BRITISH COUNCIL

ENGLISH LANGUAGE
ASSESSMENT RESEARCH GROUP

A CASE FOR THE DIAGNOSTIC USE OF THE LISTENING TEST OF APTIS FOR TEENS ADVANCED: A COGNITIVE DIAGNOSTIC ASSESSMENT STUDY

AR-G/2024/1

Shangchao Min and Hongwen Cai

ARAGs RESEARCH REPORTS ONLINE
SERIES EDITORS: RICHARD SPIBY AND CAROLYN WESTBROOK
ISSN 2057-5203

ABSTRACT

In contrast with the multitude of cognitive diagnostic assessment (CDA) studies on reading tests, CDA studies on listening tests are relatively scarce. In particular, little research has been conducted on EFL listening tests that are built for diagnostic purposes, although CDA analyses on such tests may provide more diagnostic information than tests for general proficiency purposes. This study is an initial CDA attempt to validate a general EFL listening proficiency test that is structurally compatible with diagnostic uses, comprising distinct clusters of items to represent multiple dimensions.

The data used in the study were 1,357 candidates' item-level response data of the listening subtest of Aptis for Teens Advanced. The test is built on a model of cognitive processing and limited to three kinds of target information with increasing cognitive demand, from factual information to interpretive meaning at the utterance level to meaning at discourse level. To facilitate diagnosis, the Q-matrix is specified in accordance with Aptis developers' item-attribute coding in the test specifications, with its three attributes corresponding to the three kinds of target information. CDA analyses were completed with the GDINA R package, version 2.7.8.

The results showed that:

- the cognitive attributes measured by the test could be distinguished from each other
- factual information was the easiest attribute as expected, but interpretive meaning at the utterance was more difficult than meaning at discourse level for candidates, against the test developers' assumptions
- the candidate profiles at different CEFR levels partially matched the test developers' assumptions
- the classification reliability at the test, attribute and pattern levels was generally satisfactory.

These findings support the use of Aptis for Teens Advanced listening tasks as a diagnostic tool to inform remedial learning and instruction, apart from more general uses for admissions, placement, and progress evaluation.

Authors

Dr Shangchao Min is a professor at the School of International Studies, Zhejiang University, China. Her main research interests are language assessment, especially computerised language testing, language standards, and cognitive diagnostic assessment. She serves on the editorial board of *Language Assessment Quarterly*. She has published more than 30 papers in peer-reviewed academic journals such as *Language Testing*, *Language Assessment Quarterly*, *Assessing Writing, Language Teaching*, *System*, and *Studies in Educational Evaluation*.

Hongwen Cai, Ph.D. in applied linguistics (UCLA), is a Professor of Applied Linguistics at Guangdong University of Foreign Studies, China. His research interests include language testing and assessment, educational measurement, and language education. His publications have appeared in a variety of refereed journals, including *Language Testing, Language Assessment Quarterly*, and *Modern Foreign Languages*.

CONTENTS

List of tables and figures

1. INTRODUCTION

The literature about cognitive diagnostic assessment (CDA) or diagnostic classification models (DCM) normally makes a distinction between two approaches: a diagnostic-by-design approach and a diagnostic-by-chance, or retrofitting approach (Gierl & Cui, 2008; Haberman & von Davier, 2006; Jang et al., 2019; Liu et al., 2014; Roussos et al., 2007; Stout et al., 2019). In the first approach, "a cognitive model would be developed first to specify the knowledge and skills evaluated on the test and then items would be created to measure these specific cognitive skills" (Gierl & Cui, 2008, p. 265). In contrast, the second approach chooses to "revisit existing summative assessments and their data and apply diagnostic methods in a post-hoc attempt to extract some diagnostic information from these assessments" (Roussos et al., 2007, p. 278). Stout et al. (2019) further differentiate between two types of retrofitting: "applying a DCM to a test (likely with summative intentions) initially designed to yield unidimensionally-based conclusions versus one intentionally designed to make multidimensionally based diagnostic inferences, but without…DCM playing a role at the design level" (pp. 72–73). With some rare exceptions (e.g., Bradshaw et al., 2014; Liu et al., 2014; Ranjbaran & Alavi, 2017), the diagnostic-by-chance approach has dominated the literature. However, most criticisms have also been directed toward this approach.

One major criticism concerns the appropriacy of retrofitting a multidimensional model on a test designed to measure a unidimensional construct. It is doubted that such an attempt can yield much added, yet meaningful information beyond the total score (Gierl & Cui, 2008; Haberman & von Davier, 2006). Stemming from the conceptual mismatch between the multidimensional CDA model and the unidimensional structure of the existing test are a series of practical problems, such as high attribute correlations (Liu et al., 2018; Sessoms & Henson, 2018; Toprak et al., 2019), high flat profile (i.e., master of all subskills and non-master of all subskills) rates (Lee & Sawaki, 2009a, Liu et al., 2018), and variable classification reliability of different attributes (Aryadoust, 2021).

Another issue pertains to the lack of an appropriate cognitive theory that underlies the CDA studies (Gierl & Cui, 2008; Liu et al., 2018; Ravand, 2016). Without a prespecified model, the item-by-subskill relationships in a Q-matrix are liable to changeability, due primarily to the subjectivity of content experts who specify the Q-matrix. Typically, there is a lack of agreement among the experts in the grain size of subskills (Lee & Sawaki, 2009a), the specific item-by-subskill relationships (Li & Suen, 2013), and hence high probability of incorrect mastery classifications and faulty remedial actions (Im & Corter, 2011; Kunina-Habenicht et al., 2012).

Given these limitations of the retrofitting practice, the best CDA practice is to develop an assessment informed by a cognitive theory (Alderson, 2010; Gierl & Cui, 2008; Haberman & von Davier, 2006; Jang, 2010; Lee & Sawaki, 2009b; Li & Suen, 2013; Liu, 2018; Ranjbaran & Alavi, 2017; Sinharay et al., 2010). This study attempts to focus on the use of CDA models in validating such a test, i.e., the listening test of Aptis for Teens Advanced, a test with cognitive processing built in as a springboard and intended to be used to diagnose candidates' strengths and weaknesses to inform remedial learning and instruction. The test developers make no claim to have developed the test in the diagnostic-by-design approach. However, the listening test consists of clusters of items representative of distinct cognitive operations, which renders it multidimensional in structure and compatible with diagnostic uses. Little known CDA research has been conducted on EFL listening tests that are built on a model of cognitive processing, which may show stronger diagnostic power, as one of the basic assumptions of CDA is that attributes are defined according to a systematic categorisation of skills grounded in a specific cognitive theory (Kim, 2015).

2. LITERATURE REVIEW

2.1 Issues with retrofitting CDA models

Most criticisms of diagnostic assessment are directed towards a diagnostic-by-chance approach, i.e., retrofitting a CDA model to the response data from an existing test. The key problems arise from fitting a multidimensional model to the response data from a test designed under a unidimensional model (Gierl & Cui, 2008; Haberman & von Davier, 2006; Sinharay et al., 2010). To maximise the reliability of test scores, most existing language proficiency tests are constructed to be psychometrically unidimensional within each linguistic domain (Liu et al., 2018). This indicates that items with low discriminations, or in other words, items manifesting a substantive secondary dimension such as subskills and content knowledge, are excluded (Liu et al., 2018). Thus, retrofitting existing language proficiency tests for diagnostic information is an attempt to extract information about psychologically multidimensional constructs from psychometrically unidimensional instruments.

