estimates of random intercepts and slops provided insights into the extent of score variations among participants and items relative to the fixed effects.

Specifically, the SD (0.19) for the modality-by-subject random intercepts indicated that scores for participants derived from the average intercept (-1.08) by approximately 0.19 points. The SD (0.77) for modality-by-subject random slope indicated that participants' estimated slopes deviated from the average slope of 0.71 by approximately 0.77 units. Consequently, an individual participant with a slope 1 SD below the mean (0.71 – 0.77 = -0.05) would have an estimated slope near 0, suggesting that this person's scores were not affected much by the modality in which the test items were presented. Conversely, an individual participant with a slope 1 SD above the mean (0.71 + 0.77 = 1.48) would have a steeper slope, indicating a difference of approximately 1.48 units in scores between modalities.

Similarly, the SD (0.61) for the session-by-subject random intercepts indicated that scores for participants derived from the average intercept (-1.08) by approximately 0.61 points, larger than derivation related to modality. The SD for the session-by-subject random slopes (0.73) indicated that participants' estimated slopes deviated from the average slope of 0.29 by approximately 0.73 units. Consequently, an individual participant with a slope 1 SD below the mean would have an estimated slope of 0.44, indicating that this person's scores were affected by the session in which the test was taken. Conversely, an individual participant with a slope 1 SD above the mean would have a steeper slope, indicating a difference of approximately 1.02 units in scores between the two test sessions.

The results also provided information about participants' variability across modalities and sessions, i.e., the random effect of 'subject'. The variance for the session-by-subject intercepts (0.38) with an SD of 0.61 was considerably larger than that for the modality-by-subject intercepts (0.03) with an SD of 0.19. This discrepancy suggested that differences in participants' performance in the first session were much larger than differences related to the aural modality. However, the modality-by-participant random slopes (0.77) closely matched the session-by-participant random slopes (0.73), indicating that this cohort of participants exhibited similar variability when comparing their performance across two modalities and two sessions.

In terms of the variability of test items across modalities and sessions, i.e., the random effect of 'item', the results revealed that the modality-by-item intercepts (0.30) with the SD (0.55) was close to the session-by-item intercepts (0.28) with the SD (0.53), indicating that the variance of items' scores in the aural modality were similar to those in the first session. However, the variance of modality-by-item random slopes (0.25) was larger than that of the session-by-item random slopes (0.02), indicating that scores in relation to items were more affected by modality condition than by session.

The output also included *correlations* among random effects. Specifically, the correlation between modality-by-subject intercepts and modality-by-subject slopes was -0.11, suggesting a moderate negative relationship between these two indices in the model. This result indicated that participants who had higher scores (random intercepts) in the baseline responses (the aural modality) tended to exhibit a shallower (more negative) slope in their performance when they switched to the written modality, although this relationship is not very strong due to the weak correlation coefficient. In other words, when compared to their counterparts at lower ability level in the aural modality, performance of participants with higher ability was slightly less influenced by the change of modality condition. In addition, the correlation between session-by-subject intercepts and session-by-subject random slopes was 0.11, indicating a modest positive relationship between these two indices. It suggested that, when compared to their counterparts at lower ability level in the first session, the performance of participants with higher ability was slightly more influenced by the change of session.

Furthermore, the correlation between modality-by-item intercepts and modality-by-item slopes was -0.28, indicating a moderate negative relationship between these two indices. The result suggested that items with more accurate responses in the aural modality were less influenced by the change of modality compared to items with less accurate responses. Similarly, the correlation between session-by-item intercepts and session-by-item random slopes was -0.44, indicating a negative relationship between these two components. The result suggested that items with more accurate responses in the first session were much less affected by the change of session condition. However, the magnitude of the negative influence on item responses related to the session condition was greater than that related to modality.

Model summary

Finally, the performance of the final selected model was summarised by the marginal R^2 and conditional R^2. The marginal R^2 of 0.037 indicated that only a small proportion of variance (about 3.7%) in the outcome of this subtest was explained by the fixed effects of modality and session alone. Converting this to *r*, the correlation coefficient (*r* = .19) suggested a negligible effect size according to the benchmarks proposed by Plonsky and

Oswald (2014). However, the conditional R^2 was 0.2, indicating that a larger proportion of variance (about 20%) was explained by both the fixed and the random effects, and the effect size converted (*r* = .45) was a medium effect size according to the benchmarks. This suggested that taking the fixed effects and the random effects together into account could explain a substantially larger amount of variance in the outcome than only considering the fixed effects.

In summary, GLMM on the data from TALL_VL evidenced statistically significant effects of the fixed effects (i.e., modality and test session) on participants' performance. Specifically, when tested in the written modality, or when they took the test in the second time, participants obtained significantly higher scores. In addition, participants' variability across modality and session was evidenced, and their scores varied similarly when they were tested across modality and session conditions. Conversely, the degree of variance in test items was larger across modality than session. The results also suggested that participants who had better ability in learning vocabulary items in the aural modality tended to be negatively affected by the change of modality condition but more affected by the change of session condition. However, the strength of the relationships was relatively weak. A substantial proportion of variance (approximately 20%) was accounted for by both the fixed and the random effects, in contrast to the relatively negligible proportion of variance (approximately 3.7%) explained by the fixed effects alone.

5.3.2 TALL_LA

Model selection

The model selection process (see Table 5.3) started from the comparisons of null models with random effects only. The results from AIC, BIC, and χ^2 with *p* values of the Likelihood-Ratio-Test suggested the Reduced Null 1 model the best fit among null models. Two fixed effects (modality and session) were added to this selected null model to build two main-effects models. The Main Effects 1 model included random slopes for both fixed effects, while the Main Effects 2 model introduced an interaction between these two fixed effects. Both main-effects models converged. Initially, the Main Effects 1 model was compared to the Reduced Null 1 model using a likelihood-ratio test, indicating a better model fit to the data. Subsequently, the two main-effects models were compared to each other, with the likelihood-ratio test revealing that the Main Effects 1 model exhibited lower AIC and BIC values. However, the difference between the two models was statistically insignificant. Considering the lack of statistical evidence for the interaction between the fixed effects ($\widehat{\beta}$ = -0.31, SE = 0.53, z = -0.58, *p* = .56) in the Main Effects 2 model, the final model selected was the Main Effects 1 model to allow the interpretation of the individual fixed effects as predictors in isolation.

The final selected model was:

```
accuracy ~ 1 + modality + session + (1 + modality | subject) + (1 + session
| subject) + (1 + modality | item) + (1 + session | item)
```

The output of analysis on the final selected model (see Table 5.4) showed the parameters estimates for the fixed effects and random effects, as well as the model performance.

Fixed effects

Examination of the output for fixed effects in the model showed that the regression coefficient for the intercept was 0.85, indicating the average expected value for the score when all the predictor variables (modality and session) were zero. The *p* value was below the alpha level of .001, indicating that the intercept was statistically different from zero. Furthermore, the analysis indicated that scores, on average, were estimated to be 1.27 points higher in the written modality compared to the aural modality ($\widehat{\beta}$ = 1.27, SE = 0.17, z = 7.46), with a *p*-value less than .001. Additionally, scores were estimated to be 0.61 points higher when participants took the subtest in the second session compared to their performance in the first session ($\widehat{\beta}$ = 0.61, SE = 0.17, z = 3.68), with a *p* value < .001.

Additionally, results in Table 5.4 showed an OR of 3.54 for the effect of modality on accuracy, indicating that switching from the aural to the written modality increased the odds of a correct response by a factor of 3.54 (i.e., 254%), while holding the session constant. Similarly, the OR for the effect of session on accuracy was 1.85, indicating that the odds of a correct response increased by a factor of 1.85 (i.e., 85%), while holding the modality constant. Both predictor variables had statistically significant effects on the test outcome.

Table 5.3 The model selection process (TALL_LA)

Sampling Units		N total observations = 9900 N Subjects = 165; N items = 60										
Model specification	**Model name**	**Nested / simpler Model**	**Fixed Effects added**		**Random Effects**		**Model fit**				**LRT Test against nested**	
					Subjects	**Items**	**AIC**	**BIC**	**LL**	**df**	**df**	**χ^2**
RE only	Null 1: subject + item + version/item	-	-		intercepts	intercepts	Singularity issue – simplify the model by removing version/item				-	-
	Null 2: subject x item	-	-		“	“	Unidentification issue – simplify the model by removing the subject/item interaction				-	-
	Reduced null 1: subject + item	Reduced null 2: subject + version	-		“	“	9230.2	9251.8	-4612.1	9897	0	142.27
	Reduced null 1: subject + item	Reduced null 3: subject	-		“	“	9230.2	9251.8	-4612.1	9897	1	142.27***
	Reduced null 1: subject + item	Reduced null 4: item	-		“	“	9230.2	9251.8	-4612.1	9897	6	2072.6***
FE main effects	Main effects 1	Reduced null 1: subject + item	modality + session		Slopes for 2 FEs	Slopes for 2 FEs	8205.1	8313.1	-4087.5	9885	12	1049.2***
FE two-way interactions	Main effects 2	Main effects 1	modality x session		“	“	8206.8	8322.0	-4087.5	9885	1	0.32

Notes: **RE** – Random effect; **FE** – Fixed effect; **AIC** – Aikake Information Criterion; **BIC** – Bayesian Information Criterion; **LL** – Log Likelihood; **df** – degrees of freedom; **LRT** – Likelihood Ratio Test; **χ^2** – Chi-square. A “ in the table cell indicates no changes from the previous model.

Table 5.4 Results from final selected model (TALL_LA)

Fixed Effects	Log-Odds	SE	95% CI	z	*p*
Intercept	0.85	0.16	[0.52, 1.17]	5.15	< .001
Modality	1.27	0.17	[0.93, 1.60]	7.46	< .001
Session	0.61	0.17	[0.29, 0.94]	3.68	< .001

	Odds Ratios	95% CI	*p*
Intercept	2.33	[1.69, 3.22]	< .001
Modality	3.54	[2.54, 4.94]	< .001
Session	1.85	[1.33, 2.57]	< .001

Random Effects	Variance	SD	Correlation
Modality | Subject (intercept)	1.44	1.20	
Modality | Subject (slope)	1.65	1.28	-0.23
Session | Subject (intercept)	1.08	1.04	
Session | Subject (slope)	1.86	1.36	-0.29
Modality | Item (intercept)	0.10	0.32	
Modality | Item (slope)	0.04	0.20	-0.42
Session | Item (intercept)	0.17	0.41	
Session | Item (slope)	0.01	0.09	-0.25

Model performance	Marginal	Conditional
R^2	0.089	0.433

Notes: *p*-values for fixed effects calculated using Satterthwaites approximations. Confidence Intervals have been calculated using the Wald method.

Random effects

The SD (1.20) for the modality-by-subject random intercepts indicated that scores for participants derived from the average intercept (0.85) by approximately 1.20 points. The SD (1.28) for modality-by-subject random slope indicated that participants' estimated slopes derived from the average slope of 1.26 by approximately 1.28. Consequently, an individual participant with a slope 1 SD below the mean (1.26 – 1.28 = -0.02) would have an estimated slope near 0, suggesting that this person's scores were not affected much by the modality condition. Conversely, an individual participant with a slope 1 SD above the mean (1.26 + 1.28 = 2.54) would have a steeper slope, indicating a difference of approximately 2.54 points in scores between modalities.

Similarly, the SD (1.04) for the session-by-subject random intercepts indicated that scores for participants derived from the average intercept (0.85) by approximately 1.04 points, similar to the derivation related to modality. The SD (1.36) for session-by-subject random slopes indicated that participants' estimated slopes derived from the average slope of 0.61 by about 1.36. Consequently, an individual participant with a slope 1 SD below the mean would have an estimated slope of -0.75, indicating that this person's scores were negatively affected by the session in which the test was taken. Conversely, an individual participant with a slope 1 SD above the mean score would have a steeper slope, indicating a difference of approximately 1.96 units in scores between the two test sessions.

The results also provided information about the random effect of 'subject'. The variance for modality-by-subject intercepts (1.44) with an SD of 1.20 was slightly greater than that for the session by-subject intercepts (1.08) with the SD (1.04). This discrepancy suggested that differences in participants' performance in the aural modality were slightly larger than that in the first session. Modality-by-subject random slopes (1.65) was slightly lower than session by-participant random slopes (1.86), indicating the performance of this cohort of participants exhibited slightly less variability when comparing their performance across modalities than sessions.

In terms of the information about the random effect of 'item', the results revealed that modality-by-item intercepts (0.10) was close to session-by-item intercepts (0.17), indicating the variance of items' scores in the aural modality was similar to that in the first session. Similarly, modality-by-item random slopes (0.04) was close to session-by-item random slopes (0.01), indicating scores in relation to items were minimally affected across both modality and session conditions.

The correlation between modality-by-subject intercepts and modality-by-subject random slopes was -0.23, indicating a negative relationship with a moderate strength between these two indices. This result suggested that participants who had higher scores in the aural modality tended to exhibit a steeper decline in their performance when they switched to the written modality, and this relationship is moderate. In other words, when compared to their counterparts at lower ability level in the aural modality, performance of participants with higher ability was less influenced by the change of modality condition. Similarly, the correlation between session-by-participant intercepts and session-by-participant random slopes was -0.29, indicating a moderate negative relationship. It suggested that participants who had better performance in the first session exhibited a steeper decline in their performance when they took the subtest in the second session. In other words, when compared to their counterparts achieved lower scores in the first session, the performance of participants with higher ability were less influence by session condition.

Furthermore, the correlation between modality-by-item intercepts and modality-by-item random slopes was -0.42, indicating a strong negative relationship between these two indices. The result suggested that items with more accurate responses in the aural modality were much less influenced by the change of modality compared to items with less accurate responses. Similarly, the correlation between session-by-item intercepts and session-by-item slopes was -0.25, indicating a moderate negative relationship. The result suggested that items with more accurate responses in the first session were less affected by the change of session condition. The magnitude of the negative influence on item responses related to the modality condition was greater than that related to session.

Model summary

Finally, the performance of the final selected model was summarised by the marginal R^2 and conditional R^2, suggesting that taking the fixed effects and the random effects together into account could explain a substantially larger amount of the variability in the outcome than only considering the fixed effects. Specifically, the marginal R^2 of 0.089 indicated that only a small proportion of variance (about 8.9%) in the outcome of this subtest was explained by the fixed effects of modality and session alone. The related correlation coefficient ($r = .30$) suggested a small to medium effect size according to the benchmarks proposed by Plonsky and Oswald (2014). However, the conditional R^2 was 0.43, indicating that a much larger proportion of variance (about 43%) was explained by both the fixed and the random effects, and the effect size converted ($r = .66$) was large according to the benchmarks.

In summary, GLMM on the data from TALL_LA subtest evidenced statistically significant effects of the fixed effects (i.e., modality and test session) on participants' performance. Specifically, when tested in the written modality, or when they took the test in the second time, participants obtained significantly higher scores. In addition, participants' variability across modality and session was evidenced, and their scores varied similarly when they were tested across modality and session conditions. The results also suggested that participants who had better ability in learning grammatical rules in the aural modality tended to be negatively affected by the modality and the session conditions. The strength of the relationships was moderate. A much larger proportion of variance (about 43%) was explained by both the fixed and the random effects, in contrast to the relatively small proportion of variance (about 8.9%) explained by the fixed effects only, although the effect size of the fixed effects was small to medium in magnitude.

5.3.3 TALL_CST

Model selection

The model selection process (see Table 5.5) started from the comparisons of null models with random effects only. The results from AIC, BIC, and χ^2 with *p* values of the Likelihood-Ratio-Test suggested the Reduced Null 1 model the best fit among null models. Two fixed effects (modality and session) were added to this selected null model to build five main-effects models, among which two models converged. They were the Main Effects 4 model including random slopes for modality by two random effects and the Main Effects 5 model including fixed effects interaction based on the Main effects 4. Initially, the Main Effects 4 model was compared to the Reduced Null 1 model using a likelihood-ratio test, indicating a better fit to the data. Subsequently, the two converged main-effects models were compared to each other, with the likelihood-ratio- test revealing that the Main Effects 5 model exhibited lower AIC and BIC values. However, the difference between the two models was statistically non-significant. Considering the lack of statistical evidence for the interaction between the fixed effects ($\widehat{\beta}$ = -0.02, SE = 0.03, t = -0.61, *p* = .55) in the Main Effects 5 model, the final model selected was the Main Effects 4 model to allow the interpretation of the individual fixed effects as predictors in isolation.

The final selected model was:

```
accu_rate ~ 1 + modality + session + (1 + modality | subject) + (1 +
modality | trial)
```

The output of analysis on the final selected model (see Table 5.6) showed the parameters estimates for the fixed effects and random effects, as well as the model performance.

Table 5.5 The model selection process (TALL_CST)

Sampling Units		N total observations = 4950 N Subjects = 165; N trials = 30										
Model specification	**Model name**	**Nested / simpler Model**	**Fixed Effects added**		**Random Effects**		**Model fit**				**LRT Test against nested**	
					Subjects	**Trials**	**AIC**	**BIC**	**LL**	**df**	**df**	**χ^2**
RE only	Null 1: subject + trial + version/trial	-	-		intercepts	intercepts	Failed to converge – simplify the model by removing version/trial				-	-
	Null 2: subject x trial	-	-		“	“	Could not be built – simplify the model by removing the subject/item interaction				-	-
	Reduced null 1: subject + trial	Reduced null 2: subject + version	-		“	“	-902.29	-876.26	455.15	4949	0	605.46
	Reduced null 1: subject + trial	Reduced null 3: subject	-		“	“	-902.29	-876.26	455.15	4949	1	606.11***
	Reduced null 1: subject + trial	Reduced null 4: trial	-		“	“	-902.29	-876.26	455.15	4949	1	360.79***
FE main effects	Main effects 1	-	modality + session		Slopes for 2 FEs	Slopes for 2 FEs	Singularity issue – simplify the model by removing the random slope for session by trial				-	-
	Main effects 2	-	modality + session		“	Slope for modality	Unidentification issue – simplify the model by removing the random slopes for FEs by trial				-	-
	Main effects 3	-	modality + session		“	intercepts	Failed to converge – simplify the model by removing the random slopes for session by REs				-	-
	Main effects 4	Reduced null 1: subject + trial	modality + session		Slope for modality	Slope for modality	-1227.61	-1162.54	623.81	4947	6	337.32***
FE two-way interactions	Main effects 5	Main effects 4	modality x session		“	“	-1226.0	-1154.4	623.99	4946	1	0.37

Notes: **RE** – Random effect; **FE** – Fixed effect; **AIC** – Aikake Information Criterion; **BIC** – Bayesian Information Criterion; **LL** – Log Likelihood; **df** – degrees of freedom; **LRT** – Likelihood Ratio Test; **χ^2** – Chi-square. A “ in the table cell indicates no changes from the previous model.

Table 5.6 Results from final selected model (TALL_CST)

	Fixed Effects				
	Est/Beta	**SE**	**95% CI**	**t**	***p***
Intercept	0.81	0.02	[0.77, 0.84]	46.42	< .001
Modality	0.06	0.01	[0.03, 0.09]	9.95	< .001
Session	0.02	0.01	[–0.00, 0.04]	3.45	0.074

	Random Effects		
	Variance	**S.D.**	**Correlation**
Modality | Subject (intercept)	0.01	0.12	
Modality | Subject (slope)	0.01	0.12	–0.81
Modality | Trial (intercept)	0.01	0.08	
Modality | Trial (slope)	0.00	0.04	0.06

	Model performance	
	Marginal	Conditional
R^2	0.017	0.312

Notes: *p*-values for fixed effects calculated using Satterthwaites approximations. Confidence Intervals have been calculated using the Wald method.

Fixed effects

Examination of the output for fixed effects in the model showed that the regression coefficient for the intercept was 0.81, indicating the average expected value for the score when all the predictor variables (modality and session) were zero. The *p*-value was below the alpha level of .001, indicating that the intercept was statistically different than zero. Furthermore, the analysis indicated that scores, on average, were estimated to be 0.06 units higher in the written modality compared to the aural modality ($\widehat{\beta}$ = 0.06, SE = 0.01, *t* = 9.95), with a *p* value below the alpha level of .001. Additionally, scores were estimated to be 0.02 higher when participants took the test in the second session compared to their performance in the first session ($\widehat{\beta}$ = 0.02, SE = 0.01, *t* = 3.45), with a *p* value (.074) greater than the alpha level of .05.

Random effects

The SD (0.12) for modality-by-subject random intercepts indicated that scores for participants derived from the average intercept (0.81) by approximately 0.12 points. The SD (0.12) for modality-by-subject random slopes indicated that participants' estimated slopes derived from the average slope of 0.06 by approximately 0.12. Consequently, an individual participant with a slope 1 SD below the mean (0.06 – 0.12 = -0.06) would have an estimated slope near 0, suggesting that this person's scores were not affected much by the modality in which the test items were presented. However, an individual participant with a slope 1 SD above the mean score would have a steeper slope (0.06 + 0.12 = 0.18), indicating that this person's scores were influenced by modality, with a difference of approximately 0.18 points in scores between modalities.

Similarly, the SD (0.08) for session-by-trial random intercepts indicated that scores for participants derived from the average intercept (0.02) by approximately 0.08 points. The SD for modality by-trial random slopes (0.04) indicated that participants' estimated slopes varied around the average slope of 0.02 by approximately 0.04. Consequently, an individual participant with a slope 1 SD below the mean would have an estimated slope (0.02–0.04) of -0.02, indicating that this person's scores were not affected much by the session in which the test trials were presented. Conversely, an individual participant with a slope was 1 SD above the mean score would have a slightly steeper slope (0.02 + 0.04) of 0.06, indicating this person's scores in the second session increased by only about the score of 0.06.

The variance for modality by-participant intercepts (0.01) with the SD (0.12) was very close to the by-trial intercepts (0.01) with the SD (0.08), indicating mean difference of participants' performance in the aural modality were similar to that in the first session. The SD for modality by-participant random slopes (0.12) was slightly greater than the SD for modality by-trial random slopes (0.04), indicating the variation of performance of this cohort of participants across the modality was slightly greater than that of trials across the modality.

The correlation between modality-by-subject intercepts and modality-by-subject slopes was -0.81, indicating a strong negative relationship between these two indices. This result suggested that participants who had higher ability in the aural modality exhibited a much steeper decline in their performance when they switched to the written modality. In other words, when compared to their counterparts at lower ability level in the aural modality, the performance of participants with higher ability was much less influenced by the change of modality condition.

The correlation between modality-by-trial intercepts and modality-by-trial random slopes is 0.06, indicating a negligible relationship between the two indices. This result suggested that the trials having higher accurate responses in the aural modality did not differ much compared to the trials obtaining lower accurate responses. In other words, the variability of trials was unlikely to be influenced by the modalities.

Finally, the performance of the final selected model was evaluated by the marginal R^2 and conditional R^2. The marginal R^2 of 0.017 indicated that only a very small proportion of variance (about 1.7%) in the outcome of this subtest was explained by the fixed effects of modality and session alone. The correlation coefficient (r = .13) suggested a negligible effect size according to the benchmarks proposed by Plonsky and Oswald (2014) for interpreting effect sizes in L2 research. However, the conditional R^2 was 0.31, indicating that a much larger proportion of variance (about 31%) was explained by both the fixed and the random effects, and the effect size converted (r = .56) was close to the benchmark of a large effect size. This suggested that random effects of participants and test trials in the model accounted for a larger amount of the variability in the outcome.

In summary, LMM on the data from TALL_CST evidenced the statistically significant effects of the fixed effect of modality, but not session, on participants' performance, although the magnitude of difference was very small. Specifically, when tested in the written modality but not when they took the test in the second time, participants had significantly higher scores. In addition, the variability of participants across modality was evidenced, and the results also suggested that participants who had better ability tended to be much less affected by the modality in which this subtest was administered, and the negative relationship was strong. The degree of response variance in test trials was much smaller than that in participants, and trial variability seemed to be susceptible to the modality conditions. A much larger proportion of variance (about 31%) was explained by both the fixed and the random effects, in contrast to the relatively small proportion of variance (about 1.7%) explained by the fixed effects alone, which was negligible in magnitude.

5.4 Summary of the results

The focus of this chapter was to investigate RQ2:

- To what extent does modality have effects on scores in the subtests of Vocabulary Learning, Language Analysis and Complex Span Tasks that can be administered in aural and written modality?

Correlation checks on the data of each subtest showed that the results of two modalities were significantly correlated, evidenced by the non-parametric coefficient of Kendall's *tau*. MEM was applied, which allowed the investigation of the main effect, i.e., modality, to be considered *simultaneously* with other factors (i.e., test session and the variability of participants and test items) that could be the sources of variation in the repeated design.

The adequacy of sample size and data points were confirmed, and assumptions for applying linear regression modelling were checked on all subtests respectively prior to the MEM analyses. GLMM was applied on the binary data from the subtests of TALL_VL and TALL_LA, while LMM was applied on the continuous data from TALL_CST. A systematic process of constructing and selecting models was followed, starting from the null models only including random effects to the main-effects models built by adding fixed effects and random slopes. The selection of the final appropriate model was achieved based on the comparisons of model fit indices (i.e., the values of AIC, BIC, and the results from the likelihood ratio tests). The performance of the final selected model was evaluated by the marginal and conditional R^2.

The results from MEM analyses indicated that modality as the main factor significantly differentiate participants scores in all three subtests, which provided a clear support to the hypothesis that participants' performance would benefit more if the subtests were administered in the written modality. Although test session as another main factor also had effects on participants' scores in the subtests of Vocabulary Learning and Language Analysis, its effects on participants' performance in the Complex Span Tasks were not evidenced. The variability of participants across modality were greater than the variability of test items or trials across modality in all three subtests, which was not surprising. Additionally, participants who had higher abilities in these subtests seemed to be less influenced by the modality conditions when compared to their counterparts who had lower abilities. As the models built for the purpose of investigating the effects of modality and simultaneously considering the other factors on participants' scores, the performance of these selected models showed that much larger proportion of variance was explained by all factors rather than the main factors of interest (i.e., modality and session) alone.

5.5 Discussion for RQ2

5.5.1 The fixed effect of modality

RQ2 sought to investigate whether stimuli presented in different modalities (i.e., aural and written) would have effects on scores in the subtests of TALL_VL, TALL_LA, and TALL_CST, in which test items were presented in different modalities. The scores of each subtest obtained in two modalities had a significant but weak correlation. This suggests that participants who have displayed, in general, consistent abilities of associative memory, language analytic ability and executive control capacity of WM across different input modalities of test items in these subtests. However, it is admitted that there is relatively low reliability and unidimensionality of TALL_VL in the aural suite and TALL_CST in the written suite (as mentioned in Section 4.3.2.3), and TALL_CST in the written suite lacks challenge to the participants (as mentioned in Section 4.4.2.5). These factors may potentially undermine the confidence in the findings when using these instruments.

The analyses of the (Generalised) Linear Mixed-effects Models provided strong evidence to indicate that modality, as a main fixed effect, significantly differentiate the scores in three subtests, and participants' performance benefited more if the test items were presented in the written modality. The superiority of the written modality in relation to the performance is not surprising. It is predicted that L2 learning, in general, would be easier through the visual rather than through the auditory modality, given that processing input in the written form may make the structures more salient by allowing attentional resources to focus more on the form. The untimed nature of input in the written modality as opposed to the fleeting nature of auditory modality may also facilitate deeper processing of linguistic input as L2 learners have time for self-paced processing (Gilabert et al., 2016).

Specifically, the results on the datasets of TALL_VL were in line with the finding reported by Mizumoto and Shimamoto (2008) that vocabulary size test in the aural modality was more difficult than in the written modality. A recent study about orthographic versus auditory word learning by Escudero et al. (2022) reported similar findings that recognition performance was more accurate when novel words were presented in the written forms than in the aural forms, although word learning in their study was investigated in the cross-situational learning paradigm. However, the effect size of the fixed effects (calculated from the marginal R^2) was negligible on the test scores according to the benchmarks for interpreting effect sizes in L2 research (Plonsky & Oswald, 2014), especially compared to the medium effect size of all fixed and random effects taken together.

The results of TALL_LA suggested that, compared to learning morphosyntactic rules with stimuli presented in the written modality, learning in the aural modality can be more challenging. This was in line with the findings revealed in studies comparing the modalities of tasks in task-based learning. For example, Zalbidea (2021) found that participants engaged in the L2 production and input processing in the written modality could have sustained gains of the lower-salience target structure, whereas their counterparts in the aural modality could not. The challenge of learning the morphosyntactic rules with input in the aural modality can be paramount in the cross-situational learning paradigm, as learning processes may involve detecting word boundaries, decoding the meanings of novel words, identifying lexical categories, and understanding the relations between categories established by the morphosyntactic rules (Walker et al., 2020). However, the effect size of the fixed effects (modality and test session) on test score was small (though larger when random effects were included.

The discrepancy of performance between test modalities in TALL_CST may not be exclusively related to the cognitive demand of processing stimuli presented in the aural modality. The technical design that poses longer interval between the tasks of processing and recall may add extra load in retaining the stimuli for a longer time in the aural suite than in the written suite (as discussed in Section 4.7.1.2). In this subtest, modality was the only fixed effect that differentiated the test scores significantly. However, the effects size of the fixed effect was negligible though the effect size of all fixed and random effects was medium.

The relationships between participants' abilities and the effect of modality are worth noting. It was found that in three subtests, the performance of participants who had higher ability (i.e., achieved higher scores) had a weaker association with and less affected by the modality conditions compared to their counterparts who had lower ability. This suggests that higher ability may wash out the modality effect.

The low correlations observed between the scores from the aural and written modalities across these three subtests warrant discussion. As reported in Section 5.3, the correlation coefficients of Kendall's *tau* (.16 for TALL_VL, .32 for TALL_LA, and .20 for TALL_CST) fell within the small to medium magnitude range. One plausible statistical explanation of these low correlations could be attributed to the use of the non-parametric ranking correlation coefficient, Kendall's *tau*, necessitated by the non-normal distribution of data in the current study. To complement this, an additional analysis was conducted using Pearson's *r* under the pseudo-assumption of normality. Subsequently, the correlation coefficients increased to .23 for TALL_VL, .41 for TALL_LA, and .25 for TALL_CST. This suggests that if the data had exhibited normality, the correlations might have shown an increase.

