BRITISH COUNCIL

ENGLISH LANGUAGE
ASSESSMENT RESEARCH GROUP

Technical Report

Validity and Usage of the Aptis Grammar and Vocabulary (Core) Component

TR/2020/003

Gareth McCray & Karen Dunn

ISSN 2057-7168

www.britishcouncil.org/aptis

CONTENTS

LIST OF TABLES

LIST OF FIGURES

EXECUTIVE SUMMARY

The purpose of this research paper is to report on the analysis of a large global sample of live Aptis test data to investigate to what extent the theoretical assumptions hold regarding the central role accorded to the grammar and vocabulary (henceforth "Core") component of the test.

The Aptis test offers a flexible English language assessment system in which candidates can choose to complete a range of test components, from the second language (L2) skills of Reading, Listening, Speaking and Writing. All candidates are also required to complete the grammar and vocabulary knowledge "Core" component. Scores in this component are reported to each candidate on a scale of 0–50 and also perform an important function in refining CEFR level allocation for each of the individual skill components. The findings from this research project provide empirical insights into how this scoring model functions for a large and varied global sample of participants (over 66,000 in total, with sub-analyses being carried out using a sample of at least 10,000 candidates who took all five Aptis components).

The focus of the analysis in this report is twofold. The first element of the research addresses the relationship between grammar and vocabulary as operationalised and tested within the Aptis core component. Following this, the focus broadens to incorporate all components of the Aptis test, including Listening, Reading, Speaking, and Writing. The research questions are as follows:

RQ1: To what extent are grammar and vocabulary items which constitute the Core component in Aptis separable constructs?

RQ2: To what extent does the grammar and vocabulary component represent a core construct in the Aptis test?

Under RQ1 we examine the evidence which indicates whether grammar and vocabulary components can be considered as separable. This is an important first step in the analysis since, if the grammar and vocabulary components were to be found to be separable constructs, then it would call into question the validity of using a single score from the Core component to inform the final achievement on the additional skills. Dimensionality of this component was assessed using parallel analysis and multidimensional item response theory (MIRT). The findings gave broad support for the treatment of the Core component scores as lying on a single scale.

RQ2 provides the opportunity to explore the extent to which we might consider the Core section as central to the construct of English language ability, as measured by all components (i.e. Reading, Listening, Speaking, Writing, plus the Core component). If the Core component is not "central", it calls into question the validity of using the scores on this Core component to inform the achievement on the additional components. Correlation analysis indicated that the relationship between the Core component and each of the skill areas was stronger than relationships between individual skill areas. This was complemented by MIRT analysis, drawing on evidence from the bifactor structure for assessing the viability of subscales, which indicated that the grammar and vocabulary questions in the Core component only draw minimally on latencies above and beyond the *L2 English Ability* factor hypothesised to explain Aptis test performance, in contrast to the skill-area components which rely more heavily on commonalities not captured by the general factor. Finally, a series of generalised additive model for location, shape and space (GAMLSS) indicated a broadly linear relationship between Core component performances and those of each of the skills components.

The combined evidence from each of these investigative strands provides evidence to consider the Core component as indeed "core" to the Aptis test.

The study reported in this technical document can therefore be viewed as providing empirical support for both the scoring model underpinning the role of the Core component in the Aptis test, and the cognitive processing models of L2 language proficiency upon which the Aptis test was founded, i.e., those which posit grammar and vocabulary in a core role across all language domains (Field, 2013; Khalifa & Weir, 2009). While this has been shown empirically in other contexts, the usefulness of the current study lies in its use of Aptis data, the large scale of the participant sample, and the wide range of language aptitudes incorporated. This study was able to map the relationship between grammar and vocabulary, and each of the skills of Listening, Reading, Writing, and Speaking for abilities spanning CEFR levels A0 to C[1].

With respect to the design of the Aptis test, the study findings indicate that there is no strong evidence to contradict the current approach of reporting and using the Core component scores on a single scale, i.e., amalgamating scores from grammar and vocabulary items. Additionally, the results of the investigation gave support for using the Core component scores to feed into the CEFR level allocations in the other skill areas.

Authors

Gareth McCray
Gareth McCray is a Research Fellow in Statistics at Keele University. His research interests include Psychometric Measurement, Item Response Theory and Rasch modelling, Diagnostic Testing, Eye-Tracking, and Child Development Measurement and Global Health. He has been involved with a number of national and international studies related to psychometric measurement or diagnosis including several with the World Health Organisation. His work appears in journals such as *Language Testing, BMJ Global Health, Medical Education,* and *BMC Medical Research Methods*.

Karen Dunn
Karen Dunn is a Senior Researcher in measurement and evaluation at the British Council. She holds a PhD in Applied Social Statistics and Masters in Language Studies. Karen's role involves developing the language testing and assessment research focus to provide high quality analysis for specialised interest groups. In addition, Karen contributes to a range of British Council research projects with a language testing focus, including research in conjunction with international organisations, such as the Rwandan Ministry of Education and the Pratham–ASER Centre in India.

[1] Aptis General does not discriminate between C1 and C2 level abilities; a further study using data from the Aptis Advanced would provide the opportunity.

1. INTRODUCTION

The research study described in this document explores the interpretation and use of the Aptis Grammar and Vocabulary component, also known as the “Core” component. The analytic focus is on empirically validating the theoretical assumptions underlying the design of the test scoring mechanism by which scores from this core component feed into CEFR level decisions for each of the four individual language skill areas. Close statistical scrutiny is made of a large data set comprising candidate responses to all five components of the Aptis test, first investigating the nature of the relationship between the grammar section and vocabulary section of the Core component, and subsequently the relationship of the Core component as a whole to the rest of the skills components in the Aptis test. This provides important groundwork for ongoing research into the functioning of the Aptis scoring process. In particular, the empirical evidence reported in this document is intended to form the basis of the upcoming evaluation of the functioning of live scoring mechanisms for test development purposes.

2. AIMS OF THE RESEARCH

The purpose of this research paper is to report on the analysis of a large global sample of live Aptis test data to investigate to what extent the theoretical assumptions hold regarding the central role accorded to the grammar and vocabulary (henceforth “Core”) component of the test. This section outlines the Aptis test and provides some theoretical background to the test design decisions that positioned this component at the centre of the testing system.

2.1 Core component in the Aptis testing system

The Aptis test offers a flexible English language assessment system in which candidates can choose to complete a range of test components from the second language (L2) skills of Reading, Listening, Speaking and Writing. Results are reported both on a scale between 0–50 and as a level on the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001). All candidates are also required to complete the grammar and vocabulary knowledge “Core” component.

Scores in this component are reported to each candidate on a scale of 0–50, and, in addition, perform an important function in refining CEFR level allocation for each of the individual skill components. This procedure reflects the understanding of grammar and vocabulary as key sub-processes in models of L2 language ability (Field, 2013; Khalifa & Weir, 2009). The intention in building this into the Aptis scoring model is to increase the fairness and accuracy of grade allocation in cases in which a candidate’s performance in any of the four skill areas fall just shy of a grade boundary.

The Aptis testing system was therefore designed with grammar and vocabulary at its heart, with performance on the Core component considered to be associated with some of the fundamental skills required in each of the skill components. The use of the score information to refine decisions is justified on theoretical grounds (O'Sullivan & Dunlea, 2015), as described in a little more detail below.

2.2 Theoretical assumptions underpinning the Aptis test design

The design of the Aptis test and scoring system is based on the socio-cognitive model. The technical manual states that 'tasks are designed to reflect carefully considered models of language progression that incorporate cognitive processing elements' (O'Sullivan & Dunlea, 2015, p. 6). For the receptive skills, this refers to Khalifa and Weir's (2009) cognitive model of Reading, and Field's (2013) cognitive processing framework for Listening. Both these models incorporate a central role for lexical and grammatical knowledge, with *word recognition* and *parsing* as foundational, bottom-up, sub-processes (see Figures 1 and 2).

Figure 1: The central processing core of Khalifa and Weir's (2009) cognitive model of Reading (as per adaptation for Brunfaut and McCray, 2015)

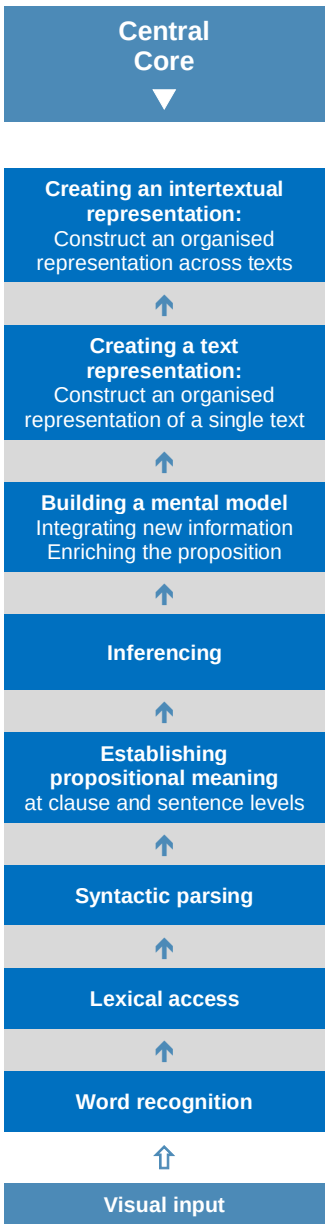
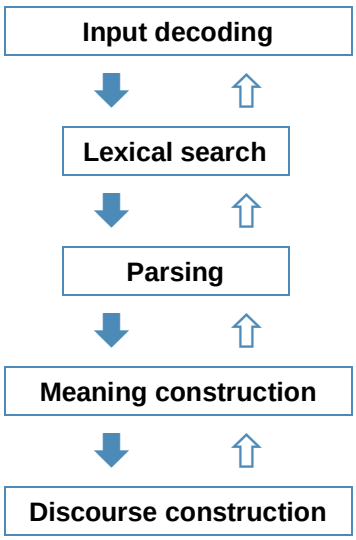


Figure 2: A simplified cognitive model of the listening process (Field, 2018)



With respect to productive skills, there is not an equivalent role for such models in test design and development, as the focus is on the language that a learner *chooses* to use in order to communicate their intended response, rather than that which they are responding to. As such, the Aptis test employs a checklist approach to task design with language functions derived from the British Council–EAQUALS Core Inventory (North, Ortega & Sheehan, 2010) and the lists for Speaking developed by (O'Sullivan, Weir & Saville, 2002) form the basis of productive skill tasks. With respect to this approach, O'Sullivan et al. (2002) state:

> We are not claiming that it is possible to predict language use at a micro level (grammatical form or lexis), but that it is possible to predict informational and interactional functions and features of interaction management.
>
> (O'Sullivan et al., 2002, p. 47)

However, in terms of understanding the process by which a learner selects language for use, models for productive language give a central role to the ability to retrieve relevant vocabulary and grammar. For example, in Levelt's (1989) model of Speaking, and Grabe and Kaplan's (1996) model of text construction in Writing, we see the lexicon, in particular, as taking a central position, with grammatical encoding and syntax closely associated.

Indeed, the majority of second language acquisition and language testing researchers endorse the view that grammar and vocabulary are foundational aspects of language ability. In their multi-componential framework of communicative language ability, (Bachman & Palmer, 1996, 2010) place 'knowledge of syntax, vocabulary, phonology, and graphology' together under the title 'grammatical knowledge' (Bachman & Palmer, 2010, p. 44) as the first element in the framework. However, it is also recognised that this knowledge will be activated differently within each language domain. For example while readers can rely on linguistic information in the text via bottom-up processes, the 'online' nature of Listening means that learners tend to draw more on top-down processes (Lund, 1991; Park, 2004). In practical terms, this means for example that the L2 listener will often compensate for lack of specific vocabulary knowledge by drawing on other, more general or metacognitive, areas of knowledge, but this is less common in L2 Reading (van Zeeland & Schmitt, 2013, p. 461). (See also Hu & Nation, 2000; Wang & Treffers-Daller, 2017). In productive skills meanwhile, the learner has more opportunity to actively decide upon the level of complexity at which they operate (e.g., Schoonen et al., 2003).

While it is not to be assumed, therefore, that grammar and vocabulary will play identical roles with respect to cognitive processing for each skill areas, these elements of language knowledge are understood to play a strong part in explaining learner performances across all language domains. This is the first time that a large-scale study has tested this empirically using Aptis test data.

3. LITERATURE REVIEW

Grammar and vocabulary have both been shown to be very good predictors of overall L2 language proficiency and correlate highly with other skills; the richest empirical evidence in this area has arisen from 'componential' studies, which are designed to assess the relative role of a range of variables in explaining L2 language ability. These often incorporate grammar or vocabulary, or both, as hypothesised explanatory factor(s). There is also a body of literature devoted solely to exploring the relationship between vocabulary and L2 proficiency, which is considered after a closer look at some relevant componential studies.