In theory, extracting information on dimensions unintended by the original test developer results in a mismatch between the substantive or psychological model underlying the diagnostic purpose and the psychometric model underlying the original test. In CDA studies, the specification of item-by-subskill relationships in a Q-matrix is usually informed by a substantive theory that assumes divisibility of the subskills (Ravand, 2016). In existing language proficiency tests, however, subskill divisibility is seldom achieved psychometrically (Gierl & Cui, 2008; Haberman & von Davier, 2006). Coercing a Q-matrix that assumes subskill divisibility on items that do not guarantee such divisibility is sure to invite misspecification, due to its post-hoc nature (Roussos et al., 2007). There is a whole host of studies on identifying and correcting Q-matrix misspecification (Chen, 2017; Chiu, 2013; de la Torre, 2008; de la Torre & Chiu, 2016; Liu, 2016; Nájera et al., 2019; Romero et al., 2014; Yu & Cheng, 2019), but if the psychologically divisible subskills have not been psychometrically operationalised in the test, any extra dimensions specified in the Q-matrix are destined to fail psychometrically.

In practice, the attributes identified from an existing test intended for nondiagnostic purposes are usually found to be highly correlated, in agreement with the unidimensional nature of most tests. A recent review of CDA studies found that most studies reported multiple attribute correlations exceeding .90, with rare exceptions (Bradshaw et al., 2014; Sessoms & Henson, 2018). The extremely high attribute correlations indicate a lack of distinction between attributes and call into question the meaning of reporting the mastery probability of separate attributes. The same could be said of flat profiles, which are found to be the most prevalent attribute profiles in some retrofitting studies (e.g., Lee & Sawaki, 2009a; Liu et al., 2018). When all-mastery and all-nonmastery profiles constitute the majority, reporting mastery profiles does not provide added information on top of the total score, as the non-flat profiles may simply be rare exceptions resulting from random errors.

Apart from the issues of high attribute correlations and high flat profile rates, a number of other problems arise in previous retrofitting studies, such as Q-matrix misspecification, lack of comparison of model and item fit between unidimensional and multidimensional models, and variable classification reliability. Q-matrix specification is problematic in most retrofitting studies. Lack of strong agreement among coders is the most common problem when deciding on the number of attributes, the grain size of attributes, and the specific attributes underlying each test item (Lee & Sawaki, 2009a, Li & Suen, 2013). A typical example that reflects this difficulty is the post-hoc combination of two attributes into a new one due to the lack of sufficient items that can be matched to one attribute (Li & Suen, 2013).

In this example, the number and the grain size of attributes are both changed, and so is the interpretation of the new attribute. This changeability may threaten to affect the model and item fit significantly. If not, it is most probably due to the indivisibility between attributes. Either case constitutes Q-matrix misspecification.

Model and item fit statistics are often used as important indicators for both Q-matrix validation and model evaluation in previous retrofitting studies (Chen, 2017; Lei & Li, 2016; Yu & Cheng, 2019). However, the comparison of model and item fit makes no substantive sense without a meaningful yardstick. Given the fact that the original test is designed based on a unidimensional psychometric model, a more meaningful way to evaluate a Q-matrix and a model is to conduct a statistical test to examine whether the multidimensional model provides a better fit than its unidimensional counterpart (Sinharay et al., 2010), i.e., treating the unidimensional IRT model as a baseline model. However, this is rarely done in practice.

Reliability is another key issue when reporting attribute mastery probability estimates and classification results in retrofitting studies (Gierl et al., 2009; Sinharay et al., 2010). The number of items measuring a certain attribute cannot be controlled when retrofitting an existing test, and oftentimes the intended attribute is measured by so few items that the estimates of the attribute mastery probabilities are subject to severe error (Sinharay et al., 2010). Furthermore, the classification reliability of different attributes may vary a lot (Aryadoust, 2021), which threatens the reliability of pattern-level classification, i.e., how accurately a test-taker is classified into a certain attribute profile.

All of the above problems are associated primarily with the retrofitting practice, and boil down to the fundamental mismatch between the divisibility assumption of a diagnostic purpose and the unidimensional nature of most existing tests. As a result, extra information from an existing test may be so scanty that it calls into question the usefulness of the extracted information. There is a dilemma here: one does not want to extract information from a low-quality test, but when a test is of high quality under the premise of unidimensionality, it would provide little information beyond the intention of the test developer (Sinharay et al., 2010). This dilemma can only be avoided with the diagnostic-by-design approach, either by developing items with reference to a pre-specified cognitive model (Gierl & Cui, 2008), or by restructuring the test to produce useful multidimensional score profiles (Sinharay et al., 2010).

2.2 Designing a CDA-based English listening test

One central assumption in designing a CDA-based English listening test is that listening skills are divisible. By nature, listening attribute divisibility is the same issue as the componentiality of language ability in general (Bachman & Palmer, 2010; Pae & Greenberg, 2014; Sawaki et al., 2009), and of various skills in particular (Alderson, 1990, 2000; Buck & Tatsuoka, 1998; Lee & Sawaki, 2009a; Lumley, 1993; Song, 2008; Tengberg, 2018; Weir et al., 2000). The belief in the multicomponential nature of language ability and specific skills such as reading and listening has often been contradicted by findings that a test intended to measure a multi-component construct is essentially unidimensional, so often that Henning (1992) emphasised the distinction between psychological dimensions and psychometric dimensions. According to Henning (1992), psychological dimensionality pertains to the psychological traits measured by a test, while psychometric dimensionality concerns the statistical properties of a test. These dimensions do not necessarily correspond to each other, i.e., psychometric unidimensionality could exist in the presence of psychological multidimensionality, and vice versa.

Similarly, Alderson (2000) made a distinction between skills that may theoretically exist and those that can be identified and realised on tests (Alderson, 2000). In Henning's terms, this distinction between theory and reality marks the tension between psychological and psychometric dimensions.

Following the reasoning of Henning (1992) and Alderson (2000), the usefulness of diagnostic information depends on whether the psychological dimensions can be realised psychometrically. To achieve psychometric multidimensionality, the construct to be measured should be based on a theoretical model that makes a clear distinction between its dimensions, preferably with support from empirical findings. With this restriction, the search for an appropriate model is soon directed to a series of cognitive models of listening, most of which propose similar key stages of comprehension in terms of word-, sentence-, and discourse-level processes (Fernández & Cairns, 2018).

An example of such models is the three-stage cognitive model of Anderson (2015), consisting of the stages of *perception*, *parsing*, and *utilization*. This model has been empirically verified in previous research with neurological evidence, i.e., psychologists have actually observed different combinations of brain regions being activated in the three stages (Anderson, 2015). The adaptation of the three stages to L2 listening consists of the processes of *decoding*, *parsing*, and *meaning construction* (Field, 2008). This model was later expanded to five processes, consisting of three lower-level processes (*input decoding*, *lexical search*, and *parsing*) and two higher-level processes (*meaning construction* and *discourse representation*). According to Field (2013), listeners first *decode* the incoming sounds with phonological knowledge, and then combine the phonemes to form individual words with *lexical* knowledge and literal understanding of the clauses and sentences with syntactic knowledge (i.e., *parsing*). The literal understanding formed at the lower level is then related to the context to *construct meaning* with pragmatic knowledge, and finally related to everything that has been said as well as external knowledge, such as world knowledge, to produce an overall meaning of the message (i.e., *discourse representation*).