Nevertheless, even with the correlations enhanced by the pseudo-assumption of normality using Pearson's *r*, they still remain relatively low. This discovery raises a fundamental inquiry concerning the nature of aptitude in aural and written modalities: Are aural and written modalities encompassing substantially distinct constructs of aptitude? This question goes beyond methodological considerations regarding modality selection in operationalising aptitude; it holds theoretical significance in understanding the potential role of aptitude in diverse learning contexts that may foreground the aural modality (naturalistic contexts) or written modality (instructed contexts). The question warrants further in-depth investigations.

The findings in this chapter offer valuable insights, indicating that modality may play a crucial role as a moderating variable that warrants thorough examination in future research on aptitude for L2 learning. For instance, if language analytic ability is exclusively assessed in the written modality, as seen in other existing aptitude batteries, it may only be relevant to classroom learning contexts employing explicit and skill acquisition-informed learning and teaching approach. Moreover, learners' abilities, such as vocabulary learning, language analytic ability and executive control in WM, may be influenced to varying degrees by modality. This highlights the importance of considering modality conditions when explain cognitive individual differences in L2 learning. Therefore, it is essential for psychometric measurements and experimental design to be sensitive to these nuances. The development of TALL into two test suites enables the exploration of modality in future research.

5.5.2 The fixed effect of test session

The results showed that test sessions significantly differentiated the scores in the subtests of TALL_VL and TALL_LA, but not in TALL_CST. This suggests a clear test-learning effect in TALL_VL and TALL_LA. A test-learning effect, also known as a test-practice effect, occurs when a test-taker performs better on a subsequent test due to familiarity with the same or similar test content, despite possessing the same level of knowledge or skills pertaining to the assessed construct (Davies et al., 1999, cited in Suga & Loewen, 2023). In this study, two aspects of the research design were implemented to address potential carry-over effects between the two test sessions: a minimum interval of 30 days was imposed, and equivalent sets of materials with counterbalanced items were used. Despite these measures, results from TALL_VL and TALL_LA still indicated that participants showed improved performance when taking the subtests for the second time. One possible explanation is that participants became familiar with the test format or procedures and developed effective strategies through repeated engagement with the same test (Suzuki & Koizumi, 2020). However, this explanation did not align with the results from TALL_CST, where the test-practice effect was not observed.

The findings revealing the presence of a test-practice effect in two language learning-related subtests (TALL_VL and TALL_LA) and the absence of such an effect in one working memory subtest (TALL_CST) are thought-provoking. Methodologically, it prompts inquiries regarding the (test-retest) reliability of specific subtests in measuring the underlying constructs of aptitude. More importantly, it touches upon the stability of the aptitude construct itself. This raises the pivotal question: is aptitude, as defined in the existing literature (Carroll, 1981; Dörnyei, 2005), an inherent and consistent trait that remains relatively stable and resistant to change through training? Or is aptitude malleable, susceptible to improvement with exposure to language learning opportunities? The scarcity of empirical evidence regarding the stability of aptitude measurements restricts comparison with past findings, especially considering the use of different aptitude batteries, some of which have faced criticism due to instrument unreliability.

A noteworthy exception is the reliability assessment of the Hi-LAB conducted in a series of studies discussed in Hughes et al. (2023). Results from their Reliability Study 3 indicated that among nine tasks in the Hi-LAB, three tasks exhibited “small” practice effects (defined as less than 1/10 of the score range between test-retest sessions) on participants’ performances, and two tasks showed “moderate” practice effects (defined as more than 1/10 of the score range instead). However, crucial details such as the interval between test-retest sessions and the use of counterbalanced test items were not included in the report. Removing methodological uncertainties concerning instrument validity is anticipated to facilitate empirical insights into the questions about stability of aptitude: whether it is an inherent trait that resists changes through training, a reflex of language experiences that can be enhanced with exposure to language learning opportunities, or perhaps a multi-faceted construct where some components are malleable, but others remain stable. Further dedicated investigation into the test-practice effect in measuring aptitude is needed and should be included in the research agenda to validate TALL.

An additional point is interesting: the interaction between the fixed effects of modality and session was not evident in all three subtests, suggesting that the effect of modality on the test scores does not vary across different sessions. In other words, the impact of modality on participants’ performance in these subtests remains consistent regardless of the specific session under consideration.

5.5.3 All fixed and random effects

The results from the (Generalised) Linear Mixed-effects Models suggested that small to medium proportions of variance can be explained by both the fixed effects and the random effects, indicating small to medium effect sizes of all effects considered together on test performance. This highlights the importance of including individual differences and task characteristics as sources of variation to provide a more comprehensive understanding of the observed effects. In linguistic research, the evaluation of MEM often relies on the conditional R^2 to reflect the predictiveness of the complete model (including both fixed and random effects) on the outcome variables, taking individual variability into account.

However, some researchers advocate using marginal R^2 rather than conditional R^2 (e.g., Barth & Kapatsinski, 2018) to evaluate the predictiveness of MEM. It is important to note that much smaller proportion of variance accounted for by fixed effects compared to all effects in MEM is frequently reported in studies in SLA (e.g., Bovolenta & Williams, 2022; Palma et al., 2022; Suarez-Rivera et al., 2022; cf. Gries, 2021). This perhaps merits a meta-analysis on marginal R^2 and conditional R^2 in the literature to determine how much of the effects we are interested in examining can actually explain L2 learning phenomena.

CHAPTER 6: RESULTS & DISCUSSION FOR RQ3 – PREDICTIVE VALIDITY

6.1 Introduction

This chapter focuses on the investigation of the predictive validity of TALL as a measure for language aptitude on learners' L2 proficiency, and hence providing the evidence of the extrapolation inference about the extent to which the scores of TALL subtests reflect participants' L2 learning success (see Section 2.2.3.1). It addresses RQ3: *To what extent do subtests of TALL predict English proficiency measured by the National Matriculation English Test (NMET)?*

The first section introduces the choice of data analytical methods for this RQ, that is, the use of statistical method of Multiple Regression Analysis (MRA) to examine subtests of TALL as the predictor variables (PVs) in relation to NMET scores of participants' L2-English proficiency as the dependent variable (DV) in the current research. It also highlights the rationale of using Dominance Analysis (DA) as a supplementary analysis method for determining the relative importance of the multiple PVs in predicting a DV (Mizumoto, 2022a). The second section presents the results from MRA and DA on the data from the Aural Suite and the Written Suite, respectively. This section is followed by a summary of the results, which sequentially leads to the final section of the discussion about the results of the predictive validity of TALL on L2 proficiency.

6.2 Multiple Regression Analysis and Dominance Analysis

MRA is a statistical analysis method to examine the relationship between a DV and multiple PVs. The main goal of MRA, in general, is to create an equation of a model, whether linear or nonlinear, that results in a line accurately representing the data while ideally attaining parsimony in the model (Jeon, 2015). The equation of MRA can yield the predicted value of the DV by summing up the coefficients representing the effect of each PV on the DV and the residual representing the variation in DV that is not explained by the PVs. Essentially, the coefficients of PVs represent the estimated changes in the DV for every one-unit increase in the corresponding PV when holding all other PVs constant.

For RQ3 in the current research, the primary purpose of analysis was to find out whether each of the five components of the aptitude construct, as measured by TALL, can stand as an individual predictor of learners' L2-English learning outcome, as represented by NMET scores, after the variances due to other aptitude components were partialled out.

The total variance explained by the model was indicated by R^2, indicating percentage of the variance of L2-English learning outcome that can be predicted from the components of aptitude in the model. To address RQ3 about the predictive validity of TALL as a battery that has five componential subtests, MRA would be an appropriate analytical method as it allows accurately describing the relationships between the DV (the proficiency score) and the PVs (the five subtests) while accounting for the shared variance across the PVs (Plonsky & Oswald, 2017).

In the current research, participants took two tests suites in a repeated design, with order of the suites and the material versions counterbalanced. The data were analysed separately for each suite, therefore, informing the extent to which the subtests, in each of the two suites, could predict L2-English proficiency scores. Data analysis code in this chapter was adapted from (Mizumoto, 2022b), rendered in an `R` markdown file (see Appendix D).

6.2.1 Data preparation and steps for assumption checks

Prior to the MRA, preparatory steps, as adapted from Jeon (2015), were followed to make sure that the data met the assumptions of MRA.

Step 1. Check the sample size of data for analysis

This step ensured that the sample size was adequate for conducting a reliable MRA according to the rules of thumb proposed in the literature. For example, the sample size should be equal or larger than $50 + 8k$ (k is the number of PVs) (Tabachnick & Fidell, 2012), or at least 15 participants for each PV (Stevens, 1996).

Step 2. Check the linearity of relationship

In this step, linear relationship between the individual PVs and the DV was checked by the scatterplots of each subtest and the NMET score.

Step 3. Check data independence

The independence assumption suggested that the observations should be independent of each other. This was checked by the autocorrelation plot conducted on the residual of the regression model. Ideally, the lags in the plot should not extend beyond the blue lines, which suggests that this assumption was met.

Step 4. Check for homoscedasticity

This assumption suggested that the variance of the errors should be constant across all levels of the variables. It was checked by the scatterplot that displayed the residuals of the regression model in relation to the fitted value predicted by the model. The residuals should spread almost equally for all values of the PVs.

Step 5. Check the normality

Data used in this section were normalised (see Section 4.5.2 and Appendix H). The assumption of the normality of the residuals was checked by plotting the histograms of residuals and the normal probability plot (Q-Q plot) on the regression model.

Step 6. Check the multicollinearity

This assumption required that PVs in MRA should not be highly correlated. It was checked by the computation of the variance inflation factor (VIF). VIF values should be less than 2.5 for each predictor to indicate that multicollinearity is not a significant issue (Allison, 1999) in the regression model.

6.2.2 Process of Dominance Analysis

As explained above, MRA is to create the best-fitting model for predicting the DV from a group of PVs. This correlation-based method examines the relationship between multiple PVs and a single DV by comparing the standardised beta (*b**) coefficients. Additionally, MRA can also determine the amount of variance in the DV that can be explained by the PVs (R^2) and the relative importance or contribution of each PV to the overall effect. However, in the L2 research community, the pervasive practice of making claims about the relative importance of individual predictors in contribution to the learning outcomes based on the magnitude of the *b** coefficients has been considered a misuse of MRA (Karpen, 2017). The core of this problem, as elaborated in detail by Mizumoto (2022a), is the existence of the suppression effect that could potentially lead to the underestimation or overestimation of the importance of PVs if the *b** coefficients are compared. The suppression effect occurs when one of the PVs is more strongly correlated with the other PVs than with the DV, which can lead to the magnitudes of the *b** coefficients being different from the magnitudes of the correlation (*r*) coefficients. This limitation cannot be overcome by computing the coefficients of partial or semipartial correlations because these methods also treated PVs as uncorrelated. Mizumoto, therefore, suggests the use of Dominance Analysis (DA) as a supplementary analysis to MRA in order to help tackle the problematic interpretation of the relative importance of PVs on DV using the *b** coefficients.

DA as one type of regression, estimates the relative importance of PVs by analysing the change in the total variance accounted by the regression model (R^2) from adding one PV to all possible combinations of the other PVs. In the current research, given that the subtests were the indispensable components of the construct of aptitude, the correlations among them should be taken into account in MRA. However, the suppression effect could potentially exist because the subtests may correlate with one another more than with the NMET as a composite score of L2 English proficiency. Therefore, to address RQ3 about the predictive validity of TALL suites for proficiency and to understand accurately the relative importance of each subtest as a predictor to the proficiency score, DA should be conducted as a complement to the results of MRA.

The following sections introduce the steps to analyse the relative importance of the subtests by conducting relative weight analysis. Relative weight analysis is an alternative to DA suggested in Mizumoto (2022a) with less computational expense. This analytical approach has been introduced in the field of L2 in Larson-Hall (2016) as a method to compute the relative importance metric. The process of conducting relative weight analysis, termed as DA by Larson-Hall (2016), started by generating multiple regression models on the data, conducting relative weight analysis using the relaimpo package (Grömping, 2006), and calculating confidence intervals for the dominant weight of all PVs based on bootstrapping procedures using the boot package (Canty & Ripley, 2021) and the yhat package (Nimon, et al., 2021). The purpose of using bootstrapped replications to obtain and report 95% confidence intervals alongside the dominance weights was to improve the interpretation of the rank order of weights. This approach was adapted because the use of bootstrapped replications helps account for sampling error variance, which can otherwise result in unstable estimates of the magnitudes and rank ordering of dominance weights (Braun et al., 2019, cited in Mizumoto, 2022a). The results of correlation coefficients between the PVs and the DV, the standardised *b** coefficients with *p* values, and the dominance weights with 95% CIs are reported in the section of results.

6.2.3 Post hoc power analysis of MRA

Other than the precision of statistics, the quality of data analysis through MRA would be largely ensured by power, the likelihood that a test would provide statistically significant findings when the relationship of interest exists in truth (Jeon, 2015). The computation of a priori power analysis for MRA should be based on previous research and/or theory, for example, using the expected value of R^2. However, given that TALL is a novel measurement that has not been used for substantive research yet to provide the reference values, the prior power analysis was not conducted before MRA. The sample size decided in the pre-data collection stage was based on the prior power analysis conducted for the mixed-effects modelling, as reported in Section 3.3.1.2, to investigate the effects of modality in measuring aptitude (to address RQ2). Therefore, after conducting MRA, a post hoc power analysis was conducted to compute power, which would help to better understand the quality through MRA in the current research and to provide preliminary evidence for prior power analysis in future research.

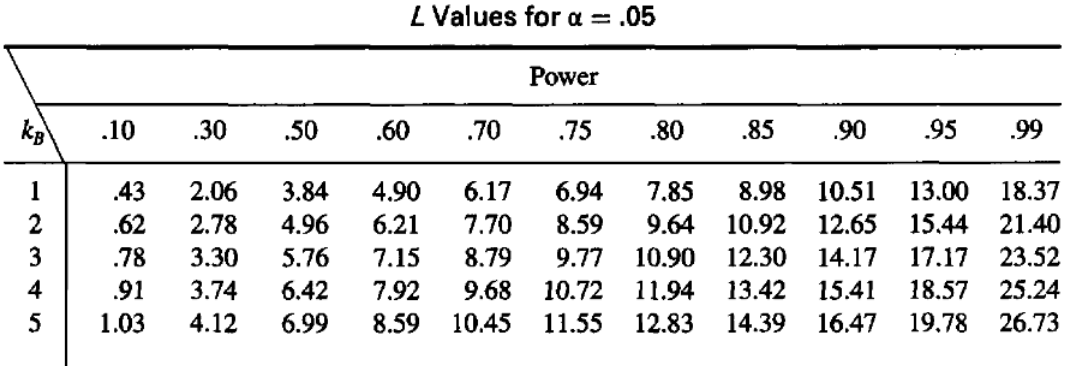
The formulas for power analysis used in this chapter are introduced by Jeon (2015). First, the population effect size (f) was computed from the total variance accounted by the regression model (R^2) in the following formula:

$$f = \frac{R^2}{1 - R^2}$$

Then the f value obtained was used to determine L, the value needed to identify power in the L table (in Cohen et al., 2003) of the selected probability level (e.g., α =.05 in the current study), using the following formula, in which N is the sample size and k is the number of predictor variables:

$$L = f^2(N - k - 1)$$

The corresponding value in the L table (see Figure 6.1) provides information about the quality of the regression model in terms of its statistical power. The expected sample size could be calculated inversely by using the target power indices of .50 as the threshold for adequacy and .80 as the threshold for the ideal (Murphy & Myors, 2004). By corresponding to these L values, substantive research can be carried out to obtain further evidence of the predictive validity of TALL on L2 learning outcomes.



L Values for α = .05

	Power										
k_B	.10	.30	.50	.60	.70	.75	.80	.85	.90	.95	.99
1	.43	2.06	3.84	4.90	6.17	6.94	7.85	8.98	10.51	13.00	18.37
2	.62	2.78	4.96	6.21	7.70	8.59	9.64	10.92	12.65	15.44	21.40
3	.78	3.30	5.76	7.15	8.79	9.77	10.90	12.30	14.17	17.17	23.52
4	.91	3.74	6.42	7.92	9.68	10.72	11.94	13.42	15.41	18.57	25.24
5	1.03	4.12	6.99	8.59	10.45	11.55	12.83	14.39	16.47	19.78	26.73

Figure 6.1 *L* values (α = .05)

(shorten from TABLE E.2 in Cohen et al., 2003, p. 651)

6.3 The results

6.3.1 Assumption checks

Multiple regression models were built on the data from the Aural Suite and the Written Suite separately to check the assumptions, following the steps introduced in the previous section. The output of assumption checks can be found in the R markdown file in Appendix D.

The final sample size of 165 in each test suite met the required sample sizes, specifically 90 according to the rule of thumb in Tabachnick and Fidell (2012) or 75 in Stevens (1996). However, this decision could potentially compromise the data's quality, given that test sessions had significant effects on participants' scores in TALL_VL and TALL_LA, as indicated by the results from GLMM in examining modality and session as the fixed effects (see Section 5.3.1 and Section 5.3.2).

The linear relationships between the subtests and NMET were plotted separately for the data from the two suites. The results from the independence assumption checks suggested that this assumption was approximately met, with only a few lags extending beyond the blue lines, which may not be a serious issue. Additionally, the results from the homoscedasticity assumption checks showed that while the scatterplot did not precisely indicate equal spreading of the residuals, they also did not exhibit highly clustered patterns. This suggested, in a conservative interpretation, that the assumption of homoscedasticity was likely satisfied. The results from the normality of residuals checks indicated that the distributions of the residuals of the regression models were generally normal. However, it is worth noting that some residuals had deviated from normality, especially at both ends. Finally, all the VIF values calculated for the predictors were below 2.5 in both suites, indicating that multicollinearity was not a significant concern in the regression models. The correlation coefficients between all the subtests and NMET scores are provided in Table 6.1 and 6.2 in the following sections. These tables also serve as evidence that the assumption of no multicollinearity was met.

6.3.2 Regression coefficients and importance weights

6.3.2.1 The aural suite

The left part of Table 6.1 showed the correlation coefficients, especially the simple correlation coefficients (see the left end column of *r*) between the DV (i.e., NMET) and each of the PVs (i.e., TALL subtests) in the aural suite. The results indicated that the coefficients were very close. Specifically, the three subtests of language learning tasks (i.e., TALL_VL, TALL_SD and TALL_LA) had coefficients of .24, .26 and .28, respectively, in relation to NMET, and the two WM subtests (i.e., TALL_SNWR and TALL_CST) had similar coefficients of .21 and .22, respectively, in relation to NMET. The standardised *b** coefficients described the direct relationship between NMET and each subtest while controlling for the indirect effects of the other subtests. The results indicated that only TALL_LA had the *b** coefficient of .20 that accounted for 20% of NMET variance with a significant *p* value = .01.

The results from applying DA to the data yielded the dominance weight of each subtest and its contribution in percentage to the total variance (R^2) accounted for by the regression model. Among these results, TALL_LA had the highest dominance weight (.052) that contributed 33.55% of the total variance, being clearly the strongest predictor to NMET. The dominance weights of all the subtests summed up to R^2 = .155, indicating that 15.5% variance was explained by the model. The column labelled '95% CI' displayed the lower and upper bounds of confidence intervals for the dominant weights of all the subtests. These intervals, calculated using bootstrapping procedures, indicated that if samples were extracted from the population 1000 times, approximately 95 out of those 100 intervals would include the dominance weight (considered as the population parameter) of the corresponding subtest. In the rightmost column labelled 'rank', all subtests were assigned a rank order based on their dominance weights. The ranking provided information about the relative importance of the subtests in the aural suite when predicting participants' L2 English proficiency, as represented by NMET scores.

Table 6.1 Coefficients of correlation and regression, and dominance weights (aural suite)

	r								Dominance		
Variables	1	2	3	4	5	6	*b**	*p*	weight (%)	95% CI	Rank
1 NMET	—										
2 TALL_VL	.26	—					.12	.15	.033 (21.29%)	[.004, .092]	2
3 TALL_SD	.24	.31	—				.09	.27	.025 (16.13%)	[.004, .075]	3
4 TALL_LA	.28	.28	.36	—			**.20**	.01	.052 (33.55%)	[.010, .128]	1
5 TALL_SNWR	.21	.31	.24	.03	—		.10	.24	.020 (12.90%)	[.002, .079]	5
6 TALL_CST	.22	.24	.13	.11	.37	—	.12	.13	.025 (16.13%)	[.002, .088]	4
Total									.155 (100%)		

Notes: Dependent variable is NMET of English proficiency. *N* = 165. R^2 = .155, 95% CI [.046, .234]. Boldface indicates the coefficient with *p* value exceeding alpha of .05.

The results also indicated the inconsistency in the order of relative importance reflected in correlation coefficients from the correlation analysis and the standardised *b** coefficients from the regression analysis. Specifically, TALL_SD which was given the lowest standardised beta coefficient (*b** = .09) was not the subtest that had the lowest correlation coefficient to NMET. This inconsistency was solved by the DA as the dominance weight of TALL_SD (.025 or 16.13%) ranked the third place among all subtests, contributing more than TALL_CST and TALL_SNWR to NMET.

Figure 6.2 visualises the dominance weights of the subtests along with their corresponding 95% confidence intervals (horizontal error bars show 95% confidence intervals computed from 1,000 bootstrapped replications). These are displayed in descending order from the subtest with the largest dominance weight (i.e., TALL_LA) to the smallest (i.e., TALL_SNWR). To determine statistical differences between pairs of subtests, the confidence intervals were considered. It was found that none of the subtests had a statistically larger dominance weight compared to another subtest, as the confidence intervals of all subtest pairs contained 0.

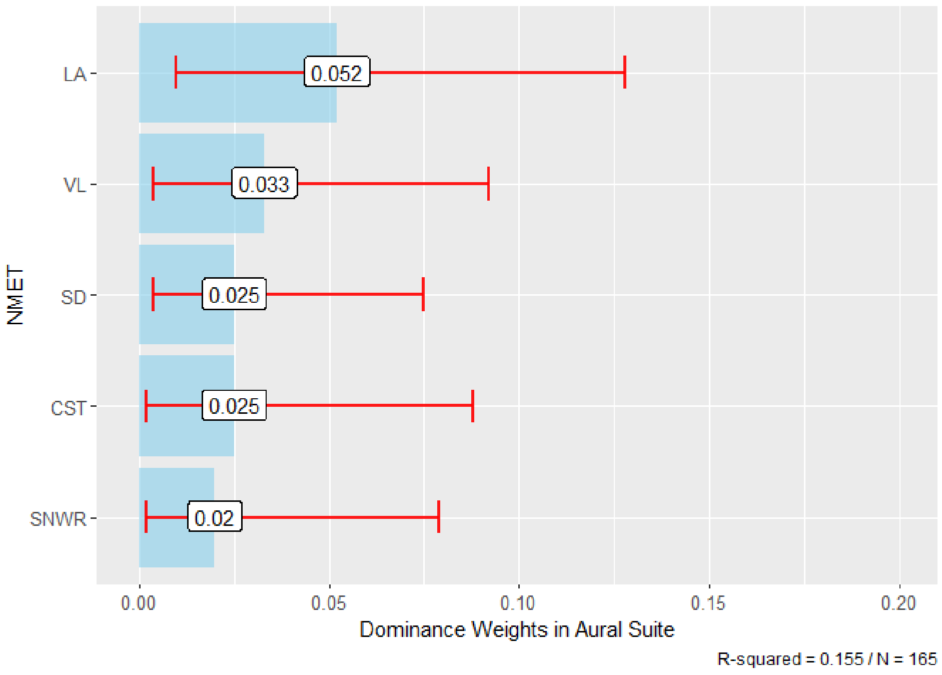


Figure 6.2 Dominance weights and 95% CI (aural suite)

6.3.2.2 The written suite

The left part of Table 6.2 showed the correlation coefficients, especially the simple correlation coefficients (the left end column of *r*) between the DV (i.e., NMET) and each of the PVs (i.e., TALL subtests) in the written suite. The results indicated that the coefficients of all the subtests except TALL_CST in relation to NMET were very close, ranging from .24 to .32. The coefficient between TALL_CST and NMET was the lowest (*r* = .16) The results of the standardised *b** coefficients indicated that two subtests, TALL_SD and TALL_LA, had *b** coefficients of .22 and .17, respectively. These coefficients accounted for 22% and 17% of NMET variance, with significant *p*-values of .004 and .03 respectively.

The results from applying DA to the data yielded the dominance weight of each subtest and its contribution in percentage to the total variance (R^2) accounted for by the regression model. Among these results, TALL_SD had the highest dominance weights (.067) that contributed 36.22% of the total variance, being the strongest predictor to NMET. The dominance weights of all subtests summed up to .185, which was a similarly low value of R^2 as that in the aural suite, indicating that 18.5% variance was explained by the model. The 95% confidence intervals displayed the lower and upper bounds of confidence intervals for the dominant weights of all the subtests. In the column of ‘rank’ provided information about the relative important of the subtests in the written suite when predicting participants’ L2 proficiency in NMET scores.

Table 6.2 Coefficients of correlation and regression, and dominance weights (written suite)

	r								Dominance		
Variables	1	2	3	4	5	6	*b**	*p*	weight (%)	95% CI	Rank
1 NMET	—										
2 TALL_VL	.24	—					.13	.07	.031 (16.76%)	[.002, .111]	4
3 TALL_SD	.32	.18	—				**.22**	.00	.067 (36.22%)	[.017, .148]	1
4 TALL_LA	.27	.15	.27	—			**.17**	.03	.043 (23.24%)	[.006, .116]	2
5 TALL_SNWR	.26	.27	.23	.16	—		.15	.06	.037 (20.00%)	[.004, .108]	3
6 TALL_CST	.16	.17	.23	.22	.30	—	.00	.99	.007 (3.78%)	[.002, .044]	5
Total									.185 (100%)		

Notes: Dependent variable is NMET of English proficiency. *n* = 165. R^2 = .185, with 95% CI [.07, .27]. Boldface indicates the coefficients with *p* values exceeding alpha of .05.

The results did not show inconsistency in the order of relative importance reflected in correlation coefficients from the correlation analysis and the standardised *b** coefficients from the regression analysis. Dominance weights of all subtests assured TALL_SD (.025 or 16.13%) ranked highest among all subtests, with TALL_LA, TALL_SNWR, and TALL_VL contributing in a descending order to the variance of NMET. TALL_CST had a negligible contribution, with its dominance weight being the least among all subtests.

Figure 6.3 presents the visualisation of the dominance weights of the subtests and the corresponding 95% confidence intervals with error bars in a descending order from the largest dominance weight (TALL_SD) to that with the smallest dominance weight (TALL_CST). The confidence intervals for all pairs of subtests indicated the statistical differences existing between the dominance weights of Sound Discrimination and Complex Span Tasks at the alpha level of .05. Other pairs of subtests did not have statistically different dominance weight in comparison with one another.

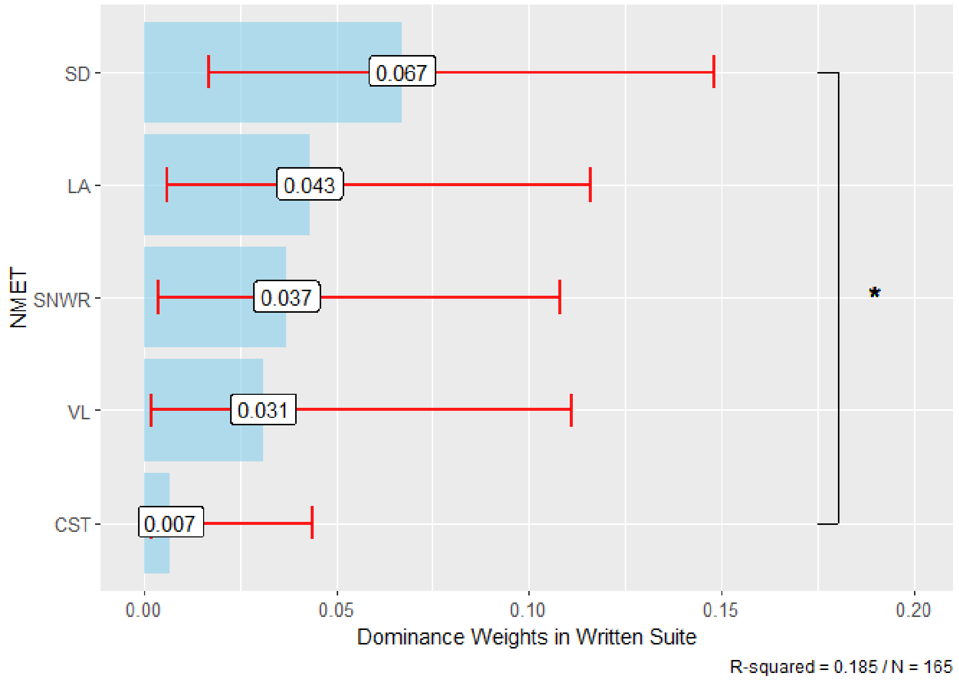


Figure 6.3 Dominance weights and 95% CI (written suite)

Note. The asterisk indicates statistically a significant difference (p < .05)

6.3.3 Post hoc power analysis

A post hoc power analysis was conducted to compute power after MRA to provide further information about the quality of the analysis in the current research and preliminary evidence for prior power analysis in future research.

Using the formulas presented in Section 6.2.3, the population effect size (f = .183) of the aural suite was obtained from the total variance accounted by the regression model (R^2 = .155). This further indicated that, based on the sample size (n = 165) in the current research, the regression analysis had statistical power (L = 5.25) of the index around .40 in the L table (see Figure 6.1), which is relatively low. Inversely, if the same regression model (R^2 = .155) could demonstrate adequate statistical power, the acceptable sample size would be n = 217 and the ideal sample size would be n = 395, calculated by using the acceptable power index of .50 and the ideal power index of .80 (Murphy & Myors, 2004) and their corresponding L values (6.99 for .50, and 12.83 for .80) in the L table.