3.1 Componential studies of L2 ability

The aim of componential studies is to tease apart L2 skills into subskills, or components, incorporating a range of different elements, often including both grammar and vocabulary as key factors. The discussion below focuses on studies looking into each language skill in turn. Overall, it serves to illustrate that while there is a general level of agreement that grammar and/or vocabulary make a substantial contribution to the prediction of L2 skills performance, there is little agreement in the *balance* in the contribution of grammar and vocabulary, even in studies investigating the same L2 skill. Some researchers emphasise the role of grammar, others vocabulary, and yet more conflate the two. These empirical studies nonetheless show that while there is a key role for both vocabulary and grammar in explaining L2 proficiency, the relationship is not necessarily straightforward. The current report will investigate the role of both grammar and vocabulary in explaining test performance in each of the four skill areas, for candidates across the ability spectrum, thus making an important contribution to this body of literature.

Listening: Joyce (2011) has demonstrated a strong role for grammatical knowledge (using an aural measure) in explaining L2 Listening performance, with an impressive correlation of .81. By contrast, Joyce's study found no significant role for an aural vocabulary breadth measure, although there was a correlation of .73 with a 'phonological modification knowledge' variable which describes the learners' ability to segment the aural input into separate lexical items (Joyce, 2011, p. 87). This highlights the importance of possessing the ability to accurately detect isolatable words in the flow of speech. See also Pan, Tsai, Huang and Liu (2018) which demonstrates the relevance of this skill at a multi-word level.

Joyce's finding of a predominant role for grammar was at odds with that of Mecartty (2000) who concluded that grammar, operationalised through a sentence completion multiple-choice task and a grammaticality judgement task, was not a significant predictor of Listening ability alongside vocabulary knowledge. The measure of vocabulary used in her study had a moderate correlation of .38 with Listening comprehension (Mecartty, 2000, p. 336).

Mecartty's findings have influenced the study design of several recent componential studies of L2 Listening (e.g., Vandergrift & Baker, 2015; Wang & Treffers-Daller, 2017). These studies do not hypothesise grammatical ability as a factor, focusing rather on metacognitive abilities and working memory capacity, alongside vocabulary knowledge. Wang and Treffers-Daller (2017) observe a correlation of .44 between vocabulary knowledge and Listening comprehension. As a result of these different approaches there is therefore no consensus on the balance in the role of grammar and vocabulary in L2 Listening.

Reading: In order to successfully parse the combinations of words encountered in a Reading text, L2 readers must be able to draw upon a suitable level of grammatical resource (Jung, 2009; Weir, 2005). Indeed, the importance of grammatical or syntactic knowledge has been long emphasised by researchers, with Grabe (2009) stating that 'the process of parsing incoming text for structural information that supports comprehension is something that happens almost every second during fluent Reading' (Grabe, 2009, p. 199). (See also Alderson, 1993 and van Gelderen et al., 2004.)

A large body of evidence has accumulated to attest to the role played by *both* grammar and vocabulary as strong predictors of Reading ability; a comprehensive overview can be found in Jeon and Yamashita's meta-analysis of the findings of component-skill studies (2014). The results of their meta-analysis indicated L2 grammar knowledge (r = .85), L2 vocabulary knowledge (r = .79) to be two of the three strongest correlates of L2 Reading comprehension, the third being 'L2 decoding' (r = .56).

Writing: The relationship between the development of L2 Writing and grammatical/vocabulary knowledge does not follow a simple trajectory (Larsen-Freeman, 2006). In a componential study comparing L1 and L2 Writing, Schoonen et al. (2003) show that linguistic variables, including lexical, grammatical and orthographic knowledge, as well as fluency of retrieval, all play a significant role in L2 Writing amongst 8th grade EFL learners. Correlations of .63 were found between vocabulary and L2 Writing proficiency, and .84 between grammar and L2 Writing (Schoonen et al., 2003, p. 193).

However, a subsequent longitudinal study, also using SEM analysis, allowed these researchers to demonstrate an increasing role for grammatical knowledge and lexical retrieval in L2 Writing between 8th and 10th grade learners, over vocabulary knowledge (Schoonen, Van Gelderen, Stoel, Hulstijn & De Glopper, 2011, p. 65).

Other studies also highlight an important role for grammar in L2 Writing (cf. McNamara, 1996; Wolfe-Quintero, Inagaki & Kim, 1998).

Speaking: Grammar plays a prominent role in research investigating the componentiality of L2 Speaking ability. McNamara (1990) and Iwashita, Brown, McNamara and O'Hagan (2008) observe a strong role for grammar in describing levels of L2 Speaking proficiency. However, both these studies base the explanation of the holistic rating scale on evidence elicited from the learner in the same Speaking test.

In response to studies set up in this way, De Jong, Steinel, Florijn, Schoonen and Hulstijn (2012) highlight the need to draw upon evidence outside of immediate Speaking test production to reference the levels of the various components hypothesised to contribute to Speaking ability. In taking a structural equation modelling approach to investigating the componentiality of L2 Speaking knowledge, these researchers found a high correlation between the latent vocabulary and grammar measures derived from written tests, with the role of vocabulary knowledge superseding grammar in the final explanatory model (De Jong et al., 2012, pp. 26–27). They also found grammar and vocabulary to play a more dominant role at lower ability levels.

3.2 Vocabulary knowledge and L2 ability

Studies designed to address the relationship between vocabulary test scores and a range of L2 skill areas include Milton, Wade and Hopkins (2010) and Zimmerman (2004) who each explored relationships between vocabulary and language performance across all four modalities, and Stæhr (2008), who examined relationships with Writing, Reading and Listening.

Findings in these studies indicate that vocabulary knowledge alone accounts for a large proportion of variance in L2 language test performance. For example, in a series of binary regression models, Stæhr's study indicated that a measure of written receptive vocabulary size can account for 72% of the variance in the ability to obtain an average score or above in the Reading test, 52% of the variance for achieving above-average Writing score, and 39% in the Listening scores. The manner of testing for vocabulary knowledge appears to have some bearing on the strength relationship with skill performance. Milton et al. (2010) employed two vocabulary test delivery modes: A-Lex, an aural vocabulary measure (Milton & Hopkins, 2005), and X-Lex, a written vocabulary measure (Meara & Milton, 2003). Correlations between the aural vocabulary measure (A-Lex) and Speaking scores were in the same range as the relationship between the written vocabulary measure (X-Lex) and the Reading and Writing scores (r=.70-.76). Zimmerman (2004) employed the PVLT, a measure of productive vocabulary (Laufer & Nation, 1999) and found this to correlate more strongly with Speaking and Listening (r=.66 for each), than Reading and Writing (r=.60 for each).

It is also worth mentioning findings from means of lexical coverage studies (Nation, 2006), which demonstrate the relationship between vocabulary knowledge and L2 receptive skills. The percentage of words known in a text can highlight the relevance of lexical knowledge in explaining comprehension of written texts (Hu & Nation, 2000; Schmitt, Jiang & Grabe, 2011) and aural texts (van Zeeland & Schmitt, 2013). With respect to current concerns, this would suggest a strong role for vocabulary in determining a candidate's level of comprehension brought to the input materials for test tasks.

3.3 Summary and issues for consideration

While confirming a key role for both vocabulary and grammar in explaining L2 proficiency, the empirical studies described in the preceding section indicate that there is not necessarily a straightforward relationship. In part, this is due to the mixed purpose and design of the existing empirical studies. However two key issues of relevance are worth discussing further: (1) the modality of testing subskills; and (2) the level of proficiency of learner sample.

Modality of testing: If the modality of testing grammar and vocabulary is matched with that of the L2 proficiency measure, it seems more likely that it will exhibit a closer relationship than unmatched modality. It would be expected that there will be a greater amount of overlapping variance in the matched modality tests. For example, if a test-taker were to complete a written measure of vocabulary followed by a Listening test, this would doubtless explain their performance to a certain degree, however only to the extent that the written vocabulary measure reflects their ability to transfer their knowledge of the words to interpreting spoken language. The degree of transfer would likely vary between individuals, thus affecting the overall variance explained by the non-matched skill test. This distinction in modality is picked up on in vocabulary studies (Milton et al., 2010; Zimmerman, 2004), but does not always show the expected pattern. For example, Zimmerman (2004) found the reverse to be the case, in that a written vocabulary measure correlated most closely with a Speaking test. The results of this single study should not lead to modality considerations being discounted. In the componential studies, one possible reason for the discrepancy in findings reported by Joyce (2011) and Mecartty (2000) relates to the distinction in the modality of the testing approach. The relevance of this commentary to the current investigations is that the grammar and vocabulary test in the Aptis testing system is delivered as a written response to visual stimulus; therefore, a key assumption underlying the use of the results from this test to refine scores across *all* language domains is that this knowledge is, to some extent, transferable across modalities.

Role of grammar and vocabulary at different points on the ability spectrum: It is possible that differences in the level of proficiency in learner samples account for some of the discrepancies in the empirical findings described above. It has been contended that grammar and vocabulary are differentially predictive of L2 proficiency at different levels of proficiency. For example, in predicting L2 speaking ability, Higgs and Clifford (1982) in their relative contribution model indicated that vocabulary played a more important role at lower ability levels; however, empirical findings in this area have been mixed (De Jong et al., 2012; Iwashita et al., 2008). In his study of L2 reading meanwhile, Purpura (1999) indicated a comparable role for lexico-grammatical knowledge across the ability spectrum, while Shiotsu (2010) found some evidence for a stronger role amongst lower-level learners. In fact, Shiotsu suggested that incorporating a standardised L2 proficiency measure into componential studies of L2 reading would be of great value in allowing research findings to be compared; this could be equally valuable for componential studies across all skill areas. Jeon and Yamashita (2014) incorporated information about proficiency in their meta-analysis of componential studies of reading ability. However, they were unable to take more than a broad-brush approach to proficiency groupings in their meta-analysis. The current study will address this by examining the relationship between grammar and vocabulary across the ability spectrum, A1–C level.

3.4 Dimensionality

Test dimensionality is a key concern in this study, firstly in order to establish whether grammar and vocabulary as measured by the Aptis test are a unitary construct, and secondly to explore the contribution of each of the individual skills tests in contributing to the hypothesised general L2 ability factor. The dimensionality of a test reflects the extent to which subsets of questions, perhaps measuring slightly different skillsets, are distinctly measurable. For example, it is likely that a test comprising 25 English grammar items and 25 algebra items given to non-native speakers of English would exhibit multidimensionality – specifically bi-dimensionality – reflecting the two dimensions of English grammar ability and algebra ability. It is not expected that there is much of a relationship between the two areas, if any.

One way of expressing this relationship could be in terms of the correlations between the scores on the two subskills, a higher correlation would indicate more ‘unidimensionality’ while a lower correlation would indicate more ‘multidimensionality’. The qualifier “more” is used here to highlight the fact that tests are not either unidimensional *or* multidimensional; indeed, all tests are multidimensional to some extent.

This report will look at the dimensionality of the Grammar and Vocabulary component in isolation, as well as the dimensionality of all five Aptis components jointly. It is thus insightful to consider previous findings in both these areas.

3.4.1 Dimensionality in English grammar and vocabulary for L2 speakers

While the example of multidimensionality is clear in the case of a test that combines grammar and algebra items, a test containing a combination of English grammar and English vocabulary items would exhibit a much less distinct split in terms of dimensionality. In all likelihood, candidates scoring highly on one set of items would have a strong tendency to score highly on the other set, given the overlap in knowledge and skills required. Data from measures with closely related constructs, which tap into very closely related knowledge domains, very often result in item responses that are consistent with both unidimensional and multidimensional interpretations (Reise, Moore & Haviland, 2010).

Varying statistical methods and diverse judgements of the empirical evidence made by researchers can lead to different interpretations of the dimensionality (or more technically "factor structure") of construct(s). Indeed, taking examples from studies that aim to describe the components of Reading ability, the decision to incorporate grammar and vocabulary as indicators of combined or distinct factors is variable. For instance, Purpura (1999) draws on both vocabulary and grammar measures to form a single 'lexico-grammatical ability' factor, while Shiotsu and Weir (2007) and Shiotsu (2010) maintain a distinction between the two constructs in order to explore their relative contribution to Reading ability. In a study comparing the role of grammar and vocabulary in explaining L2 Reading compared to L2 Listening, Mecartty (2000) advises caution in making a complete distinction between grammatical and lexical knowledge:

> Because of the complex interplay between lexical and grammatical knowledge, it is still difficult to ascertain their overall contributions to L2 comprehension. In the acquisition of new vocabulary, learners also need to know their orthographic, phonetic, morpho-syntactic, and conceptual properties. As a result, the research instruments purported to measure lexical and grammatical knowledge constantly overlap making these knowledge sources difficult to isolate and produce clear-cut results. (Mecartty, 2000, p. 337)

Separating the constructs of grammar and vocabulary is therefore not as clear-cut a distinction as it may first appear. This notion is expanded upon by Alderson and Kremmel (2013) who explore the capacity of expert judges to distinguish whether test items assess 'syntactic' or 'lexico-semantic' knowledge. They conclude that it is very difficult to draw a definitive division between the two areas, observing that while it is convenient to separate grammar and vocabulary knowledge for diagnostic purposes, for instance, this may well be an artificial and ultimately misleading distinction to make from a construct perspective. These researchers caution that 'testers and applied linguists need to recognise the slipperiness of the slope between the constructs and need to qualify or describe their dichotomies' (Alderson & Kremmel, 2013, p. 550). In a later study, Kremmel, Brunfaut and Alderson (2017) demonstrate a strong role for a 'phraseological knowledge' factor in explaining Reading test scores that is not entirely distinct from either grammar or vocabulary, indicating a bridging role for phraseological knowledge which might render the full separation of grammar and vocabulary constructs difficult to operationalise in practice. Support for this perspective is also found in corpus analyses, which highlight the primacy of phraseology in both spoken and written language data (e.g., Römer, 2009; Römer, 2017; Sinclair, 2004).