It should be noted that these five processes do not follow a linear order from the lowest level to the highest level; rather, different processes may be activated simultaneously, with compensatory mechanism among different processes (Harding et al., 2015). That is to say, these processes do not perform in isolation but in various combinations when listeners are listening to speakers' words or answering listening items. It therefore follows naturally that diagnostic listening assessments built with Field's model as a frame of reference should not only include items targeting specific cognitive processes, but also incorporate "an additional layer of diagnosis targeting process interactions" (Harding et al., 2015, p. 329).

The listening test of Aptis for Teens Advanced is such a test. Built on Field's (2013) model of cognitive processing, the test consists of clusters of items tapping into clearly defined cognitive operations, including single cognitive processes and multiple process interactions (O'Sullivan et al., 2020). The listening test, for example, consists of three clusters of items. The first cluster focuses on meaning construction, targeting factual information; the second cluster focuses on discourse construction, targeting meaning at discourse level; while the third cluster also focuses on discourse construction, its target is twofold—interpretive meaning at the utterance level as well as meaning at discourse level. From the psychometric perspective, the clear distinction between clusters is conducive to a multidimensional structure. Furthermore, limiting the test to three focuses has the potential to improve the reliability of diagnostic feedback, as each dimension is represented by a sufficient number of items. Although Aptis for Teens Advanced is initially intended as a general proficiency test, we believe this treatment conforms essentially to the diagnostic-by-design recommendation mentioned above (Gierl & Cui, 2008; Sinharay et al., 2010).

2.3 Validating a CDA-based English listening test

The potential problems in retrofitting studies provide a ready checklist for conducting a validation study on tests with diagnostic potential, including Q-matrix specification, model selection, model and item fit, attribute mastery probabilities, and classification reliability. There are other types of validity problems worth exploring (Gierl & Cui, 2008; Sessoms & Henson, 2018; Sinharay et al., 2010), but these problems highlight the difference between the diagnostic-by-design and diagnostic-by-chance approaches. All these issues are heavily influenced by model selection; therefore, model selection becomes a key choice in the validation effort.

Previous listening CDA applications used multiple models to provide diagnostic information (e.g., Aryadoust, 2021; Lee & Sawaki, 2009a; Liu et al., 2018; Yi, 2017), including compensatory, noncompensatory and saturated models. Compensatory models assume that the lack of competence in one attribute can be compensated for by the presence of another attribute. Examples of compensatory models widely used in language assessment are the additive cognitive diagnostic model (ACDM; de la Torre, 2011) and the deterministic inputs, noisy "or" gate (DINO; Templin & Henson, 2006) model. The ACDM assumes that each attribute additively increases the probability of a correct response, and mastery of one attribute can compensate for the nonmastery of another attribute (Ravand & Robitzsch, 2018). The DINO model is the most compensatory model in that the mastery of one attribute can completely compensate for the nonmastery of all other attributes (Yi, 2017).

Noncompensatory models posit that the lack of competence in one attribute cannot be compensated for by the presence of another attribute. Examples of commonly used noncompensatory models are the reduced reparameterized unified model (RRUM; DiBello et al., 2007), and the deterministic inputs, noisy "and" gate (DINA; Junker & Sijtsma, 2001) model. The RRUM is a noncompensatory counterpart of the ACDM and the DINA model is a noncompensatory counterpart of the DINO model (Li et al., 2016). The DINA model is the most restrictive noncompensatory model in that candidates can answer the item correctly only if they have mastered all the attributes measured by the item (Liu et al., 2018).

Saturated models can be used when the relationships among attributes are unknown (Javidanmehr & Sarab, 2019; Li et al., 2016). Saturated models are flexible enough to capture different kinds of relationships among attributes. A commonly used example of saturated models is the generalised DINA model (G-DINA; de la Torre, 2011), which is a general model that subsumes many CDA models including the above-mentioned DINO, ACDM, RRUM and DINO models.

In listening CDA research, noncompensatory models have been used more frequently (e.g., Aryadoust, 2021; Buck & Tatsuoka, 1998; Sawaki et al., 2009) than compensatory models (e.g., Yi, 2017). However, the results of previous listening CDA research on the comparison of different models (e.g., Lee & Sawaki, 2009a; Liu et al., 2018; Yi, 2017) reported no obvious superiority of one type of model over another. This suggests that both compensatory and noncompensatory relationships may exist among different listening attributes. The practical way to select CDA models, therefore, is to examine the absolute and relative model fit indices of different models and find the model that fits the data best.

Attribute mastery probabilities can then be computed. In this regard, previous findings are mixed as to whether there is a hierarchy of difficulty among attributes in previous listening CDA retrofitting studies. Some supported the hypothesis that higher-level attributes are more difficult than lower-level attributes. For instance, Liu et al. (2018) reported that for a mock TOEFL iBT listening test, basic comprehension was generally easier than pragmatic understanding, and connecting information and understanding organisation, although there was some variation of the attribute mastery probability across easy, medium and hard item banks. On the other hand, some refuted the hypothesis that higher-level attributes are more difficult. For example, Aryadoust (2021) showed that catching surface details was even more difficult than making inferences. The mixed findings are expected, since the difficulty of the attribute lies not so much in the level of the attribute, but in the difficulty of the language and the density of information of the tasks to which the attribute is applied (Alderson, 2007). Nevertheless, if a diagnostic listening test is built with the hierarchy of attributes as a premise, empirical evidence needs to be collected to demonstrate that the empirical difficulty of attributes is in alignment with the test developers' assumptions.

The psychometric quality of CDA diagnosis should also be evaluated if classification decisions are to be made based on test results from diagnostic assessment in the real world (Iaconangelo, 2017; Wang et al., 2015). Previous CDA research on listening seldom reports classification reliability, with the exception of Aryadoust (2021), Liu et al. (2018) and Min & He (2022). This is partly due to the fact that all of them are retrofitting efforts to extract diagnostic information from existing proficiency tests, with no immediate intention to provide the diagnostic classification results to real candidates. Nevertheless, if classification decisions are to be made in the real world, information about the psychometric quality of such diagnosis at the test, attribute and pattern levels should be presented (Iaconangelo, 2017; Wang et al., 2015). Otherwise, candidates may be given misleading diagnostic feedback and potentially take the wrong remedial actions (Sinharay et al., 2019).

3. RESEARCH QUESTIONS

The target test of this study is the listening test of Aptis for Teens Advanced, which is designed following Field's (2008, 2013) listening processing model. One of the intended uses of the test result is to diagnose candidates' strengths and weaknesses to inform remedial learning and instruction within a specific institution (Fairbairn, Spiby & Dunlea, 2017). This study addresses four research questions:

1. To what extent are the cognitive attributes measured by the diagnostic listening test distinguishable from each other?
2. To what extent are the empirical difficulty levels of the cognitive attributes in line with the test developers' assumptions?
3. To what extent are the cognitive profiles of candidates at different CEFR levels in line with the test developers' assumptions?
4. To what extent does the test produce reliable diagnostic information about candidates' attribute mastery statuses?

Relating to the sources of problems in retrofitting existing proficiency tests for diagnostic purposes discussed above, RQ1 pertains to model selection, attribute correlations, and profile distribution, most notably proportion of flat profiles, all of which are associated with attribute distinctness; RQ2 pertains to assumptions about the relative difficulty levels of different attributes in the cognitive model that guides the design of the test; RQ3 is concerned with consistency between the attribute mastery statuses and the proficiency levels of the candidates. RQ4 is related to model and item fit, and reliability of classification results.