Similarly, the population effect size (f = .23) of the written suite was calculated from the total variance accounted by the regression model (R^2 = .185). This further indicated that, based on the sample size (n = 165), the regression analysis had statistical power (L = 8.27) of the index close to .60 in the L table (Figure 6.1), which is above the acceptable threshold of power index of 0.50. Inversely, if the same regression model (R^2 = .185) could demonstrate acceptable statistical power, the acceptable sample size would be n = 140 and the ideal sample size would be n = 253, calculated by using the corresponding L values in the L table.

6.4 Summary of the results

The focus of this chapter was to investigate RQ3:

- To what extent do subtests of TALL predict English proficiency measured by the National Matriculation English Test (NMET)?

MRA with supplementary DA was used, as these facilitated the examination of the total variance of the dependent variable, i.e., NMET of L2 English proficiency, being accounted for by five subtests of TALL in the aural and the written suites, respectively. Prior to conducting MRA, assumptions were checked and warranted. The results of MRA indicated that the subtests in both suites could only explain about 16 – 19 % variance of NMET. Dominance weights computed through DA indicated that TALL_LA was the strongest predictor of NMET in the aural suite, though the differences between any pairs of subtests in dominance weights were not statistically significant at the alpha level of .05. The rank order of dominance weights of the subtests in the written suite had a different pattern from that in the aural suite. Specifically, TALL_SD had the highest contribution in explaining the variance of NMET and it had statistically significant difference compared to the subtest of TALL_CST (which had a negligible contribution) in their dominance weights.

Post hoc power analysis indicated that the regression analyses had statistical power of indices of .40 for the aural suite and .60 for the written suite approximately on the sample size (*n* = 165) in the current study. The results suggested that if the population effect sizes were taken as the preliminary evidence for prior power analysis in future research, larger sample sizes (ideally, 395 for the aural suite and 253 for the written suite) would be necessary to warrant the acceptable statistical power of the regression model computing the predictive validity of the subtests of TALL on L2 proficiency represented by NMET.

6.5 Discussion for RQ3

RQ3 sought to investigate the predictive validity of TALL in participants' L2-English proficiency represented by their self-reported scores in the National Matriculation English Test (NMET). To answer RQ3, MRA with supplementary DA were used. This 'macro' approach, as explored in the current study, examines the correlations of aptitude constructs with language achievement. This approach has received less attention in recent decades compared to the 'micro' approach, which involves investigating aptitude through experimental or quasi-experimental research designs focused on instructional comparisons or intervention processes (Skehan, 2019). Nevertheless, the current results of correlations between TALL (in two suites) and L2 proficiency scores are informative, providing insights into how aptitude measures relate to L2 achievement that has been developed overtime.

6.5.1 Predictive validity of TALL for global aptitude

The results of MRA indicated that participants' scores obtained in the aural suite of TALL could explain about 16% of the total variance of their scores on NMET, while their scores in the written suite could explain about 19% of the total variance of their scores on NMET. These results aligned with what reported by Granena and Long (2013): aptitude typically accounts for 10%–20% of the variance in predicting learners' ultimate L2 attainment. Converting the percentages of the total variance (i.e., R^2) to correlation coefficients, participants' performance in TALL's aural suite correlated to their L2 proficiency (as measured by NMET) at *r* (165) = .39, close to a medium effect size according to the benchmarks suggested by Plonsky and Oswald (2014), and their performance in TALL's written suite had a medium effect size of *r* (165) = .43 in terms of its correlation to their NMET scores. The results suggested that other variables of individual differences, e.g., learners' motivation and previous learning experience, that are not captured by TALL may also predict the L2 proficiency, which is not surprising. In general, the current study suggested that aptitude had a small to medium effect size on participants' L2 proficiency. Compared to using measures for aptitude in the aural suite, aptitude may correlate better with L2 proficiency—at least when measured by assessments such as NMET—when measured in the written suite.

The findings of the moderate correlations between aptitude (written suite) and L2 proficiency was slightly lower than a meta-analytic correlation (*r* = .49) between aptitude/aptitude components (measured by the MALT, the PLAB, the LLAMA and VORD (Child, 1998)) and general L2 proficiency (represented by course/exam grades or TOEFL) in Li (2016). Furthermore, findings in the current study align with correlations between the composite LLAMA scores and L2 outcomes, reported in the synthesis study by Bokander (2023). The author found that (drawing on only a small number of studies that have used composite scores of the LLAMA tests), all 7 studies in the synthesis reported significant correlations. Specifically, Artieda and Muñoz (2016) yielded a moderate correlation of *r* (88) = .39 between the LLAMA tests and the end-of-year official school L2 proficiency tests, with its largest sample of participants at the intermediate level of L2-English. Their findings were in line with the correlations revealed in the current study.

It needs to be noted that the research plan was affected by the pandemic (as indicated in the COVID-19 Impact Statement). As a result, participants were recruited based on their self-reported L2 proficiency scores, which were obtained after at least 6 years of instructed English-as-a-foreign-language learning. Given that TALL's design was built upon Skehan's (2016) Stages Approach that draws upon aptitude componential abilities in the *early* stages of language learning, it is expected that the correlations between TALL and L2 attainments after long-term learning experience would be weak, given that so many other factors come into play over time during language learning. However, the results of the current study still evidence moderate correlations between aptitude and L2 attainment.

Further research is essential to explore the predictive validity of TALL in L2 learning outcomes using the 'micro' approach. This entails obtaining evidence regarding the role of TALL in explaining aspects of language learning within diverse contexts through experimental or quasi-experimental studies. Such studies should maintain stringent control over the nature of the learning measures, considering factors such as the written/aural modality, explicit or metalinguistic knowledge involved, timed/untimed conditions, and spontaneous/controlled elicitation. It is important to recognise that predictive power of aptitude batteries is likely to be influenced by these factors in the learning measures. Additionally, it is also needed to investigate the role of aptitude in naturalistic learning environments, particularly where learning predominantly occurs in the aural modality, as seen in the case of learners with low literacy or (im)migrant learners. Such research will contribute to the evaluation of the underlying structure of aptitude upon which TALL has been constructed and will enhance our understanding of the relationship between the global aptitude construct and various aspects of L2 learning.

6.5.2 Individual predictors of the subtests

The results of DA provided the breakdown of dominance weights of the subtests as individual predictors in explaining L2 proficiency scores. The results indicated that, in the aural suite, TALL_LA (Language Analysis) was the strongest predictor and the only subtest that significantly predicted the scores of NMET, followed by TALL_VL, TALL_SD, TALL_CST, and TALL_SNWR sequentially. The subtests in the written suite, however, had a different rank order of dominance weights: TALL_SD was the strongest predictor, followed by TALL_LA, both predicted NMET significantly. TALL_SNWR, TALL_VL and TALL_CST ranked sequentially in a descending order of the dominance weights as non-significant predictors. Furthermore, TALL_SD and TALL_CST in the written suite was the only pair that had significantly different dominance weight. This suggests that, in general, the subtests of TALL play roles that are not significantly different in explaining L2-English proficiency. These findings indicated that modality has effects on the predictive validity of TALL, as different suites of TALL show varying levels of explanatory power on learning gains.

The predictive role of each subtest will be discussed sequentially below.

6.5.2.1 Predictive role of TALL_LA

The finding of TALL_LA predicting L2 proficiency scores significantly is not surprising. It is in line with the results reported by Bokander's (2023) synthesis of predictive validity of LLAMA_F (measuring language analytical ability) on general L2 learning. The author reported that LLAMA_F has contributed more correlational evidence than other LLAMA subtests and has the highest proportion of significant correlations, although individual studies investigating predictive validity of LLAMA_F on general L2 learning reported mixed findings. For example, LLAMA_F had a significant positive correlation of *r* (88) = .39 with participants at L2-English intermediate level in Artieda and Muñoz's (2016) study. However, LLAMA_F was not found to significantly predict the proficiency of learners of Swedish (*n* = 48) at the beginner level who had non-Germanic L1 background in Bokander's (2020) study. Bokander explained that the lack of correlation between LLAMA_F and learning outcomes could possibly be due to the way the L2 was measured: The global measure of Swedish using C-test does not target any particular linguistic features but rather on the general language proficiency tapping into textual, grammatical, and lexical knowledge. Although NMET used in the current study is also a global proficiency test (albeit it is different from the

C-test), it was significantly predicted by TALL_LA in both modalities. This provides strong support to the notion that language analytic ability plays important role in L2 learning (Skehan, 1998; Roehr-Brackin, 2022).

6.5.2.2 Predictive role of TALL_SD

As shown in the results, TALL_SD for phonetic coding ability predicted NMET scores significantly in the written suite, and this subtest had the highest dominance weights among all subtests. This finding is intriguing because the coefficient and dominance weight from the regression model for TALL_SD is lower in the aural suite than in the written suite, although the subtest remains the same in both suites (albeit experienced in different versions by individual participants). It suggests that TALL_SD, as a measure for the ability of encoding and differentiating unfamiliar sounds, explains more variance of a composite score of NMET when it is used in the written suite than in the aural suite. In addition, another subtest measuring sound-related ability, that is TALL_SNWR, also has a lower coefficient and dominance weight in the aural suite relative to in the written suite, although it does not explain the variance of NMET scores at a significant level. Taking these findings together, it seems that these two subtests, administered in the aural modality but included in the written suite may predict L2 proficiency in NMET better compared to their predictive power in the aural suite when other subtests are *also* administered in the aural modality. The possible explanation could be that the mix of modalities in the written suite of TALL (with some aural and some written tests) has closer alignments with the test composition of the NMET: the score of NMET reflects the combination of performance from the sections involving listening, reading, knowledge use and writing that are in both aural and written modalities. This finding also provide evidence to support that the ability to handle unfamiliar sounds plays a fundamental role in learning a new language, and it can be an important predictor of L2 proficiency (Skehan, 1998, 2016).

This finding was in line with the results reported in Li's (2016) synthesis of empirical studies investigating correlations between L2 learning achievement and aptitude components measured by the MLAT and the PLAB, in which phonetic coding ability was found to be the strongest predictor. It also aligned with the findings reported by Artieda and Muñoz (2016) that LLAMA_E (measuring sound-symbol pairing in the written modality) had a significant correlation (r = .26) with L2-English proficiency scores of 88 participants at intermediate level, although LLAMA_D (measuring sound sequency recognition in the aural modality) did not correlate significantly with participants' L2 proficiency. However, as noted above, the results in Bokander (2020) from the multiple regression analysis showed that LLAMA_E and LLAMA_D displayed non-significant coefficients in explaining the variance of proficiency of Swedish in the subgroup of participants (*n* = 48) who had non-Germanic L1 background. The discrepancy of findings in the role of sound-related aptitude subtests can perhaps be explained by the different proficiency tests used in these studies. Specifically, the English proficiency test in Artieda and Muñoz (2016) measured use of language, reading, listening, writing, and speaking of the participants, which were similar to the NMET in the current study. The Swedish proficiency in Bokander (2020) was tested in the C-test involving textual, grammatical, and lexical knowledge, which was in the written format.

6.5.2.3 Predictive role of TALL_VL

The results in the current study indicated that TALL_VL, in both suites, eliciting the construct of associative memory, plays a small but non-significant role to predict NMET. This role does not appear to vary when tested in different modalities.

The limited role of associative memory on explaining L2 proficiency in the current study aligns with Bokander's (2020) finding of a non-significant regression coefficient of LLAMA_B (measuring associative memory) in explaining L2-Swedish proficiency among the non-Germanic L1 subgroup. However, it's worth noting that Artieda and Muñoz (2016) reported a weak but significant positive correlation coefficient (*r* (88) = .21)) between LLAMA_B and L2-English proficiency participants at intermediate level. Notably, this correlation coefficient was slightly lower than Kendall's tau coefficients (.*26 in aural suite*, .24 in written suite) of TALL_VL and NMET.

In summary, while MRA did not establish a statistically significant predictive relationship between TALL_VL and NMET, there does appear to be some small statistical association between associative memory and L2 proficiency. This could possibly be explained by the relatively advanced stage of the participants in this study. Their relatively large and still expanding lexicons may not accurately reflect the very early stages of word learning that TALL_VL is designed to assess. Another possible explanation could be related to the issue of relatively lower reliability and unidimensionality of TALL_VL in the aural modality (as mentioned in Section 4.3.2.3), which might undermine the confidence in the findings when using this particular instrument.

6.5.2.4 Predictive role of WM subtests

The results indicated non-significant role played by the two WM subtests on predicting NMET scores. Their predictive powers ranked lower in both suites when compared to the subtests of TALL_VL, TALL_SD, and TALL_LA that involve language learning tasks (except that TALL_SNWR contributed slightly higher than TALL_VL in the written suite). TALL_CST in the written suite, in particular, was the predictor that explained the least amount of variance in NMET scores.

The finding that executive control capacity in WM contributes the least to the prediction of L2 proficiency aligns with the non-significant (and some unexpected non-positive) relationships between executive functions and high-attainment outcomes of advanced foreign language learners reported in Linck et al. (2013) when executive functions were assessed by the domain-general measures in the Hi-LAB. However, PSTM in WM, as measured by domain-specific span tasks in their study, was found to significantly predict learning outcomes, which did not align with the finding that TALL_SNWR fails to predict NMET in the current study. Possible reasons for this disparity could be the different L2 proficiency tests used and variations in the proficiency levels between the two studies.

It's noteworthy that WM has been found to have significant associations with L2 proficiency in studies using correlation analysis. For example, in Kormos and Sáfár's (2008) research in the classroom-based foreign language learning context, the authors explored the relationship between WM (measured by a nonword repetition test and a backward digit span test) and participants' performance in a global English proficiency test, encompassing reading, listening comprehension, composition, language use, and oral exam. For Hungarian-speaking teenage learners at the pre-intermediate level (*n* = 21), the weighted average scores of the nonword span tasks exhibited a moderate and significantly positive correlation (*r* =.47) with the global scores of the proficiency test. Regarding general WM (as termed by the authors) measured by the backward digit span test, a significantly positive correlation (*r* =.55) was observed between the total scores of L2 proficiency and the backward digit span test scores for participants (*n* = 45) at both beginner and pre-intermediate levels. Interestingly, the two WM tests were found to be uncorrelated with each other.

Given the heterogeneity in sampling, scoring, and statistical methods used, direct comparison between the results of the current study and Kormos and Sáfár's (2008) study can be challenging. However, the correlation and regression coefficients of TALL_SNWR concerning L2 proficiency in the current study may still suggest the potential role of PSTM in explaining the L2 proficiency of learners beyond the beginner level. However, the predictive power of PSTM may be relatively small. In contrast to Kormos and Sáfár's findings, the two WM subtests in the current study were significantly correlated. This result is plausible since the subtests were constructed based on the P/E model (Wen, 2016) and both used domain-specific verbal stimuli.

Admittedly, the holistic L2 proficiency score in the current study may not offer further evidence regarding the role of WM in predicting language learning outcomes involving specific competencies. Understanding the relationships between WM and various L2 learning processes merits extensive research, as elucidated in the discussion of the relationships between WM and specific learning processes by Juffs and Harington (2011). Additionally, it's crucial to acknowledge the relatively lower reliability and unidimensionality (as mentioned in Section 4.3.2.3), as well as the lack of challenge in assessing participants' ability in TALL_CST in the written modality (as mentioned in Section 4.4.2.5). This could potentially undermine the confidence in the findings when using this subtest in the current study.

6.5.3 Statistical power of regression analyses

The results of post hoc power analysis indicated that the regression analyses in the current study had relatively low statistical power indices, approximately .40 for the aural suite and .60 for the written suite, given the sample size of 165 participants in this study. These post hoc power analyses suggest that the sample size could be considered almost acceptable, as it is above the threshold of power index of 0.50 for the written suite, but not for the aural suite. These findings can sever as preliminary evidence for the need to conduct power analysis in subsequent research endeavours. Ideally, to ensure sufficient statistical power for the regression model with an *ideal* index of .80 based on the population effect sizes found in the current study, future studies should recruit a sample size of 395 participants for the aural suite, and 253 participants for the written suite.

The findings from this study contribute to the ongoing debate in the field, as highlighted by Isbell et al. (2022), Nicklin and Vitta (2021), and Plonsky (2013), regarding the need to enhance the methodological rigour of quantitative research. These results underscore the importance of conducting prior power analysis, which is essential to ensure the robustness of results when investigating the predictive validity of aptitude measures on L2 learning outcomes.

CHAPTER 7: CONCLUSION

7.1 Summary of the study

This thesis has reported the development of an internet-based aptitude battery, known as *Tests of Aptitude for Language Learning* (TALL), and the preliminary checks of its internal validity as an aptitude battery and its predictive validity for explaining participants' scores on an L2-English proficiency test.

The development of TALL has been motivated by the theoretical inquires and methodological concerns regarding the aptitude batteries used in L2 learning research. First, the development of aptitude measurements has not generally kept pace with changes in theoretical frameworks that characterise the multi-faceted construct of aptitude pertaining to L2 learning. Second, it is crucial to ascertain the reliability and validity of aptitude batteries *prior* to conducting aptitude-learning research, yet this step has been surprisingly neglected to date (cf. Bokander & Bylund, 2020). Based on the theoretical frameworks of the Stages Approach (Skehan, 2016) and the Phonological/Executive (P/E) Model (Wen, 2016), TALL has been developed to measure four facets of aptitude that represent cognitive abilities considered to be involved in the early stages of L2 learning and development. These abilities are associative memory, phonetic coding ability, language analytic ability, and working memory (specifically, phonological short-term memory and executive control capacity), measured by five subtests, i.e., Vocabulary Learning (TALL_VL), Sound Discrimination (TALL_SD), Language Analysis (TALL_LA), Serial Nonwords Recall (TALL_SNWR), and Complex Span Tasks (TALL_CST), respectively.

In addition, TALL has been designed to address unresolved concerns in aptitude measurement to date. First, it has employed domain-specific verbal stimuli in all subtests, including WM subtests. Second, TALL has two separate test suites to differentiate the aural and the written modalities in which test items are presented in the subtests of TALL_VL, TALL_LA, and TALL_CST. Additionally, the potential confounds of L2 knowledge have been mitigated by using participants' L1 as the instructional language of the entire battery and the encoding language for the stimuli in WM subtests. Finally, items for language learning tasks have been created in a semi-artificial language to ensure the novelty for participants in the current research.

Importantly, TALL has been developed into an internet-based battery to enable remote data collection during the Covid-19 pandemic. Specific technical considerations served to minimise problems that may threaten its internal validity (Newman et al., 2021).

For example: (i) archival techniques for recording response time allows researchers to identify anomalous (too fast) responses; (ii) explicit instructions and warnings displayed on the screen throughout the test aims to reduce dishonest test behaviours; (iii) assigned single-use test codes prohibit participants from reattempting the test. The efforts of developing TALL into an internet-based instrument merited making the instrument openly accessible to other researchers (Pan & Marsden, under review).

The current study was conducted to answer three research questions. RQ1 sought to examine the reliability and internal validity of TALL, based on a validation plan that provide evidence for making (1) a generalisation inference about all subtests being representative as measures for their intended constructs; (2) a scoring inference about items in each subtest being efficient to assess participants' abilities; and (3) an explanation inference about TALL aligning with the theoretical frameworks underpinning its construction. RQ2 investigated the extent to which modalities had effects on the scores in the three subtests that had test stimuli presented in the aural and the written modalities. RQ3 concerned the predictive validity of the subtests of TALL in explaining participants' L2-proficiency represented by self-reported scores on the National Matriculation English Test (NMET).

The final data for analyses were collected from 165 participants who were Year-one undergraduates from eleven universities in China and had taken the NMET within six months prior to the study recruiting date. They took two rounds of TALL in a within-subject design, with two test suites (one 'aural' suite, with only aural subtests, and the other 'written' suite, with written and aural subtests) and two versions of items in each subtest being counterbalanced over two sessions. A minimum 30-day interval between the two sessions was imposed to reduce any carry-over effect that participants would have in taking TALL repeatedly. After completing tests, participants received 50-yuan cash through online payment and a report of the scores in all subtests.

7.2 Summary of the findings

Analysis 1

Analysis 1 of the reliability and internal validity of TALL was conducted on the data at the subtest level, the item level, and the battery level.

The results showed that, in general, all subtests had satisfactory reliability according to the coefficients of omega and Cronbach's alpha, except that three datasets, i.e., Version A of TALL_VL in the aural suite, Version A and Version B of TALL_CST in the written suite, had coefficients of alpha lower than .74, the field-specific acceptable threshold proposed by

Plonsky and Derrick (2016). The indices used to explore the unidimensionality of the subtests revealed that these three datasets had displayed lower unidimensionality compared to other datasets, suggesting that these subtests may not measure a single underlying dimension or construct.

The data at the item level were analysed used the Rasch model (for dichotomous datasets of TALL_VL, TALL_SD, and TALL_LA) and the Generalised Partial Credit model (for polytomous datasets of TALL_SNWR and TALL_CST) based on the Item Response Theory. The results did not reveal clear evidence to suggest that any items were of poor quality that may threaten the validity of the subtests, and so deletion of items was not necessary. However, the amount of information provided by the instruments varied between the subtests, with TALL_CST in the written suite providing least total information about participants' executive control capacity in WM and performed particularly poor at discriminating those who had ability above the average. This suggested a ceiling effect of TALL_CST administered in the written modality.

The internal validity of TALL as a comprehensive aptitude battery was investigated at the battery level on the aggregated datasets of the aural suite and the written suite, respectively. The results of Principal Component Analysis revealed that neither suite had much redundancy and so there was no reduction (removal) of any subtests. The results of the Confirmatory Factor Analysis provided strong evidence that the data of both suites fitted well to the hypothesised four-primary-factor model based on the theoretical frameworks (the Stages Approach and the P/E Model) on which TALL was constructed.

Analysis 2

Analysis 2 about the effects of modality on subtest scores presented in the aural and the written modalities (whereby each participant experienced different stimuli in the different modalities) was conducted using generalised linear mixed-effect modelling on the dichotomous data of TALL_VL and TALL_LA, and linear mixed-effect modelling on the polytomous data of TALL_CST. The results provided clear support to the hypothesis that participants' performance in these three subtests would benefit more if the items or stimuli are presented in the written form than in the aural form. Although test session as another main factor was also found to have effects on participants' scores in TALL_VL and TALL_LA, which suggested a clear test-learning effect (Davies et al., 1999), the effect of test session on participants' performance in the Complex Span Tasks was not evidenced. The results of model performance evaluation indicated that a much larger proportion of the variance in test performance was explained by all factors, which included both the fixed effects and the random effects of individual factors into account, than by the main factors of interest (i.e., modality and session) alone. The results from the (Generalised) Linear Mixed-effects Models indicated that substantial proportion of variance can be explained by both the fixed effects (i.e., modality and test session) and the random effects (i.e., participants and test items), and the effect sizes of all effects taken together were medium to large on explaining performance. This highlights the importance of including individual differences and task characteristics as the sources of variation to provide a more comprehensive understanding of the observed effects.

Analysis 3

Analysis 3 of the validity of the subtests of TALL in predicting L2-English proficiency measured by the NMET involved multiple regression analysis with supplementary dominance analysis. The results of multiple regression analysis showed that all subtests could only explain about 16 – 19 % variance of NMET (which converted to a small to medium effect size of the correlations), which is not surprising as it suggested that other variables of individual differences, e.g., learners' motivation and previous learning experience, that are not captured by TALL may also predict the L2 proficiency scores. Dominance weights computed through dominance analysis indicated that TALL_LA in the aural suite was the strongest predictor of the NMET, while TALL_SD in the written suite had the highest contribution in explaining the variance of the NMET.

7.3 Limitations and implications for future research

7.3.1 … related to implicit-explicit language learning and knowledge

TALL has been developed to measure different components of the aptitude construct and tailored to investigate effects of modality on aptitude measures. However, it has a limitation—it is not designed to differentiate implicit and explicit aptitude constructs that are involved in implicit and explicit learning. The main reason for this limitation is the challenge of developing domain-specific measures that are valid for assessing implicit aptitude.

A special issue of *Studies in Second Language Acquisition* was dedicated to implicit language aptitude (Li & DeKeyser, 2021) with the aim of distinguishing implicit aptitude from the "traditional" concept of explicit aptitude in explaining the nature of L2 learning or knowledge. Methodological concerns regarding the validity of measurements used to test language aptitude claimed to be implicit persist. For example, Perruchet (2021) reviews four possible reasons to explain the unexpected general pattern in which implicit learning tasks (supposed to measure implicit aptitude) did not, or only weakly, correlated with each other and demonstrated inconsistent predictive power on L2 learning, as reported by the studies published in the special issue. These reasons are: (i) an inadequate choice of tasks; (ii) a low reliability of measures used; (iii) the profound impact of the initial ability level on the supposedly ‘new’ implicit tasks; and (iv) the fact that L2 learning may also rely on the automatic utilisation of knowledge initially gained through explicit learning, which may not be fully captured by the implicit learning tasks used in the experimental settings.

These methodological limitations, in turn, raise fundamental questions about the operationalisation of implicit aptitude. A practical challenge in distinguishing between implicit and explicit aptitude lies in the lack of convincingly established methods for capturing implicit/explicit learning or knowledge (Isbell and Rogers, 2021). Recent studies using a fine-grained research paradigm have compared neurocognitive and behavioural data, suggesting that advanced learners dynamically utilise both types of knowledge (Suzuki, et al., 2022, 2023). Thus, there is a pressing need for a valid experimental paradigm to investigate implicit (versus explicit) aptitude and its predictive role in L2 learning. This direction will be an integral part of a long methodological journey to refine and iterate TALL.

7.3.2 … related to sampling bias

While the current research recruited participants from eleven colleges and universities to include samples that could reflect wide level of individual abilities, it is not immune to the systematic sampling bias in SLA research (Andringa & Godfroid, 2020) which tends to sample from WEIRD (Western, Educated, Industrialised, Rich, and Democratic; Henrich et al., 2010) populations. Even though the sample is from the Chinese (non-Western) context, it represents only a narrow slice of learner populations, and it shares many characteristics with other WEIRD samples, e.g., most of the participants are young college students who are expected to have high levels of cognitive ability and L1 literacy.

The superior aptitude of college students for language learning has been evidenced. For instance, strong L1 and L2 relationships were reported between individual differences (measured by the MLAT) in L1 attainment and in L2 achievement (Sparks et al., 2023), and Sparks and Dale (2023) argued that the predictive power of the MLAT for L2 achievement is primarily due to its functioning as a measure of L1 ability. Thus, aptitude research relying on learners with high L1 literacy level impedes generalising findings to wider learner populations. Additionally, WM measures and its functions are found to vary across the lifespan (see Gajewski et al., 2018), thus, college students, being self-selecting and then triaged through admission criteria do not fully reflect the full spectrum of working memory capacities (Wen et al., 2021).

Making TALL an IBR instrument should, in principle, unlock the potential to diversify participant sample beyond the traditional pool of college undergraduates (Casler et al., 2013; Newman et al., 2021). However, in practice, IBR platforms may still contribute findings that represent a narrow sample of the entire learner population (Newson et al., 2021; De Oliveira & Baggs, 2023), as they only reach those who have access to online digital facilities, as highlighted in the guidelines by the British Psychological Society on IBR (BPS, 2021). In the current study, efforts were made to address issues related to accessibility, such as using participants' L1-Chinese as the language in recruiting communication, consent form, and test instructions. These efforts also helped address potential conflation with L2-English proficiency. Additionally, asynchronous technical support was offered through chat function of a social media group to engage participants who lacked sufficient digital literacy. However, despite these efforts, some obstacles related to accessibility were not fully addressed.

Various issues need to be addressed in future research especially when wider learner populations are sought (e.g., participants from low L1 literacy and less social-economically privileged backgrounds). For example, including an automatic 'read aloud' feature (see Shepperd, 2022) in participants' background questionnaire and consent form could address low levels of print literacy. However, the test instructions in TALL need to be presented in different modalities aligned with test suites. Digital literacy (participants' knowledge about the set-up of the computer), internet access, and access to an appropriate environment for testing are also essential to collecting quality data. One solution is to provide a test venue that is distraction-free with stable internet access.

7.3.3 … related to aptitude–treatment interaction research

The existence of a valid aptitude battery can help to diversify instructional methods to optimise learning outcomes based on individual differences (Cronbach & Snow, 1977, cited in Hughes et al., 2023; DeKeyser, 2019; Roehr-Brackin, 2020). Research on aptitude–treatment interaction has been ongoing since the MLAT's inception. However, its progress has been hindered by the limited availability and validation of aptitude measures. These measures should be underpinned by theoretical frameworks and supported by accumulated empirical evidence that links aptitude components to L2 learning.

In particular, two directions for aptitude–treatment interaction research merit exploration.

First, within instructed L2 learning contexts, little research has compared learning outcomes across different aptitude components using approaches, such as immersion, task-based instruction, and traditional grammar-based instruction (Li & Zhao, 2021).

Moreover, it is important to consider optimal strategies for managing cognitive demands during classroom activities and aptitude assessments, particularly when dealing with vulnerable populations, which also relates to the concern about sampling bias discussed in Section 7.3.2.

Second, in naturalistic (such as long-term immersion) L2 learning contexts, limited research has investigated the role of aptitude in explaining learning outcomes (e.g., Bolibaugh & Foster, 2021). In untutored settings, for any explicit learning to take place, learners must direct their own attention to discern meaningful patterns from the input. Thus, high aptitude or specific components like language analytic ability may enhance learning success. To investigate this, an aptitude battery in the aural modality, such as TALL, may be useful. However, Roehr-Brackin (2020) highlights that the existing studies in naturalistic learning environments have predominantly depended on short-term interventions or the use of artificial or mini-languages. This raises questions about the depth of our understanding regarding aptitude–treatment interaction in such contexts.