While the discussion here primarily focuses on Reading, it is highly plausible that a similar blurring between lexis, phrase, and grammar knowledge would occur in Listening, especially in view of Joyce's insights into the role of learners' ability/inability to separate individual lexical items in the flow of speech (Joyce, 2011). Indeed, a recent study (Pan et al., 2018) has indicated that the ability to recognise multi-word expressions in Listening input makes a significant contribution to comprehension and, hence, language test performance. With respect to productive skills, a command of phraseology is integral to applying word knowledge.

Where the constructs under scrutiny are closely related in this way, investigating the underlying latent structure of a test will not lead to a straightforward interpretation. It is therefore unsurprising that several analyses of the same test of grammar and vocabulary do not necessarily give a clear answer as to whether it is valid to interpret scores on two related but distinct subskills, or on the same scale. In order to account for both the interrelated nature of the constructs of grammar and vocabulary, as well as their potential for separability, a range of factor analytic methods are employed in the current study, including the bifactor model which is specifically designed account for this issue. More details on this approach are included below.

3.4.2 Dimensionality in English subskills for L2 learners

An understanding of L2 language tests as both uni- and multi-dimensional is a generally accepted perspective within the academic language testing community. Harsch's (2014) state of play summary emphasises that 'language proficiency can be conceptualised as unitary *and* divisible, depending on the level of abstraction and the purpose of the assessment and score reporting' (Harsch, 2014, p. 153). This is based on considerable investigations into the latent structure of English language tests. In'nami and colleagues investigated the factor structure of several tests: TOEIC Reading and Listening (In'nami & Koizumi, 2012); TEAP four-skills test (In'nami, Koizumi & Nakamura, 2016) and Aptis (In'nami & Koizumi, *forthcoming*) finding evidence for separable but strongly related subskills in all cases. A higher-order factor model with skill-specific factors predicted by a general language factor was also shown to best represent L2 ability in TOEFL four skills test (Sawaki, Stricker & Oranje, 2008, 2009).

Dimensionality investigations therefore play an important role in understanding how the general construct of L2 language ability is reflected in a multi-skill language test. While a narrower test is most likely to provide the most satisfactory measure of a single dimension, in order to meaningfully test a construct as complex as L2 language ability, we need to sample from various content domains. If we had a test that comprised solely of, say, Reading items and we claimed that the score on this test represented an individual's overall L2 English language proficiency, we would be open to criticism of the "content validity" of our test, in that it was insufficient and sampled from too restrictive a range of abilities to truly call it a test of English language. Thus, in order to widen the domain content, and make the scores on our test a more accurate representation of proficiency in English, we might include items of different types, i.e., Listening, Speaking and Writing, vocabulary knowledge, grammatical knowledge, etc.

All these domains draw on highly related but conceptually distinct knowledge and skills. The wider we cast the net of what we consider to be a content domain of knowledge in our construct, the more multidimensional our measure becomes. This illustrates a fundamental tension in test construction, specifically, the less unidimensional the construct our score represents, the more difficult it is to interpret the score meaningfully, but at the same time, the more unidimensional the construct the more narrow the real world behavioural inferences we can make from the score. Thus, the question "Is this measure unidimensional?" is not the question we should ask, rather we should be asking "To what extent can I interpret the scores from these two related but distinct constructs on the same scale without loss of information on either skill?" (Reise et al., 2010).

These thoughts feed into the investigations carried out in the current study, as we explore the extent to which the four skills, plus grammar and vocabulary, relate to one another with respect to explaining overall L2 language ability.

3.5 Research questions and project outline

The focus of this technical report is twofold. The first element of the research addresses the relationship between grammar and vocabulary as operationalised and tested within the Aptis Core component. Following this, the focus broadens to incorporate all components of the Aptis test, including Listening, Reading, Speaking and Writing. The research questions are as follows.

RQ1: To what extent are grammar and vocabulary items which constitute the Core component in Aptis separable constructs?

Under this research question we examine the evidence which indicates whether grammar and vocabulary components can be considered as separable. This is an important first step in the analysis since if the grammar and vocabulary components were to be found to be separable constructs, then it would call into question the validity of using a single score from the Core component to inform the final achievement on the additional skills.

Three related but distinct analyses are used to respond to this question from various angles. Exploratory approaches using parallel analysis (Horn, 1965) and principal components on the polychoric correlation matrix provide information about whether there is a clear distinction between (and within) groups of grammar and vocabulary items, while multidimensional item response theory (MIRT) addresses the dimensionality of the Core component, with statistical comparisons between unidimensional, multidimensional correlated trait and bifactor models giving an indication of how best to account for the latent structure of the test scores.

RQ2: To what extent does the grammar and vocabulary component represent a core construct in the Aptis test?

This question provides the opportunity to explore the extent to which we might consider the Core section as central to the construct of English language ability, as measured by all components (i.e., Reading, Listening, Speaking, Writing plus the Core component). If the Core component is not "central", it calls into question the validity of using the scores on this Core component to inform the achievement on the additional components.

In order to investigate this, a correlational analysis of the raw score data matrix is followed by MIRT analysis, which provides a more sophisticated indication of the interrelationships by allowing us to scale the responses and correlate the resulting ability factors/dimensions. Finally under this research question, a Generalised Additive Model for Location, Scale and Shape (GAMLSS) is fit to show the bivariate relationships between the Core component scores and each of the other skill areas (Stasinopoulos, Rigby, Heller, Voudouris & De Bastiani, 2017). This enables us to assess the association between the scores at all points along the score continuum.

4. METHOD

4.1 Description of the data

Data, numbering 66,847 unique cases, were taken from past administrations of the Aptis General tests between April 2016 to March 2017. Each row of data contained information from the Core and any, or all, of the Listening, Reading, Speaking and Writing sections for a given candidate, depending on which combination of skills they elected to take. These scores are all recorded on a scale of 0–50. Core component scores are given on every row of the dataset, reflecting the fact that this component is compulsory for all candidates, whereas the combination of additional skills is elective and varies between candidates. Data came from 263 testing centres around the world where Aptis is offered.

The total numbers of candidates who sat each component are given in Table 1. The most popular skills were Reading, followed by Listening, then Speaking, then Writing. However, all components are very well represented in the dataset.

Table 1: Numbers of candidates responding to each component

Skill	Core	Reading	Listening	Writing	Speaking
Number	66,847	51,873	48,302	39,955	36,050

When undertaking Aptis, all candidates are required to take the Core component, they can then additionally take between one and four individual skills components. Figure 3 provides a visualisation of the number of skill components taken by each test-taker (inclusive of the Core component). The majority of test-takers took either three components or all five components. Candidates taking only a single component did not complete the test, since two is the minimum viable option.

Figure 3: Numbers of skill components undertaken by participants

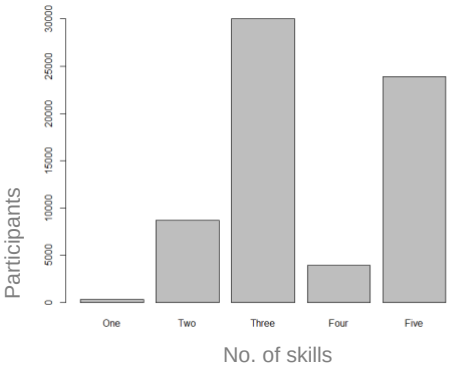


Table 2 gives the breakdown of numbers of individuals taking particular combinations of skill components. The most frequent combination is to take all five components (n=23,909), followed by taking the Core, Reading and Listening (n=18,438), followed by Core, Speaking and Writing (n=7,320). Interestingly, very few participants (n=18) opt to take a combination of Core, Listening and Writing. The candidates taking solely the Core component (n=297) can also be considered anomalous. It may be that this group were in the process of undertaking additional components when the data were extracted from the system. Since they provide valid information for the Core component analysis, they are however retained in the dataset.

Table 2: Skills breakdown for each candidate

Skills mix	Number of candidates
Core	297
Core + Reading	2,591
Core + Listening	1,493
Core + Writing	2,721
Core + Speaking	1,877
Core + Reading + Listening	18,438
Core + Reading + Writing	3,093
Core + Reading + Speaking	397
Core + Listening + Writing	18
Core + Listening + Speaking	749
Core + Writing + Speaking	7,320
Core + Reading + Listening + Writing	2,146
Core + Reading + Listening + Speaking	1,050
Core + Reading + Writing + Speaking	249
Core + Listening + Writing + Speaking	499
Core + Reading+ Listening + Writing + Speaking	23,909

For each component undertaken, candidates will have been presented with one of a number of test forms (or versions). Within each skill area, the test forms are fully equated (O'Sullivan & Dunlea, 2015) however, there are no linking items across versions to provide an anchor for this analysis. This issue was addressed in several different ways. For some purposes, test forms are analysed separately, given the scope afforded by the sample size. In other instances, such as the MIRT analysis, the assumption of random equivalence between groups is invoked, an acceptable supposition given that the test forms were randomly assigned to individuals in the same sitting (Kolen & Brennan, 2014).

Figures 4–8 show the score distributions for each component for the full dataset. Zero-inflation in the Speaking test is known to be related to technical issues and non-starters and thus zero scores were cleaned from the dataset. A small incidence of zero inflation is also observed in the Core component score distribution, however, this is not substantial enough to disrupt the analysis.

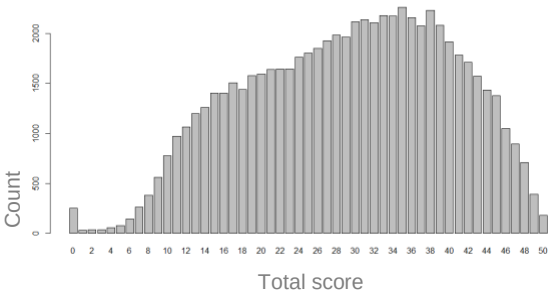


Figure 4: Core score distribution

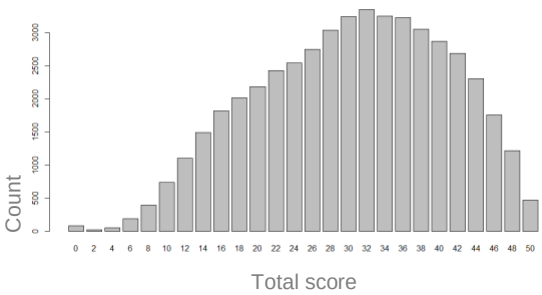


Figure 5: Listening score distribution

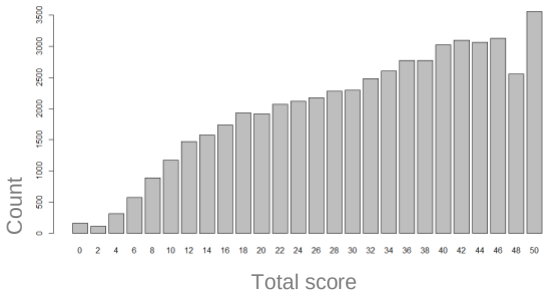


Figure 6: Reading score distribution

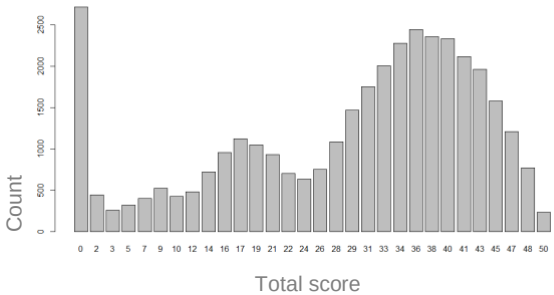


Figure 7: Speaking score distribution

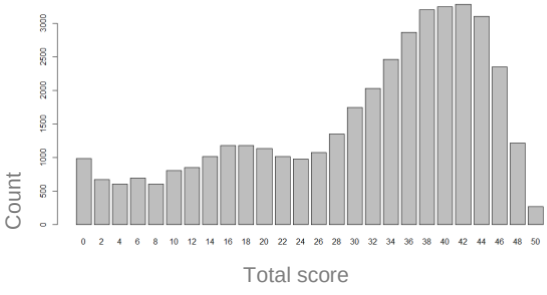


Figure 8: Writing score distribution

4.2 Missing data and other manipulations

There was a very small amount of missing data at the item level for the Core component (1.7% of the responses), and the Listening component (0.2%), recorded as '-1'. There were no missing data recorded in the other components. For the Listening data there was no discernible pattern in the missingness, and given these were missing values for test questions, a missing response can reasonably be understood to indicate a failure to meet the requirements of the test question, therefore these were recoded as incorrect. For the Core component, the missingness was found to cluster towards the end of the test, when the candidates were presented with the items/tasks focusing on vocabulary knowledge. It is likely that some candidates ran out of time. This is illustrated in Figure 9, in which the grammar scores are plotted against the vocabulary questions[2] (i.e. the two 'halves' of the Core component). There is a clear positive correlation between the two scores, as would be expected. However, along the bottom of the plot we can see a band of participants who score zero on the vocab section but score reasonably high on the grammar part of the component.