4. METHODS

4.1 Data

The data used in the study were 1,357 candidates' item-level responses to the listening subtest of Aptis for Teens Advanced in a test administration in April 2018. Aptis for Teens Advanced is designed to test general English proficiency of young EFL/ESL learners. The test is administered on computer, comprising subtests of listening, reading, speaking, writing and grammatical and vocabulary knowledge. The information presented in the score report includes a numeric score within the range of 0 to 50 as well as a CEFR level for the four skill subtests (Fairbairn et al., 2017).

The present study focused on the listening subtest, which is a 45-minute test consisting of 25 multiple-choice-question (MCQ) and matching items targeting different processes of listening comprehension at different CEFR levels (Fairbairn et al., 2017). In relation to the cognitive processing levels delineated in Field's (2008, 2013) model, the Aptis listening test specifications also outline the type of information targeted by each item. This serves the test developers' purpose to provide diagnostic information apart from a total score. An overview of the listening subtest is given in Table 1.

Table 1: Listening test structure of Aptis for Teens Advanced

CEFR level	Cognitive processing	Information targeted	No. of items per task type	Interaction format
B1	Meaning construction	Factual information	4 matching items	Monologues
B1	Meaning construction	Factual information	2 MCQ items	Dialogues
B2	Discourse construction	Meaning at discourse level	11 MCQ items	Dialogues and monologues
C1	Discourse construction	Interpretative meaning at the utterance level and meaning at discourse level	8 matching items	Dialogues

A notable feature of the test is that the items are arranged in three distinct clusters. As shown, the listening subtest of Aptis for Teens Advanced comprises four matching items and two MCQ items targeting factual information (cluster 1), 11 MCQ items targeting meaning at discourse level (cluster 2), and eight matching items targeting interpretive meaning at the utterance and meaning at discourse level simultaneously (cluster 3).

4.2 Data analysis

The definition of attributes was the first step in CDA analysis. For this purpose, the researchers employed the attributes that were clearly defined in test specification. They were (1) factual information, (2) interpretive meaning at the utterance level, and (3) meaning at discourse level, corresponding to the three clusters of items.

The second step was Q-matrix construction. The three clusters serve as a ready specification of a Q-matrix. The researchers did not revise the Q-matrix because for one thing, Q-matrix construction is theory-driven rather than data-driven, and for another, this study is a validation study of the theorised item-skill matrix by employing CDA. The structure of the Q-matrix constructed is already summarised in Table 1, but is reshaped in the more familiar binary notation in Table 2, in which the number 1 indicates that an attribute is represented by a given item, while 0 indicates that the attribute is not represented by the item.

Table 2: Q-matrix for the Aptis listening test

Item No.	Attribute 1 (Factual information)	Attribute 2 (Interpretive meaning at the utterance)	Attribute 3 (Meaning at discourse level)
1	1	0	0
2	1	0	0
3	1	0	0
4	1	0	0
5	1	0	0
6	1	0	0
7	0	0	1
8	0	0	1
9	0	0	1
10	0	0	1
11	0	0	1
12	0	0	1
13	0	0	1
14	0	0	1
15	0	0	1
16	0	0	1
17	0	0	1
18	0	1	1
19	0	1	1
20	0	1	1
21	0	1	1
22	0	1	1
23	0	1	1
24	0	1	1
25	0	1	1

Preliminary validation was conducted to establish the psychometric quality of the Aptis listening test. Specifically, IRT analyses were performed using the IRTPRO 4.2 computer program (Cai et al., 2011). The 1-parameter logistic model (1PLM; Rasch, 1960) and 2-parameter logistic model (2PLM; Birnbaum, 1968) were first fit to the data. The 3-parameter logistic model (3PLM; Birnbaum, 1968) was ruled out as an alternative because it has been well documented that unstable estimation of the pseudo-guessing parameter in the 3PLM tends to occur (Lord, 1980), which consequently leads to poor estimation of other item parameters (Baker, 1987; Swaminathan & Gifford, 1985). The model with better model-data fit was then selected for item parameter estimation, based on which the psychometric quality of all the items in the listening test was evaluated.

To address RQ1, CDA analyses were run by employing the GDINA R package, version 2.7.8 (Ma et al., 2020). Five CDA models were compared to choose an appropriate model that best fit the data, including the G-DINA model (de la Torre, 2011), the DINA model (Junker & Sijtsma, 2001), the ACDM (de la Torre, 2011), the RRUM (DiBello et al., 2007), and the DINO model (Templin & Henson, 2006). Both relative and absolute fit of these models were evaluated to identify the best model to provide diagnostic information on subskill mastery. Polychoric correlation analyses of the subskill mastery were then conducted to examine the distinctness of cognitive attributes measured by the test. Additionally, the proportions of candidates classified into flat profiles and jagged profiles were examined to shed light on the distinctness of cognitive attributes.

To address RQ2, the attribute mastery probabilities for all candidates were first estimated to gauge the difficulty of each attribute. Candidates were then classified into different CEFR levels based on their raw scores according to the cut scores yielded from a full formal standard setting procedure with a panel of experts (Spiby, personal communication, August 18, 2020), and the subskill mastery probabilities of different proficiency groups were calculated and compared to examine whether the order of difficulty of attributes is the same for candidates at different CEFR levels.

To address RQ3, the number of candidates classified into each latent pattern for all five proficiency levels was summarised to identify the representative patterns at each CEFR level. Then the probability estimates on the attributes were subject to mixed factorial ANOVA, with CEFR level as the between-subject factor and attribute as the within-group factor. These analyses provided information about whether candidates' cognitive profiles differed systematically across CEFR levels, and whether such distribution was in line with the test developers' assumptions.

To address RQ4, the test-, pattern- and attribute-level classification accuracy indices were estimated to establish the classification reliability of the CDA based on the best-fitting model, using approaches proposed by Iaconangelo (2017) and Wang et al. (2015). These analyses provided information about the psychometric quality of CDA diagnosis at different levels.

5. RESULTS

This section starts with the presentation of some descriptive statistics of the test. Then, IRT analyses including model fit of different IRT models and item parameter estimates are reported. Next, model fit statistics of different CDA models are compared to select the best one to fit the data. Finally, diagnostic information is obtained from the best fitting CDA model and further analysed to address the four research questions, including attribute distinctness (RQ1), attribute difficulty at overall and different CEFR levels (RQ2), candidate profiles across CEFR levels (RQ3), and classification reliability at attribute, pattern and test levels (RQ4).