7.3.4 … related to Open Research practice

The investment of time and funding to develop TALL into an internet-based instrument merited making the instrument openly accessible to other researchers to address the restriction issues around most of the existing aptitude batteries. With this goal in mind, a roadmap has been drawn to first deploy TALL on an openly accessible test platform (https://tall-webtest.com), enabling separate access for researchers and invited test takers. To be specific, researchers can self-administrate data collection on the platform by generating test codes, download and upload datasets at the item level, and navigate the test manuals and try out demo tests for their own research interests or for pedagogical (e.g., research training) purposes. Meanwhile, participants are allowed to access via the "Test-taker Entry" only, using the test codes they receive from researchers; hence, they remain naive to the full test, as they cannot access it without a code. The effort to develop functionality in access control and self-administration is particularly necessary, not only to secure the quality of the data collected, but also to maintain the sustainability of TALL as a resource that can be used by the community without relying on any individual to administer it. Furthermore, in the longer term, TALL will be developed into an open data tool that can amass data collected by using this uniform battery and offer open access to the accumulated data pool, collected across different sites by different teams, to the research community.

TALL, functioning as a shared open research infrastructure for data collection and accumulation, holds potential aligned with the vision that is crucial for the advancement of SLA theory and research practices (as argued by MacWhinney, 2017).

First, TALL can be used as a reliable once some minor revisions have been implemented following the findings of the current study. These revisions may involve adjusting test items to improve the alignment of item difficulty and discrimination between the two versions in TALL_VL and TALL_SD (as discussed in Section 4.7.2.1), revising verbal stimuli for recall in TALL_CST, and implementing technical refinements to record stimuli recall in TALL_SNWR and TALL_CST (as discussed in Section 4.7.2.2). TALL has the potential to serve as a convenient measure of aptitude that allows for remote data collection at no cost. This may facilitate better sampling practices and multi-site studies to obtain larger and more diverse samples.

Second, TALL reduces duplication of effort where researchers 'reinvent the wheel'. Third, it constrains researcher degrees of freedom (Simmons et al., 2011) caused by methodological variation, which may adversely affect (comparability of) results. This should facilitate replication and reproducibility in research (Bolibaugh et al., 2021; Marsden et al., 2018). Finally, in the long run, TALL could amass a cumulative open data pool; that is, aggregated data collected by using a uniform battery, which helps reduce heterogeneity between studies, a prerequisite for high-quality syntheses of research findings (Plonsky & Ziegler, 2016).

7.4 Contributions of the current research

In conclusion, the present study has made significant contributions to the field of foreign language aptitude research.

First, despite the presence of aptitude batteries such as the MLAT (Carroll & Sapon, 1959), the LLAMA tests (Meara, 2005; Meara & Rogers, 2019), and the Hi-LAB (Linck, 2013) for over 60 years, 18 years and 10 years respectively, they had encountered theoretical and methodological challenges: (i) not keeping pace with evolving theoretical frameworks of aptitude constructs pertaining to L2 learning; and (ii) lacking comprehensive validation evidence from research conducted independently of the authors of the aptitude measures, engaging diverse learner populations. The present study addressed the first challenge by developing a new aptitude battery, TALL, with theoretically conceptualised componential constructs. It then provided preliminary evidence of TALL's internal validity and predictive validity in explaining L2 proficiency. Furthermore, TALL has been developed openly accessible for further validation checks by other researchers, taking the initial step to address the second challenge. By creating TALL test suites in aural and written modalities and incorporating language domain-specific items, this study bridges empirical and methodological gaps by systematically exploring the effects of input modality and reconciling domain generality–specificity disparity (Wen et al., 2017; Wen & Skehan, 2021) in measuring aptitude by including domain specific measures in the WM subtests.

The second contribution of the present study is the demonstration of a series of practices aimed at preventing misconduct and questionable research practices. These practices are designed to safeguard the methodological rigour of quantitative research within the field of applied linguistics. They encompass various aspects, including the development of a validation plan for a thorough scrutiny of reliability and internal validity, conducting prior power analysis to ensure an adequate sample size and post hoc power analysis to inform sample size requirements for future research. Additionally, datasets were careful prepared with considerations about removing outliers, while assumptions for statistical analyses were rigorously examined. Selection of indices was based on their ability to capture factorial dimensions and item quality nuances. Furthermore, to promote research transparency, all materials, data and R code are openly available on the OSF site (https://osf.io/czqxt/) and field-specific IRIS repository (www.iris-database.org) upon the approval of this thesis. This accessibility encourages replications and further scrutiny of TALL for future refinements. TALL has been developed into a tangible product with the potential to evolve into an open research infrastructure. This infrastructure aims to facilitate less heterogeneous data collection and data access for high-quality synthesis research. The development of this aptitude battery prototype underscores the importance of empirical testing. Such testing can be achieved on a large-scale data collected across time and space with the open availability of the battery (Pan & Marsden, under review).

This research, while primarily a methodological endeavour involving the development of a new aptitude battery and the provision of initial evidence for its validation, contributes to our understanding of aptitude itself. TALL effectively operationalises the componential constructs of aptitude as proposed in the Stage Approach and the P/E model. The findings indicate that these components are distinct from one another, and most of them exhibit significant positive correlation. Moreover, all five subtests in both suites load almost evenly on the first principal component, displaying no redundancy and explaining substantial proportions of the total variances. The data align well with the four-primary-factor structure informed by theoretical frameworks.

The research also demonstrates the significant influence of modality on measuring aptitude components (associative memory, language analytic ability, and executive control in WM). It indicates that performance related to these constructs is consistently better when tested in the written modality compared to the aural modality. Importantly, the effects of modality may vary in their impact on learners at different ability levels, emphasising the importance of considering modality as a moderating variable when examining the role of aptitude in diverse L2 learning outcomes among wider populations in various learning conditions.

Furthermore, test-practice effects are evidenced in certain language learning subtests (i.e., TALL_VL and TALL_LA). This finding warrants further investigation, especially when learners take an aptitude test more than once. The question of whether aptitude can be improved through such repetitions is intriguing and relevant to whether aptitude is a stable trait that is unsusceptible to training.

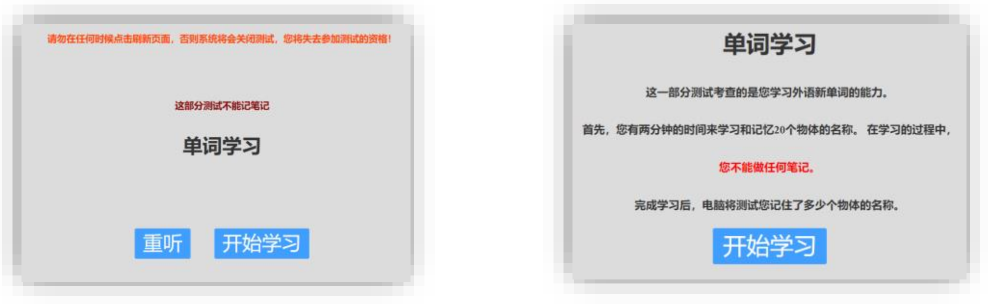
Appendix A: Subtest manuals

TALL_VL (v1.0): Vocabulary Learning

TALL_VL is the subtest designed to measure participants' ability to learn new vocabulary items that they are exposed to. This subtest is intended to assess the retrieval memory component of aptitude.

This subtest can be administered in either the aural modality, using the TALL aural suite, or the written modality, using the TALL written suite. Participants will have **two minutes** to learn 20 new vocabulary items that correspond to 20 familiar nouns in a novel language. They will then be tested on their ability to match the given vocabulary items to the corresponding pictures of the nouns.

The opening pages of TALL_VL provide participants with test instructions, as shown in Figure VL.1.



请勿在任何时候点击刷新页面，否则系统将会关闭测试，您将失去参加测试的资格！

这部分测试不能记笔记

单词学习

(in the aural suite)

单词学习

这一部分测试考查的是您学习外语新单词的能力。

首先，您有两分钟的时间来学习和记忆20个物体的名称。 在学习的过程中，

您不能做任何笔记。

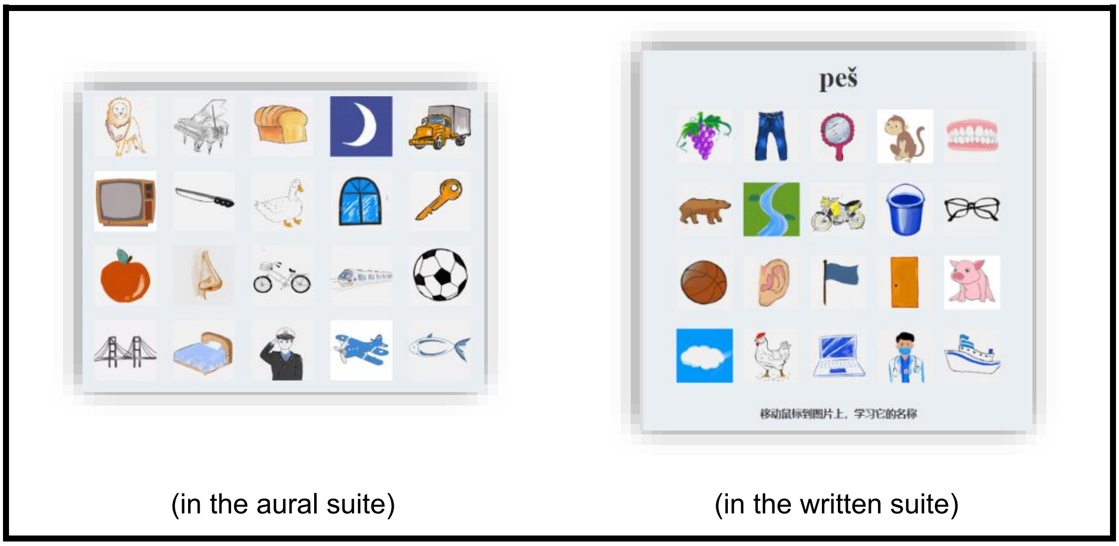
完成学习后，电脑将测试您记住了多少个物体的名称。

(in the written suite)

Figure VL.1 Opening pages

TALL_VL has two sequential phases.

In the learning phase, participants will be exposed to 20 pictures arranged in a fixed layout on the screen (see Figure VL.2). In the aural suite, participants can listen to the name of the picture by clicking on the corresponding picture with the mouse. In the written suite, participants can read the name of the picture by hovering over the corresponding picture with the mouse cursor.



移动鼠标到图片上，学习它的名称

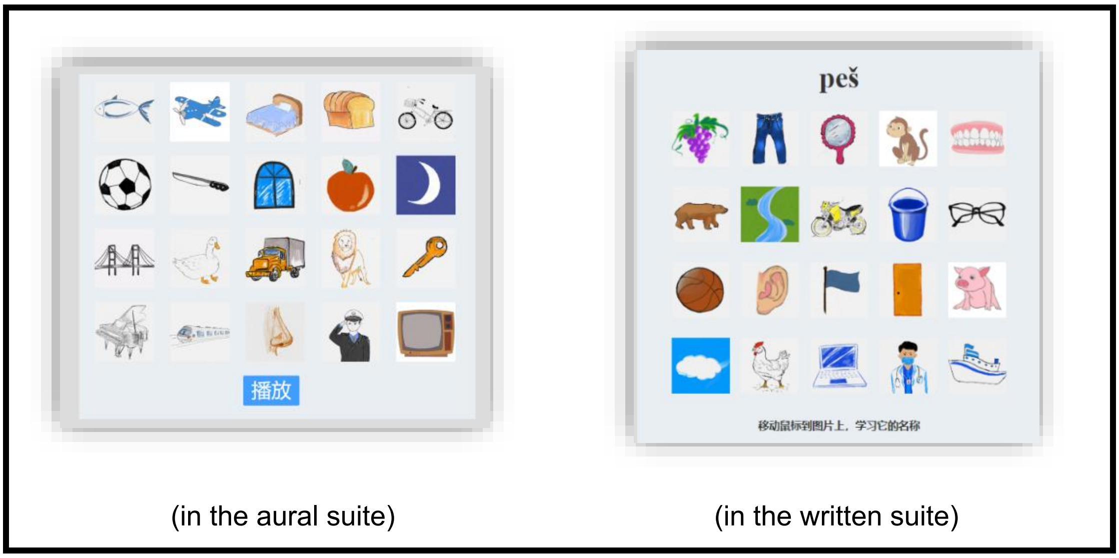
(in the aural suite)

(in the written suite)

Figure VL.2 The learning phases of TALL_VL

Participants are allowed to click on or hover over any picture as many times as they wanted, with a two-minute countdown bar displayed on the screen to help them manage their learning pace.

In the testing phase, participants will be presented with the same 20 pictures on the screen as in the learning phase, but the layout of the pictures will be randomly arranged. They will then be presented with the acoustic or written forms of the 20 vocabulary items (see Figure VL.3), one at a time in random order. Participants will be tested on their ability to identify the correct picture that matches the form they hear or read by clicking on the corresponding picture.



移动鼠标到图片上，学习它的名称

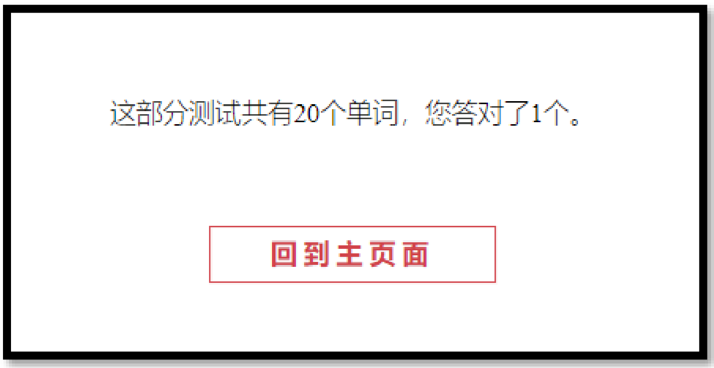
(in the aural suite)

(in the written suite)

Figure VL.3 The testing phases of TALL_VL

Participants will be allowed to take the test at their own pace without a time limit to complete the testing phase. Their performance will be scored based on the number of vocabulary items they can correctly match with the corresponding pictures, with a total score of 20.

Participant’s final score will be displayed immediately upon completion of the subtest (as in Figure VL.4). Participants can click on “回到主页面”（*Back to the homepage*）to move onto the next subtest.



这部分测试共有20个单词，您答对了1个。

Figure VL.4 The score report page of TALL_VL

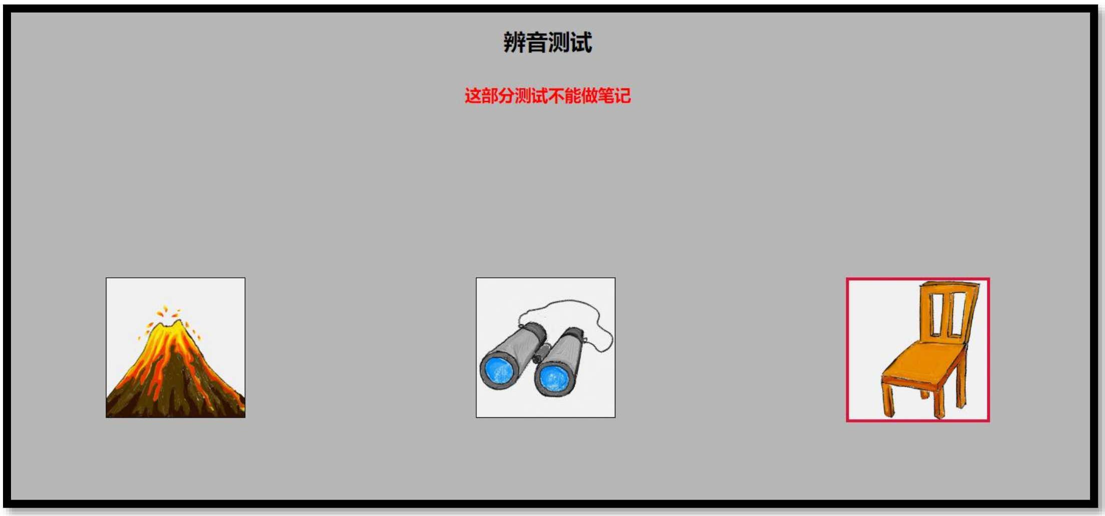
TALL_SD (v1.0): Sound Discrimination

TALL_SD is the subtest designed to measure participants' phonemic coding ability to identify and retain unfamiliar sounds in a new langauge.

This subtest is administered in the aural modality in both the TALL aural suite and the written suite. Participants are exposed to three basic sounds, either in isolated form or embedded in short sentences. Their phonemic coding ability is tested by discriminating these sounds correctly when they are embedded in 30 sentences.

TALL_SD has two sequential phases.

In the learning phase, participants listen to three isolated basic sounds while seeing the corresponding pictures on the screen (see Figure SD.1).

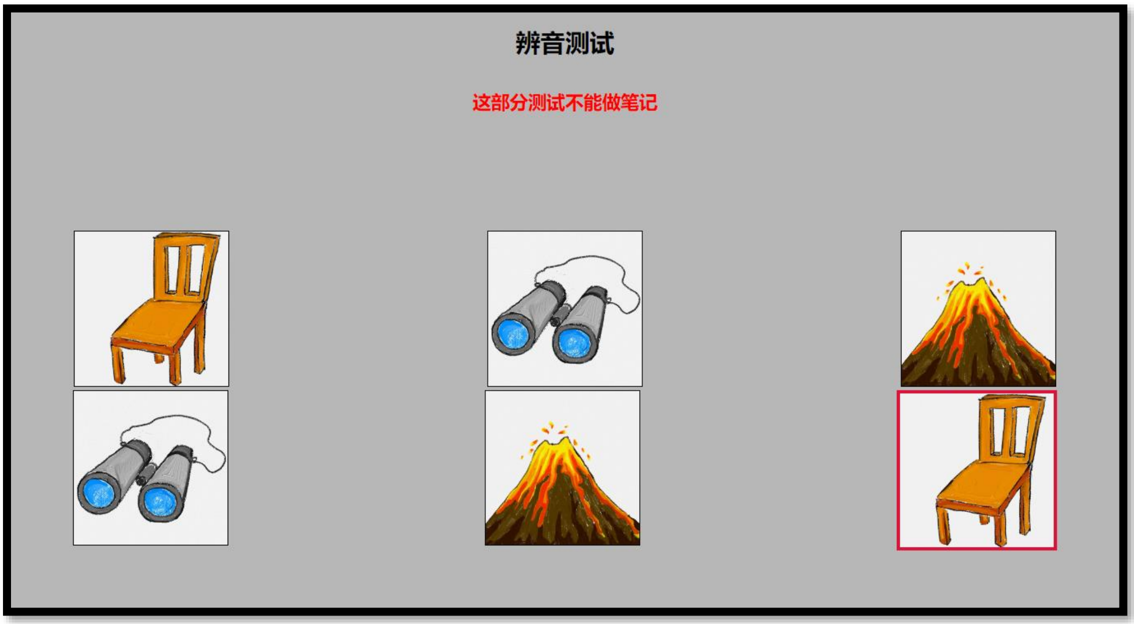


辨音测试

这部分测试不能做笔记

Figure SD.1 Learning three basic sounds in TALL_SD

Then, they listen to four sets of phrases, each containing three phrases with one of the three basic sounds embedded. While a phrase is played, the corresponding picture of the basic sound embedded in this phrase is highlighted to show the match between the basic sound and its meaning, as shown in Figure SD.2.

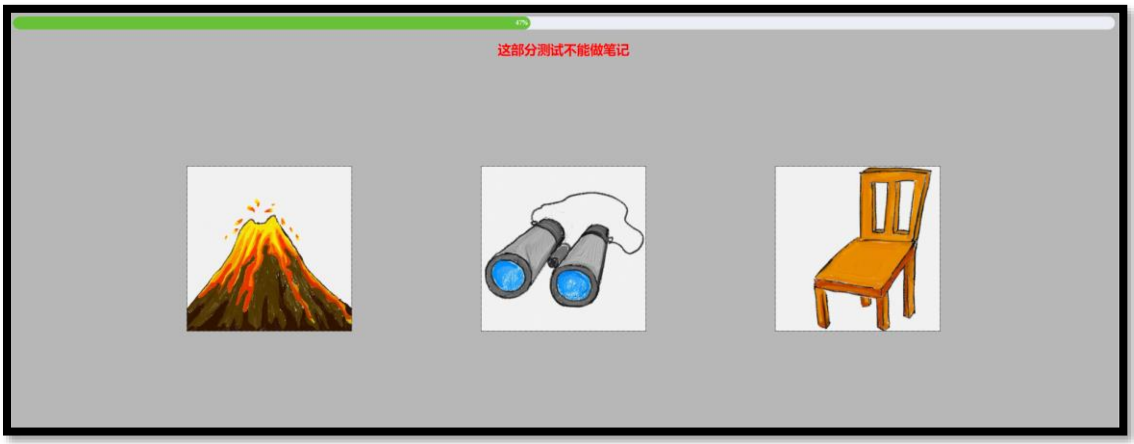


辨音测试

这部分测试不能做笔记

Figure SD.2 Learning phrases containing the basic sounds in TALL_SD

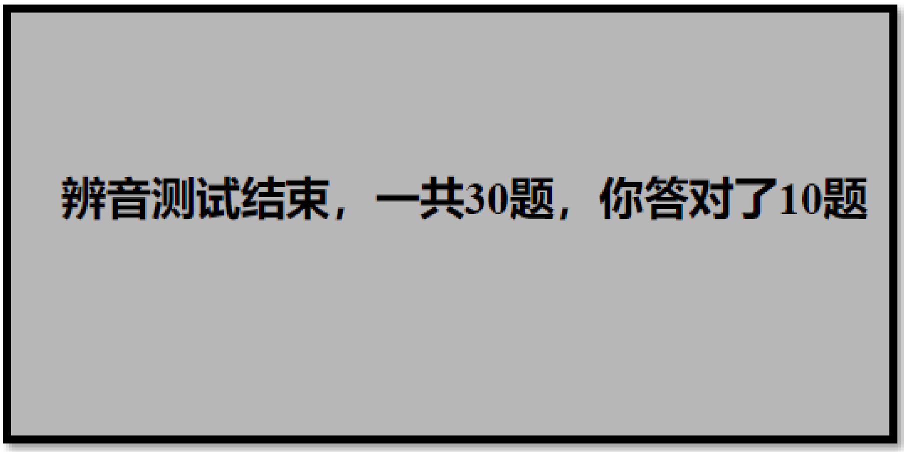
In the testing phase, participants are presented with the same three pictures that correspond to the basic sounds as in the learning phase. They listen to 30 test stimuli, one at a time, and are required to discriminate which of the three sounds is embedded in the stimuli by clicking on the corresponding picture of the sound. Participants have a maximum of 15 seconds to make the choice for each stimulus. There are 30 stimuli in total in the testing phase, and each basic sound is presented equally 10 times. The order of all stimuli is randomly assigned, and a bar on the screen shows the testing progress, indicating the proportion of total testing items participants have completed (see Figure SD.3).



这部分测试不能做笔记

Figure SD.3 The testing phase in TALL_SD

Participants’ performance in this subtest is scored by the number of the correct choices they make, with a total score of 30. A participant’s final score is displayed immediately upon the completion of this subtest (as in Figure SD.4).



辨音测试结束，一共30题，你答对了10题

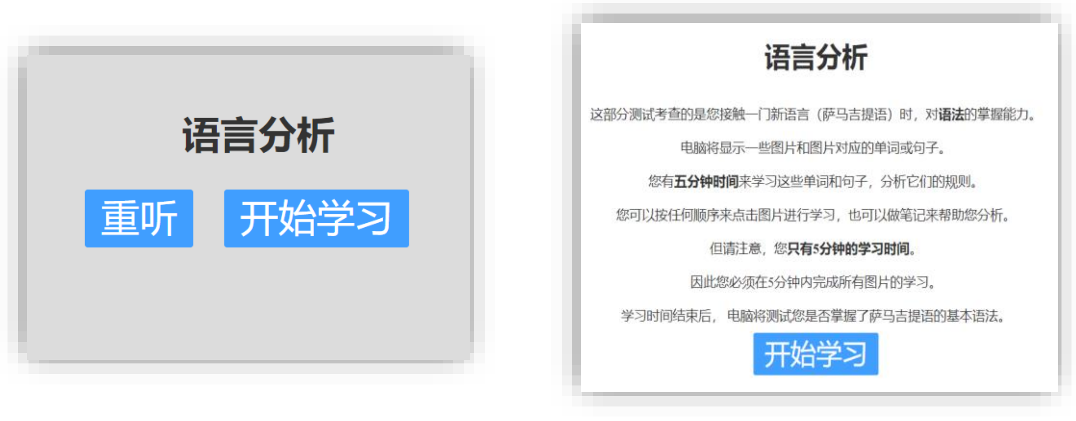
Figure SD.4 The score report page of TALL_SD

TALL_LA (v1.0): Language Analysis

TALL_LA is the subtest designed to measure participants' language analytic ability in learning grammatical features in a miniature language adapted from Lithuanian.

This subtest can be administered in either the aural modality, using the TALL aural suite, or the written modality, using the TALL written suite. Participants will have **five minutes** to learn new vocabulary items and grammatical rules in a novel language. Their language analytic ability will be tested by choosing the correct sentences that corresponding to the given pictures.

The opening pages of TALL_LA provide participants test instructions are shown in Figure LA.1.



(in the aural suite)

语言分析

这部分测试考查的是您接触一门新语言（萨马吉提语）时，对**语法**的掌握能力。

电脑将显示一些图片和图片对应的单词或句子。

您有**五分钟时间**来学习这些单词和句子，分析它们的规则。

您可以按任何顺序来点击图片进行学习，也可以做笔记来帮助您分析。

但请注意，您**只有5分钟的学习时间**。

因此您必须在5分钟内完成所有图片的学习。

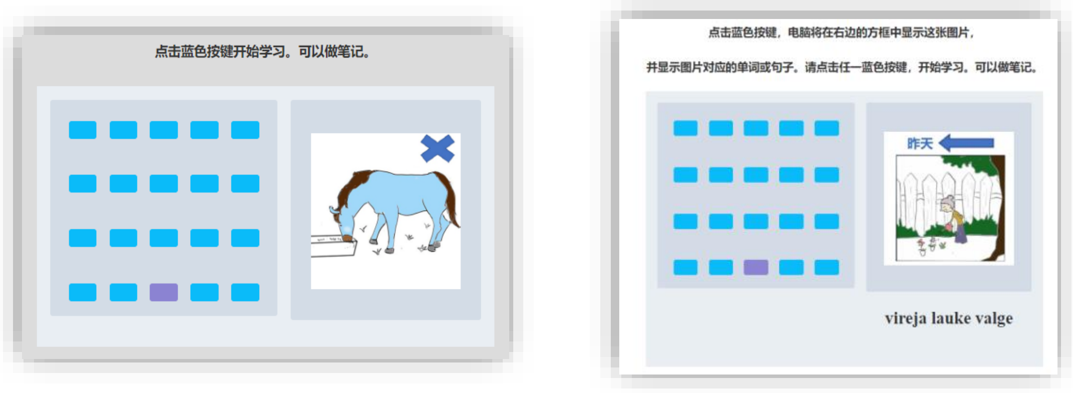
学习时间结束后，电脑将测试您是否掌握了萨马吉提语的基本语法。

(in the written suite)

Figure LA.1 The opening pages of TALL_LA

TALL_LA has two sequential phases.

In the learning phase, participants will click one of the blue buttons arranged in a grid on the screen and then be presented with a picture. As shown in Figure LA.2, in the aural modality, participants will hear verbal phrases or sentences that describe the meaning of the displayed pictures, while in the written modality, they will read the verbal phrases or sentences on the screen.



点击蓝色按键开始学习。可以做笔记。

(in the aural suite)

点击蓝色按键，电脑将在右边的方框中显示这张图片，

并显示图片对应的单词或句子。请点击任一蓝色按键，开始学习。可以做笔记。

vireja lauke valge

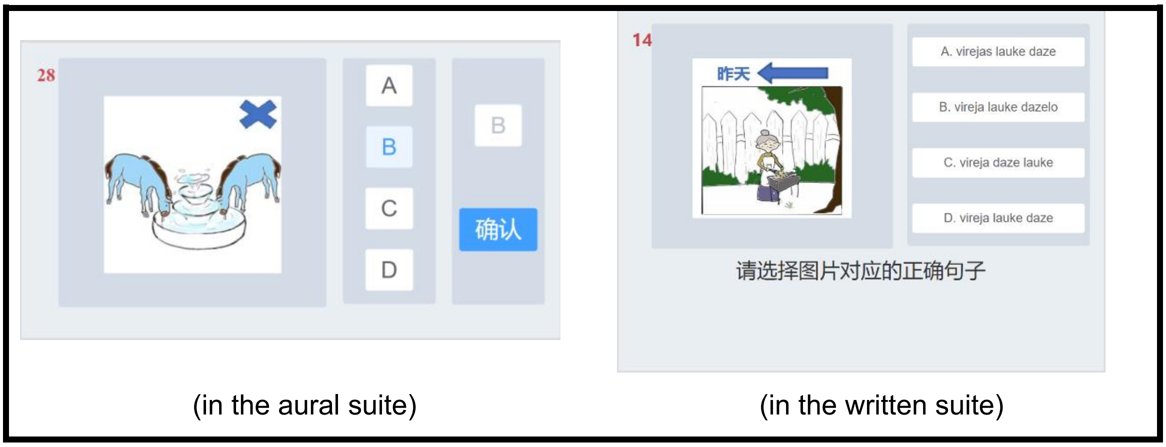
(in the written suite)

Figure LA.2 The learning phases of TALL_LA

Participants have **five minutes to explore 20 pictures** corresponding to 20 phrases and sentences consisting of vocabulary items of two nouns, three verbs, and two adjectives or adverbs. Morphological and syntactic properties are presented in the sentences, including three morphosyntactic rules (i.e., nominal endings, verbal inflections, and word order) in the target language. Participants are allowed to click the pictures in any order and as many times as they wanted in the learning phase, with a five-minute countdown bar displayed on the screen to help them manage their learning pace.