It was thought that this was due to participants not reaching the end of the test, and thus having full row missing data. This pattern does not reflect the ability of the candidates *per se* and would negatively impact the integrity of the statistical procedures aiming to explore dimensionality. After investigation, it was decided to exclude all participants with six or more missing responses from the dataset to avoid this issue. Figure 10 shows what happens when we exclude all participants with six or more missing responses. We can see the 'line' at the bottom of Figure 9 has disappeared in this plot. These are the data used moving forward in the analysis.

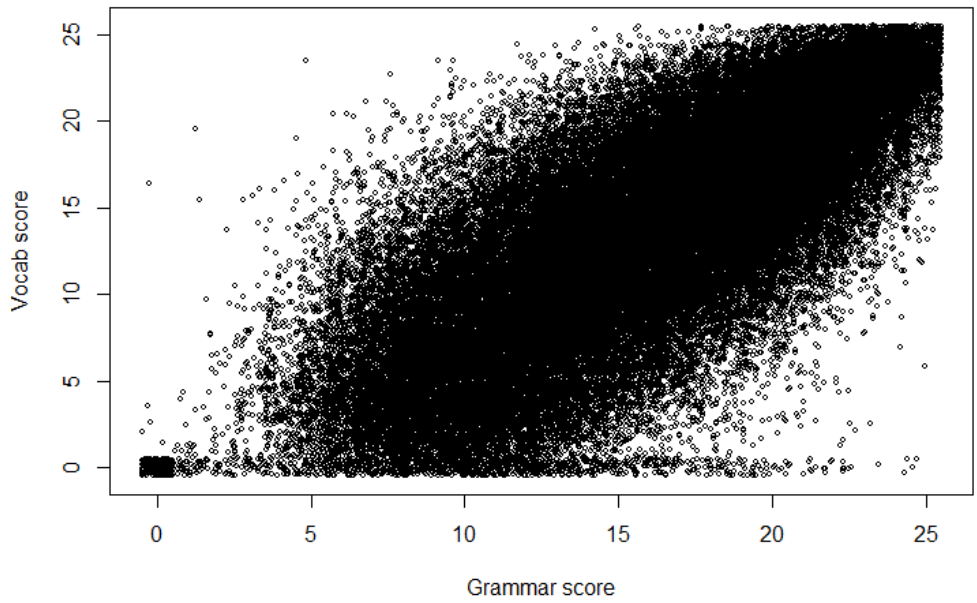


Figure 9: Scatterplot of grammar and vocabulary scores in the Core component before removal of missing item responses

[2] A *jitter*, i.e. a small amount of random movement around the location of each point, has been added to this plot to allow us to more clearly see the underlying pattern of correlation.

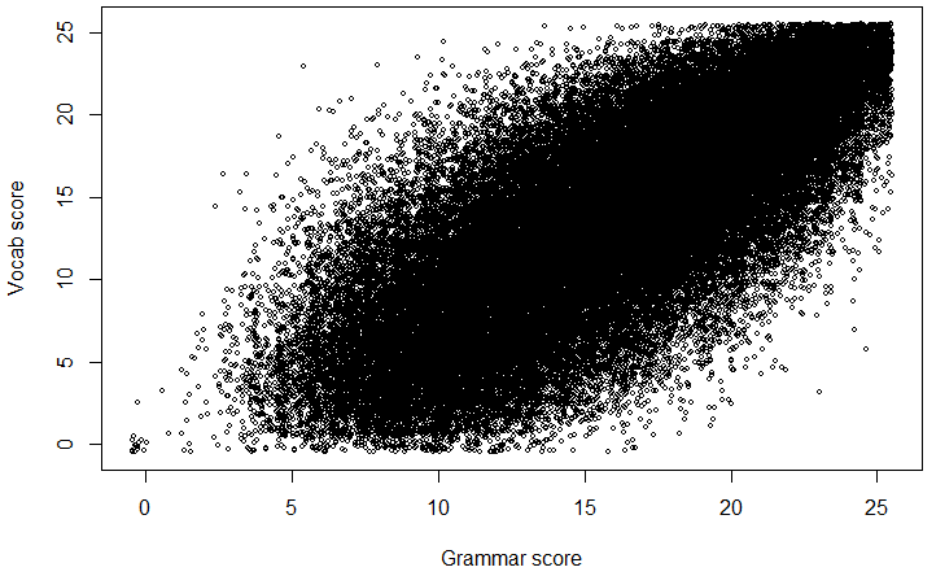


Figure 10: Scatterplot of grammar and vocabulary scores in the Core component following removal of missing data

4.3 Parallel analysis for dimensionality assessment

To answer research question 1, which addresses the dimensionality of the grammar and vocabulary component of the test, we will initially use parallel analysis (Hayton, Allen & Scarpello, 2004). Parallel analysis is an *exploratory* technique, meaning that the researcher does not pre-specify which factor the items load upon, rather, the items are free to load on any factors. This contrasts to a *confirmatory* methodology in which the researcher pre-specifies which items factors load on to. Parallel analysis adjusts the commonly used Kaiser's criterion, i.e., accepting all principal components/factors with an eigenvalue of > 1 (Kaiser, 1960) to take account of the fact that the sample used for in the analysis is finite (Hayton et al., 2004). Parallel analysis has been found to be the most accurate diagnostic tool for the assessment of dimensionality and specifying the correct number of factors under simulation (Slocum-Gori & Zumbo, 2010). This analysis was carried out using the 'psych' package (Revelle, 2017) in R (R Development Core Team, 2018).

4.4 Multidimensional IRT (MIRT) and bifactor models for dimensionality assessment

To gain insights into the dimensionality of the data, Multidimensional Item Response Theory (MIRT) models, a factor analysis type model that can be fitted to binary and ordinal testing data, were fit under the assumption of random equivalence (Kolen & Brennan, 2014). This assumption was necessary so that responses to all test forms could be included in spite of the fact that the lack of overlap prevented an adequately robust correlation matrix to be extracted as is required for traditional analysis of this nature. MIRT models can be interpreted similarly to factor analysis models, i.e., MIRT items load onto "factors" with different weights, these MIRT factors can be rotated using traditional FA methods (e.g., varimax, oblimin, etc.), and individuals have "factor scores" which represent their abilities in the modelled constructs. Note that, similar to factor analysis, MIRT is simply a method of modelling latent structures and can be either exploratory or confirmatory – in this report we are using confirmatory MIRT.

Given the size of the dataset, the complexity of the models and current computing power, it was only possible to fit the models described here to a subset of 10,000 participants – still, a very large sample size. The three models fit and compared in this report are shown in Figures 10, 11 and 12, namely, the unidimensional model, the correlated factors model and the bifactor model, respectively. This analysis was carried out using 'mirt' package (Chalmers, 2012) in R (R Development Core Team, 2018).

Figure 11: Representation of a unidimensional IRT model

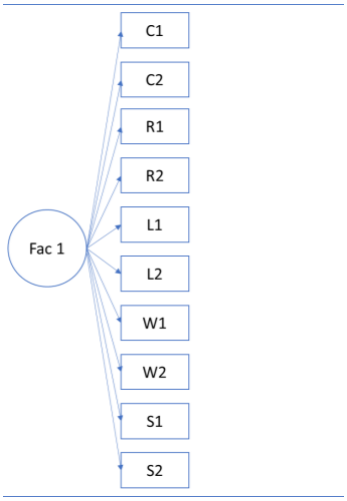


Figure 12: Representation of a multidimensional IRT model (correlated factors)

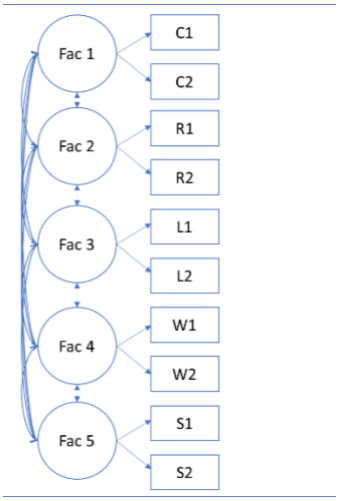
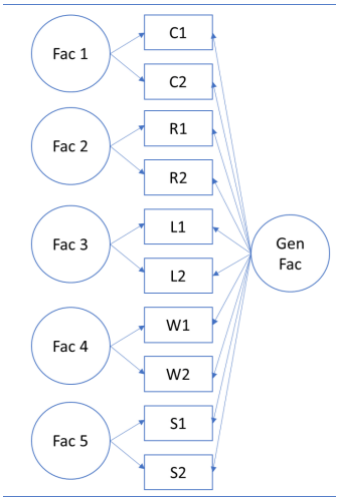


Figure 13: Representation of a bifactor IRT model



Note that a further modelling possibility not explored in this study would have been a *higher order* factor model, e.g., as per Sawaki et al.'s analysis of the TOEFL test (Sawaki et al., 2009) in which all subskill factors are hypothesised to load onto a general factor. This model is closest to that shown in Figure 12 but distinguished by having a general factor latent variable loading onto all skill area factors, as opposed to a series of correlations. This general factor captures the shared variance between the subskill factors. For the purposes of the current analysis, the correlated factor model was used over the higher order factor model because of practical complexities caused by the lack of crossover between items within a subskill (i.e. across different test forms).

It is also useful to contrast the bifactor model shown in Figure 13 with the higher order model, as the interpretation of these can be confused. The bifactor model was applied in the current modelling exercise as it is particularly suitable for investigating the plausibility of subscales (Reise et al., 2010). In the bifactor model, the common variance across all items is directly modelled by a general factor (as per the unidimensional model shown in Figure 11). The residual variance for each subskill is then modelled by a grouping factor. These subskill group factors in the bifactor model represent variance that is specifically *not* accounted for by the general factor; in other words, they represent the variance exclusive to each skill area. For a full discussion of the differences and similarities between the bifactor and the higher order model see e.g., Chen, West, and Sousa (2006), Dunn and McCray (2020).

The interpretation of the bifactor model can be illustrated with reference to modelling Aptis test scores. In Aptis we are measuring Core, Reading, Listening, Speaking and Writing, with the assumption that underlying these is a general factor representing *L2 English ability*, picked up to varying extents by each of the five components. Rather than model this as a higher-order latency, in the bifactor model, it is hypothesised that the general factor loads directly on the observed variables, and additionally each item also loads onto one, and only one, grouping factor associated with other items in the same subskill.

Importantly, in this implementation of the bifactor model, each grouping factor is hypothesised not to be correlated with any of the other skill-specific grouping factors, or indeed the general factor. In this way, the bifactor model separates the latent structure into a shared general underlying ability that influences performance across all skill areas, and skill-specific traits that impact only on performance in a single skill area. For example, in the Aptis Listening component, it may be that a distinct share of the variance in scores is related to the test-takers' ability to distinguish individual words in a flow of speech (cf. Joyce, 2011). This skill is independent of the general factor, in as much that it does not impact on the ability of the candidates to give a correct or incorrect answer for any of the other skills components but is a latency that has a direct impact on performance for all Listening items. A more detailed explanation of the bifactor model with reference to modelling Aptis data, is given in Dunn and McCray (2020).

When a confirmatory bifactor model has been fit to multidimensional data, it is possible to assess the magnitude of the loadings of the items on both general and grouping factors. A *low* loading on the general and a *high* loading on the grouping factor would indicate that a construct is providing little information on the general factor and much that is independent of the general factor – as such it might be best modelled as a subscale. Conversely, a *high* loading on the general factor and a *low* loading on the specific factor indicates that a construct is providing little information beyond the general factor and is therefore "safe" to consider unidimensional.