5.1 Descriptive statistics

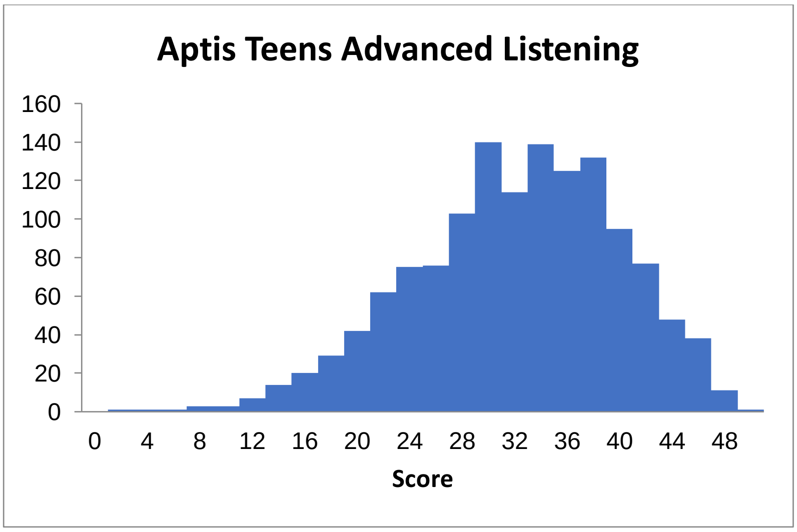
Table 3 presents the descriptive statistics of candidates' raw scores in the listening test. As is shown, the majority of candidates were classified at B1 through C1, with some rare exceptions at A0–A2 and C2, using the cut scores yielded from a formal standard setting procedure (Spiby, personal communication, August 18, 2020).

Table 3: Descriptive statistics of the listening test

Level	Number	Min	Max	Mean	SD
A0–A2	9	2	10	7.33	2.83
B1	112	12	20	17.52	2.51
B2	709	22	34	29.05	3.82
C1	515	36	46	39.63	3.09
C2	12	48	50	48.17	0.58
All	1,357	2	50	32.14	7.88

Figure 1 depicts the raw score distributions of the listening test, showing a negative skewness. However, this is not considered problematic for answering the research questions, as the target levels of this test are B1, B2 and C1, all of which were well represented by a large number of candidates (Table 3), so that reliable results may be expected of the subsequent CDA analyses.

Figure 1: Score distributions of the listening test



5.2 IRT analyses

Table 4 summarises the fit indices of the two IRT models for the listening test. A variety of criteria provided by the IRTPRO 4.2 program were employed to evaluate the model-data fit, including -2log-likelihood (-2LL; Neyman & Pearson, 1992), Akaike information criterion (AIC; Akaike, 1974), Bayesian information criterion (BIC; Schwarz, 1978), and RMSEA (root mean square error of association; Browne & Cudeck, 1993).

Table 4: Summary of fit indices of IRT models

IRT models	-2LL	AIC	BIC	RMSEA
1PLM	38,123	38,175	38,310	0.07
2PLM	37,070	37,170	37,430	0.05

The fit indices showed that the 2PLM was a better model for the listening test, as it offered a better solution in terms of model fit, indicated by the lower -2LL value, and model parsimony, as indicated by the lower AIC and BIC values (Rijmen, 2010). It should be noted that the 1PLM is nested within the 2PLM. Therefore, the likelihood ratio test, also named χ^2-difference test (du Toit, 2003), was carried out to test the significance of the difference in model fit. The results of the χ^2 -difference test showed that the 2PLM fit the data significantly better than the 1PLM (χ^2 (29) = 1,053, p < .05). The 2PLM was therefore selected as the final model for item parameter estimation.

Table 5 provides the item parameter estimates obtained from the 2PLM analysis, including discrimination (*a*) and difficulty (*b*) estimates with their standard errors. The results showed that for most items (i.e., 22/25), the discrimination parameter estimates were above 0.35, a threshold value for acceptable discriminating power according to Baker (1985). This indicates that the majority of items can effectively discriminate between different levels of candidates in terms of their listening ability. The difficulty parameter estimates ranged from -2.27 to 1.31, with the exception of one inflated estimate for Item 5, possibly due to unstable estimation of the discrimination parameter. This wide range of difficulty estimates indicates that the test could measure candidates' listening ability at different proficiency levels.

Table 5: Item parameter estimates

Item	Label	*a*	*s.e.*	*b*	*s.e.*
1	L1a	2.09	0.18	-0.69	0.05
2	L1b	2.16	0.2	-1.31	0.07
3	L1c	1.9	0.25	-2.27	0.17
4	L1d	2.34	0.2	-0.58	0.04
5	L2	-0.06	0.06	8.97	10.39
6	L3	1.37	0.16	-2.26	0.19
7	L4	0.61	0.07	-0.57	0.11
8	L5	0.77	0.09	-1.73	0.18
9	L6	1.37	0.13	-1.43	0.1
10	L7a	0.52	0.07	0.5	0.13
11	L7b	0.5	0.07	1.31	0.21
12	L8a	0.78	0.08	0.11	0.08
13	L8b	1	0.09	0.51	0.07
14	L9a	0.62	0.08	-1.96	0.24
15	L9b	1.03	0.1	-1.63	0.14
16	L10a	0.19	0.06	-1.96	0.72
17	L10b	0.53	0.07	0.06	0.11
18	L11a	0.7	0.08	-1.33	0.15
19	L11b	0.47	0.07	-0.31	0.13
20	L11c	1.02	0.09	-0.75	0.08
21	L11d	-0.39	0.07	-1.43	0.27
22	L12a	0.77	0.08	0.16	0.08
23	L12b	1.21	0.1	-0.54	0.06
24	L12c	0.61	0.08	-1.21	0.16
25	L12d	1.93	0.2	-1.66	0.1

5.3 CDA analyses

Five CDA models, including a saturated model (G-DINA), two compensatory models (DINO, ACDM), and two noncompensatory models (DINA, RRUM) were fitted to the candidates' observed responses. Relative and absolute fit of these models were evaluated to select the best model for follow-up CDA analyses. Diagnostic information obtained from the best model was then subject to further analyses to address the four research questions.

5.3.1 Model-data fit

Table 6 summarises the relative and absolute model fit of the five models. Three relative model fit indices were used, including -2LL (Neyman & Pearson, 1992), AIC (Akaike, 1987), and BIC (Schwarz, 1978). For -2LL, AIC and BIC, smaller values indicate better relative fit. SRMSR (standardised root mean square root of squared residuals; Maydeu-Olivares & Joe, 2014) is an absolute model fit index, with a value below 0.05 suggesting acceptable absolute fit.

Table 6: Relative and absolute model fit

Model	−2LL	Npars	AIC	BIC	SRMSR	χ^2	df	p
G-DINA	36,857	73	37,003	37,383	0.050			
DINA	37,108	57	37,222	37,519	0.053	251	16	.000
ACDM	36,895	65	37,025	37,364	0.050	38	8	.000
RRUM	36,909	65	37,039	37,378	0.050	52	8	.000
DINO	37,060	57	37,173	37,470	0.053	203	16	.000

As shown in the table, the G-DINA model had the lowest -2LL and AIC values, suggesting that the G-DINA model is the best fitting model. This is expected as the G-DINA model is a saturated model, which always fits the data better than the reduced models (Chen et al., 2013) owing to its complex parameterization. ACDM had the lowest BIC value, suggesting that the ACDM fit the data as well as the G-DINA model. This is probably due to the fact that the BIC imposes a larger penalty on more complex models than AIC (Li et al., 2016). However, according to a simulation study conducted by Lei and Li (2014), AIC performed the best among the three relative fit indices. AIC was therefore considered a more crucial criteria for model selection in this study. In addition, a series of likelihood ratio tests showed that the saturated G-DINA model fit the data significantly better than the four nested models, as can be seen in the last three columns of Table 6.

In terms of absolute fit, the SRMSR value of the G-DINA, ACDM, and RRUM models were below the 0.05 rule of thumb (Maydeu-Olivares & Joe, 2014), indicating that these three models fit the data well. However, the two overly restrictive models (i.e., the DINA and DINO model) did not fit the data well. Based on the relative and absolute fit indices, the G-DINA model was selected as the final model for diagnostic classifications.