In the testing phase, participants are presented with pictures one at a time and are required to choose the correct sentence from four given options displayed on the screen to describe the meaning of the picture.

As shown in Figure LA.3, in the aural modality, participants are required to click each button of the four optional choices to listen to the sentences and then decide which option is the correct description. In the written modality, the four options are displayed on the screen for them to choose by clicking the correct optional sentence that describe the picture.



A. virejas lauke daze

B. vireja lauke dazelo

C. vireja daze lauke

D. vireja lauke daze

请选择图片对应的正确句子

(in the aural suite) (in the written suite)

Figure LA.3 The testing phases of TALL_LA

TALL_LA has no time limit for participants to complete all 30 testing items that are presented in random order, and the testing progress is displayed by a bar on the screen showing the proportion of total testing items participants have completed.

Participants' performance is scored by the number of the correct choices they make, with the total score of 30. The final score is displayed immediately upon the completion of the subtest (Figure LA.4). Participants can click on "结束"（*Completed*）to move onto the next subtest.

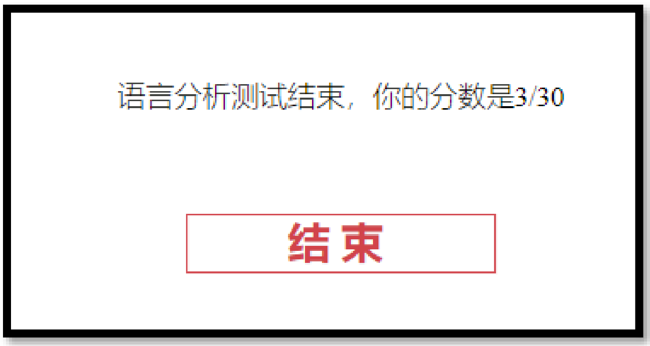

语言分析测试结束，你的分数是3/30

Figure LA.4. The score report page of TALL_LA

TALL_SNWR (v1.0): Serial Nonwords Recall

TALL_SNWR is the subtest designed to measure participants' phonological short-term memory in working memory by requiring participants to repeat a series of nonwords in the order they are presented.

This subtest is administered in the aural modality in both the aural suite and the written suite.

The opening page of TALL_SNWR provide participants test instructions is shown in Figure SNWR.1.

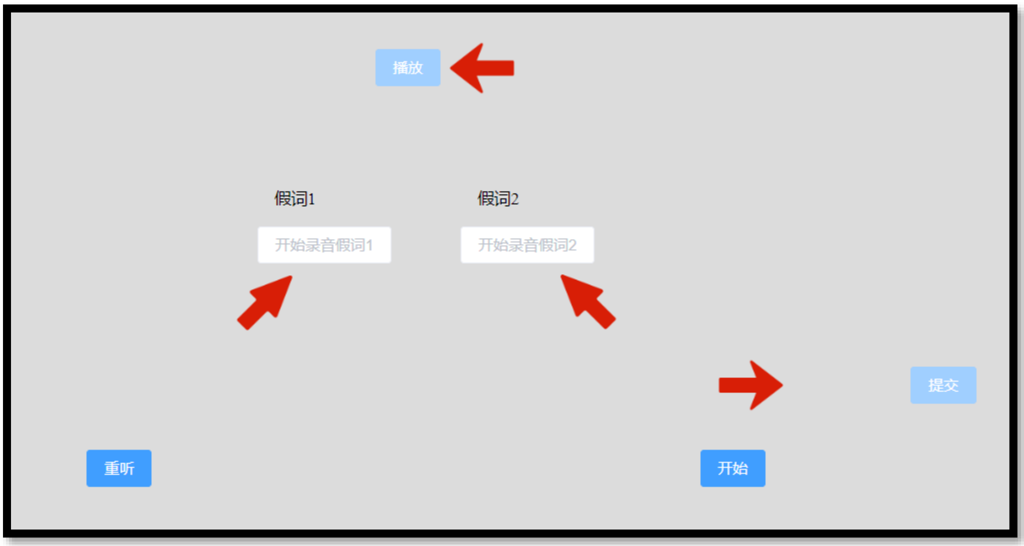
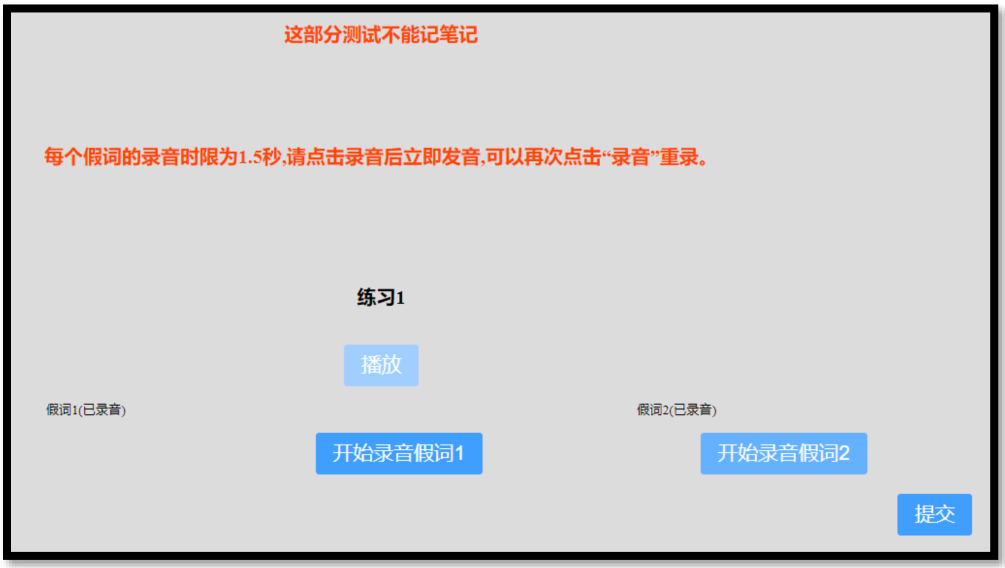


Figure SNWR.1. Test instruction page of TALL_SNWR

TALL_SNWR has two sequential phases.

In the practice phase (see Figure SNWR.2), participants are provided with three practice trials to become familiar with the experimental format. These trials have two, three, and five nonwords, respectively, and they are different from the trials in the testing phase.



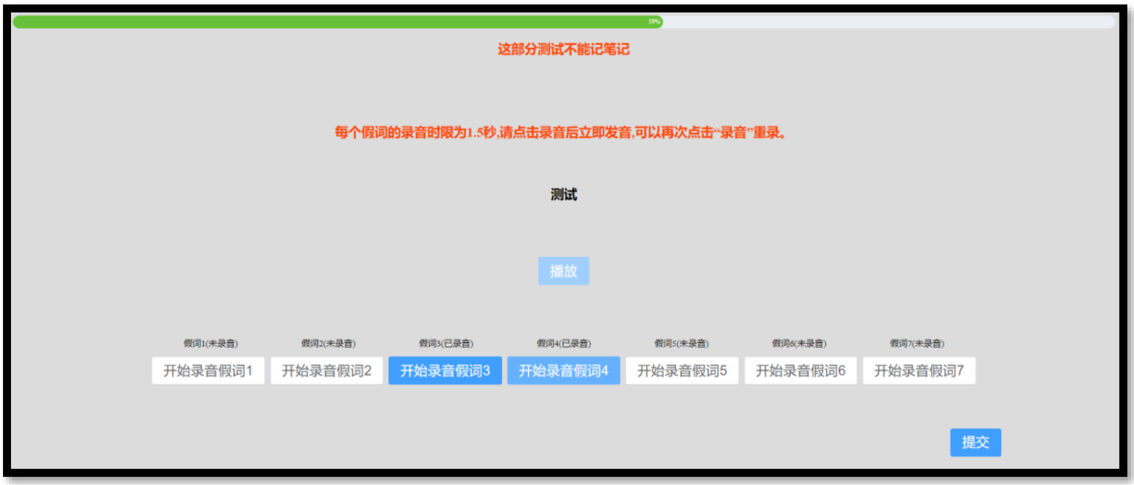
这部分测试不能记笔记

每个假词的录音时限为1.5秒,请点击录音后立即发音,可以再次点击“录音”重录。

Figure SNWR.2 The practice phase of TALL_SNWR

As shown in Figure SNWR.2, participants will listen to a trial containing a series of nonwords presented sequentially with identical speed (1000 ms) and intervals (1500 ms). They are then required to repeat the nonwords in the display order by clicking the corresponding ‘开始录音假词 1’ (*start recording nonword No.1*) buttons on the screen. After completing the recordings of the nonwords in each trial, participants must submit their recordings by clicking the ‘提交’ (*submit*) button on the screen.

The testing phase follows the same procedure as the practising phase (see Figure SNWR.3): participants first listen to a trial with a series of nonwords, then repeat the nonwords in the presented order by clicking the recording buttons, submit their recordings to complete the trial, and move on to the next trial until the end of the subtest.



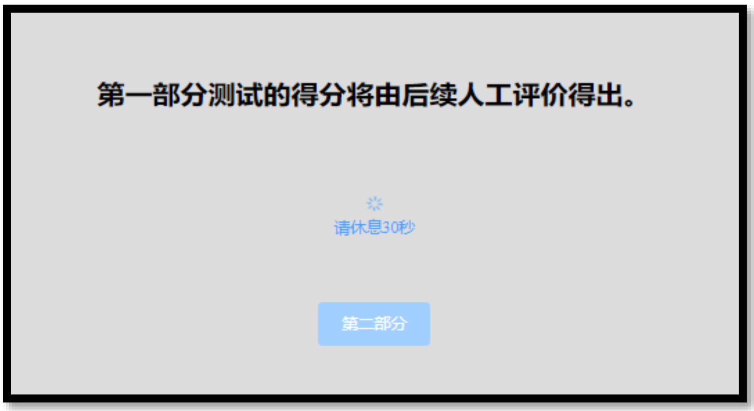
这部分测试不能记笔记

每个假词的录音时限为1.5秒,请点击录音后立即发音,可以再次点击“录音”重录。

Figure SNWR.3 The testing phase of TALL_SNWR

Seventeen trials with 74 nonwords are randomly presented in the testing phase, with each trial having nonwords ranging from two to seven. Participants are allowed to record and submit their recall of each trial within 30 seconds. If they fail to submit the recall of a trial within the time limit, the test program will automatically move on to present the next trial. The progress of the testing phase is displayed by a bar on the screen that shows the proportion of the total testing trials a participant has completed.

The final score of TALL_SNWR is the number of nonwords that are assessed manually as being correctly articulated in the correct display order, with a total score of 74. Participants will be notified about the scoring plan (see Figure SNWR.4). They can take a short break (min. 30 seconds) and click on "第二部分" (*Part Two*) to move onto the second part of working memory test, that is, the subtest of Complex Span Tasks in TALL.



第一部分测试的得分将由后续人工评价得出。

请休息30秒

Figure SNWR.4 The closing page of TALL_SNWR

TALL_CST (v1.0): Complex Span Task

This subtest is used to measure participants' executive control capacity in working memory. Essentially, the subtest comprises two intertwined components of storage and processing. The processing compoenet is inserted between the stimuli to be retained by participants, as a distractor, to prevent their rehearsal. The primary objective of participants in the test is to retain the presented stimuli despite the interference caused by the distractor.

TALL_CST uses verbal stimuli in both processing and recalling tasks. The stimuli for sentence meaning processing are developed either in the aural modality in the TALL aural suite, or in the written modality in the written suite.

Figure CST.1 and Figure CST.2 show the experimental paradigms of TALL_CST that have dual tasks design to engaged participants' processing the meaning of sentence stimuli and recalling the letters in the correct display order.

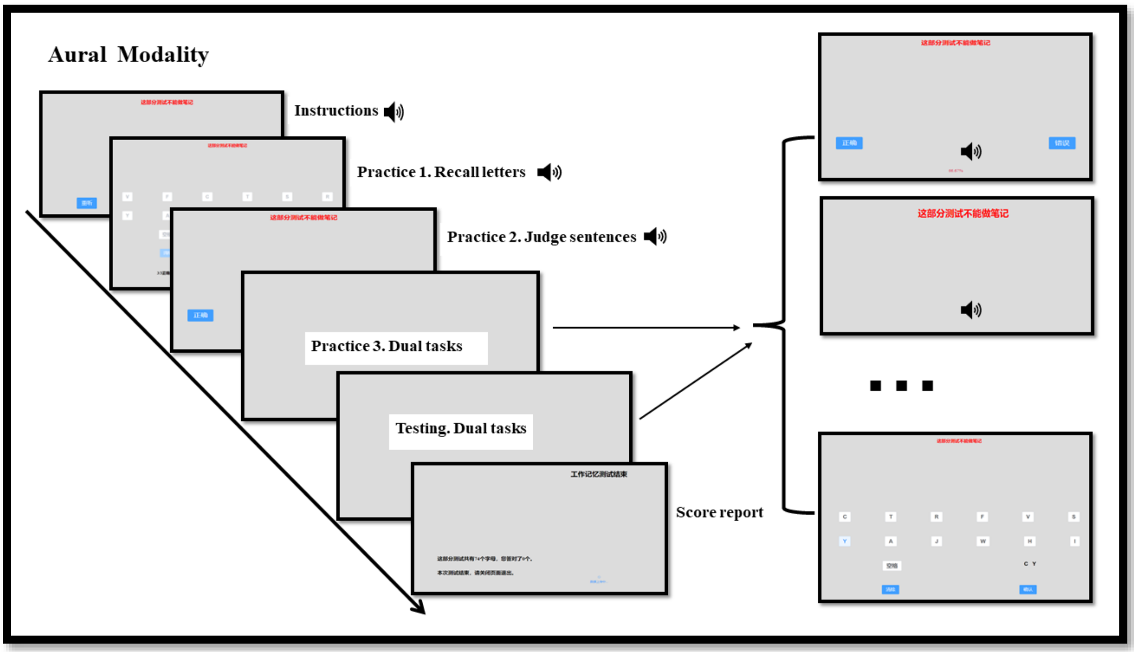


Figure CST.1 Experimental paradigm of TALL_CST (in the aural suite)

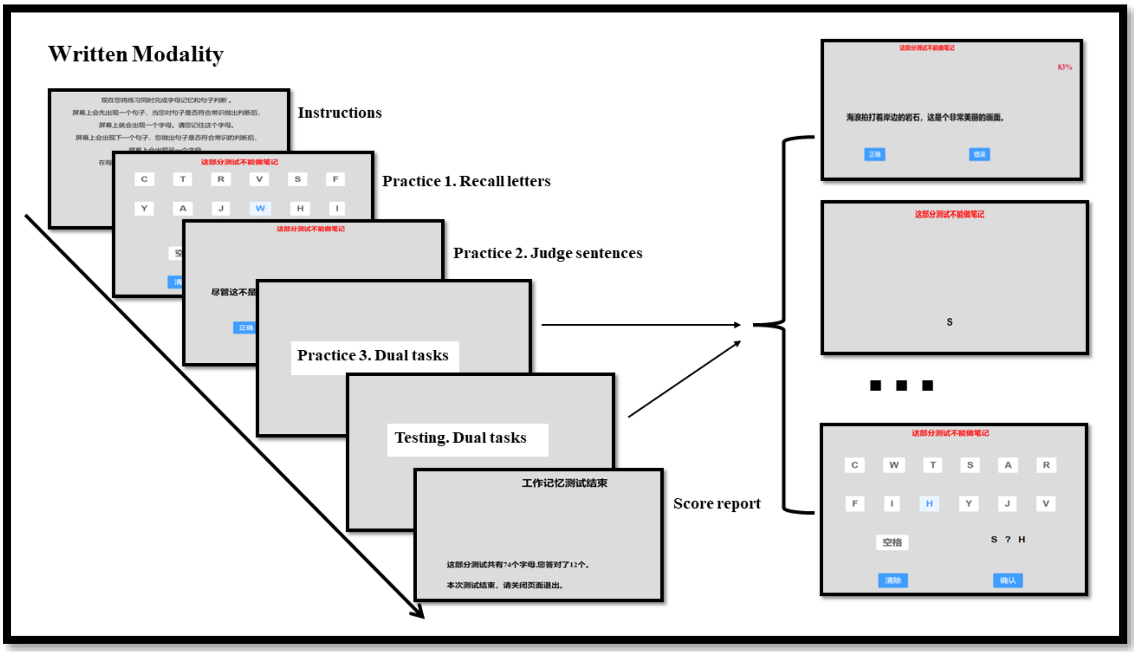

Written Modality
现在您将练习同时完成字母记忆和句子判断。
屏幕上会先出现一个句子，当您对句子是否符合常识做出判断后，
屏幕上就会出现一个字母。请您记住这个字母。
屏幕上会出现下一个句子，您做出句子是否符合常识的判断后，
Instructions
这部分测试不能做笔记
C T R V S F
Y A J W H I
Practice 1. Recall letters
这部分测试不能做笔记
尽管这不是
正确
Practice 2. Judge sentences
Practice 3. Dual tasks
Testing. Dual tasks
工作记忆测试结束
这部分测试共有74个字母,您答对了12个。
本次测试结束，请关闭页面退出。
Score report
这部分测试不能做笔记
83%
海浪拍打着岸边的岩石，这是个非常美丽的画面。
正确
错误
这部分测试不能做笔记
S
这部分测试不能做笔记
C W T S A R
F I H Y J V
空格
S ? H
清除
确认

Figure CST.2 Experimental paradigm of TALL_CST (in the written suite)

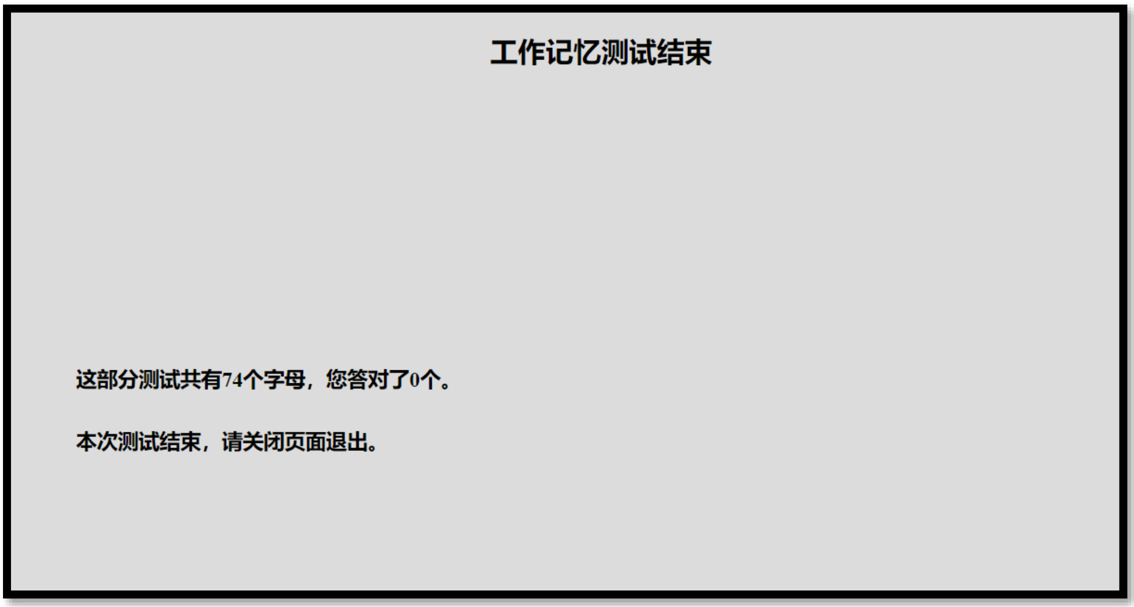
The practice phase is design in three steps to help participants become familiar with the testing procedure. First, participants will practise the sentence processing task by listening to a sentence (in the aural modality) or reading a sentence on the screen (in the written modality) in Chinese, then make a semantic judgement on whether the sentence is sensible in terms of meaning by clicking the button '正确' (*correct*) or '错误' (*incorrect*). Second, participants will practise the recall task, in which they listen to or read a string of letters in English, then click the corresponding letters on the screen based on the correct display order of the letters. Third, participants will have the opportunity to practise the testing format of the combination of sentence processing task and letter recall task with a trial. That is, they make a judgement of the plausibility of the meaning of a sentence first, then they are presented with a letter followed by another sentence judgement, then a letter to recall, until the end of the trial when they are required to recall all the letters in the correct presentation order.

There are 15 trials in the testing phase, each containing 3 to 7 sentences. The total number of letters to be recalled, 74, is the same as the total number of sentences. Participants follow the same procedure as the third step in the practice phase, that is, the combination of processing and recall tasks. They will be randomly presented with a trial containing sentence stimuli for meaning processing and letter stimuli for recall. They are required to recall the letters in the correct order at the end of the trial. The test program will proceed to the next sentence stimulus if participants fail to make a semantic judgment within the time limit based on their individual performance in the sentence processing tasks during the practice phase. A stimulus without a response to the judgment will be recorded as incorrect. Participants are also required to complete the recall of the letter string at the end of each trial within 30 seconds. Twelve English letters are randomly assigned in all trials and presented (aurally in the aural modality and on the screen in the written modality) with identical speed (800 ms) and interval (200 ms), while the sentence stimuli in each trial are presented in a constant order, and the sequence of the trials is randomised. Participants' judgements of the semantic plausibility of the sentence stimuli and their recall of the string of letters are stored for data analysis.

For self-monitoring purposes, participants will be provided with the percentage of accuracy in sentence processing throughout the testing phase, and a progress bar on the screen shows the proportion of the total testing trials a participant has completed.

Participants' performance in this subtest is scored by the total number of correct letters that are recalled in the correct order, with a total score of 74.

The final score will be displayed immediately upon the completion of the subtest (see Figure CST.3). Participants can exit TALL by closing this page.



工作记忆测试结束

这部分测试共有74个字母，您答对了0个。

本次测试结束，请关闭页面退出。

Figure CST.3 The closing page with the final score of TALL_CST

Appendix B: Consent forms

[English version]

PhD Project: Adult learners' development of online predictive processing

This project is being conducted by Junlan Pan, a PhD researcher at the Department of Education of the University of York (UK). **This web-based test** (Chinese adult learners' cognitive individual differences) is a part of the PhD project "Adult learners' development of online predictive processing", conducted from 2019 at University of York. The informed consent form is produced in the Chinese version for participants' use and remains identical in content with the English version for departmental ethical approval.

Dear Participant,

My name is Junlan Pan. I am a PhD researcher at the Department of Education, University of York. I am currently carrying out a research project about "Adult learners' development of online predictive processing". I would like to invite you to take part in a web-based test of cognitive individual differences, which is a part of this research project.

Before agreeing to take part, please read this information sheet carefully and let me know if anything is unclear or you would like further information.

What is the purpose of this study?

The study is designed to validate the web-based measurement of Chinese learners' cognitive individual differences in learning a foreign language.

What would this mean for you?

As I would like to measure learners' differences in learning a foreign language in different modes, participating in this test has two phases: (1) completing the aural version of the test before the end of 2020, and (2) completing the written version of the test in at least four weeks after the completion of phase (1).

It will take you up to 60 minutes to complete each phase, and each phase must be completed in one go. All participants who complete both versions of the test will receive 50 Yuan as a Thank You. At the end of the second phase of this test, I will provide you a summary of your testing results with descriptions of your related cognitive abilities in language learning.

At the end of this test, I will ask if you would be willing to participate in an online interview to explore some themes from the test. I will select only 10 participants for this part.

Participation is voluntary

Participation is optional. If you do decide to take part, you will be asked to complete a consent form online and I will email you a copy of information sheet for your records. You may feel a manageable degree of stress when you work on the testing items. If you change your mind at any point during the study, you will be able to withdraw your participation without having to provide a reason. Once the test is completed you can request for the data to be withdrawn by emailing jp1763@york.ac.uk up to 2 weeks after the data is collected.

Storing and using your data

The data that you provide (e.g., test results) will be stored anonymously by code number. Any information that identifies you will be stored separately from the data.

The data will be stored in a password protected file and will only be accessible to me and my supervisor, Prof. Emma Marsden, who is involved in the project. The anonymous data may be used in presentations, online, in research reports, in project summaries or similar. Your individual data will not be identifiable but if you do not want the data to be used in this way please do not participate in this test. In addition, anonymous data may be used for further analysis. The data will be kept for indefinitely in the digital repository of IRIS (www.iris-database.org) for the practice of Open Science.

I will process personal data for research purposes under Article 6(1) (e) of the General Data Protection Regulation (GDPR): Processing is necessary for the performance of a task carried out in the public interest. Special category data is processed under Article 9 (2) (j): Processing is necessary for archiving purposes in the public interest, or scientific and historical research purposes or statistical purposes.

For information about General Data Protection Regulation (GDPR) please follow the link:

https://www.york.ac.uk/education/research/gdpr_information/

Questions or concerns

If you have any questions about this participant information sheet or concerns about how your data is being processed, please feel free to contact Junlan Pan by email jp1763@york.ac.uk, or the Chair of Ethics Committee via email education-research-admin@york.ac.uk. If you are still dissatisfied, please contact the University's Data Protection Officer at dataprotection@york.ac.uk

I hope that you will agree to take part in. If you are happy for you to participate, please complete the form attached and click "submit" to return the form.

Many thanks for your help with this research.

Yours sincerely

Junlan Pan

Chinese learners' cognitive individual differences in FL learning

Consent Form

Please click each box if you are happy to take part in this research.

Statement of consent	Click each box
I confirm that I have read and understood the information given to me about the above-named research project and I understand that this will involve me taking part as described above.	
I understand that participation in this study is voluntary.	
I understand that my data will not be identifiable and the anonymous data may be used in publications, presentations and online.	
I confirm that I have read the information about GDPR	

Name:

Signature:

Date:

[Chinese version]

UNIVERSITY of York Department of Education

博士研究项目: 成人学习者在线预测加工能力的发展

项目由英国约克大学教育系博士研究生盘峻岚执行。本在线测试（中国成人学习者的认知个体差异）是博士项目“成人学习者在线预测加工能力的发展”（2019 年起）的一部分。中文知情同意书供测试参加者使用，与教育系伦理审查通过的英文版内容一致。

尊敬的测试参加人:

您好!

我叫盘峻岚, 是英国约克大学教育系的博士研究生, 正进行一项关于“成人学习者在线预测加工能力的发展”的研究项目。很荣幸地邀请您参加这个项目中的认知个体差异在线测试。

在您同意参加测试前，请仔细阅读以下信息，如果您有任何疑问或需要了解更多信息，请联系我。

本研究的目的是什么?

本研究旨在设计中国外语学习者的认知个体差异测试, 并证实其内在效度。

本研究对您意味着什么?

由于我们将使用不同的方式测试学习者的差异, 您需要分两次参加测试: （1）首先完成视听版或文字版的测试, （2）在前一次测试后至少四周后完成另一版本的测试。

每一次测试至少需要一个小时, 不中断地一次性完成。完成两次测试后您将收到 50 元人民币作为酬谢。本研究完成了第二阶段测试后, 我们将为您提供一份有关您学习外语的认知能力的报告。

自愿参加原则

本研究系自愿参加。如果您愿意参加, 我们将提供相关信息和知情同意书。测试中您会有适度的紧张感。如果您在测试过程中不想继续参加了, 您可以选择退出, 不需要解释理由。

匿名与保密

您提供的数据（比如测试的分数）将编码保存。任何可以辨识您身份的信息都将与数据分开保存。您可以在数据收集的过程中放弃参加本研究。在数据收集完毕后的两周内, 您也可以发邮件至 jp1763@york.ac.uk,要求撤销您的数据。

数据保存与使用

数据将加密保存, 这意味着在关于本研究的演讲、成果上线、研究报告、项目总结中的数据都将匿名。匿名数据将无法识别您的身份。另外, 匿名数据可能会用于后续的分析, 并

有可能通过开放数据库分享给其他研究人员进行分析。如果您不同意我们用这种方式使用您的数据，请您不要参加本测试。

我们将按照《通用数据保护条例》(GDPR)第 6(1)(e)条规定进行个人数据保护：处理数据是从事以公益为目的的工作的需要。

特殊目录数据的处理根据第 9(2)(j)条：处理数据是公益性的存档，或科学和历史研究目的，或统计目的的需要。

关于《通用数据保护条例》(GDPR)的信息，请登陆查询：
https://www.york.ac.uk/education/research/gdpr_information/

问题或顾虑

如果您有任何关于测试信息和数据处理的问题或顾虑，敬请联系盘峻岚：jp1763@york.ac.uk，或者联系伦理委员会主席：education-research-admin@york.ac.uk。如果您对他们提供的信息还有不满意之处，敬请联系约克大学数据保护官员：dataprotection@york.ac.uk。

我真诚地希望您能同意参加本测试。如果您愿意参加，请完成所附的知情同意书，点击"确认"发回。

非常感谢您对本研究的帮助！

此致

敬礼

盘峻岚

请点击"继续"确认您已经理解了上述信息，并同意参加本次测试。

中国外语学习者认知个体差异测试

知情同意书

如果您同意参加测试，请点击每一项确认。

同意内容	点击确认
我确认我已经阅读并理解了关于这一研究项目的信息，我理解研究需要我参与其中。	
我理解参加研究是自愿的。	
我理解本研究保存的我的数据将无法辨识我的身份，我的匿名数据将被用于发表、演示、上线或分析。	
我确认我已经阅读了关于 GDPR 的信息。	

姓名（签字）：

日期:

Appendix C: Participant’s background questionnaire

学习背景调查

您的性别:□ 男　　□ 女

1）您的最高学历:

□ 初中
□ 高中

□ 高职高专

□ 大学本科

□ 研究生

您的专业属于：

□ 人文门类 （文学、历史学、哲学、艺术学）

□ 社会科学门类（教育学、法学、经济学、管理学、军事学）

□ 理工门类（理学、工学）

□ 农医门类（医学、农学）

□ 以上学科的交叉

学科为：

2）您的外语学习经历

您学习（参加课程学习）了___ 年英语

1. 0-3 年以内

2. 3-6 年以内

3. 6-10 年以内

4. 10 年以上

您是否有海外学习/生活的经历? □ 是　　□ 否

您在哪个（或哪些）国家学习/生活过? （　　　）

您在这个（这些）国家学习/生活了多长时间? （　　）

您的高考英语成绩是__________分。

您参加过的其它英语水平考试有（可多选）

□ 大学英语四级，成绩为

□ 大学英语六级，成绩为

□ 雅思考试，成绩为

□ 托福考试，成绩为

□ 以上考试都未参加过

除英语之外，您学习（参加课程学习）过其它外语吗？（可多选）

□ 俄语，学习了____年

□ 日语，学习了____年

□ 法语，学习了____年

□ 德语，学习了____年

□ 西班牙语，学习了____年

□ 其他语种，学习了____年

您的其它外语水平考试成绩（可多选）：

□ 俄语，考试名：___ ；分数：

□ 日语，考试名：___ ；分数：

□ 法语，考试名：___ ；分数：

□ 德语，考试名：___ ；分数：

□ 西班牙语，考试名：___ ；分数：

□ 其他语种，考试名：___ ；分数：

□ 没有参加过水平考试

Appendix D: List of R markdown files

R code and output from analyses were documented in R markdown files and accessible on the OSF page (https://osf.io/3nxaw/) of the project. The list below explains these files and in which section they provide results.