Reise et al. (2010) offer a cutting-edge technique for assessing the dimensionality of a test via their *comparative method* methodology. This method involves fitting a straightforward unidimensional model and a bifactor model to the same data. The loadings on the single factor of the unidimensional model can then be compared with the equivalent loadings on the general factor on the bifactor model. The magnitude of differences between the loadings of items on these two modelling procedures indicates the strength of the case for unidimensionality, as this represents the extent to which the scale is distorted by forcing multidimensional data onto a unidimensional scale. If there is little difference between the loadings on the unidimensional versus bifactor model, the data can be considered "unidimensional enough" to be reported on a single scale. In addition to this, Reise et al. (2010) recommend examining the size of the specific factor loadings. Higher grouping factor loadings are indicative that it might be useful to create subscales.

Regarding the explorations of dimensionality in test data, Reise et al. (2010) assert that "by judicious selection of fit statistics and rules-of-thumb, and by deciding whether to parcel items or allow correlated residuals, informed researchers basically can conclude whatever they wish regarding dimensionality" (Reise et al., 2010, p. 17). In other words, if you go looking for multidimensionality you will probably find it. Rather than asking the question *Is this measure unidimensional (yes/no),* Reise et al. (2010) suggest a more useful question to ask would be *If we model this data as uinidimensional can we live with the impact of the distortion of the scale on validity?* Thereby asking whether the multidimensional constructs are sufficiently correlated such that interpretations of the score are not greatly affected by this multidimensionality? Put simply: is the scale 'unidimensional enough' to serve the measurement purpose to which it is being put?

4.5 Generalised Additive Models for Location, Scale and Shape (GAMLSS)

To investigate the nature of the relationship between the Core component and the other skill areas, Generalised Additive Models for Location, Scale and Shape (GAMLSS) were employed (Stasinopoulos et al., 2017). GAMLSS are a flexible class of models which allows, among other things, to find nonlinear best fit lines and centile estimates for pairs of variables. Essentially, this sidesteps the need to specify a function, linear or otherwise, in describing a bivariate relationship. The nonlinear best fit lines in GAMLSS are determined by smoothing, i.e., using data surrounding a point to influence the estimate of the current data point. In a GAMLSS model using a normal distribution, the mean (location), variance (scale), skew (shape) and kurtosis (shape) can be simultaneously estimated across a reference variable. The advantage of GAMLSS over other techniques for our purpose is that, with very large datasets such as the one used in this paper, they allow very accurate estimation of fit lines and centiles. In this study, we are interested in using GAMLSS models to look at the nonlinear relationships between the scores in the Core component and those on the skills components.

Models were fit in the 'gamlss' package (Rigby & Stasinopoulos, 2005) in R (R Development Core Team, 2018).

5. RESULTS

The results are discussed for each research question in turn over the following two sub-sections.

5.1 Research Question 1: To what extent are grammar and vocabulary separable constructs in Aptis?

5.1.1 Parallel analysis

The findings from the parallel analysis on the item response data indicated that for each test form included in the analysis, more than one component should be extracted (NB: "component" here is used to refer to the statistically derived components, not the test section). Table 3 shows the numbers recommended are between 4 and 6, with a mean of 5.37, indicating that there is multidimensionality across all versions of the Core component.

The sections below describe analysis focused on uncovering further the patterns of multidimensionality evidenced at this exploratory stage.

Table 3: Number of components suggested for extraction by parallel analysis

Version included in analysis	#1	#2	#3	#4	#5	#6	#7	#8
Number of components	5	6	5	6	4	6	5	6

5.1.2 Core component loadings

A PCA was run on the tetrachoric correlation matrix of a single version of the Core component and *oblimin* rotation applied. This oblique rotation method allows the extracted components to correlate, resulting in a set of transformed components (TCs) (Revelle, 2017). The results of this analysis can be seen in Table 4, with estimated loadings under .35 suppressed for ease of interpretation. The picture is mixed, with no clear split in the item loadings by component[3]. If there were a clear separation between grammar and vocabulary items, we would see many grammar items loading highly on a factor and many vocabulary items loading highly on another component. The dominant component from the analysis is TC1, with 14 grammar items and 9 vocabulary items loading most heavily onto it, although this only accounts for just under half of the 50 Core items. There is, however, some small indication that the components could be related to item-type specific clusters, as TC2 is strongly oriented to vocabulary items, with 10 of these loading onto this component and only one of the grammar items. TC3 meanwhile sees a reversal of this pattern, with 6 grammar items and 2 vocabulary items loading onto this component. Unfortunately, there was not full content information on the items across the full test. However in order to investigate further the structure of the Core component, we take a closer look at the vocabulary items below.

[3] NB: "Component" in this sense refers to the name of the mathematically derived relationships, rather than the component of the Aptis test.

Table 4: Estimated component loadings from PCA of a single version of the Core component following oblimin rotation

ITEM TYPE	ITEM NUMBER	TC1	TC2	TC3	TC4	TC5
GRAM	Item 1	0.7				
GRAM	Item 2	0.8				
GRAM	Item 3			0.43		
GRAM	Item 4			0.57		
GRAM	Item 5	0.56				
GRAM	Item 6			0.54		
GRAM	Item 7			0.58		
GRAM	Item 8	0.49				
GRAM	Item 9	0.39				
GRAM	Item 10					0.55
GRAM	Item 11	0.4				0.48
GRAM	Item 12		0.46			
GRAM	Item 13					
GRAM	Item 14			0.57		
GRAM	Item 15	0.79				
GRAM	Item 16	0.89				
GRAM	Item 17	0.92				
GRAM	Item 18					-0.37
GRAM	Item 19			0.68		
GRAM	Item 20	0.56				
GRAM	Item 21	0.37				
GRAM	Item 22	0.39			-0.54	
GRAM	Item 23				-0.5	-0.45
GRAM	Item 24	0.58				
GRAM	Item 25	0.73				
VOCAB	Item 26		0.56			
VOCAB	Item 27		0.83			
VOCAB	Item 28		0.93			
VOCAB	Item 29		0.68			
VOCAB	Item 30					
VOCAB	Item 31	0.73				
VOCAB	Item 32	0.44				
VOCAB	Item 33		0.36			
VOCAB	Item 34	0.49				
VOCAB	Item 35	0.78				
VOCAB	Item 36	0.88				
VOCAB	Item 37	0.9				
VOCAB	Item 38	0.66				
VOCAB	Item 39	0.54				
VOCAB	Item 40	0.5				
VOCAB	Item 41					
VOCAB	Item 42			0.54		
VOCAB	Item 43		0.46			0.42
VOCAB	Item 44			0.49		
VOCAB	Item 45		0.51			
VOCAB	Item 46		0.62			
VOCAB	Item 47					
VOCAB	Item 48					
VOCAB	Item 49		0.67			
VOCAB	Item 50	0.41	0.45			

5.1.3 Examination of vocabulary item type and factor loadings

The vocabulary section of the Core component consists of four item types: *Synonyms (S), Understanding meaning form context (U)*, *Collocation (C)* and *Definitions (D)*. It was hypothesised that these different item types may play some part in the multidimensional structure of the Core component, and thus be a source of non-construct related variance. In other words, previous experience or specific ability to respond to particular types of items may be a driver of the multidimensionality.

To assess this source of non-construct-related variance, the vocabulary items (n=25) from a single version of Core component were subject to PCA in relation to item type[4]. In total, the number of test-takers analysed was 8,326. A parallel analysis was run on the tetrachoric correlation matrix of the items which suggested the extraction of three components. Table 5 provides the results of the extraction of three components, again a component loading cut-off of 0.35 is used to aid interpretation.

Interpreting the component loadings, there is little evidence that a major cause of the multidimensionality in the vocabulary items is due to item type. The latent structure captured by the components appear to reflect the progressive difficulty of the items, with the easier items loading on the first component and harder items on the second, with a transition point at the B1 level. Indeed, the Synonym items at B1 have moderate-to-high loadings across both TC1 and TC2, while the Synonym items at A1 level are firmly related to TC1. In summary, there is no clear evidence of different subskills driving multidimensionality that can be found in the data.

Table 5: Estimated component loadings from PCA of vocabulary items from a single version of the Core component

Task Type	CEFR Level	TC1	TC2	TC3
S	A1	0.88		
S	A1	0.88		
S	A1	0.37		0.38
S	A1	0.67		
S	A1			0.82
U	A2	0.86		
U	A2	0.77		
U	A2	0.85		
U	A2		0.51	
U	A2	0.88		
D	A2	0.85		
D	A2	0.84		
D	A2	0.85		
D	A2	0.86		
D	A2	0.9		
S	B1	0.35	0.41	
S	B1	0.49	0.35	
S	B1	0.8		
S	B1		0.42	
S	B1		0.53	
C	B2		0.56	
C	B2	0.62		
C	B2		0.86	
C	B2		0.4	0.4
C	B2		0.71	

[4] A different version than that analysed in Section 5.1.2 above.

5.1.4 Confirmatory MIRT analysis of grammar versus vocabulary items

Given the fact that it appears the major sources of dimensionality in the Core component are not due to a separation between grammar and vocabulary items, it was decided to fit confirmatory MIRT models to force the assessment of this kind of dimensionality.

Below are the results of the specification of three models (please refer back to Figures 11 to 13 in Section 4.4. for graphic illustration of these models):

a) a unidimensional model which assumes that the vocabulary and grammar items load onto a single factor
b) a two-correlated-factor model, in which vocabulary items load on one factor and grammar items on another
c) a bifactor model which has a general factor, one vocabulary-specific factor and one grammar-specific factor, all of which are uncorrelated with one another.

With respect to model comparisons, given the models are not all nested, comparative fit statistics AIC (Akaike, 1974), BIC (Schwartz, 1978), AICc (Hurvich & Tsai, 1989) (Hurvich & Tsai, 1989) (sample size corrected AIC) and SABIC (Enders & Tofighi, 2008) (sample size corrected BIC) were used to assess relative model fit. Table 6 provides these statistics – the interpretation of these is that the lower the value of the statistic, the more parsimonious a fit the model is to the data. The comparative fit statistics all indicate that the unidimensional model is the poorest fit to the model, the two-correlated-factor model is a slightly better fit, and the bifactor model is a distinctly better fit than both.

Table 6: Model fit statistics for the three MIRT models fit to grammar and vocabulary data

	Unidimensional model	Two correlated factor model	Bifactor model
No. parameters	400	801	1200
LL	-1588868	-1584599	-1571181
AIC	3179336	3170800	3144761
BIC	3186593	3178065	3155646
AICC	3179356	3170820	3144807
SABIC	3184050	3175520	3151833

Having established the bifactor model provides the closest reflection of the data, we now need to ascertain whether splitting grammar and vocabulary items would bring a practically significant improvement in score reporting. According to Reise et al.'s comparative method (Reise, Cook & Moore, 2015), we can assess the impact of reporting what is ostensibly a multidimensional scale on a unidimensional scale by: a) assessing the difference in the discrimination parameters (loadings) between the unidimensional model and the general factor of the confirmatory bifactor model; and b) additionally assessing the factor loadings on the orthogonal non-general factors.

To this end, Table 7 shows, by domain, the means and SDs for: 1) differences in loading between the unidimensional and bifactor general factor loadings; 2) the loadings on the general factor; and 3) the loadings on the grouping factor associated with each skill.

Table 7: Mean loadings for the bifactor model of core component

	Grammar	Vocabulary
General factor	1.15 (0.54)	1.81 (0.81)
Grouping factor	0.32 (0.42)	0.47 (0.57)
Absolute difference between general factor of bifactor model and single factor of unidimensional model	0.05 (0.07)	0.09 (0.11)

An interpretation of the summary estimates shown in Table 7 is that while there are some items which could benefit from a multidimensional model (evidenced by non-zero mean loadings on skill-specific grouping factors), overall the unidimensional model performs very similarly to the bifactor model. This is demonstrated by the low difference in loadings on the general factor when compared with the unidimensional model. It can be concluded therefore that the unidimensional model does not distort the overall construct of *grammar and vocabulary ability* to a great degree[5]. In other words, the construct is *unidimensional enough* to be reported on a single scale, and splitting the construct will provide negligible improvement in measurement**.**

5.1.5 Joint interpretation of analyses responding to RQ1

In summary, some evidence for multidimensionality within the Core component item response matrix has been discerned (in the parallel analysis and exploratory PCA). However, the distinction between dimensions does not appear to rest on items focusing on grammar versus vocabulary (as shown by the loadings for the rotated PCA solution, and the MIRT analysis). In fact, the strongest indication of a pattern in the dimensionality relates to the item difficulty. Unfortunately, full details of item content were not available for the whole dataset beyond top-line identification of grammar and vocabulary items. However, analysis on a subset of vocabulary items from a single version indicated this to be the case. This is strongly recommended as an area for further investigation using a dataset with more comprehensive item-level details.

With respect to the findings of the MIRT analysis, a statistically significant better fit for a two-correlated-factor model over the unidimensional model was noted, and considerably better fit for the bifactor model over both of these. This finding indicates that there is skill-specific variance that may need to be taken into consideration. However, a closer examination of the incremental information that the bifactor model would bring implies that the combination of grammar and vocabulary items is "unidimensional enough" (Reise et al., 2015) to be reported as a single score.

5.2 Research Question 2: To what extent does the grammar and vocabulary component represent a core construct in the Aptis test?