5.3.2 Attribute distinctness

To examine the distinctness of attributes measured by the test (RQ1), we computed the polychoric correlation coefficients among attributes and the proportion of candidates in flat and jagged profiles. Table 7 presents the polychoric correlation coefficients among the attributes estimated from the G-DINA model. The correlation coefficients among the three attributes were all significant, ranging from 0.364 to 0.855, suggesting that the attributes are meaningfully correlated, yet distinguishable from each other, except that the attributes “factual information” (A1) and “meaning at discourse level” (A3) are not quite distinct from each other.

Table 7: Polychoric correlation coefficients among attributes

	A1: Factual information	A2: Interpretive meaning at the utterance level	A3: Meaning at discourse level
A1	1		
A2	.390**	1	
A3	.855**	.364**	1

Note: **p < 0.01

Another source of evidence on the distinctness of cognitive attributes that can be obtained from CDA analyses is the proportion of candidates classified into flat profiles (e.g., 000, 111) and jagged profiles (e.g., 001, 010, 011). Table 8 shows the overall class probabilities for each attribute profile. Candidates in this study were classified into 8 latent patterns (2^3). For instance, the second latent class "100" indicates that candidates classified into this pattern are expected to be a master of Attribute 1 (factual information), and a non-master of Attribute 2 (interpretive meaning at the utterance) and Attribute 3 (meaning at discourse level).

Table 8: Attribute profile probabilities

Latent class	Attribute profile	Class probability
1	000	0.243
2	100	0.086
3	010	0.097
4	001	0.023
5	110	0.072
6	101	0.197
7	011	0.016
8	111	0.267

As shown in Table 8, the attribute profile "111" had the highest class probability (i.e., 26.7%), indicating that 26.7% of the candidates were classified as a member of this latent class, namely, a master of all the three subskills. The attribute profile "000" had the second highest class probability of around 24.3%, indicating that 24.3% of the candidates had not mastered any of the three subskills. Following these two flat profiles, the jagged profile 101 had the third highest class probability of around 19.7%, which may be related to the somewhat high correlation between the first and third attributes (i.e., 0.855).

The proportion of flat profiles totalled 51%, slightly higher than the proportion of jagged profiles. This suggests that the diagnostic classifications on attributes can provide certain additional information over what a total score can offer, and that the attributes measured in the test can be generally distinguished from each other, with room for improvement for the distinctness between Attributes 1 and 3.

5.3.3 Attribute difficulty

The difficulty of attributes can be obtained from the mastery probabilities of the attibutes in CDA analyses (RQ2). To examine whether the difficulty of attributes is in alignment with the test developers' assumptions, we computed candidates' attribute mastery probabilities at overall and group levels. Table 9 displays the mastery probability of each attribute for all candidates, and for candidates at different CEFR levels. Candidates were classified into different CEFR levels using the cut scores yielded from a standard setting procedure (Spiby, personal communication, August 18, 2020).

Table 9: Mastery statistics of each attribute at overall and group levels

Level	Number of candidates	A1: Factual information	A2: Interpretive meaning at the utterance	A3: Meaning at discourse level
Overall	1,357	0.621	0.452	0.502
A0–A2	9	0.000	0.000	0.000
B1	112	0.027	0.128	0.000
B2	709	0.479	0.355	0.266
C1	515	0.947	0.654	0.934
C2	12	0.998	0.817	1.000

As shown in the first row of Table 9, the mastery probability of all candidates ranged from 0.452 to 0.621. That is, approximately 45.2% of the candidates mastered interpretive meaning at the utterance level, indicating it is the most difficult attribute, followed by meaning at discourse level (50.2%), and then factual information (62.1%).

In addition to attribute-level mastery probabilities for all candidates, Table 9 presents the mastery probabilities for candidates at different CEFR levels. It can be seen that the mean mastery probabilities on all three attributes increase monotonically by CEFR level, conforming to the test developers' expectations. Following Rupp, Templin, and Henson's (2010) practice, we classified an attribute mastery probability equal to or greater than 0.5 as mastery on that attribute, and a probability less than 0.5 as non-mastery. Therefore, averagely speaking, A0-A2, B1 and B2 students can be identified as non-masters of all three attributes. C1 and C2 students excel at all three attributes.

It should be noted, however, that the order of difficulty of attributes varied for candidates at different CEFR levels. Although interpretive meaning at the utterance level was relatively the most difficult attribute for candidates at C1 (i.e., 0.654) and C2 levels (i.e., 0.817), meaning at discourse level posed more challenge to candidates at A0–A2 (i.e., 0.000), B1 (i.e., 0.000), and B2 levels (i.e., 0.266).

5.3.4 Candidate profiles

In order to examine whether the candidates' cognitive profiles at different CEFR levels match the test developers' assumptions (RQ3), the number of candidates classified into each latent pattern for all five proficiency levels was summarised to identify the representative patterns at each CEFR level. Moreover, the probability estimates on the attributes were subject to mixed factorial ANOVA to see whether they differed systematically across CEFR levels, with CEFR level as the between-subject factor and attribute as the within-group factor.

Table 10 shows the number of candidates classified into each latent pattern for all five proficiency levels. As shown, A0–A2 students were non-masters of all attributes (000) whereas C2 students were masters of all attributes (111) or at least two attributes (101). Most B1 candidates (97/112) were identified as a non-master of all attributes. B2 students displayed a wide range of subskill profiles, while C1 students were generally diagnosed as mastering at least two attributes.

Table 10: Number of candidates classified into each pattern across proficiency groups

Subskill profile	A0-A2	B1	B2	C1	C2	Total
000	9 (100%)	97 (87%)	210 (30%)	1 (0%)	0 (0%)	317 (23%)
010	0 (0%)	14 (13%)	134 (19%)	1 (0%)	0 (0%)	149 (11%)
001	0 (0%)	0 (0%)	19 (3%)	6 (1%)	0 (0%)	25 (2%)
100	0 (0%)	0 (0%)	96 (14%)	0 (0%)	0 (0%)	96 (7%)
011	0 (0%)	0 (0%)	4 (1%)	11 (2%)	0 (0%)	15 (1%)
101	0 (0%)	0 (0%)	112 (16%)	166 (32%)	2 (17%)	280 (21%)
110	0 (0%)	1 (1%)	84 (12%)	10 (2%)	0 (0%)	95 (7%)
111	0 (0%)	0 (0%)	50 (7%)	320 (62%)	10 (83%)	380 (28%)
Total	**9 (100%)**	**112 (100%)**	**709 (100%)**	**515 (100%)**	**12 (100%)**	**1,357 (100%)**