Please note that these files must be downloaded to view; they are in HTML format and do not need to be opened in R.

RQ1 in Section 4.3.2.3:

1. Reliability and Unidimensionality checks (https://osf.io/ctpvd)

RQ1 in Section 4.4.2:

2. IRT (Rasch model) analyses for TALL_VL, TALL_SD, and TALL_LA (https://osf.io/uy4mn)
3. IRT (Generalised partial credit model) analyses for TALL_SNWR and TALL_CST (https://osf.io/brsjy)

RQ1 in Section 4.5.1:

4. Principal Component Analysis for both suites (https://osf.io/wafe6)

RQ1 in Section 4.5.2:

5. Confirmatory Factor Analysis for both suites (https://osf.io/ba2yh)

RQ2 in Section 5.3:

6. (Generalise) Linear Mixed-effects Modeling for TALL_VL, TALL_LA, and TALL_CST (https://osf.io/9q4ax)

RQ3 in Section 6.3:

7. Multiple Regression and Dominance Analysis (https://osf.io/h5n4g)

Appendix E: Data preparation protocol

Step 1: Data from the second round of TALL_CST, which was the last subtest participants took, were cleaned to remove erroneous or outlying data points. This step resulted in the data of participants who took two rounds of the test and completed all five subtests in the second session.

Step 2: Data of TALL_SNWR in the second session were manually checked to ensure that the recordings were available for scoring. This quality control step was crucial, as some participants might not have followed the test instructions or failed to set up the recording function on their PCs. The clean data from this step were then merged with the data obtained in Step 1, thereby returning the data of participants who had matched data in two WM subtests (TALL_SNWR and TALL_CST) from both test sessions.

Step 3: Data from TALL_LA were initially cleaned by removing erroneous data points and data from behavioural outliers. The resulting clean dataset was merged with the data obtained in Step 2, thereby returning the data of participants who had matched data from TALL_LA, TALL_SNWR and TALL_CST in two test sessions.

Step 4: Data of TALL_SD were cleaned, involving the removal of erroneous data points. The clean dataset was then merged with the data obtained in Step 3, consequently returning the data of participants who had matched data from TALL_SD, TALL_LA, TALL_SNWR and TALL_CST in two test sessions.

Step 5: Data from TALL_VL underwent cleaning process that involved removing erroneous data points and data from behavioural outliers. The clean dataset was merged with the data obtained in Step 4, thus returning the data of participants who had matched data from all five subtests in two test sessions.

Appendix F: IRT models and analysis protocol

This file documents the detailed information (in Section 3.3.5.2 and Section 4.4.1) about introducing Item Response Theory (IRT), the debated about using Rasch modelling in language testing research, the reasons for applying different statistic packages for Rasch modelling in the current study, and the stepwise IRT analyses performed on dichotomous data (from TALL_VL, TALL_SD, and TALL_LA) and polytomous data (from TALL_SNWR and TALL_CST).

1. Item Response Theory (IRT)

IRT is a psychometric assessment method used to investigate the relationship between a participant's response to a single test item and the overall performance on a latent trait the item is designed to measure. van der Linden (2016) defines IRT by three key principles. First, the focus is on how human subjects respond to individual test items rather than a predetermined score for the whole test, as is the case with the traditional test theory. Second, IRT recognises that responses are random and require a probabilistic model to explain their distribution. Third, IRT separates the parameters for test takers' abilities and the properties of the test, which distinguishes it from Classical Test Theory (CTT) that represents a test taker's expected score on a set of items as a linear combination of a true score and a test error, without separating the true score according to the separate effects further based on subject's ability and the property of the test.

As a model-based technique, item analyses based on IRT has gained several advantages over that on Classical Test Theory (CTT), which can make the inference of the abilities of a group of test takers according to the values of the indices, but the values may not be stable if the test is trailed on another group with variable ability. The considerable merit of IRT is that it enables more sophisticated analysis of test items independent from the whole test. It does do by measuring latent traits while considering the relative abilities of the trial group to understand the difficulty of the items for a broader range of test takers than those represented in one trial group. Additionally, IRT provides a more complete picture of how items function. Therefore, it is also known as Latent Trait theory (Crock & Algina, 2008), on which analytical approaches are allowed to calculate estimate parameters on both the item and the individual test taker that can provide nuanced information about person ability level, item difficulty, item discrimination, etc. As such, each item in a measurement can be treated as a separate and independent entity, which is fundamentally different from CTT that each item is assumed to have the same level of difficulty or the same amount of contribution to the overall test score (Draheim et al., 2018). In other words, the crucial difference between analyses on CTT and IRT can be compared to the difference between descriptive and inferential statistics (Knoch & McNamara, 2015). With descriptive statistics, the raw scores (e.g., counts of items a candidate got right) and item related values (simple counts of how many candidates got an item right) describe the characteristics of a specific testing sample. They do not draw inferences about the characteristics of the population from which the sample is taken, nor make any claim as to the representativeness of the sample. With inferential statistics, on the other hand, the characteristics of the sample are used to make estimates of the population from which the sample has taken. This comparison suggests that IRT analyses can provide inferences on how items are related to person ability, which is usually represented by the score on the test and estimated by a mathematical modelling about the chances of a candidate with certain ability achieving certain scores on items at given difficulty. As such, generalisations beyond the performance of a sample of test takers on a sample of testing items can be made about the ability of the entire population in relation to the measurement consisting of such items, as well as the level of difficulty of the items for the prospective test takers.

In the current research, the results from IRT analyses may provide statistical inferences on the extent to which each subtest of TALL is a valid measure to the component of the aptitude that it is intended to measure.

2. The debate about the use of Rasch modelling

Rasch modelling based on IRT has become generally accepted and widely used in language testing and assessment research despite the debates and disagreements in two decades before 2000 about choosing Basic Rasch model over other models, such as 2- and 3-parameter IRT models, for the analysis of dichotomous data in the test. According to McNamara and Knoch's (2012) review of the history of Rasch analysis, advocates for more complicated IRT models criticised the Rasch assumption of equal item discrimination. However, proponents of the Basic Rasch model argue that its strength lies in its capability to detect significant deviations in item discrimination from the assumed values, which can potentially impact measurement accuracy. This demonstrates the model's careful consideration of its underlying assumptions. In the 1980s, applied linguists seemed to hold more reservations than psychometricians about the use of Rasch analysis in language testing. They were concerned about the assumption of unidimensionality in Rasch, as applying a single dimension to analyse language test data would be inappropriate given the complex nature of language proficiency. However, these objections were challenged by defenders of Rasch analysis with empirical evidence that Rach analyses could confirm the unidimensionality of a language test that appeared to test multiple dimensions, and hence it could be a powerful means of examining underlying construct issues in communicative language tests (e.g., McNamara, 1990). The disputes about the appropriateness of using Rasch analyses seem to be resolved in language testing research since the first decade of the twenty-first century as Rasch measurement, particularly multi-faceted Rasch measurement has been uncontroverially accepted as a useful tool (McNamara & Knoch, 2012).

3. Reasons for applying different statistic packages for Rasch modelling

The reason for applying two statistic packages in the current study was to explore the possibility that different statistic approaches may yield inconsistent results to diagnose the item quality. As reviewed in Nicklin and Vitta (2022), the major difference between these two statistic packages in terms of Rasch analysis is that they have applied different maximum likelihood estimation methods, which is the statistical processes for calculating the item difficulty and person ability parameters. Specifically, the `ltm` package uses Marginal Maximum Likelihood Estimation (MMLE) to estimate data based on the assumption that individual person parameters conform to a specific distribution. This assumed distribution is then used in estimating item parameters, allowing individual parameters to be "marginalised" from the likelihood. As such, MMLE is considered more computationally efficient than other methods that estimate items independently of individuals. In contrast, the `eRm` package uses Conditional Maximum Likelihood Estimation (CMLE) to assess data without considering individual differences. In this method, estimates of individual parameters are removed from item parameter equations because they may not be consistent and could introduce bias or disagreement with the model's expectations. By avoiding the calculation of individual parameters until the item difficulty parameters are determined, CMLE may produce more precise residuals and fit statistics, while satisfying the objective of not relying on a particular sample for item estimates.

However, as asserted by Nicklin and Vitta, if individual differences are important to the analysis, caution should be taken when interpreting the results. To provide evidence about the differences between the results produced by different software and `R` packages applying different estimation methods, Nicklin and Vitta analysed 1000 simulated dichotomous datasets for Rasch models assessments using six software or `R` packages with three different estimation methods, that is, MMLE, CMLE, and Joint Maximum Likelihood Estimation (JMLE). They concluded that the differences between results produced with estimation methods were negligible, and the discrepancies observed in fit statistic estimations were attributable to the software choice. Therefore, they echoed the recommendation by Linacre (2021) that at least two packages are applied to obtain the confirmation of estimate values when conducting `R`-based Rasch analysis. The exploratory approach of using two packages in the current study for Rasch model analysis followed this recommendation.

4. Stepwise analyses

1 PL, 2 PL, and 3PL models for dichotomous data (TALL_VL, TALL_SD, and TALL_LA)

Step 1. Model comparisons

The initial step was to use `rasch` function in the `ltm` package to generate the basic form of the Rasch models, assuming equal discrimination parameters with the constraint argument specifying the value of the parameter as 1. The data were then be fitted to the unconstrained version of the Rasch model by calling the same `rasch` function without specifying the constraint argument. These two Rasch models were compared using the results of a likelihood ratio test, which was performed using the `anova` function. Specifically, the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC) and the *p*-value of the log-likelihood ratio were provided as evidence of the model fitness statistics. Lower values in the AIC and BIC indicate better model fit, while *p*-values suggest the null hypothesis, stating that the two models have the same model fit, should be rejected or not.

The extensions of the unconstrained Rasch model were further explored using functions in the `ltm` package. First, the 2PL model was generated using the `ltm` function, which assumes a different discrimination parameter for each item. This function accommodated up to two latent variables to represent the underlying structure of the data. An added feature of the `ltm` function allowed for an interaction between these two latent variables. Consequently, three optional 2PL models were compared to the previously identified Rasch model, which had exhibited a better fit in prior analysis, to assess their relative performance.

The 3PL model was subsequently applied to the data using the `tpm` function, which introduced a fixed guessing parameter to the unconstrained Rasch model. Two options (‘`rasch`’ and ‘`latent.trait`’) were specified to construct the optional 3PL models.

Model comparisons between the better-fitted Rasch model and the 2PL and 3PL models provided evidence regarding whether extensions of discrimination and guessing parameters were necessary. If model comparison results indicated the 2PL or 3PL model exhibited a better fit to the data, item analyses would be consistently conducted in the `ltm` package. However, if the results favoured the Rasch model as having a superior fit, the `RM` function in the `eRm` package was employed to generate supplementary model estimations. These estimations served to compare the fitness of items and persons fitness obtained from different statistics packages (as explained in the previous section). It is important to note that since the `eRm` package does not support the estimation of the 2PL and 3PL models, no comparisons between these two model estimations were made within the R packages.

Step 2. Model Fitness

Fit statistics of the chosen model to the data were obtained in this step. In the Rasch model estimations, fit statistics are used to determine whether a subtest has been developed with sufficient quality that values for both persons and items can be represented using the measure. These statistics are essential for investigating whether the assumption of unidimensionality for the Rasch analysis has been met. To put in another way, fit indices help to identify items that do not fit the unidimensional construct and hence diverge significantly from the expected ability/difficulty pattern. It determines whether item estimations can be considered as meaningful quantitative summaries of the observations, indicating the extent to which each item contributes to the measurement of the only construct (Bond & Fox, 2015).

For the Rasch model, `GoF.rasch` function through the `ltm` package was used to perform a parametric bootstrap goodness-of-fit test using χ^2 statistic. Based on 200 iterated datasets, the non-significant *p*-value > .05 would suggest an acceptable fit of the model (Rizopoulos, 2006). `LRtest` function through the `eRm` package also provided statistics of model fitness by splitting the dataset into two parts for conducting the Andersen likelihood ratio test. Similarly, the non-significant *p*-value > .05 would evidence the fitness of the model (Mair & Hatzinger, 2007).

If the 2PL or 3PL model was the model with better fit than the Rasch model, `margins` function through the `ltm` package was used to investigate the model fitness by examining the two-way χ^2 residuals. To calculate these residuals, the 2 x 2 contingency tables for all available items were created, and the goodness of fit in each cell was evaluated by means of Pearson's χ^2 statistics, using the value 3.5 as a rule of thumb.

Step 3. Model estimations and item/person fitness

The descriptive statistics of the chosen model with better fitness were analysed by the summary() function in the ltm and eRm packages, providing coefficients of difficulty values and standard errors of all items. In the ltm package, descript() function also provided output containing the χ^2 *p*-values for pairwise associations between the items, with non-significant results, i.e., *p*-values > .05 informing the potential problematic items. This interpretation assumed that in the latent variable models, latent variables can account for the high associations between items. If pairs of items did not reject the null hypothesis of independence, the assumption of the models could be violated (Rizopoulos, 2006). The parameter estimates through descript() were transformed to probability estimates using coef(). The results showed the probability of a correct response to a specific item for the average test taker.

In the ltm package, functions of item.fit() and person.fit() can be used to compute item and person fit statistics for 1PL, 2PL and 3PL models. The χ^2 (displayed as X^2) statistic tested the null hypothesis that the item responses follow the chosen model against the alternative hypothesis that they did not. A large χ^2 value indicated poor fit of the item to the model, while a small χ^2 value indicated good fit. The associated Pr(>X^2) *p*-values indicated the probability of observing the corresponding X^2 values as large as or larger than the values under the null hypothesis. The significant p-values < .05 would inform the poor fitness of the items to the model.

For the Rasch model, itemfit() and personfit() in the eRm package were used, which produced output easier to interpret. Several cut-off values of the item and person model fit were applied as the criteria to ascertain the adherence of the items and participants to the model's expectations. Mean-square (MNSQ) fit statistics, which are calculated from the residuals to indicate the distance between an observed data point and the model's expectation, were analysed and reported. MNSQ value is expected to be close to 1.0 if the item or person fits well to the model. Values greater than 1.0 suggest noise unfit to the model or other unclear variance in the data, which can degrade the measurement, while values less than 1.0 suggest the data points' overfit to the model, indicating that the data are 'too good to be true' that may cause the report of inflated statistics (Bond, et al., 2020). Two MNSQ fit statistics, *outfit* and *infit*, are reported in the eRm package. As explained by Nicklin and Vitta (2022), Outfit MNSQ statistics are susceptible to being influenced by outliers in the dataset, while Infit MNSQ statistics are weighted to address the issue of outliers, but are more prone to being influenced by data points that fit the model too well. The current research took the reference of the acceptable ranges of MNSQ values [0.50, 1.50] (Wright & Linacre, 1994), and [0.70, 1.30] with the infit *t* statistics ranging of [-2, 2] (Bond & Fox, 2015). If the MNSQ values of any items and persons were greater than 1.50, further examinations were conducted to investigate the reasons of the misfitting.

It is important to highlight that, in the current research, items that were answered correctly by every participant were considered for removal. This was informed by Nicklin and Vitta's (2022) item deletion strategy, as this research shared the same goal of determining which items to retain in the test. The authors' speculation that poor model fit might not necessarily indicate poor item quality but could instead be attributed to careless mistakes by high-ability persons or lucky guessed by low-ability participants appeared reasonable and was therefore taken into consideration in the current research.

The outputs of item and person statistics from both packages were consulted to provide comparable results that aimed to make more reliable interpretations.

Step 4. Plots of model estimations

Plots of IRT model estimations provides visualisations of the items and persons characteristics and the test information.

Items or Persons Pathway Maps plotted the location of each item or each person against its infit t-statistic. They provided clear visualisations to identify misfitting items or persons if the items or persons lying outside of the range between -2 and +2.

A person-item map presented the location of item parameters and the distribution of person parameters along the horizontal axis of the latent dimension, with the easiest items at the top and the most difficult items at the bottom, and the most skilled test takers at the right-hand side and the least skilled at the left-hand side. This map helped to compare the distribution and position of items to those of persons. If the instrument measures the latent trait accurately, items should be located along the entire scale, with smooth transitions in-between. In an ideal scenario in which adequate test takers represent the population, their ability levels reflected by the person parameters should be spread out across the entire range of the latent dimension.

An Item Characteristic Curve (*ICC*) plots the probability (P_i) that a test taker with ability value (θ) answers the item *i* correctly. The slope of the ICC indicates how steeply the probability of a correct response increased as ability level increased, and hence a steep ICC indicated that the item discriminated participants' ability well, while a shallow slope of the curve showed that the item discriminated ability poorly. In addition, a positive slope showed that when the test takers' ability level increased, the probability of correctly answering the item increased as well, which indicated that the item was effective to measure the construct being targeted and to differentiate test takers having various levels of latent ability.

An Item Information Curve (*IIC*), on the other hand, shows how much information about the latent ability an item provided at each level of the latent ability. It helps to indicate how well an item can discriminate between test takers at different level of ability and provides an estimate of the precision of an item at each level of ability. The location of the IIC on the axis of ability represented the level of ability at which the item is most informative.

Similarly, Test Information Curve (*TIC*) was plotted for the whole instrument, showing how much information about the latent ability the instrument provided at each level of ability. The range of ability levels measured by the instrument were represented by the width of the TIC. This information was used to evaluate the appropriateness of the test for the participants and to identify whether the test was too difficult or lacked challenges.

GPCM for dichotomous data (TALL_VL, TALL_SD, and TALL_LA)

The `ltm` package was used, as it applies Marginal Maximum Likelihood Estimation (MMLE) to estimate data based on the assumption that individual person parameters conform to a specific distribution (Nicklin & Vitta, 2022). Generalised Partial Credit model (GPCM) (Muraki, 1997) in the this package was used to examine the model fit of the polytomous data in the subtests of Serial Nonwords Recall and Complex Span Task.

It needs to be noted that although Partial Credit model (PCM) is available in the `eRm` package for the analysis of the polytomous data, it differs from GPCM in a key aspect. PCM assumes item discrimination remains consistent across different set sizes (the number of stimuli in one trial), which may not hold true for the current research data. This is because the trials with varying span sizes, ranging from 2 or 3 to 7 stimuli, can provide differing levels of discrimination evidence related to latent ability. As a result, the GPCM, which permits different constraint options on the discrimination parameter, appears to be the more suitable IRT model for analysing the polytomous data in the current research.

Step 1. Model comparisons

To build the GPCM, `gpcm` function was used in the `ltm` package that allows three constraint options, i.e., gpcm option assuming each trial having an estimated discrimination parameter, 1PL option assuming the discrimination parameter being equal for all trials, and Rasch option assuming the equal discrimination parameter being fixed at one. All three options were applied to the split data sets separately, and the models were then compared with each other to ensure that the chosen model had the best fit to each dataset.

Step 2. Model Fitness

Fitness statistics of the chosen model to the data were obtained in this step. For the Rasch model, the GoF.gpcm() function through the ltm package was used to perform a parametric bootstrap goodness-of-fit test using χ^2 statistic. Based on 50 iterated datasets, the non-significant *p*-value > .05 would suggest an acceptable fit of the model (Rizopoulos, 2006).

Step 3. Model estimations

The descriptive statistics of the chosen model with better fitness were analysed by the `summary` function, providing coefficients of the category threshold parameters and the discrimination parameter. The category threshold parameters represented the points on the latent trait scale that determine when the test takers were equally likely to endorse one answer option versus the next. The lower values of the category threshold parameters indicated that the item was easier to endorse, while the higher values indicated that the item was more difficult to endorse. The discrimination parameter provided information about the how well the item distinguished between individuals with different level of the latent ability. In the outputs, the z-value for each coefficient was also displayed, which was obtained by the coefficient divided by the standard error and indicated whether the coefficient was statistically significant. As a rule of thumb, a z-value with the absolute value greater than 1.96, indicating the statistical significance of the coefficient at the 5% level, suggested that the parameter was unlikely to have arisen by chance.

Step 4. Plots of model estimations

An Item Characteristic Curve (*ICC*) graphically represents the probabilities that a test taker with certain ability on the latent ability scale responds correctly to all categories of the trial correctly. A positive slope shows that when a test takers' ability level increases, the probability of correctly endorsing the category increases as well, which indicates that the category is effective to evidence the construct being targeted and to differentiate test takers with various levels of latent ability.

An Item Information Curve (*IIC*), on the other hand, shows how much information about the latent ability an item provides at each level of the latent ability. It helps to indicate how well an item can discriminate between test takers at different level of ability and provides an estimate of the precision of an item at each level of ability. The location of the IIC on the axis of ability represents the level of ability at which the item is most informative.

Similarly, the Test Information Curve (*TIC*) is plotted for the whole instrument, which shows how much information about the latent ability the instrument provides at each level of ability. The range of ability levels measured by the instrument is represented by the width of the TIC. This information is used to evaluate the appropriateness of the test for the participants and to identify whether the test is too difficult or the test lacks challenges. In addition, the TIC also displays how well the tests measure individuals of different abilities based on the location of the ability level the information curve peak. As asserted by Draheim et al. (2018), the TIC is not affected by the distribution of test takers' ability levels and is invariant to the ability levels of the participants in the test.

Appendix G: A stepwise protocol of PCA

PCA in the current research was conducted through functions in the psych package (Revelle, 2022), the `FactoMineR` package (Husson, et al., 2017), and `factoextra` package (Kassambara & Mundt, 2017) in R, following the steps below.

Step 1. Preliminary analysis

The first step was to check the correlations of the subtests as PCA was conducted based on the assumption that TALL measured the same underlying factor dimension(s) that were correlated with each other. Prior to the calculation of correlation coefficients, normality checks on the data of the test suites were conducted, informing the choice of the appropriate type of correlation coefficients.

The analysis of correlation aimed to detect two potential problems: (1) the correlations were not high enough, usually with coefficients lower than .3, which may violate the assumption of PCA; (2) the correlations were too high, with coefficients greater than .9, which suggested extreme multicollinearity and singularity that should be avoided for factor analysis. However, these cut-off values of the coefficients may be subjective (Field, et al., 2012). Bartlett's test of sphericity was used as an objective method to examine whether the correlation matrix resembled an identity matrix, suggesting all correlation coefficients were close to zero. The statistically significant *p*-value < .05 would indicate that the variables somehow correlated with all other variables. In addition, the determinant of the R-Matrix was used to provide information about whether the correlation matrix was singular (determinant is 0), or whether all subtests were completely unrelated (determinant is 1), or somewhere in between. The value greater than the necessary value of 0.00001 would suggest the determinant not being problematic.

In this step, the sample size related to the stability of factor solutions was also examined, using the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy (Kaiser, 1970). The acceptable value of KMO would be greater than .5, with values between .5 and .7 being mediocre, values between .7 and .8 being good, values between .8 and .9 being great, and values above .9 being superb (Hutcheson & Sofroniou, 1999).

Step 2. Factor extraction

Given that the purpose of conducting PCA in the current research was to explore whether the subtests could measure the principal component without much redundancy rather than to reduce or consolidate variables, in this step, solutions without reducing the number of factors as the same of the number of subtests were conducted, using `PCA` function in the `FactoMineR` package. Eigenvalues of the solutions were inspected against the rules of retaining factors having eigenvalues greater than 1 (Kaiser, 1960), greater than 0.70 (Jolliffe, 1972) and greater than 0.512 for sample size of 100 (Stevens, 2002). Meanwhile, to improve the interpretation of the solution, the oblimin method of oblique rotation was used, which allowed factors to correlate.

Another criterion for deciding the proper solutions of retaining the number of factors is the cumulative percentage of variance that the extracted factors explain. Field et al. (2012) suggests that the minimum cumulative percentage of explained variance should be around 55–65%, while a field-specific thresholds in factor analytic L2 research is suggested to be approximately 60% (Plonsky & Gonulal, 2015), which means that it may be appropriate to retain the number of factors until they account for at least 60% of the total variance (Loewen & Gonulal, 2015).

The strength of the association between the subtests and each factor dimension can be examined by the factor loadings. However, the interpretation of what establishes a high loading is subjective and may follow different optimal factor loading scores (Loewen & Gonulal, 2015). In this section, factor loadings > .30 were considered significant, following the practice of a field-specific sample study, Loewen et al. (2009), suggested by Plonsky and Gonulal (2015). The results are presented in the tables of factor loadings.

Step 3. Plots

Scree plots provided visualisation of the eigenvalues ordered from largest to the smallest, which informed the number of principal components. The graphs of variables (i.e., subtests) were also plotted to describe the relationships between the variables and the factor dimensions. In specific, the correlation between a variable and the factor dimensions was described as the coordinate of the variable on the factor dimensions. The quality of representation of the variables on the factor dimensions was calculated as cos2, which was the squared coordinate. The contribution (in percentage) of a variable to a given factor dimension was calculated by the cos2 of the variable divided by the total cos2 of the dimension. In addition, a correlation circle plotted showed the variable correlations on the dimensions of the first two principal components. The plot could be interpreted in the following way: positively correlated variables were grouped together, while negatively correlated variables were positioned on opposite sides of the plot origin (opposed quadrants). The distance between variables and the origin showed the quality of the variables on the factor map, with variables away from the origin being well represented on the factor map (Kassambara & Mundt, 2017).

Appendix H: A stepwise protocol of CFA

The objective of using CFA in the current study was to evaluate how well the hypothesised four-factor model fitted the empirical data. CFA was conducted using the `lavaan` package (Rosseel, 2012) in R.

Step 1. Data preparation

Prior to the CFA, data preparation was required for having a reliable parameter estimation. The prerequisites of CFA involved the check on skewness and kurtosis of variables and outliers that affect the analysis (Schoonen, 2015). The results of normality checks are reported in Section 4.5.1.1 and 4.5.1.2, indicating that the data of all subtests in both suites were not normally distributed. Therefore, data transformation was applied on the individual data set of each subtest in each suite.

Initially, extreme observations within the dataset were identified and handled as outliers in the scores. In the aural suite, nine extreme observations from TALL_VL, eight from TALL_SD, one from TALL_LA, five from TALL_SNWR, and nine from TALL_CST. In the written suite, three from TALL_VL, seven from TALL_SD, nine from TALL_LA, six from TALL_SNWR, and four from TALL_CST.

To manage these outliers, the extreme observations were replaced with scores that were recalculated by reverting them from a z-score of 2. These extreme observations were defined as those with absolute z-scores exceeding 2 standard deviations from the mean. The replacement scores were calculated as the mean plus two standard deviations.

The decision to modify extreme scores stemmed from the observation that the extreme scores achieved by participants were not treated as statistical outliers in the current research (as explained in Section 3.3.5.1). While retaining these values in the dataset had its advantages, such as preventing an inflation of the Type I error rate, it was evident that these extreme values were unrepresentative and could introduce bias into the model estimations. Therefore, the adjustment of scores was considered a more favourable option, despite the potential drawback of altering the original score distribution to enhance normality, as noted by Field, et al. (2012). This choice aimed to ensure that extreme values did not unduly influence subsequent analyses, thus leading to more robust and reliable results.

Subsequently, the dataset with the extreme values replaced underwent a transformation using the arcsine-square-root transformation, a standard procedure for handling proportional data (Sokal & Rohlf, 1995). The aim of this transformation was to achieve normality in the data distribution. To ensure that the transformed datasets of variables met the criteria for normality, skewness and kurtosis were assessed. An absolute skew.2SE or kurt.2SE value greater than 1, following Field, et al. (2012) for significance at p < .05, indicated significant skewness or kurtosis. These checks were crucial for verifying data assumptions in subsequent analyses.

Step 2. Model building and model fit check

A four-factor model was specified using the structure of factors postulated by the theoretical framework of aptitude construct, that is, the subtests of TALL_VL, TALL_SD, TALL_LA load to three factors respectively and the subtests of TALL_SNWR and TALL_CST load to one factor. Model fit was checked to evaluate how well the data of both suites fit the four-factor model by comparing the observed data with the predicted data based on the model.

The model fit can be evaluated using various goodness-of-fit indices to assess the fit of specified model to the data. Specifically, the statistical results under "Model Test User Model" in the output are relevant for assessing the fit of the model to the data, in which a non-significant *p*-value above .05 may suggest a good fit between the model and the data. The results under "Model Test Baseline Model" indicate the comparison of the user model to a baseline model (i.e., a simpler model that assumes no relationships or associations between the variables), which is to determine if the specified user model provide a significantly better fit to the data compared to the baseline model with the p-value < .05. The Chi-square test statistics, degrees of freedom, and p-value in the output of both "Model Test User Model" and "Model Test Baseline Model" are reported. In addition, the approximate fit indices are also reported to indicate the model fit in CFA. Specifically, Comparative Fit Index (CFI) and the Tucker-Lewis Index (TLI) above 0.9 and close to 1, Root Mean Square Error of Approximation (RMSEA) below .05, and Standardized Root Mean Square Residual (SRMR) below .08 would evidence the good fit of the data to the theoretical model (Brown, 2015).