5.2.1 Correlation of raw scores

A simple method of assessing the extent to which the Core component is central to Aptis is by the assessment of the raw score correlations between each component. Table 8 provides the Pearson correlation matrix for the raw scores for each component (using pairwise complete observations, i.e., all available data) and a mean value of correlation. It can be seen from the table that, on average, the Core component correlates highest with the others (r = 0.77). The lowest average correlation with other components is for the Listening component (r = 0.73). All other components correlate similarly with the other components (r = 0.75). The lowest correlation in raw scores is between Listening and Writing (r = 0.69), while the highest is between Core and Writing (r = 0.79). Given the very large

[5] As a point of comparison, Reise et al. (2014) present an example of a measure they do not consider to be appropriate to report on a multidimensional scale whose mean difference between discrimination parameters between the one factor and bifactor model general factor is 0.11 logits.

sample size, this table provides evidence that the Core component is the best predictor component of the other skills, however, the difference in mean correlations with other components is not large.

Table 8: Raw score correlations between components

	Core	Reading	Listening	Writing	Speaking	Overall	Mean
CORE	1					**Core**	**0.77**
READING	0.79***	1				**Reading**	**0.75**
LISTENING	0.76***	0.73***	1			**Listening**	**0.73**
WRITING	0.77***	0.76***	0.69***	1		**Writing**	**0.74**
SPEAKING	0.75***	0.71***	0.74***	0.75***	1	**Speaking**	**0.74**

*** indicates that $p < 0.001$

5.2.2 Confirmatory MIRT analysis of four skill components and the Core component

A more nuanced method for assessing the centrality of the Core component is to use Multidimensional IRT. MIRT allows us to better correlate the underlying trait ability (i.e. Core, Reading, Writing, etc.) than does simply correlating the overall scores. That said, both methods should tell approximately the same story.

The three models described in Section 4.4 above were fit to a subset of 10,000 participants who had completed all five components. This decision to limit the analysis to the participants with a full set of response data stems from the need to maximise the available information in the dataset, plus to allow this large model to fit some of the response categories for which there were low responses to be collapsed. An algorithm was written to allow categories to be collapsed until there either there was no category with fewer than 20% of the total number of responses, or the variable was binary.

Table 9 provides the fit statistics for the three models fitted. On all measures, the bifactor model is the best fit, followed by the correlated factor model, followed by the unidimensional model.

Table 9: Model fit statistics for the three MIRT models fit to all components

	Unidimensional model	Five-correlated-factor model	Bifactor model
No. parameters	1771	1781	2564
LL	-451175	-449941	-444079
AIC	905892	903444	893286
BIC	918662	916286	911773
AICC	906655	904216	895055
SABIC	913034	910626	903625

Table 10 shows the correlations between factors in the five-factor model. Note that the Core component still has the highest correlation with the other skill factors, although, Reading is now close behind.

Table 10: Correlations between factors in five-factor model

	Core	Reading	Listening	Writing	Speaking	Overall	Mean
CORE	1					**Core**	0.842
READING	.866	1				**Reading**	0.838
LISTENING	.830	.824	1			**Listening**	0.817
WRITING	.875	.866	.789	1		**Writing**	0.837
SPEAKING	.797	.799	.826	.816	1	**Speaking**	0.809

Table 11 meanwhile shows: 1) the mean values and standard deviations for the loading on the general factor; and 2) the mean loading for the grouping factors in the bifactor model of all skill components.

Table 11: Mean loadings for the bifactor model on full dataset

	Core	Reading	Listening	Writing	Speaking
General factor	1.37 (0.67)	2.16 (0.61)	1.27 (0.59)	2.02 (0.67)	2.40 (0.56)
Grouping factor	0.15 (0.55)	0.62 (0.74)	0.63 (0.39)	0.79 (0.33)	1.29 (0.52)

The loading on the general factor is related to how well the items in the subskills correlate, on average, with the general construct, i.e. *L2 English ability*. These loadings are related to some extent to the number of response categories in the observed variables, so items with more response categories tend to have a higher loading; this explains the higher values for Reading, Writing and Speaking, all of which have polytomous response scales, while Core and Listening item responses are binary. The loading on the grouping factors on the other hand represent the unique information that is measured by a specific subskill, distinct from the general factor (Dunn & McCray, 2020).

A key point of interest to note in this analysis is the comparison between loading values for the grouping factors. The mean loading for the grouping factor associated with the Core component is 0.15. This indicates that the Core component carries little additional information over the general factor. Compare this with the much higher mean loadings on the grouping factors for the other language domains, which are shown to be carrying much more information than is measured by the general factor. In other words, this can be cited as evidence to show that the grammar and vocabulary component actually is “core” to the construct of L2 English as measured by Aptis.

5.2.3 GAMLSS analysis of Core component scores against other components

The analysis described above provides point estimates to indicate the strength of relationships between the various skill components of Aptis with the Core component. In view of moving to consider the practical application of using the score information from the Core component to adjust the CEFR level boundaries, we had an interest in assessing the predictive value of the Core component score at all points along the score continuum. GAMLSS analysis allows us to address the question: *Is the association between score on the Core component and score on the skills components stronger at higher scores?* If, for example, the scores on the Core and skills components are more closely associated at higher overall scores, then it may make sense to only adjust CEFR level decisions for the higher scoring candidates.

On the x-axes of plots Figures 14–18, we have individuals' scores on the skills components while on the y-axes we have their score on the core component. The solid black line, in the plots, is the 50th centile line (i.e., a line of best fit derived from the GAMLSS), and the red and blue dashed lines are 25th/75th and 10th/90th centile lines, respectively. As an example, examining Figure 14, a test-taker scoring 20 on the Core score component would have: i) an expected score of ~21 on the Listening component (i.e. the average score for those who score 20 on the core); ii) 50% chance of scoring between ~18 and ~27 on the Listening component; and iii) 90% chance of scoring between ~14 and ~31 on the Listening component. Given our Core score, we can therefore predict the likely score range of a test-taker on a subskill. When the confidence band is narrower, we can better predict the subskill score from the Core score. When the gradient of the 50% quantile (black) line is steeper, it means that the association between changes in scores on the core and subskills is stronger.

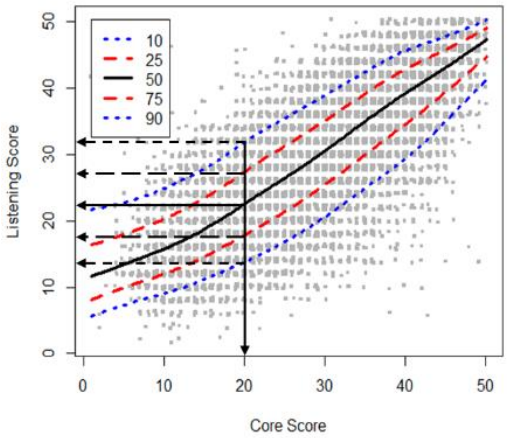


Figure 14: Example GAMLSS

Looking to Figures 15–18, we can see that there is a clear positive relationship between Core scores and subskill scores, meaning that knowing the Core score of an individual would allow one to predict, with some accuracy, their score on any of the subskills. Figures 15, 17 and 18, and to a lesser extent 14, show an 's' shaped 50% quantile line. At the top end of the plots, this 's' shape and accompanying narrowing of the upper 10 (blue dashed) and 25% (red dashed) quantile lines, is simply due to a ceiling effect. At the bottom end, the lessening of the association between Core and subskills is due to floor effects (clear in Figures 17 and 18) and perhaps to the fact that at scores less than 10, candidates are scoring worse than chance. In order to attempt to deal with these floor and ceiling effects as well as possible, unidimensional IRT models were fit and the ability estimates used to map the relationships between Core and subskills.

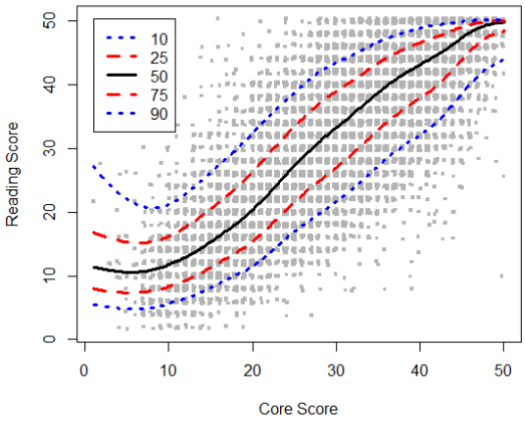


Figure 15: Raw Core vs Reading

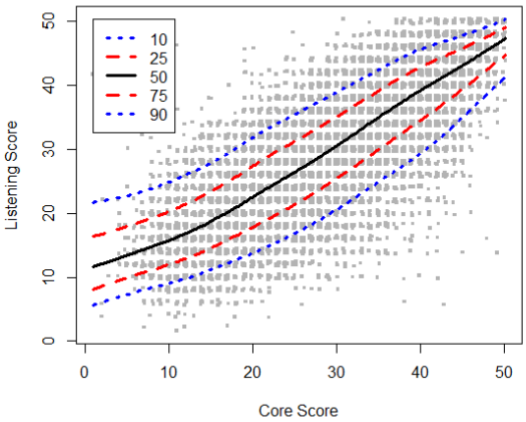


Figure 16: Raw Core vs Listening

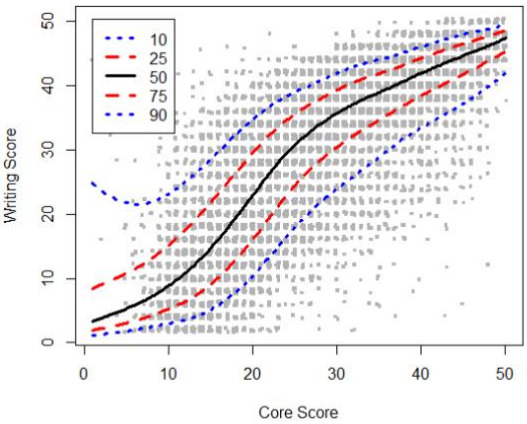


Figure 17: Raw Core vs Writing

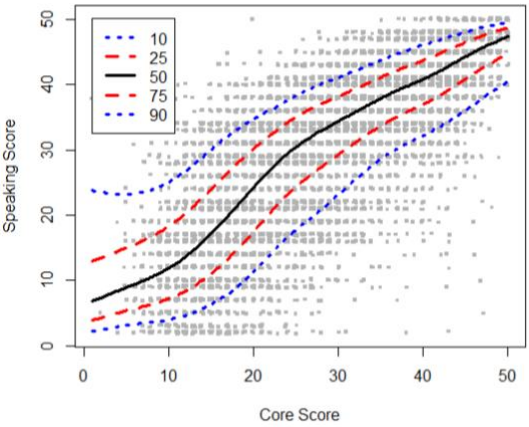


Figure 18: Raw Core Vs Speaking

Figures 19–22 represent the GAMLSS model fitted to the IRT scaled scores on the test. For components with dichotomously scored items (Core and Listening), the model employed was the three-parameter logistic model (Birnbaum, 1968). For those with polytomously scored items (Reading, Speaking, Writing), the graded response model was used (Samejima, 1969). The rationale for this is that the IRT models allow us to non-linearly scale the raw scores to an interval scale to better reflect changes in underlying ability.