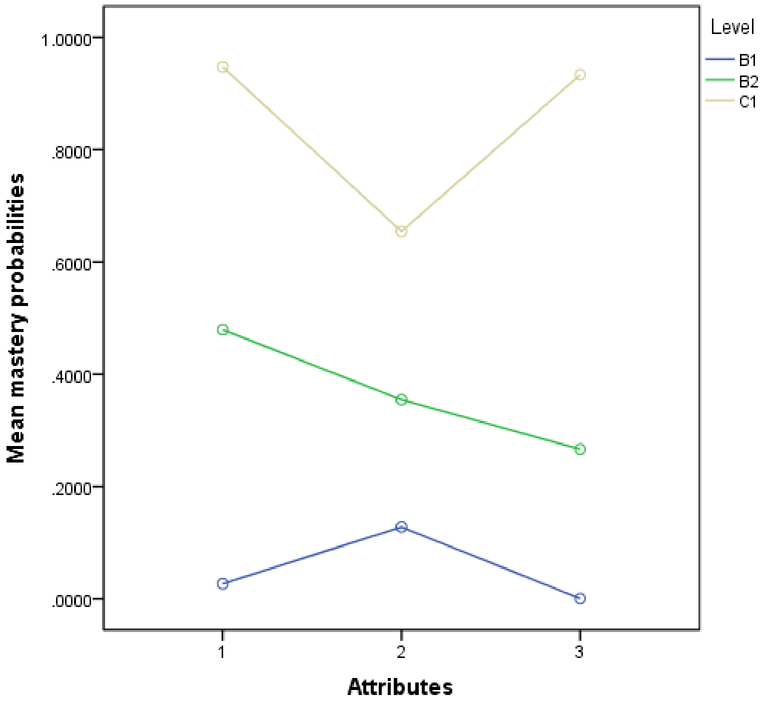
Given the limited number of candidates classified at A0–A2 (N = 9) and C2 (N = 12) levels, only data for three proficiency levels (i.e., B1, B2 and C1) were subject to mixed factorial ANOVA. The results yielded two significant main effects, proficiency level (F (2, 1,333) =980.08, $p < 0.001$, partial $\eta^2 = 0.60$) and subskill (F (2, 1,333) =53.97, $p < 0.001$, partial $\eta^2 = 0.04$), and a significant subskill× proficiency (F (4, 1,333) = 66.97, $p < 0.001$, partial $\eta^2 = 0.09$) interaction effect. A simple effect analysis was therefore conducted.

In terms of the simple effect of proficiency level, the results indicated that the probabilities of the three groups were significantly different (see Table 9) and post-hoc tests suggested that each of the three probabilities differed significantly from each of the others, with C1 outperforming B2 and B1, and B2 outperforming B1 (p = 0.000 for all pairwise comparisons) on all attributes.

Similar significant results have been obtained with regards to attributes (see Table 9). For candidates at the CEFR level B1, their mastery of Attribute 1 was significantly better than that of Attribute 3 (p = 0.006), both of which were significantly worse than that of Attribute 2 (p = 0.000). Candidates at the CEFR level B2 mastered Attribute 1 significantly better than Attribute 2 (p = 0.000), both of which were significantly better than that of Attribute 3 (p = 0.000). For candidates at the CEFR level C1, their mastery probabilities on Attributes 1 and 3 were significantly better than on Attribute 2 (p = 0.000 for both comparisons); however, no significant difference was observed between mastery of Attributes 1 and 3 (p = 0.129).

Figure 2 shows the mastery probabilities of the three proficiency groups on three attributes. It could be observed from the figure that the mastery discrepancy among three groups was smaller for Attribute 2 (Interpretive meaning at the utterance) than that for Attribute 1 (Factual information) and Attribute 3 (Meaning at discourse level). One possible explanation is that Attribute 2 is measured together with Attribute 3 all the time, which may have masked part of its unique variance. However, it is unknown whether and/or to what extent such test design (i.e., the overlapping measurement of A2 and A3) has some bearing with the high polychoric correlation between A1 and A3 (i.e., 0.855).

Figure 2: Mastery probabilities of three proficiency groups



5.3.5 Classification reliability

The classification reliability was estimated to establish the psychometric quality of the diagnostic classification (RQ4). Table 11 displays the classification accuracy at test, attribute and pattern levels. The value for the overall test was 0.836, suggesting that the test has an 83.6% probability of classifying a randomly selected test-taker into his/her true latent class. The classification accuracy estimates for all the attributes were all above 0.90, indicating that candidates can be accurately classified as masters or non-masters at the attribute level. The classification accuracy at the pattern level ranged from 0.499 to 0.926, suggesting some variability in diagnosis accuracy. The pattern-level classification accuracy is positively correlated with the number of candidates involved, where patterns with more candidates tend to have higher accuracies, and those with fewer candidates tend to have lower accuracies. This is expected as accurate classification of test-takers into less frequent latent classes has been reported to be challenging, probably due to the scarcity of data (Iaconangelo, 2017).

Table 11: Classification accuracy at test, attribute, and pattern levels

Level	Attributes and patterns	Accuracy	No. of candidates
Test	All patterns	0.836	1,357
Attribute	A1: Factual information	0.954	1,357
	A2: Interpretive meaning at the utterance	0.947	1,357
	A3: Meaning at discourse level	0.916	1,357
Pattern	P1: 000	0.862	317
	P2: 100	0.629	97
	P3: 010	0.813	149
	P4: 001	0.569	25
	P5: 110	0.631	96
	P6: 101	0.919	278
	P7: 011	0.499	15
	P8: 111	0.926	380

6. DISCUSSION

The present study attempted to validate an EFL listening test which was structurally compatible with the diagnostic-by-design approach by employing CDA analyses. Four types of validity evidence were gathered to support the diagnostic use of the Aptis for Teens Advanced listening test. The four types of evidence were: (1) the distinctness of listening attributes; (2) the alignment between the empirical difficulty of attributes and the test developers' assumptions; (3) the alignment between candidate profiles and the test developers' assumptions; and (4) the reliability of diagnostic classifications. These are issues frequently raised in criticisms of CDA studies, retrofitting studies in particular. It is hoped that the results of this study could shed some light on how a test targeting distinct cognitive operations helps resolve some of these issues.

6.1 RQ1: Distinctness of listening attributes

The results of the study showed that the two flat profiles were the most prevalent patterns in this study. This is in line with the findings of the majority of CDA retrofitting studies in listening assessment (e.g., Aryadoust, 2021; Lee & Sawaki, 2009a; Toprak et al., 2019) and reading assessment (e.g., Javidanmehra & Sarab, 2019; Li et al., 2016; Ravand, 2016), which reported a higher proportion of flat profiles than jagged profiles. However, the proportion of flat profiles in this study is just around 51%, much lower than that for previous listening CDA retrofitting studies, such as the Cambridge-Michigan Language Assessment (CaMLA) listening study (i.e., 89%; Toprak et al., 2019), the Singapore–Cambridge General Certificate of Education (GCE) listening study (i.e., 67%; Aryadoust, 2021), and the TOEFL listening study (i.e., ranging from 59% to 76% in different CDA analyses; Lee & Sawaki, 2009a). This indicates that tests targeting distinct cognitive operations can provide meaningful added information beyond the total score.

The lower proportion of flat profiles in this study may be attributed to the lower correlation coefficients among attributes in the present study. Except for the correlation between "factual information" and "meaning at discourse level" ($r = 0.855$), the other two correlation coefficients were estimated to be as low as 0.364 and 0.390. These are generally lower than those reported in previous retrofitting studies. For example, the listening subskill correlations reported in the TOEFL listening retrofitting study in Lee and Sawaki (2009a) ranged from .58 to .76. In the CaMLA listening study (Toprak et al., 2019), all correlation coefficients among listening subskills were found to be above 0.9. This indicates that tests based on a systematic categorisation of distinct subskills tend to be more multidimensional than existing tests that are designed under the unidimensional scoring paradigm. Since the utility of fine-grained information extracted from CDA analysis is closely associated with the psychometric dimensionality of the test (e.g., Lee & Sawaki, 2009a; Liu et al., 2018), diagnostic tests built with multidimensional attributes in test specifications serve as a more viable alternative for diagnostic uses than existing unidimensional tests.