References

Abu-Rabia, S. (2001). Testing the interdependence hypothesis among native adult bilingual Russian– English students. *Journal of Psycholinguistic Research, 30*(4), 437–455.

Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods, 16*(2), 270-301. doi:10.1177/1094428112470848

Al-Hoorie, A. H., & Vitta, J. P. (2019). The seven sins of L2 research: A review of 30 journals' statistical quality and their CiteScore, SJR, SNIP, JCR Impact Factors. *Language Teaching Research, 23*(6), 727–744. https://doi.org/10.1177/1362168818767191

Allison, P. D. (1999). *Multiple regression.* Thousand Oaks, CA: Pine Forge Press.

Andringa, S., & Godfroid, A. (2020). Sampling bias and the problem of generalizability in applied linguistics. *Annual Review of Applied Linguistics, 40*, 134–142. https://doi.org/10.1017/s0267190520000033

Anthony, J. L., & Francis, D. J. (2005). Development of phonological awareness. *Current Directions in Psychological Science, 14*(5), 255–259. https://doi.org/10.1111/j.0963-7214.2005.00376.x

Artieda, G., & Muñoz, C. (2016). The LLAMA tests and the underlying structure of language aptitude at two levels of foreign language proficiency. *Learning and Individual Differences, 50*, 42–48. https://doi.org/10.1016/j.lindif.2016.06.023

Aruguete, M.S., Huynh, H., Browne, B.L., Jurs, B., Flint, E., & McCutcheon, L.E. (2019). How serious is the "carelessness" problem on Mechanical Turk? *International Journal of Social Research Methodology, 22*(5), 441–449. https://doi.org/10.1080/13645 579.2018.1563966

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Bachman, L. F., & Cohen, A. D. (1998). Language testing – SLA interfaces: An update. In L. F. Bachman and A. D. Cohen (eds.), *Interfaces between Second Language Acquisition and Language Testing Research* (pp. 1–31). Cambridge, England: Cambridge University Press.

Baddeley, A. D. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences, 4*, 417–423.

Baddeley, A. (2003). Working memory and language: An overview. *Journal of Communication Disorders, 36*(3), 189–208.

Baddeley A. (2017). Modularity, working memory, and language acquisition. *Second Language Research, 33*(3), 299–311.

Baddeley, A. (2022). Working memory and the challenge of language. In J. W. Schwieter, & E. Z. Wen (Eds.), *The Cambridge Handbook of Working Memory and Language.* Cambridge: Cambridge University Press

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. A. Bower (Ed.), *Recent Advances in Learning and Motivation* (Vol. 8, pp. 47-89). Academic Press

Baddeley, A. D., Gathercole, S., & Papagno, C. (1998). The phonological loop as a language acquisition device. *Psychological Review, 105*, 158–173.

Baddeley, A., Lewis, V., & Vallar, G. (1984). Exploring the articulatory loop. *Quarterly Journal of Experimental Psychology A, 36*,233252.

Baddeley, A. D., Papagno, C., & Vallar, G. (1988). When long-term learning depends on short-term storage. *Journal of Memory and Language, 27*, 586–595.

Bakker, M., & Wicherts, J. M. (2014). Outlier removal, sum scores, and the inflation of the type i error rate in independent samples t tests: The power of alternatives and recommendations. *Psychological Methods, 19*(3), 409–427. https://doi.org/10.1037/met0000014

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (3rd ed.). Chichester, England: Wiley

Barth, D., & Kapatsinski, V. (2018). Evaluating logistic mixed-effects models of corpus-linguistic data in light of lexical diffusion. In D. Speelman, K. Heylen, & D. Geeraerts (Eds.), *Mixed-Effects Regression Models in Linguistics* (pp. 99–116). Springer.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1–48. doi:10.18637/jss.v067.io1.

Bohannon, J. (2016). Mechanical Turk upends social sciences. *Science Magazine, 352*(6291), 1263–1264. https://doi.org/10.1126/scien ce.352.6291.1263

Bokander, L. (2020). Language aptitude and crosslinguistic influence in initial L2 learning. *Journal of the European Second Language Association, 4*(1), 35. https://doi.org/10.22599/jesla.69

Bokander, L. (2023). Exploring the predictive validity of the LLAMA (v1) Language Aptitude Tests. In Z. Wen, P. Skehan, & R. Sparks (eds.), *Language Aptitude Theory and Practice* (pp. 94–116). Cambridge University Press.

Bokander, L., & Bylund, E. (2020). Probing the internal validity of the LLAMA language aptitude tests. *Language Learning, 70*(1), 11–47. https://doi.org/10.1111/lang.12368

Bolibaugh, C., & Foster, P. (2013). Memory-based aptitude for nativelike selection. In G. Granena, & M. Long (Eds.). *Sensitive Periods, Language Aptitude, and Ultimate L2 Attainment* (pp. 179–204). John Benjamins.

Bolibaugh, C., & Foster, P. (2021). Implicit statistical learning in naturalistic and instructed morphosyntactic attainment: an Aptitude-Treatment interaction design. *Language Learning, 71*(4), 959–1003. https://doi.org/10.1111/lang.12465

Bolibaugh, C., Vanek, N., & Marsden, E. (2021). Towards a credibility revolution in bilingualism research: Open data and materials as stepping stones to more reproducible and replicable research. *Bilingualism: Language and Cognition, 24*(5), 801–806. https://doi.org/10.1017/s1366728921000535

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental Measurement in the Human Sciences* (3rd ed). Routledge.

Bond, T., Yan, Z., & Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the Human Sciences* (4th ed.). Routledge. https://doi.org/10.4324/9780429030499

Bovolenta, G., & Marsden, E. (2021). Expectation Violation Enhances the Development of New Abstract Syntactic Representations: Evidence from an Artificial Language Learning Study. *Language Development Research*. https://doi.org/10.31219/osf.io/zyegf

Bovolenta, G., & Williams, J. N. (2022). Implicit learning in production: productive generalization of new Form–Meaning connections in the absence of awareness. *Language Learning, 73*(3), 723–758. https://doi.org/10.1111/lang.12551

BPS. (2021). *Ethics guidelines for internet-mediated research.* https://www.bps.org.uk/news-and-policy/ethics-guidelines-internet-mediated-research

Braun, M. T., Converse, P. D., & Oswald, F. L. (2019). The accuracy of dominance analysis as a metric to assess relative importance: The joint impact of sampling error variance and measurement unreliability. *Journal of Applied Psychology, 104,* 593–602. https://doi.org/10.1037/apl0000361

Brown, J. D. (2014). Classical theory reliability. In A. J. Kunnan (Ed.), *The Companion to Language Assessment* (pp.1165–1181). Oxford, UK: Wiley–Blackwell.

Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research* (2nd ed). Guilford.

Brown, V. A. (2021). An Introduction to Linear Mixed-Effects Modeling in R. *Advances in Methods and Practices in Psychological Science, 4*(1), 251524592096035. https://doi.org/10.1177/2515245920960351

Buhrmester, M. D., Talaifar, S., & Gosling, S. D. (2018). An Evaluation of Amazon's Mechanical Turk, Its Rapid Rise, and Its Effective Use. *Perspectives on Psychological Science, 13*(2), 149–154. https://doi.org/10.1177/1745691617706516

Cai, R., & Dong, Y. (2012). 信息类型、编码通道与编码语言对工作记忆广度的影响——支持层级观的证据［Effects of information type, encoding modality, and encoding language on working memory span: Evidence for the hierarchical view]. *外语教学与研究, 44*(3), 376–388.

Canty, A., & Ripley, B. (2021). *boot: Bootstrap R (S-PLUS) functions. R package* (Version 1.3-28) [Computer software]. Retrieved from https://CRAN.R-project.org/package=boot

Carroll, J. B. (1962). The prediction of success in intensive Foreign language training. In R. Glaser (Ed.), *Training Research and Education* (pp. 87–136). Pittsburgh, PA: University of Pittsburgh Press.

Carroll, J. (1973). Implications of aptitude test research and psycholinguistic theory for foreign language teaching. *International Journal of Psycholinguistics, 2*, 5–14.

Carroll, J. B. (1981). Twenty-five years of research on foreign language aptitude. In K. C. Diller (Ed.), *Individual Differences and Universals in Language Learning Aptitude* (pp. 83-118). Rowley, MA: Newbury House.

Carroll, J. B. (1990). Cognitive abilities in Foreign language aptitude: Then and now. In T. Parry & C.W. Stansfield (Eds.), *Language Aptitude Reconsidered* (pp.11–29). Englewood Cliffs, NJ: Prentice-Hall.

Carroll, J. B. (1993). *Human Cognitive Abilities: A survey of Factor-analytic Studies.* Cambridge University Press. https://doi.org/10.1017/CBO9780511571312

Carroll, J. B. & Sapon, S. (1959). *Modern Language Aptitude Test (MLAT)*. The Psychological Corporation.

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon’s MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior, 29*(6), 2156–2160.

Chalmers, J., Eisenchlas, S. A., Munro, A., & Schalley, A. C. (2021). Sixty years of second language aptitude research: A systematic quantitative literature review. *Language and Linguistics Compass, 15*(11). https://doi.org/10.1111/lnc3.12440

Chan, E., Skehan, P., & Gong, G. (2011). Working memory, phonemic coding ability and foreign language aptitude: Potential for construction of specific language aptitude tests – the case of Cantonese. *Ilho do Desterro: A Journal of English Language, Literatures and Cultural Studies, 60*(1), 45–73.

Chandler, J., Paolacci, G., Peer, E., Mueller, P., & Ratliff, K.A. (2015). Using nonnaive participants can reduce effect sizes. *Psychological Science, 26*, 1131–1139. https:// doi.org/10.1177/09567 97615 585115

Chang, L. Y., Plaut, D. C., & Perfetti, C. A. (2016). Visual complexity in orthographic learning: Modeling learning across writing system variations. *Scientific Studies of Reading, 20*, 64–85. https://doi.org/10.1080/10888438.2015.1104688

Cheung, J.H., Burns, D.K., Sinclair, R.R., & Sliter, M. (2017). Amazon Mechanical Turk in organizational psychology: An evaluation and practical recommendations. *Journal of Business and Psychology, 32*, 347–361. https://doi.org/10.1007/s1086 9-016-9458-5

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Erlbaum.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Cohen, A. D., & Macaro, E. (2013). Research methods in second language acquisition. In E. Macaro (Ed.), *The Bloomsbury Companion to Second Language Acquisition* (pp. 107–133). New York, NY: Bloomsbury Academic.

Cohen, L., Manion, L., & Morrison, K. (2011). *Research Methods in Education* (7th ed.). London: Routledge.

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user’s guide. *Psychonomic Bulletin and Review, 12*(5), 769–786. https://doi.org/10.3758/BF03196772

Conway, J. M., & Huffcutt, A. I. (2003). A Review and Evaluation of Exploratory Factor Analysis Practices in Organizational Research. *Organizational Research Methods, 6*(2), 147–168. https://doi.org/10.1177/1094428103251541

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. *Psychological Bulletin, 104*(2), 163–191.

Cowan, N. (1999). An embedded-processes model of working memory. In A. Miyake & P. Shah (Eds.), *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control* (pp. 62–101). Cambridge University Press.

Cowan, N. (2017).The many faces of working memory and short-term storage. *Psychonomic Bulletin & Review, 24*, 1158–1170. doi:10.3758/s13423-016-1191-6.

Crocker, L., & Algina, J. (2008). *Introduction to Classical and Modern Test Theory*. Cengage Learning.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology, 41*, 137–163.

Cronbach, L., & Snow, R. (1977). *Aptitudes and Instructional Methods: A Handbook for Research on Interactions.* New York: Irvington.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamera, T. (1999). *Dictionary of Language Testing.* Cambridge University Press.

Degani, T., & Goldberg, M. (2019). How individual differences affect learning of translation-ambiguous vocabulary. *Language Learning, 69*(3), 600–651. https://doi.org/10.1111/lang.12344

Dennis, S.A., Goodson, B.M., & Pearson, C.A. (2020). Online worker fraud and evolving threats to the integrity of MTurk data: A discussion of virtual private servers and the limitations of IP-based screening procedures. *Behavioral Research in Accounting, 32*(1), 119–134. https://doi.org/10.2308/bria-18-044

DeKeyser, R. M. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition, 22*(4), 499–533.

DeKeyser, R. (2019). The future of language aptitude research. In Wen, Z., Skehan, P., Biedroń, A., Li, S., & Sparks, R. L. (Eds). *Language Aptitude: Advancing Theory, Testing, Research and Practice*. Rouledge.

DeKeyser, R., & Koeth, J. (2011). Cognitive aptitudes for L2 learning. In E. Hinkel (Ed.), *Handbook of Research in Second Language Teaching and Learning, Volume II* (pp. 395–406). New York, NY: Routledge.

De Oliveira, G. S., & Baggs, E. (2023). *Psychology's WEIRD problems.* Cambridge University Press. https://doi.org/10.1017/9781009303538

Derrick, D. J. (2016). Instrument Reporting Practices in Second Language Research. *TESOL Quarterly, 50*(1), 132–153. https://doi.org/10.1002/tesq.217

DeVoe, S.E., & House, J. (2016). Replications with MTurkers who are naïve versus experienced with academic studies: A comment on Connors, Khamitov, Moroz, Campbell, and Henderson (2015). *Journal of Experimental Social Psychology, 67*, 65–67. https://doi.org/10.1016/j.jesp.2015.11.004

Dörnyei, Z. (2005). *The Psychology of the Language Learner: Individual Differences in Second Language Acquisition*. Mahwah, NJ: Lawrence Erlbaum.

Doughty, C. J. (2019). Cognitive Language Aptitude. *Language Learning, 69* (S1), 101–126. https://doi.org/10.1111/lang.12322

Doughty, C. J., Campbell, S. G., Bunting, M., Bowles, A., & Haarmann, H. (2007). *The development of the high-level language aptitude battery* (Technical Report). College Park, MD: University of Maryland Center for Advanced Study of Language.

Doughty, C. J., & Mackey, A. (2021). Language aptitude: Multiple perspectives. *Annual Review of Applied Linguistics*. https://doi.org/10.1017/s0267190521000076

Draheim, C., Harrison, T. L., Embretson, S. E., & Engle, R. W. (2018). What item response theory can tell us about the complex span tasks. *Psychological Assessment, 30*(1), 116–129. https://doi.org/10.1037/pas0000444

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition, 24*, 143–188.

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science, 11*,19–23.

Escudero, P., Smit, E. A., & Angwin, A. J. (2022). Investigating orthographic versus auditory Cross-Situational word learning with online and Laboratory-Based testing. *Language Learning, 73*(2), 543–577. https://doi.org/10.1111/lang.12550

Farmer, T. A., Fine, A. B., Misyak, J. B., & Christiansen, M. H. (2016). Reading span task performance, linguistic experience, and the processing of unexpected syntactic events. *Quarterly Journal of Experimental Psychology, 70*, 413–433.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2016). G power 3. *Journal of Materials and Environmental Science, 7*(10), 3759–3766.

Field, A. (2013). *Discovering statistics using SPSS*. Sage.

Field, A., Miles, J., & Field, Z. (2012). *Discovering Statistics Using R*. Sage.

Flake, J. K., & Fried, E. I. (2020). Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. *Advances in Methods and Practices in Psychological Science, 3*(4), 456–465. https://doi.org/10.1177/2515245920952393

Follmer, D.J., Sperling, R.A., & Suen, H.K. (2017) The role of MTurk in education research: Advantages, issues, and future directions. *Educational Researcher, 46*(6), 329–334. https://doi.org/10.3102/00131 89X17 725519

Fougnie, D. & Marois, R. (2011). What limits working memory capacity? Evidence for modality-specific sources to the simultaneous storage of visual and auditory array. *Journal of Experimental Psychology: Learning, Memory, and Cognition. 37*(6), 1329-1341.

French, L. M. and O'Brien, I. (2008). Phonological memory and children's second language grammar learning. *Applied Psycholinguistics, 29*, 463–487.

Gajewski, P. D., Hanisch E., Falkenstein M., Thönes S., & Wascher, E. (2018). What does the N-back task measure as we get older? Relations between working-memory measures and other cognitive functions across the lifespan. *Frontiers in Psychology, 9*:2208. https://doi.org/10.3389/fpsyg.2018.02208

Ganschow, L., & Sparks, R. L. (1996). Anxiety about Foreign language learning among high school women. *Modern Language Journal, 80*(2), 199–212.

Gass, S. M., & Lee, J. (2011). Working memory capacity, inhibitory control, and proficiency in a second language. In M. S. Schmid, & W. Lowie (Eds.), *Modeling Bilingualism: From Structure to Chaos. In Honor of Kees de Bot* (pp. 59–84). John Benjamins.

Gass, S., Winke, P., Isbell, D. R., & Ahn, J. (2019). How captions help people learn languages: A working-memory, eye-tracking study. *Language Learning and Technology, 23*(2), 84–104. https://doi.org/10125/44684

Gathercole, S. E. (1995). Is nonword repetition a test of phonological memory or long-term knowledge? It all depends on the nonwords. *Memory & Cognition, 23*, 83–94.Gathercole, S. E. (2006). Nonword repetition and word learning: The nature of the relationship. *Applied Psycholinguistics, 27*(4), 513–543. https://doi.org/10.1017/s0142716406060383

Gathercole, S. E., & Baddeley, A. D. (1989). Evaluation of the role of phonological STM in the development of vocabulary in Children: A longitudinal study. *Journal of Memory and Language, 28*, 200–213.

Gathercole, S., & Baddeley, A. (1993). *Working memory and language.* Lawrence Erlbaum Associates.

Gathercole, S., Willis C., Baddeley, A., & Emslie, H., (1994). The children's test of nonword repetition: A test of phonological working memory. *Memory, 2*(2), 103–127.

Gilabert, R., Manchón, R., & Vasylets, O. (2016). Mode in theoretical and empirical TBLT research: Advancing research agendas. *Annual Review of Applied Linguistics, 36*, 117–135. https://doi.org/10.1017/S0267190515000112

Gleibs, I.H. (2017). Are all "research fields" equal? Rethinking practice for the use of data from crowdsourcing market places. *Behavior Research Methods, 49*, 1333– 1342. https://doi.org/10.3758/s1342 8-016-0789-y

Gleser, G., Cronbach, L., & Rajaratnam, N. (1965). Generalizability of scores influenced by multiple sources of variance. *Psychometrika, 30*(4), 395–418. doi:10.1007/BF02289531

Granena, G. (2013). Cognitive aptitudes for L2 learning and the LLAMA language aptitude test. In G. Granena, & Long, M. H. (Ed.), *Sensitive Periods, Language Aptitude, and Ultimate L2 Attainment* (pp. 105–130). John Benjamins.

Granena, G. (2019). Cognitive aptitudes and L2 speaking proficiency. *Studies in Second Language Acquisition, 41*(2), 313–336. https://doi.org/10.1017/s0272263118000256

Granena, G., & Long, M. H. (2013). *Sensitive Periods, Language Aptitude, and Ultimate L2 Attainment*. John Benjamins Publishing Company.

Gray, M.L., Suri, S., Ali, S.S., & Kulkarni, D. (2016). The crowd is a collaborative network. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing* (pp. 134–147). New York: ACM.

Green, S. B., & Yang, Y. (2015). Evaluation of dimensionality in the assessment of internal consistency reliability: coefficient alpha and omega coefficients. *Educational Measurement: Issues and Practice, 34*(4), 14–20. https://doi.org/10.1111/emip.12100

Gries, S. T. (2021). (Generalized Linear) Mixed-Effects Modeling: a learner corpus example. *Language Learning, 71*(3), 757–798. https://doi.org/10.1111/lang.12448

Grigorenko, E. L., Sternberg, R. J., & Ehrman, M. E. (2000). A theory-based approach to the measurement of foreign language learning ability: The CANAL-F theory and test. *Modern Language Journal, 84*(3), 390–405. https://doi.org/10.1111/0026-7902.00076

Grömping, U. (2006). Relative importance for linear regression in R: The package relaimpo. Journal of Statistical Software 17(1), 1–27. https://doi.org/10.18637/jss.v017.i01

Hair, J.F., Black, W. C., Babin, B.J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). Englewood Cliffs, NJ: Prentice Hall.

Hauser, D., Paolacci, G., & Chandler, J.J. (2019). Common concerns with MTurk as a participant pool: Evidence and solutions. In F.R. Kardes, P.M. Herr, & N. Schwarz (Eds.), *Handbook of Research Methods in Consumer Psychology* (pp. 319–337). New York: Routledge.

Havik, E., Roberts, L., van Hout, R., Schreuder, R., and Haverkort, M. (2009). Processing subject-object ambiguities in the L2: A self-paced reading study with German L2 learners ofDutch. *Language Learning, 59*, 73–112.

Hayes-Harb, R., & Barrios, S. (2021). The influence of orthography in second language phonological acquisition. *Language Teaching, 54*(3), 297–326. https://doi.org/10.1017/s0261444820000658

Henning, G. (1987). *A Guide to Language Testing: Development, Evaluation, Research.* Rowley, Massachusetts: Newbury House.

Henrich, J., Heine, S., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2–3), 61–83. https://doi.org/10.1017/s0140525x0999152x

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement, 66*(3), 393–416. https://doi.org/10.1177/0013164405282485

Hicks, K.L., Foster, J.L., & Engle, R.W. (2016). Measuring working memory capacity on the web with the Online Working Memory Lab (the OWL). *Journal of Applied Research in Memory and Cognition, 5*, 478–489.

Huang, J.L., Liu, M., & Bowling, N.A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology, 100*(3), 828–845. https://doi.org/10.1037/a0038510

Hughes, M., Golonka, E., Tseng, A., & Campbell, S. (2023). The High-Level Language Aptitude Battery (Hi-LAB). In Z. Wen, P. Skehan, R. L. Sparks (Eds), *Language Aptitude Theory and Practice* (pp.73–93). Cambridge University Press.

Hunt, N. C., & Scheetz, A. M. (2019). Using MTurk to distribute a survey or experiment: Methodological considerations. *Journal of Information Systems, 33*(1), 43–65. https://doi.org/10.2308/isys-52021

Husson, F., Josse, J., Le, S., & Mazet, J. (2017). *FactoMineR: Multivariate Exploratory Data Analysis and Data Mining.* https://CRAN.R-project.org/package=FactoMineR.

Hutcheson, G., & Sofroniou, N. (1999). *The Multivariate Social Scientist: Introductory Statistics Using Generalized Linear Models.* Thousand Oaks, CA: Sage.

Iizuka, T., & DeKeyser, R. (2023). Scrutinizing LLAMA D as a measure of implicit learning aptitude. *Studies in Second Language Acquisition*, 1–23. https://doi.org/10.1017/s0272263122000559

Isbell, D. R., Brown, D., Chen, M., Derrick, D. J., Ghanem, R., De Lourdes Arvizu, M., Schnur, E., Zhang, M., & Plonsky, L. (2022). Misconduct and Questionable Research Practices: The Ethics of Quantitative Data Handling and Reporting in Applied Linguistics. *The Modern Language Journal, 106*(1), 172–195. https://doi.org/10.1111/modl.12760

Isbell, D. R., & Kim, J. (2023). Developer involvement and COI disclosure in high-stakes English proficiency test validation research: A systematic review. *Research Methods in Applied Linguistics, 2*(3), 100060. https://doi.org/10.1016/j.rmal.2023.100060

Isbell, D.R., & Rogers, J. (2021). Measuring implicit and explicit learning and knowledge. In P. Winke & T. Brunfaut (eds.) *The Routledge Handbook of Second Language Acquisition and Language Testing.* Routledge.

Jeon, E. H. (2015). Multiple regression. In L. Plonsky (Ed.), *Advancing Quantitative Methods in Second Language Research* (pp. 131–158). New York, NY: Routledge.

Jolliffe, I. T. (1972). Discarding variables in a principal component analysis 1: Artificial data. *Appl. Statist., 21,* 160–173.

Jolliffe, I. T. (2002). *Principal Component Analysis* (2nd ed.). New York: Springer.

Juffs, A., & Harrington, M. (2011). Aspects of working memory in L2 learning. *Language Teaching, 44*(2), 137–166. https://doi.org/10.1017/S0261444810000509

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*, 141–151.

Kaiser, H. F. (1970). A second generation little jiffy. *Psychometrika, 35*(4), 401–415. https ://doi.org/10.1007/BF022 91817

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika, 39*, 31–36.

Kan, I.P., & Drummey, A.B. (2018). Do imposters threaten data quality? An exam- ination of worker misrepresentation and downstream consequences in Amazon’s Mechanical Turk workforce. *Computers in Human Behavior, 83*, 243–253. https:// doi.org/10.1016/j.chb.2018.02.005

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education and Praeger.

Kane, M. J. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kang, T., Cohen, A. S., & Sung, H. J. (2005). *IRT model selection methods for polytomous items*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal, Canada.

Karpen, S. C. (2017). Misuses of regression and ANCOVA in educational research. *American Journal of Pharmaceutical Education, 81*, 6501. https://doi.org/10.5688/ajpe6501

Kassambara, A., & Mundt, F. (2017). *Factoextra: Extract and Visualize the Results of Multivariate Data Analyses.* http://www.sthda.com/english/rpkgs/factoextra.

Knoch, U., & McNamara, T. (2015). Rasch analysis. In L. Plonsky (Ed.), *Advancing Quantitative Methods in Second Language Research* (pp. 275–304). New York, NY: Routledge.

Kormos, J., & Sáfár, A. (2008). Phonological short-term memory, working memory and foreign language performance in intensive language learning. *Bilingualism: Language and Cognition, 11*(2), 261–271. https://doi.org/10.1017/s1366728908003416

Larson-Hall, J. (2016). *A guide to doing statistics in second language research using SPSS and R* (2nd ed.). Milton Park, UK: Routledge.

Leeser, M., & Sunderman, G. (2016). Methodological issues of working memory tasks for L2 processing research. In G. Granena, D. O. Jackson, & Y. Yilmaz (Eds.), *Cognitive individual differences in second language processing and acquisition* (pp. 89–104). John Benjamins.

Li, L., & Luo, S. (2019). Development and preliminary validation of a foreign language aptitude test for Chinese learners of foreign languages. In Z. Wen, P. Skehan, A. Biedroń, S. Li, & R. Sparks (Eds.), *Language Aptitude: Advancing Theory, Testing, Research and Practice* (pp. 33–55). Routledge.

Li, S. (2015). The associations between language aptitude and second language grammar acquisition: A meta-analytic review of five decades of research. *Applied Linguistics, 36*, 385–408.

Li, S. (2016). The construct validity of language aptitude. *Studies in Second Language Acquisition, 38*, 801–842.

Li, S. (2019). Six decades of language aptitude research: A comprehensive and critical review. In Z. Wen, Skehan, P., Biedroń, A., Li, S., & Sparks, R. L. (2019). *Language Aptitude: Advancing Theory, Testing, Research and Practice* (pp. 78–96). New York: Routledge.

Li, S. (2022). Explicit and implicit language aptitudes. In S. Li, P. Hiver, & M. Papi (Eds.). *The Routledge Handbook of Second Language Acquisition and Individual Differences* (pp. 37–53). New York: Routledge.

Li, S., & DeKeyser, R. (2021). Implicit language aptitude: Conceptualizing the construct, validating the measures, and examining the evidence. *Studies in Second Language Acquisition, 43*(3), 473–497. https://doi.org/10.1017/s0272263121000024

Li, S., Ellis, R., & Zhu, Y. (2019). The associations between cognitive ability and L2 development under five different instructional conditions. *Applied Psycholinguistics, 40*(3), 693–722. https://doi.org/10.1017/S0142716418000796

Li, S., & Prior, M. T. (2022). Research methods in applied linguistics: A methodological imperative. *Research Methods in Applied Linguistics, 1*(1), 100008. https://doi.org/10.1016/j.rmal.2022.100008

Li, S., & Zhao, H. (2021). The methodology of the research on language aptitude: A systematic review. *Annual Review of Applied Linguistics, 41*, 25–54. https://doi.org/10.1017/s0267190520000136

Lieder, F., & Griffiths, T. L. (2020) Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences, 43*, e1, 1–60. doi:10.1017/S0140525X1900061X.

Linacre, J. M. (2021). R statistics Rasch packages: A survey. *Rasch Measurement Transactions, 34*(1), 1805–1807. https://www.rasch.org/rmt/rmt341.pdf

Linck, J. A., Hughes, M. M., Campbell, S. G., Silbert, N. H., Tare, M., Jackson, S., Smith, B. K., Bunting, M. F., & Doughty, C. J. (2013). Hi-LAB: a new measure of aptitude for High-Level Language Proficiency. *Language Learning, 63*(3), 530–566. https://doi.org/10.1111/lang.12011

Linck, J. A., Osthus, P., Koeth, J. T., & Bunting, M. F. (2014). Working memory and second language comprehension and production: A meta-analysis. *Psychonomic Bulletin & Review, 21*(4), 861–883. https://doi.org/10.3758/s13423-013-0565-2

Loewen, S., & Gonulal, T. (2015). Exploratory factor analysis and principal components analysis. In L. Plonsky (Ed.). *Advancing Quantitative Methods in Second Language Research* (pp. 182–212). New York: Routledge.

Loewen, S., Li, S., Fei, F., Thompson, A., Nakatsukasa, K., Ahn, S., & Chen, X. (2009). Second language learners' beliefs about grammar instruction and error correction. *Modern Language Journal, 93,* 91–104.

Logie, R. H., Camos,V., & Cowan, N. (2021). *Working Memory: State of the Science*. Oxford: Oxford University Press.