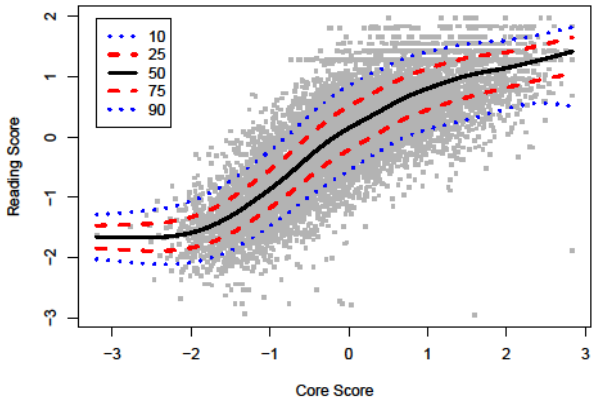


Figure 19: IRT Core vs Reading

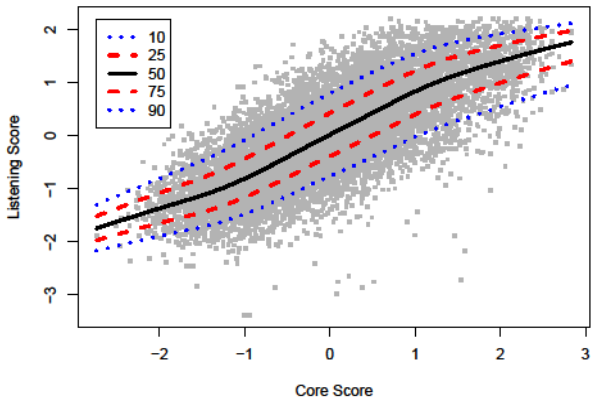


Figure 20: IRT Core vs Listening

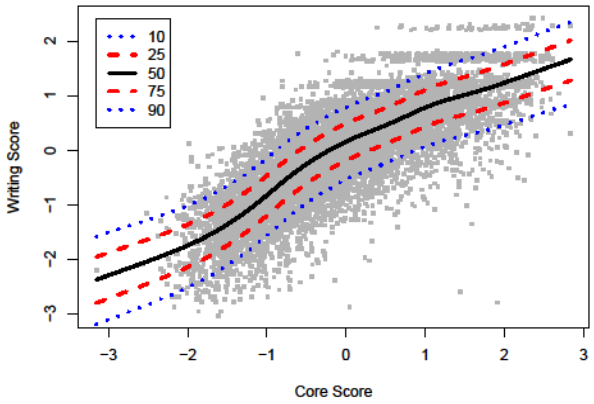


Figure 21: IRT Core vs Writing

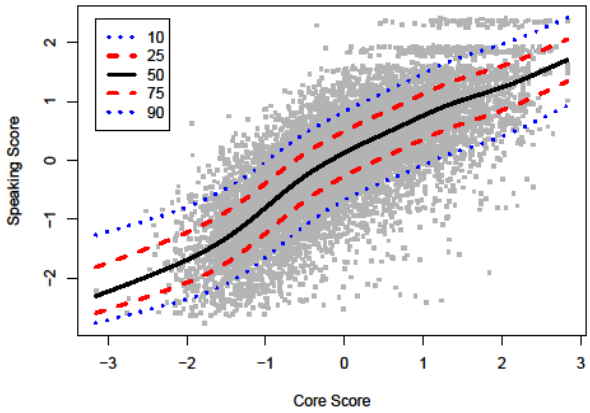


Figure 22: IRT Core vs Speaking

Interpreting Figures 19–22, we can immediately see that below -2 logits on the Core component there is very little data, and what there is bears little relationship to the skill scores; this is particularly marked for Reading (Figure 19). Looking to the estimated curves after this -2 mark, it is apparent that, overall, the relationships between Core and each of the skills has been smoothed somewhat compared to the models using the raw scores, and there is an approximately linear relationship in each case.

The most consistent pattern of association can be seen for Listening (Figure 20), with quantile lines being roughly parallel. At the top right of the plot, however, the 10th quantile line curves inwards, this is evidence of a slight ceiling effect, a situation which is echoed to a more severe degree in Reading, where the best fit line itself starts to flatten off slightly at the higher end of the ability spectrum. This can be understood as a ceiling effect in the Reading test in light of the truncated distribution in the Reading scores observed in the raw data (Figure 6, Section 4.1).

With respect to the relationship between Core and each of the skill areas, it can be seen that there is a discernible positive relationship in each domain across the ability spectrum (as shown by the upward slope from left to right in each of the figures). The GAMLSS modelling approach is useful in that the best fit line does not assume linearity (or other function) and can therefore be useful in showing changes in the steepness of the slope along the axes.

In the case of Speaking, Writing and Reading there is a slightly steeper slope below 0 logits, indicating that the relationship is in fact *strongest for the candidates that are performing lower than average*. The slight "s" shape apparent in these curves indicates that the association between these skills and Core score is slightly weaker at the lower and higher ends of the core ability distribution and slightly stronger in the middle. This is perhaps attributable to features of the test scoring mechanism.

For Listening, there is more of a consistent relationship across the spectrum. One point to note is that Listening is the only test component to consist of dichotomously scored items (as is the Core component), the other skills components are scored polytomously, which might influence the IRT ability estimates. Additionally, it is clear from the slightly fractured pattern in the scatter plots for Reading, Writing and Speaking at the top right-hand side of the plots that the grain of information about ability is not so fine as in Core or Listening. It is therefore recommended that the exact shape of these slopes is not over-accentuated in the interpretation. It can be concluded that there is not strong evidence of significant departures from an approximately linear relationship between the trait measured by core and that measured by the skill area components.

5.2.4 Joint interpretation of studies responding to RQ2

Overall, we have some evidence that, from a statistical perspective, performance on the grammar and vocabulary items in the Core component can viewed as reflecting a fundamental element of the candidates' L2 English ability that is applicable across the four skill areas.

This is demonstrated in the strong positive correlations observed between the Core component and skill area scores (both for raw score correlations and correlations between the scaled factors using the five-factor MIRT model). This analysis indicated that the relationship between the Core component and each of the skill areas was stronger than relationships between individual skill areas. While these correlational approaches assume linearity in the relationship, the GAMLSS modelling exercise enabled us to show this assumption largely holds, with a similar relationship between IRT calibrated ability estimates across the spectrum for the Core component and each of the individual L2 skills. This was, however, less distinct at both extremes of the score scale. With respect to the higher end of the scale, ceiling effects in some of the components were observed. The overall picture could be strengthened by conducting a complementary study using data from the Aptis Advanced variant, which challenges candidates to a greater extent at the higher end of the ability spectrum.

The bifactor models incorporating observed score data from all Aptis components clearly indicated that the hypothesised *L2 English ability* factor accounted well for the variance in the Core component responses as well as those of each L2 skill. However, while the Core component had virtually no loading on a grouping factor, each of the L2 skills respectively demonstrated considerably higher grouping factor loadings. This indicates that: a) the L2 English ability factor is highly relevant in predicting item responses across the full test; and b) the grammar and vocabulary questions only draw minimally on an additional skill-specific grouping factor, in comparison to the other language domains which rely more heavily on commonalities not captured by the general factor. This pattern of relationships provides evidence to consider the Core component as "core" in the test.

As is consistent with the aims of the study, therefore, we have accrued various pieces of statistical evidence to support the theoretical basis on which the Aptis test was structured at development phase. This study is valuable because of the size of the sample population, and the huge range of abilities and English language learning backgrounds represented in the dataset. While it is ultimately a judgement call as to whether the Core component is central to the collection of all the items across each language domain, we have certainly found evidence to support the theoretical assumption, and nothing to refute it outright.

6. DISCUSSION

This study addresses two research questions seeking to establish the extent to which empirical evidence supports the scoring approach taken in the Aptis test, whereby the scores from the Core component feed in to refine CEFR level decisions across the four skill areas. In the first investigation of its kind using Aptis data, answering these questions involved exploring patterns in the score data gathered from a large global sample of Aptis test candidates using sophisticated statistical tools. The study provides empirical support for the cognitive processing models of L2 language proficiency upon which the Aptis test was founded, i.e., those which posit grammar and vocabulary in a core role across all language domains (Field, 2013; Khalifa & Weir, 2009). While this has been shown empirically in other contexts, the usefulness of the current study lies in its use of Aptis data, the large scale of the participant sample, and the wide range of language aptitudes incorporated. This study was able to map the relationship between grammar and vocabulary, and each of the skills of Listening, Reading, Writing and Speaking for abilities spanning CEFR levels A0 to C[6].

With respect to the design of the Aptis test, the study findings indicate that there is no strong evidence to contradict the current approach of reporting and using the Core component scores on a single scale, i.e., amalgamating scores from grammar and vocabulary items. Additionally, the results of the investigation gave support for using the Core component scores to feed into the CEFR level allocations in the other skill areas. Although one caveat highlighted in the introduction to the analysis bears repeating here, which is that investigations into the dimensionality of tests and testing systems rarely point to an unequivocal answer. The interpretation of dimensionality analysis will always be linked to the context of measurement. With this in mind, the key findings under each research question are summarised below.

Firstly, under RQ1 there was an investigation into the dimensionality of the Core component with a view to understanding whether the constituent grammar and vocabulary items were *unidimensional enough* to treat scores on a single scale. While there was some evidence of multidimensionality detected in the Core component scores, there was no compelling indication that the source of this structure lying in the content of the items, i.e., grammar versus vocabulary items. Interpretation of these findings was given in the light of the Reise et al's (2014) observation that all tests will exhibit signs of multidimensionality to some extent, it is only problematic if the extent of this multidimensionality is degrading to the overall scale. Analysis comparing a unidimensional and a bifactor model gave no indication that the grammar and vocabulary elements of the Core component test exhibit an excessive degree of "skill-specific" variance to merit their interpretation on separate scales.

Investigations under RQ2 involved statistical modelling of response data for all components using a multidimensional IRT approach in order to understand the role of grammar and vocabulary in explaining L2 language ability in the context of the other skills. This provided information about the contribution of each skill area to an underlying trait hypothesised to represent general L2 language ability. It was found that the general L2 language ability factor (in the bifactor MIRT model) accounted well for the variance in Core component scores, with virtually no loading on its specific factor. This was interpreted as evidence that, from a statistical perspective, performance on the grammar and vocabulary items in the Core component reflect a fundamental element of the candidates' L2 English ability that is applicable across the four skill areas. The GAMLSS modelling exercise enabled us to show that the assumption of a linear relationship between Core and L2 skill areas largely holds, with a relatively constant relationship between IRT calibrated ability estimates across the ability spectrum for the Core component and each of the individual L2 skills.

[6] Aptis General does not discriminate between C1 and C2 level abilities; a further study using data from the Aptis Advanced would provide the opportunity.

In addition to the findings directly related to the design and functioning of the Aptis test, the findings here also highlight some interesting points from a second language acquisition perspective, and thus contribute to the wider literature in this area. While there is evidence for a strong positive association between grammar and vocabulary knowledge and the different language domains across the ability spectrum, the findings in this study tally with studies which indicate that the *very closest* relationship exists at the slightly lower levels of ability (e.g., in L2 speaking De Jong et al. (2012); in L2 reading Shiotsu (2010)), and may explain Jeon and Yamashita's failure to reject their null hypothesis of a stronger effect at higher levels of proficiency in reading (Jeon & Yamashita, 2014, p. 195). Additionally, the insights accorded by this analysis regarding the nature of the relationship between the Core component and the skill-specific components can be used in future to refine and develop the use of Core scores in the determination of the final scores for each skill area. This however is beyond the scope of the current report, for which it was sufficient to illustrate the relationship in general, since the aim was to investigate whether the Core component could be viewed as playing a genuinely "core" role in the test. However, in future it would be valuable to carry out further detailed investigations. This could, for example, involve looking further into the variation at different ability levels, or perhaps utilising data from Aptis Advanced to give a more complete picture of the relationship across the L2 English ability spectrum.

Another more general interest finding to arise from this study relates to the modality of the testing. The analysis presented here shows that the modality of testing grammar and vocabulary does, to some extent, bear a relationship with the statistical explanation of skill-area scores. The grammar and vocabulary questions in the Core component are presented in a written multiple choice, and while there is a strong positive correlation between test performances and all the L2 skill areas, the relationships are weaker for the non-matched skills. This pattern is also replicated in the MIRT analysis. In a five-correlated factor model, the highest correlations with the Core component loadings are with the loadings on the Writing and Reading components. The bifactor model meanwhile shows the Reading component to have the lowest loading on its specific factor compared to the other skill areas, second only to the Core component. This indicates that the ability underpinning the responses to the Reading component and Core component are most similar, i.e., rely less on additional abilities, compared to the other skill areas. It would be of interest to see if the balance changed if the stimulus for the core component were delivered aurally (thus giving it more in common with the Listening test)[7]. Again, this is beyond the scope of the current study, however it provides potential inspiration for further research and potential test development options.

In summary, the results of this in-depth study provided satisfactory detail to answer RQs 1 and 2 in a manner affirmative to the structure of the Aptis testing system. The analysis presented also provides some additional insights that could form the basis of future research topics, and aid the ongoing refinement and development of the Aptis testing system.

[7] Note: This would not be possible using data from the current delivery model.

7. LIMITATIONS

This study involved modelling huge disjointed data sets, made possible in many cases using the assumption of random equivalence. While the randomised allocation of test versions across test centres globally means that we are confident that this assumption holds, it has stretched the statistical estimation demands during the analysis. This meant that the MIRT analysis was limited to a sub-sample of 10,000 candidates, and that there were also some limitations as to which MIRT models it was possible to employ. For example, the higher-order factor model was not estimable with the concurrent data structure. This was not crucial to the research questions being asked in the current study, however, the widespread use of this model in other studies (e.g., In'nami & Koizumi, 2012; In'nami et al., 2016; Sawaki et al., 2008, 2009) means that it would have been useful to employ this model for the purposes of direct comparison.

A further limitation of the study is that it only employed data from the Aptis General test, which is designed to provide information about the English language ability of candidates ranging between A0–C on the CEFR scale. While it was a strength of this study to have inbuilt information from a standardised test about the ability level of the participants, it would have been of further value to be able to discriminate candidates at the very high end of the ability spectrum. This would have been particularly relevant in gaining insights into the relationship between skill areas and grammar and vocabulary for the more able L2 users. This is highlighted in the report as a possible area for further investigation, as it would be a possibility to use data from the Aptis Advanced variant for this purpose.