Mackey, A., Adams, R., Stafford, C. A., & Winke, P. (2010). Exploring the relationship between modified output and working memory capacity. *Language Learning, 60*(3), 501–533. https://doi.org/10.1111/j.1467-9922.2010.00565.x

MacInnis, C.C., Boss, H.C.D., & Bourdage, J.S. (2020). More evidence of participant misrepresentation on MTurk and investigating who misrepresents. *Personality and Individual Differences, 152*, 109603. https://doi.org/10.1016/j.paid.2019.109603

Mair, P., & Hatzinger, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software, 20*(9), 1–20. https://doi.org/10.18637/jss. v020.i09

Mair, P., & Wilcox, R. (2020). Robust statistical methods in R using the WRS2 package. *Behavior Research Methods, 52*, 464–488. https://doi.org/10.3758/s13428-019-01246-w

Marsden, E., Morgan-Short, K., Thompson, S., & Abugaber, D. (2018). Replication in second language research: Narrative and systematic reviews, and recommendations for the field. *Language Learning, 68*(2), 321–391. https://doi.org/10.1111/lang.12286

Masrai, A. (2019). Exploring the impact of individual differences in aural vocabulary knowledge, written vocabulary knowledge and working memory capacity on explaining L2 learners' listening comprehension. *Applied Linguistics Review, 11*(3), 423–447. https://doi.org/10.1515/applirev-2018-0106

Mattys, S. L., Baddeley, A., & Trenkic, D. (2017). Is the superior verbal memory span of Mandarin speakers due to faster rehearsal? *Memory & Cognition, 46*(3), 361–369. https://doi.org/10.3758/s13421-017-0770-8

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, N.J.: L. Erlbaum Associates.

McNamara, T. (1990). Item response theory and the validation of an ESP test for health professionals. *Language Testing, 7*(1), 52–75.

McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing, 29*(4), 555–576. https://doi.org/10.1177/0265532211430367

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23*(3), 412–433. https://doi.org/10.1037/met0000144

Meara, P. (2005). *LLAMA Language Aptitude Tests*. Swansea: Lognostics.

Meara, P. M., & Rogers, V. E. (2019). *The LLAMA Tests v3.* Lognostics. https://www.lognostics.co.uk/tools/LLAMA_3/index.htm

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational Measurement* (3rd ed.) (pp. 13–103). New York: Macmillan.

Meteyard, L., & Davies, R. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language, 112*, 104092. https://doi.org/10.1016/j.jml.2020.104092

Miller, G. A., Galanter, E., & Pribram, K. H. (1960). *Plans and the Structure of Behavior*. New York, NY: Holt, Rinehart & Winston.

Milton, J., & Hopkins, N. (2006). Comparing phonological and orthographic vocabulary size: Do vocabulary tests underestimate the knowledge of some learners? *Canadian Modern Language Review, 63*(1), 127–147.

Mislevy, M., Linck, J., Campbell, S., Jackson, S., Bowles, A., Bunting, M., & Doughty, C. J. (2010). *Predicting high-level foreign language learning: A new aptitude battery meets reliability standards for personnel selection tests*. Center for Advanced Study of Language Technical Report. College Park: University of Maryland.

Miyake, A., & Friedman, N. P. (2012). The nature and organization of individual differences in executive functions: Four general conclusions. *Current Directions in Psychological Science, 21*(1), 8–14. http://doi.org/10.1177/0963721411429458

Miyake, A., & Shah, P. (1999). *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control*. Cambridge: Cambridge University Press.

Mizumoto, A. (2023). Developing and disseminating data analysis tools for open science. In L. Plonsky (Ed.), *Open Science in Applied Linguistics*. Routledge.

Mizumoto, A. (2022a). Calculating the relative importance of multiple regression predictor variables using dominance analysis and random forests. *Language Learning, 73*(1), 161–196. https://doi.org/10.1111/lang.12518

Mizumoto, A. (2022b). Analysis code. *Datasets from "Calculating the relative importance of multiple regression predictor variables using dominance analysis and random forests"* [Analysis code]. IRIS Database, University of York, UK. https://www.iris-database.org/iris/app/home/detail?id=york:940279

Mizumoto, A., & Shimamoto, T. (2008). A comparison of Aural and Written Vocabulary Size of Japanese EFL University Learners. *Language Education and Technology*, 35–51.

Mujtaba, S. M., Gol, A. K., & Parkash, R. (2021). A study on the relationship between language aptitude, vocabulary size, working memory, and L2 writing accuracy. *Foreign Language Annals, 54*(4), 1059–1081. https://doi.org/10.1111/flan.12584

Mulaik, S. A. (1972). *The Foundations of Factor Analysis*. New York: McGraw-Hill.

Murphy, K., & Myors, B. (2004). *Statistical Power Analysis.* Mahwah: Lawrence Erlbaum.

Newman, A., Bavik, Y. L., Mount, M., & Shao, B. (2021). Data Collection via Online Platforms: Challenges and Recommendations for Future Research. *Applied Psychology, 70*(3), 1380–1402. https://doi.org/10.1111/apps.12302

Newson, M., Buhrmester, M. D., Xygalatas, D., & Whitehouse, H. (2021). Go WILD, not WEIRD. *Journal for the Cognitive Science of Religion, 6*(1–2). https://doi.org/10.1558/jcsr.38413

Nicklin, C., & Plonsky, L. (2020). Outliers in L2 Research in Applied Linguistics: A Synthesis and Data Re-Analysis. *Annual Review of Applied Linguistics, 40*, 26–55. https://doi.org/10.1017/s0267190520000057

Nicklin, C., & Vitta, J. P. (2021). Effect-Driven Sample Sizes in Second Language Instructed Vocabulary Acquisition Research. *Modern Language Journal, 105*(1), 218–236. https://doi.org/10.1111/modl.12692

Nicklin, C., & Vitta, J. P. (2022). Assessing Rasch measurement estimation methods across R packages with yes/no vocabulary test data. *Language Testing, 39*(4), 513–540. https://doi.org/10.1177/02655322211066822

Nimon, K. F., Oswald, F. L., & Roberts, K. J. (2021). *yhat: Interpreting regression effects. R package* (Version 2.0-3) [Computer software]. Retrieved from https://CRAN.R-project.org/package=yhat

Norouzian, R. (2020). Sample size planning in quantitative L2 research. *Studies in Second Language Acquisition, 42*(4), 849–870. https://doi.org/10.1017/S0272263120000017

Norris, J. M., & Ortega, L. (2012). Assessing learner knowledge. In S. M. Gass & A. Mackey (Eds.), *The Routledge Handbook of Second Language Acquisition* (pp. 573–589). Routledge.

Oberauer, K., Lewandowsky, S., Awh, A., Brown, G. D. A., Cowan, N., Donkin, C., Farrell, S., Hitch, G. J., Hurlstone, M., Ma, W., Morey, C. C., Nee, D. E., Schweppe, J., Vergauwe, E., & Ward, G. (2018). Benchmarks for models of working memory. *Psychological Bulletin, 144*(9), 885–958. doi:10.1037/ bul0000153.

Ophir, Y., Sisso, I., Asterhan, C.S.C., Tikochinski, R., & Reichart, R. (2020). The Turker blues: Hidden factors behind increased depression rates among Amazon’s mechanical Turkers. *Clinical Psychological Science, 8*(1), 65–83. https://doi. org/10.1177/21677 02619 865973

O’Reilly, D. R., & Marsden, E. (2020). Eliciting and measuring L2 metaphoric competence: Three decades on from Low (1988). *Applied Linguistics, 42*(1), 24–59. https://doi.org/10.1093/applin/amz066

Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why researchers should ALWAYS check for them). *Practical Assessment, Research & Evaluation, 9*(6), 1-8.

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experi- ments. *Journal of Behavioral and Experimental Finance, 17*, 22–27. https://doi. org/10.1016/j.jbef.2017.12.004

Palma, P., Marin, M., Onishi, K. H., & Titone, D. (2022). Learning, inside and out: Prior linguistic knowledge and learning environment impact word learning in bilingual individuals. *Language Learning, 72*(4), 980–1016. https://doi.org/10.1111/lang.12501

Pan, J., & Marsden, E. (under review). Developing internet-based *Tests of Aptitude for Language Learning* (TALL): An open research endeavour.

Perruchet, P. (2021). Why is the componential construct of implicit language aptitude so difficult to capture? A commentary on the special issue. *Studies in Second Language Acquisition, 43*(3), 677–691. https://doi.org/10.1017/s027226312100019x

Pimsleur, P. (1966). *Pimsleur Language Aptitude Battery (PLAB)*. New York: The Psychological Corporation.

Pittman, M., & Sheehan, K. (2016). Amazon's Mechanical Turk a digital sweatshop? Transparency and accountability in crowdsourced online research. *Journal of Media Ethics, 31*(4), 260–262. https://doi.org/10.1080/23736 992.2016.1228811

Plonsky, L. (2013). Study quality in SLA. *Studies in Second Language Acquisition, 35*(4), 655–687. https://doi.org/10.1017/S0272263113000399

Plonsky, L., & Derrick, D. J. (2016). A Meta-Analysis of reliability coefficients in second language research. *The Modern Language Journal, 100*(2), 538–553. https://doi.org/10.1111/modl.12335

Plonsky, L., & Gass, S. (2011). Quantitative Research Methods, Study Quality, and Outcomes: The Case of Interaction Research. *Language Learning, 61*(2), 325–366. https://doi.org/10.1111/j.1467-9922.2011.00640.x

Plonsky, L., & Gonulal, T. (2015). Methodological Synthesis in Quantitative L2 Research: A Review of Reviews and a Case Study of Exploratory Factor Analysis. *Language Learning, 65*(S1), 9–36. https://doi.org/10.1111/lang.12111

Plonsky, L., Marsden, E., Crowther, D., Gass, S., & Spinner, P. (2020). A methodological synthesis and meta-analysis of judgment tasks in second language research. *Second Language Research, 36,* 583–621*.* https://doi.org/10.1177/0267658319828413

Plonsky, L., & Oswald, F. L. (2014). How Big Is 'Big'? Interpreting Effect Sizes in L2 Research. *Language Learning, 64*(4), 878–912. https://doi.org/10.1111/lang.12079

Plonsky, L., & Oswald, F. L. (2017). Multiple regression as a flexible alternative to ANOVA in L2 research. *Studies in Second Language Acquisition, 39*(3), 579–592. https://doi.org/10.1017/s0272263116000231

Purpura, J. E., Brown, J. D., & Schoonen, R. (2015). Improving the validity of quantitative measures in Applied Linguistics research. *Language Learning, 65*(S1), 37–75. https://doi.org/10.1111/lang.12112

Plonsky, L., & Ziegler, N. (2016). The CALL-SLA interface: Insights from a second-order synthesis. *Language Learning & Technology, 20*(2), 17 – 37. http://dx.doi.org/10125/44459

Raykov, T., & Marcoulides, G. A. (2017). Thanks coefficient Alpha, we still need you! *Educational and Psychological Measurement, 79*(1), 200–210. https://doi.org/10.1177/0013164417725127

R Core Team (2022). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.*

Revelle, W. (2022) *psych: Procedures for Personality and Psychological Research.* Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 2.2.5.

Revelle, W., & Condon, D. M. (2019). Reliability from α to ω: A tutorial. *Psychological Assessment, 31*(12), 1395–1411. https://doi.org/10.1037/pas0000754

Révész, A. (2012). Working memory and observed effectiveness of recasts on different L2 outcome measures. *Language Learning, 62,* 93–132.

Rizopoulos, D. (2006). ltm: An R Package for Latent Variable Modeling. *Journal Of Statistical Software, 17*(5), 1–25. http://rsirt.googlecode.com/files/ltmPackage.pdf

Robinson, P. (1997). Individual differences and the fundamental similarity of implicit and explicit adult second language learning. *Language Learning, 47*,45–99.

Robinson, P. (2002a). Learning conditions, aptitude complexes and SLA: A framework for research and pedagogy. In P. Robinson (Ed.), *Individual Differences and Instructed Language Learning* (pp. 113–133). Amsterdam: John Benjamins.

Robinson, P. (2002b). Effects of individual differences in intelligence, aptitude and working memory on adult incidental SLA: A replication and extension of Reber, Walkenfeld and Hernstadt, 1991. In P. Robinson (Ed.), *Individual Differences and Instructed Language Learning* (pp. 211–266). Amsterdam: John Benjamins.

Robinson, P. (2005). Aptitude and second language acquisition. *Annual Review of Applied Linguistics, 25,* 46–73.

Robinson, P. (2007). Aptitudes, abilities, contexts, and practice. In R. M. DeKeyser (Ed.), *Practice in Second Language* (pp. 256–286)*.* Cambridge: Cambridge University Press.

Robinson, P. (2012). Individual differences, aptitude complexes, SLA processes, and aptitude test development. In M. Pawlak (ed.), *New Perspectives on Individual Differences in Language Learning and Teaching* (pp. 57–75)*.* Berlin Heidelberg: Springer-Verlag.

Roehr, K. (2008). Metalinguistic knowledge and language ability in university-level L2 learners. *Applied Linguistics, 29,* 173–199.

Roehr, K. (2012). Aptitude treatment interaction (ATI) research. In P. Robinson (Ed.), *The Routledge Encyclopaedia of Second Language Acquisition* (pp. 31–35). Routledge.

Roehr-Brackin, K. (2021). Measuring aptitude. In P. Winke & T. Brunfaut (Eds.), *The Routledge Handbook of Second Language Acquisition and Testing* (pp. 147–156). Routledge.

Roehr-Brackin, K. (2022). Investigating explicit and implicit L2 knowledge and learning: Replications of Erlam (2005) and Roehr-Brackin and Tellier (2019). *Language Teaching, 55*(2), 271–283. https://doi.org/10.1017/s0261444820000415

Rogers, V., & Meara, P. (2019). Turning a LLAMA into an ALPACAA and back again: An initial revised attempt at assessing aptitude. *Language Aptitude Roundtable.* https://viviennerogers.info/wp-content/uploads/2020/09/alpacaa_2019_macau.pdf

Rogers, V., Meara, P., Barnett-Legh, T., Curry, C., & Davie, E. (2017). Examining the LLAMA aptitude tests. *Journal of the European Second Language Association, 1*(1), 49–60. https://doi.org/10.22599/jesla.24

Rogers, V., Meara, P., & Rogers, B. (2023). Testing language aptitude: LLAMA validation and refinement. In Z. Wen, P. Skehan, R. L. Sparks (Eds), *Language Aptitude Theory and Practice* (pp.47–72). Cambridge University Press.

Roque-Gutierrez, E., & Ibbotson, P. (2023). Working memory training improves children ' s syntactic ability but not vice versa: A randomized control trial. *Journal of Experimental Child Psychology, 227*, 105593–105593. https://doi.org/10.1016/j.jecp.2022.105593

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Ruiz, S., Rebuschat, P., & Meurers, D. (2019). The effects of working memory and declarative memory on instructed second language vocabulary learning: Insights from intelligent CALL. *Language Teaching Research, 25*, 510–539. https://doi.org/10.1177/1362168819872859

Sáfár, A., & Kormos, J. (2008). Revisiting problems with foreign language aptitude. *International Review of Applied Linguistics in Language Teaching, 46*, 113–136.

Saffran, J.R., & Thiessen, E.D. (2007). Domain-general learning capacities. In E. Hoff & M. Shatz (Eds.), *Blackwell Handbook of Language Development* (pp. 68–86). Malden, MA: Blackwell. doi:10.1002/9780470757833.ch4

Sagarra, N. (2007). From CALL to face-to-face interaction: The effect of computer-delivered recasts and working memory on L2 development. In A. Mackey (Ed.), *Conversational Interaction in Second Language Acquisition: A series of Empirical Studies* (pp. 229–248). Oxford: Oxford University Press.

Saito, K. (2019). The role of aptitude in second language segmental learning: The case of Japanese learners' English /ɹ/ pronunciation attainment in classroom settings. *Applied Psycholinguistics, 40*(1), 183–204. https://doi.org/10.1017/s0142716418000528

Saito, K., Sun, H., & Tierney, A. (2019). Explicit and implicit aptitude effects on second language speech learning: Scrutinizing segmental and suprasegmental sensitivity and performance via behavioural and neurophysiological measures. *Bilingualism: Language and Cognition, 22*, 1123–1140.

Saito, K., Suzukida, Y., Tran, M., & Tierney, A. (2021). Domain-General Auditory Processing Partially Explains Second Language Speech Learning in Classroom Settings: A Review and Generalization Study. *Language Learning, 71*(3), 669–715. https://doi.org/10.1111/lang.12447

Sasaki, M. (1996). *Second language proficiency, foreign language aptitude, and intelligence.* Baltimore: Peter Lang.

Schoonen, R. (2015). Structural equation modeling in L2 research. In L. Plonsky (Ed.), *Advancing Quantitative Methods in Second Language Research* (pp. 213–242). Routledge.

Schwieter, J. W. & Wen, E. Z. (2022). T*he Cambridge Handbook of Working Memory and Language.* Cambridge: Cambridge University Press.

Schwieter, J. W., Wen, E. Z., & Bennett, T. (2022). Working memory and language: An overview of key topics. In J. W. Schwieter, & E. Z. Wen (Eds.), *The Cambridge Handbook of Working Memory and Language.* Cambridge: Cambridge University Press

Shank, D.B. (2016). Using crowdsourcing websites for sociological research: The case of Amazon Mechanical Turk. *American Sociologist, 47*(1), 47–55. https://doi. org/10.1007/s1210 8-015-9266-9

Shepperd, L. (2022). Including underrepresented language learners in SLA research: A case study and considerations for internet-based methods. *Research Methods in Applied Linguistics, 1*(3), 100031. https://doi.org/10.1016/j.rmal.2022.100031

Singleton, D. (2017). Language aptitude: Desirable trait or acquirable attribute? *Studies in Second Language Learning and Teaching, 7*(1), 89–103. https://doi.org/10.14746/ssllt.2017.7.1.5

Skehan, P. (1998). *A cognitive Approach to Language Learning*. Oxford: Oxford University Press.

Skehan, P. (2002). Theorising and updating aptitude. In P. Robinson (ed.), *Individual Differences and Instructed Language Learning* (pp. 69-95). Amsterdam, Netherlands: John Benjamins.

Skehan, P. (2012). Language aptitude. In S. Gass & A. Mackey (eds.), *The Routledge Handbook of Second Language Acquisition* (pp. 381–395). New York: Routledge.

Skehan, P. (2016). Foreign language aptitude, acquisitional sequences, and psycholinguistic processes. In G. Granena, D. Jackson & Y. Yilmaz (eds.), C*ognitive Individual Differences in L2 Processing and Acquisition* (pp.15–38). Amsterdam: John Benjamins. https://doi.org/10.1075/bpa.3

Skehan, P. (2019). Language aptitude implicates language and cognitive skills. In Z. Wen, P. Skehan, A. Biedroń, S. Li, & R. L. Sparks (eds.), *Language Aptitude: Advancing Theory, Testing, Research and Practice* (pp. 56–77). New York: Rouledge.

Skehan, P. (2023). Testing language aptitude: A commentary on batteries and reanalysis of constructs. in Z. Wen, P. Skehan, & R. Sparks (eds.), *Language Aptitude Theory and Practice* (pp.208–245). Cambridge University Press.

Shin, J., & Hu, Y. (2022). A methodological synthesis of working memory tasks in L2 research. In J. W. Schwieter, & Z. Wen (Eds.), *The Cambridge Handbook of Working Memory and Language* (pp. 722–745). Cambridge: Cambridge University Press.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive psychology. *Psychological Science, 22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Slocum-Gori, S. L., & Zumbo, B. D. (2010). Assessing the unidimensionality of psychological scales: Using multiple criteria from factor analysis. *Social Indicators Research, 102*(3), 443–461. https://doi.org/10.1007/s11205-010-9682-8

Slocum-Gori, S. L., Zumbo, B. D., Michalos, A. C., & Diener, E. (2009). A note on the dimensionality of quality of life scales: An illustration with the Satisfaction With Life Scale (SWLS). *Social Indicators Research: An International Interdisciplinary Journal for Quality of Life Measurement, 92*, 489–496.

Snow, R. E. (1991). Aptitude-treatment interaction as a framework for research on individual differences in psychotherapy. *Journal of Consulting and Clinical Psychology, 59*, 205–216.

Sok, S., & Shin, H. W. (2021). Investigating the role of cognitive variables in second language learners' listening comprehension: Aptitude and metacognitive awareness. *International Journal of Listening*, *36*(2), 138–151. https://doi.org/10.1080/10904018.2021.1954926

Sokal, R.R., & Rohlf, F.J. (1995). *Biometry: The Principles and Practice of Statistics in Biological Research*. (3rd ed.). New York: W .H. Freeman and Co.

Sparks, R. L., & Dale, P. S. (2023). The prediction from MLAT to L2 achievement is largely due to MLAT assessment of underlying L1 abilities. *Studies in Second Language Acquisition*, 1–25. https://doi.org/10.1017/s0272263123000037

Sparks, R. L., Dale, P. S., & Patton, J. (2023). Individual differences in L1 attainment and language aptitude predict L2 achievement in instructed language learners. *The Modern Language Journal, 107*(2), 479–508. https://doi.org/10.1111/modl.12841

Sparks, R., & Ganschow, L. (2001). Aptitude for learning a foreign language. *Annual Review of Applied Linguistics, 21*, 90–111.

Sparks, R., Patton, J., Ganschow, L., & Humbach, N. (2011). Subcomponents of L2 aptitude and L2 proficiency. *The Modern Language Journal, 95*, 253–273.

Speciale, G., Ellis, N. C., & Bywater, T. (2004). Phonological sequence learning and short-term store capacity determine second language vocabulary acquisition. *Applied Psycholinguistics, 25*, 293–320.

Sternberg, R. J. (1997). *Successful intelligence*. New York: Plume.

Sternberg, R. J. (2002). The theory of successful intelligence and its implications for language aptitude testing. In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 13–44). Amsterdam: Benjamins.

Stevens, J. (1996). *Applied multivariate statistics for the social sciences* (3rd ed.).Mahwah: Erlbaum.

Stevens, J. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Mahwah: Erlbaum.

Stone, A.A., Walentynowicz, M., Schneider, S., Junghaenel, D.U., & Wen, C.K. (2019). MTurk participants have substantially lower evaluative subjective wellbe- ing than other survey participants. *Computers in Human Behavior, 94*, 1–8. https:// doi.org/10.1016/j.chb.2018.12.042

Stritch, J.M., Pedersen, M.J., & Taggart, G. (2017). The opportunities and limitations of using Mechanical Turk (MTurk) in public administration and management scholarship. *International Public Management Journal, 20*(3), 489–511.

Suarez-Rivera, C., Linn, E., & Tamis-LeMonda, C. S. (2022). From play to language: infants' actions on objects cascade to word learning. *Language Learning, 72*(4), 1092–1127. https://doi.org/10.1111/lang.12512

Suga, K., & Loewen, S. (2023). Potential test-learning effects of an oral elicited imitation test: Methodological considerations for form-focused instruction studies. *Research Methods in Applied Linguistics, 2*(1), 100035. https://doi.org/10.1016/j.rmal.2022.100035

Suzuki, Y. (2021a). Probing the construct validity of LLAMA_D as a measure of implicit learning aptitude. *Studies in Second Language Acquisition, 43* (S2), 663 - 676. https://doi.org/10.1017/s0272263120000704

Suzuki, Y. (2021b). Individual differences in memory predict changes in breakdown and repair fluency but not speed fluency: A short-term fluency training intervention study. *Applied Psycholinguistics, 42* (4), 969 - 995. https://doi.org/10.1017/S0142716421000187

Suzuki, Y., Jeong, H., Cui, H., Okamoto, K., Kawashima, R., & Sugiura, M. (2022). An fMRI validation study of the word-monitoring task as a measure of implicit knowledge: Exploring the role of explicit and implicit aptitudes in behavioral and neural processing. *Studies in Second Language Acquisition, 45*(1), 109–136. https://doi.org/10.1017/s0272263122000043

Suzuki, Y., Jeong, H., Cui, H., Okamoto, K., Kawashima, R., & Sugiura, M. (2023). fMRI reveals the dynamic interface between explicit and implicit knowledge recruited during elicited imitation task. *Research Methods in Applied Linguistics, 2*(2), 100051. https://doi.org/10.1016/j.rmal.2023.100051

Suzuki, Y., & Koizumi, R. (2020). Using equivalent test forms in SLA pretest-posttest design research. In P. Winke, & T. Brunfaut (Eds.), *The Routledge Handbook of Second Language Acquisition and Language Testing* (pp. 457–467). New York: Routledge.

Tabachnick, B. G., & Fidell, L. S. (2012). *Using multivariate statistics* (6th ed.). Boston, MA: Pearson.

Tseng, A., Greiner, J., Levitas, D., et al. (2015). *The Hi-LAB Longitudinal Predictive Validity Study: Second Year Accomplishments and Next Steps* (Technical Report). College Park, MD: University of Maryland Center for Advanced Study of Language.

Unsworth, N., Redick, T. S., Heitz, R. P., Broadway, J. M., & Engle, R. W. (2009). Complex working memory span tasks and higher-order cognition: A latent-variable analysis of the relationship between processing and storage. *Memory, 17*(6), 635–654. https://doi.org/10.1080/09658210902998047

van der Linden, W. J. (2016). *Handbook of Item Response Theory: Volume 1: Models.* Chapman & Hall/CRC.

Vuong L. C. & Wong P. C. M. (2019). From individual differences in language aptitude to personalized learning. In Z. Wen, A. Biedroń, P. Skehan, S. Li, & R. Sparks (Eds.), *Language Aptitude: Advancing Theory, Testing, Research and Practice* (pp. 330–342). New York: Routledge.

Walker, N., Monaghan, P., Schoetensack, C., & Rebuschat, P. (2020). Distinctions in the acquisition of vocabulary and grammar: An individual differences approach. *Language Learning, 70*(52), 1–34. https://doi.org/10.1111/1467-923X.12837

Wen, Z. (2016). *Working Memory and Second Language Learning: Toward an Integrated Approach*. Multilingual Matters. https://doi.org/10.21832/9781783095735

Wen, Z. (Edward), Biedroń, A., & Skehan, P. (2017). Foreign language aptitude theory: Yesterday, today and tomorrow. *Language Teaching, 50*(1), 1–31. https://doi.org/10.1017/s0261444816000276

Wen, Z., & Jackson, D. O. (2022). Working memory. In S. Li, P. Hiver, & M. Papi (Eds.), *The Routledge Handbook of Second Language Acquisition and Individual Differences* (pp. 54–66). New York: Routledge

Wen, Z., Juffs, A., & Winke, P. (2021). Measuring working memory. In P. Winke & T. Brunfaut (Eds.), *The Routledge Handbook of Second Language Acquisition and Testing* (pp. 167–76). Routledge.

Wen, Z., & Skehan, P. (2021). Stages of Acquisition and the P/E Model of Working Memory: Complementary or contrasting approaches to foreign language aptitude? *Annual Review of Applied Linguistics, 41*, 6–24. https://doi.org/10.1017/S0267190521000015

Wen, Z., Skehan, P., Biedroń, A., Li, S., & Sparks, R. L. (2019). *Language aptitude: Advancing Theory, Testing, Research and Practice.* Rouledge:Taylor & Francis Group.

Wen, Z., Skehan, P., & Sparks, R. L. (2023a). *Language Aptitude Theory and Practice.* Cambridge University Press.

Wen, Z., Skehan, P., & Sparks, R. L. (2023b). Language aptitude research: From testing to theory and practice. In Z. Wen, P. Skehan, & R. Sparks (eds.), *Language Aptitude Theory and Practice* (pp. 1–21)*.* Cambridge University Press.

Wickham et al., (2019). Welcome to the tidyverse. *Journal of Open Source Software, 4*(43), 1686, https://doi.org/10.21105/joss.01686

Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist, 53,* 300–314.

Williams, J. N. (2012). Working memory and SLA. In S. Gass & A. Mackey (Eds.), *The Routledge Handbook of Second Language Acquisition* (pp. 427–441). Abingdon/New York: Routledge.

Williams, J. N., & Lovatt, P. (2003). Phonological memory and rule learning. *Language Learning, 53*(1), 67–121.

Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications. *arXiv (Cornell University)*. https://arxiv.org/pdf/1308.5499

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370–371. https://rasch.org/rmt/rmt83b.htm

Xiao, R., Rayson, P., & McEnery, T. (2009). *A Frequency Dictionary of Mandarin Chinese*. In Routledge eBooks. https://doi.org/10.4324/9780203883075

Yalçın, Ş., & Spada, N. (2016). Language aptitude and grammatical difficulty: An EFL classroom-based study. *Studies in Second Language Acquisition, 38*, 239–263.

Yi, W. (2018). Statistical sensitivity, cognitive aptitudes, and processing of collocations. *Studies in Second Language Acquisition, 40*, 831–856.

Zalbidea, J. (2017). 'One Task Fits All'? The Roles of Task Complexity, Modality, and Working Memory Capacity in L2 Performance. *Modern Language Journal, 101*(2), 335–352. https://doi.org/10.1111/modl.12389

Zalbidea, J. (2021). On the scope of output in SLA. *Studies in Second Language Acquisition, 43*(1), 50–82. https://doi.org/10.1017/s0272263120000261

Zalbidea, J., & Sanz, C. (2020). Does learner cognition count on modality? Working memory and L2 morphosyntactic achievement across oral and written tasks. *Applied Psycholinguistics, 41*(5), 1171–1196. https://doi.org/10.1017/s0142716420000442

Zinbarg, R. E., Revelle, W., & Yovel, I. (2007). Estimating ωh for structures containing two group factors: Perils and prospects. *Applied Psychological Measurement, 31*(2), 135–157. doi:10.1177/0146621605278814

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and Mcdonald's ωH: their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70*(1), 123–133. https://doi.org/10.1007/s11336-003-0974-7

Zinbarg, R. E., Yovel, I., Revelle, W., & McDonald, R. P. (2006). Estimating generalizability to a latent variable common to all of a scale's indicators: A comparison of estimators for ωh. *Applied Psychological Measurement, 30*(2), 121–144. doi: 10.1177/0146621605278814