A final limitation of the study worth mentioning was the lack of available information on item content. This would have been especially useful in exploring the dimensionality of the Core component, as we could have more fully investigated the origin of the multidimensionality detected.

REFERENCES

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control, 19*, 716–723.

Alderson, J. C. (1993). The Relationship between Grammar and Reading in English for Academic Purposes Test Battery. In D. Douglas & C. Chappelle (Eds.), *A New Decade of Language Testing Research: Selected Papers from the 1990 Language Testing Research Colloquium*. Alexandria, VA: TESOL.

Alderson, J. C., & Kremmel, B. (2013). Re-examining the content validation of a grammar test: The (im)possibility of distinguishing vocabulary and structural knowledge. *Language Testing, 30*(4), 535-556. Retrieved from: http://journals.sagepub.com/doi/10.1177/0265532213489568. doi:10.1177/0265532213489568

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. Lord & M. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Brunfaut, T., & McCray, G. (2015.) Looking into test-takers' cognitive processes while completing reading tasks: A mixed-method eye-tracking and stimulated recall study. *ARAGS Research Reports Online, No. 1*. London: British Council. (Retrieved from: https://www.britishcouncil.org/sites/default/files/brunfaut_and_mccray_report_final_0.pdf)

Chalmers, P. R. (2012). MIRT: A Multidimensional Item Response Theory Package for the Environment. *Journal of Statistical Software, 48*, 1/29.

Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality-of-life. *Multivariate Behavioral Research, 41*, 189–225.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press.

De Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition, 34*(1), 5-34. Retrieved from https://www.cambridge.org/core/product/identifier/S0272263111000489/type/journal_article. doi:10.1017/S0272263111000489

Dunn, K. J., & McCray, G. (2020). The place of the bifactor model in confirmatory factor analysis investigations into construct dimensionality in language testing. doi: 10.3389/fpsyg.2020.01357

Enders, C. K., & Tofighi, D. (2008). The impact of misspecifying class-specific residual variances in growth mixture models. *Structural Equation Modeling: A Multidisciplinary Journal, 15*, 75–95. doi:https://doi.org/10.1080/10705510701758281

Field, J. (2013). Cognitive validity. In L. T. A. Geranpayeh (Ed.), *Examining listening* (pp. 77–151). Cambridge: Cambridge University Press.

Field, J. (2018). *Rethinking the second language listening test: From theory to practice*. Bristol, UK: Equinox Publishing.

Grabe, W. (2009). *Reading in a second language: moving from theory to practice*. Cambridge: Cambridge University Press.

Grabe, W., & Kaplan, R. B. (1996). *Theory and practice of writing: an applied linguistic perspective*. London: Longman.

Harsch, C. (2014). General language proficiency revisited: Current and future issues. *Language Assessment Quarterly, 11*(2), 152-169. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/15434303.2014.902059. doi:10.1080/15434303.2014.902059

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: a tutorial on parallel analysis. *Organizational Research Methods, 7*, 191–205.

Higgs, T. V., & Clifford, R. (1982). The push toward communication. In T. V. Higgs (Ed.), *Curriculum, competence and the foreign language teacher* (pp. 243–265). Skokie, IL: National Textbook Company.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*, 179–185.

Hu, M., & Nation, P. (2000). *Unknown Vocabulary Density and Reading Comprehension* (Vol. 13).

Hurvich, C. M., & Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika, 76*, 297–307.

In'nami, Y., & Koizumi, R. (2012). Factor structure of the revised TOEIC® test: A multiple-sample analysis. *Language Testing, 29*(1), 131–152. Retrieved from http://journals.sagepub.com/doi/10.1177/0265532211413444. doi:10.1177/0265532211413444

In'nami, Y., Koizumi, R., & Nakamura, K. (2016). Factor structure of the Test of English for Academic Purposes (TEAP®) test in relation to the TOEFL iBT® test. *Language Testing in Asia, 6*(1). Retrieved from http://www.languagetestingasia.com/content/6/1/3. doi:10.1186/s40468-016-0025-9

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics, 29*(1), 24–49. Retrieved from https://academic.oup.com/applij/article-lookup/doi/10.1093/applin/amm017. doi:10.1093/applin/amm017

Jeon, E. H., & Yamashita, J. (2014). L2 reading comprehension and its correlates: A meta-analysis. *Language Learning, 64*(1), 160–212. Retrieved from http://doi.wiley.com/10.1111/lang.12034. doi:10.1111/lang.12034

Joyce, P. (2011). Componentiality in L2 listening. In B. O'Sullivan (Ed.), *Language testing: Theories and practices*. Basingstoke: Palgrave Macmillan.

Jung, J. (2009). Second Language Reading and the Role of Grammar. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics, 9*.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*(20), 141–151.

Khalifa, H., & Weir, C. J. (2009). *Examining Reading: Research and practice in assessing second language reading*. Cambridge: Cambridge University Press.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices*. New York: Springer-Verlag.

Kremmel, B., Brunfaut, T., & Alderson, J. C. (2017). Exploring the role of phraseological knowledge in foreign language reading. *Applied Linguistics, 38*(6), 848–870. Retrieved from https://academic.oup.com/applij/article-lookup/doi/10.1093/applin/amv070. doi:10.1093/applin/amv070

Larsen-Freeman, D. (2006). The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Applied Linguistics, 27*, 590–619.

Laufer, B., & Nation, P. (1999). A vocabulary size test of controlled productive ability. *Language Testing, 16*, 33-51.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA, US: The MIT Press.

Lund, R. J. (1991). A comparison of second language listening and reading comprehension. *The Modern Language Journal, 75*(2), 196-204. Retrieved from https://www.jstor.org/stable/328827?seq=1#page_scan_tab_contents. doi:10.2307/328827

McNamara, T. (1990). Item response theory and the validation of an ESP test for health professionals. *Language Testing, 7*, 52–76.

McNamara, T. (1996). *Measuring second language performance*. London/New York: Longman.

Meara, P. M., & Milton, J. (2003). *X-lex: the Swansea levels test*. Newbury. UK: Express Publishing.

Mecartty, F. H. (2000). Lexical and grammatical knowledge in reading and listening comprehension by foreign language learners of Spanish. *Applied Language Learning, 11*(2), 323–348. Retrieved from https://www.researchgate.net/publication/312922557_Lexical_and_grammatical_knowledge_in_reading_and_listening_comprehension_by_foreign_language_learners_of_Spanish.

Milton, J., & Hopkins, N. (2005). *Aural Lex*. Swansea: Swansea University.

Milton, J., Wade, J., & Hopkins, N. (2010). Aural word recognition and oral competence in English as a foreign language. In *Insights into Non-native Vocabulary Teaching and Learning* (pp. 83–98).

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review, 63*(1), 59–82.

North, B., Ortega, A., & Sheehan, S. (2010). *EAQUALS Core Inventory for General English*: British Council.

O'Sullivan, B., & Dunlea, J. (2015). *Aptis General Technical Manual Version 1.0.* London: British Council

O'Sullivan, B., Weir, C. J., & Saville, N. (2002). Using observation checklists to validate speaking-test tasks. *Language Testing, 19*(1), 33–56. doi:10.1191/0265532202lt219oa

Pan, Y.-C., Tsai, T.-H., Huang, Y.-K., & Liu, D. (2018). Effects of expanded vocabulary support on L2 listening comprehension. *Language Teaching Research, 22*(2), 189–207. Retrieved from http://journals.sagepub.com/doi/10.1177/1362168816668895. doi:10.1177/1362168816668895

Park, G.-P. (2004). Comparison of L2 listening and reading comprehension by university students learning English in Korea. *Foreign Language Annals, 37*(3), 448–458. Retrieved from http://doi.wiley.com/10.1111/j.1944-9720.2004.tb02702.x. doi:10.1111/j.1944-9720.2004.tb02702.x

Purpura, J. (1999). *Learner Strategy Use and Performance on Language Tests: A Structural Equation Modeling Approach*. Cambridge: Cambridge University Press.

R Development Core Team. (2018). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Reise, S. P., Cook, K. F., & Moore, T. M. (2015). Evaluating the Impact of Multidimensionality on Unidimensional Item Response Theory Model Parameters. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 13–40). New York, NY: Routledge.

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: exploring the extent to which multidimensional data yield univocal scale scores. *J Pers Assess, 92*(6), 544–559. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/20954056. doi:10.1080/00223891.2010.496477

Revelle, M. W. (2017). *psych: Procedures for Personality and Psychological Research*. Evanston, IL. Retrieved from https://cran.r-project.org/package=psych.

Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape (with discussion). *Applied Statistics, 54*, 507–554.

Römer, U. (2009). The inseparability of lexis and grammar: Corpus linguistic perspectives. *Annual Review of Cognitive Linguistics,, 7*, 140–162.

Römer, U. (2017). Language assessment and the inseparability of lexis and grammar: Focus on the construct of speaking. *Language Testing, 34*(4), 477–492. Retrieved from http://journals.sagepub.com/doi/10.1177/0265532217711431. doi:10.1177/0265532217711431

Samejima, F. (1969). *Estimation of Latent Ability Using a Response Pattern of Graded Scores*. Richmond, VA]: Psychometric Society. Retrieved from http://www.psychometrika.org/journal/online/MN17.pdf.

Sawaki, Y., Stricker, L. J., & Oranje, A. H. (2008). *Factor structure of the TOEFL® Internet-based test (iBT): Exploration in a field trial sample*. Retrieved from Princeton, NJ: http://doi.wiley.com/10.1002/j.2333-8504.2008.tb02095.x

Sawaki, Y., Stricker, L. J., & Oranje, A. H. (2009). Factor structure of the TOEFL Internet-based test. *Language Testing, 26*(1), 5–30. Retrieved from http://journals.sagepub.com/doi/10.1177/0265532208097335. doi:10.1177/0265532208097335

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal, 95*, 26–43.

Schoonen, R., Van Gelderen, A., De Glopper, K., Hulstijn, J., Simis, A., Snellings, P., & Stevenson, M. (2003). First language and second language writing: The role of linguistic knowledge, speed of processing, and metacognitive knowledge. *Language Learning, 53*(1), 165–202. Retrieved from http://doi.wiley.com/10.1111/1467-9922.00213. doi:10.1111/1467-9922.00213

Schoonen, R., Van Gelderen, A., Stoel, R. D., Hulstijn, J., & De Glopper, K. (2011). Modeling the development of L1 and EFL writing proficiency of secondary school students. *Language Learning, 61*(1), 31–79. Retrieved from http://doi.wiley.com/10.1111/j.1467-9922.2010.00590.x. doi:10.1111/j.1467-9922.2010.00590.x

Schwartz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics, 6*, 461–464.

Shiotsu, T. (2010). *Components of L2 reading: Linguistic and processing factors in the reading test performances of Japanese EFL learners*. Cambridge: Cambridge University Press and Cambridge ESOL.

Shiotsu, T., & Weir, C. J. (2007). The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance. *Language Testing, 24*(1), 99–128. Retrieved from http://journals.sagepub.com/doi/10.1177/0265532207071513. doi:10.1177/0265532207071513

Sinclair, J. M. (2004). *Trust the text: Language, corpus and discourse*. London: Routledge.

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. doi:10.1007/s11205-010-9682-8

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal, 36*(2), 139–152. Retrieved from http://dx.doi.org/10.1080/09571730802389975. doi:10.1080/09571730802389975

Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V., & De Bastiani, F. (2017). *Flexible Regression and Smoothing: Using GAMLSS in R*. New York: Chapman and Hall/CRC.

van Gelderen, A., Schoonen, R., Glopper, K., Hulstijn, J., Simis, A., Snellings, P., & Stevenson, M. (2004). *Linguistic Knowledge, Processing Speed, and Metacognitive Knowledge in First- and Second-Language Reading Comprehension: A Componential Analysis* (Vol. 96).

van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics, 34*(4), 457–479. Retrieved from https://academic.oup.com/applij/article-lookup/doi/10.1093/applin/ams074. doi:10.1093/applin/ams074

Vandergrift, L., & Baker, S. (2015). Learner variables in second language listening comprehension: An exploratory path analysis. *Language Learning, 65*(2), 390–416. Retrieved from http://doi.wiley.com/10.1111/lang.12105. doi:10.1111/lang.12105

Wang, Y., & Treffers-Daller, J. (2017). Explaining listening comprehension among L2 learners of English: The contribution of general language proficiency, vocabulary knowledge and metacognitive awareness. *System*, 139–150. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S0346251X16304560. doi:10.1016/j.system.2016.12.013

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Oxford: Palgrave Macmillan.

Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second language development in writing: Measures of fluency, accuracy & complexity*. Honolulu: University of Hawaii Press.

Zimmerman, K. J. (2004). *The role of vocabulary size in assessing second language vocabulary.* (Master of Arts), Brigham Young University, Retrieved from https://scholarsarchive.byu.edu/etd/578 All Theses and Dissertations database. (578)

BRITISH COUNCIL
APTIS TECHNICAL REPORTS

ISSN 2057-7168