The finding of this study has provided supporting evidence for the divisibility of listening ability into multiple subskills. The finding refutes an earlier claim that the listening skill is unidimensional (Oller, 1983), but resonates with previous research which shows that a number of psychometrically distinct factors can be identified within EFL listening comprehension (e.g., Freedle & Kostin, 1996; Song, 2008). The distinctness of the higher-level cognitive processes specified in Field's (2008, 2013) listening processing model is empirically supported in this study, in line with Harding et al.'s (2015) proposal that Field's (2008, 2013) model may provide a theory of diagnosis for the design of diagnostic listening tests.

Another factor that may have contributed to the distinctness of attributes in this study was the number of items representing each attribute. The minimum number of items corresponding to an attribute in this version of the listening test was six, which helped avoid the problem of having too few items to represent a target attribute (e.g., Li & Suen, 2013). All these have highlighted the importance of planning in advance and the advantage of the diagnostic-by-design approach. It should be noted, however, that follow-up research on other forms of the test is warranted to examine whether the high correlation between Attribute 1 and 3 is merely a statistical artefact of this test form or a manifestation of the psychological indistinctiveness of the two attributes in the current Aptis test design.

Overall, the findings suggest that the cognitive attributes measured by Aptis listening tasks could be distinguished from each other, yet further revisions need to be made to better differentiate between the attributes "factual information" and "meaning at discourse level".

6.2 RQ2: Hierarchy of attribute difficulty

The finding that understanding factual information was the easiest attribute converges with previous research that showed extracting explicit information is easier to master than higher-level attributes such as making inferences and summarising (e.g., Lee & Sawaki, 2009a; Rost, 2016; Vandergrift, 2007). However, they are in contrast with another body of research (e.g., Alderson, 2005; Buck, 1990; Brindley, 1997) that reported no significant difference between the difficulty of higher-level attributes and lower-level attributes. One possible explanation for the discrepancy is that the difficulty of the attribute is not only contingent upon the nature of the information, but also the difficulty of the language to which the attribute is applied (Alderson, 2007).

In the present study, the nature of information associated with the attribute "factual information" is mostly concrete and the lexical level of the task stays within the K3 (i.e., K1, K2, K3) range, whereas information related to the other two attributes (i.e., interpretive meaning at the utterance, meaning at discourse level) is fairly abstract and the lexical level is within the K5 range (i.e., K1, K2, K3, K4, K5). It is therefore difficult to conclude whether it is the difficulty of cognitive processing or the difficulty of the language that makes the attribute "understanding factual information" the easiest one. That said, the finding that understanding factual information was the easiest attribute is in alignment with the test developers' intentions, as understanding factual information is the subskill targeting B1 students, while the other two attributes targeted B2 students according to the test specification (Fairbairn et al., 2017).

In contrast to meaning at discourse level, interpretive meaning at the utterance was found to be more difficult to master. This finding contradicts the postulation of Field's (2008, 2013) cognitive processing model that meaning at discourse level lies on the top of the hierarchy of subskills. In addition, the order of difficulty of the two attributes varied for candidates at different CEFR levels. Although interpretive meaning at the utterance level was the most difficult attribute for the high proficiency groups (i.e., C1), meaning at discourse level was the most difficult for the intermediate proficiency group (i.e., B1 and B2[1]). These findings lend support to previous claims that the same subskill could exert differential cognitive loads for candidates at different proficiency levels (Aryadoust, 2021; Ravand & Robitzsch, 2018). Another tenable explanation is that the developmental trajectory of subskills varies over proficiency levels, which may be further complicated by individual variations.

A useful approach to examining the hierarchy of attribute difficulty may be to distinguish between a strong claim and a weak claim. The strong claim, which strictly identifies higher difficulties with subskills on higher levels of the hierarchy, is usually an unattainable goal, as item difficulty is contingent on many other factors, particularly task characteristics such as the setting, test rubrics, input, expected response, and the relationship between input and response (Bachman & Palmer, 2010). A more plausible approach is to resort to the weak claim, which only requires the subskill difficulties to agree with the test design. As concerns this test form, the test design stipulates that items measuring Attribute 1 are targeted to level B1, while those measuring Attributes 2 and 3 are indiscriminately targeted to level B2. The results reported above showed that this is obviously a more attainable goal.

It should be noted, however, that Attribute 2 (i.e., interpretive meaning at the utterance level) overlapped with Attribute 3 (i.e., meaning at discourse level) in this study, such that any item measuring Attribute 2 also measured Attribute 3. It is unknown to what extent such a design may have confounded the difficulty measures of the two attributes, and/or exerted certain influence on the high correlation between A1 and A3. Although it is desirable to include items targeting interactions among attributes or processes (Harding et al., 2015), the classification accuracy may be negatively impacted if one attribute is always measured together with another attribute (Cai et al., 2018; Liu et al., 2018; Madison & Bradshaw, 2015). A more preferable design might be to measure each attribute in isolation as well as in combination with different attributes if applicable.

Overall, the findings provide partial support to the claim that the mastery probability of each target attribute is in alignment with its expected difficulty level hypothesised by the cognitive process models that guide the specification of the listening test.

[1] The number of test-takers at A0–A2 and C2 is too small to generate any convincing claim about the difficulty of attributes for those groups.

6.3 RQ3: Candidate profiles across CEFR levels

The finding that B1, B2 and C1 students displayed a wide range of subskill profiles indicates that the test may function well in diagnosing the strengths and weaknesses of B1, B2 and C1 candidiates. This is consistent with the test developers' intention that the listening test primarily targets B1, B2 and C1 students. In addition, the mastery probabilities of the three proficiency groups (i.e., B1, B2, C1) on the three attributes differed significantly from each other, indicating that the test could discriminate candidates from different proficiency levels on all three attributes.

The most representative latent subskill profiles at all CEFR levels are flat profiles, such as "000" (a non-master of all attributes) at A0–A2, B1 and B2, and "111" (a master of all attributes) at C1 and C2 levels. This finding not only converges with findings from previous studies (e.g., Aryadoust, 2021; Javidanmehra & Sarab, 2019; Lee & Sawaki, 2009a; Li et al., 2016; Ravand, 2016; Toprak et al., 2019) that report a higher proportion of flat profiles than jagged profiles, but also suggests that the subskill profiles are compatible with general listening proficiency levels, as most of the high proficiency group students (i.e., C1 and C2) were identified as mastering the three attributes (i.e., 111), and most of the low and intermediate proficiency group students (i.e. A0-A2, B1 and B2) were classified as non-masters of all the three attributes (i.e., 000).

The most representative jagged profiles differ across proficiency levels. They are "010" at B1, "010" and "101" at B2, and "101" at C1. This indicates that the most typical jagged profile for lower proficiency students is "010", while the most typical one for higher proficiency students is "101". In substantive terms, the attribute "understanding interpretive meaning at the utterance level" has the highest chance of being mastered among B1 and B2 students, while C1 students may slip on items measuring "understanding interpretive meaning at the utterance level". This, to some extent, provides rebuttal to the presupposition underlying the Aptis listening task construction that lower-level candidates can make use of lower-level cognitive processing like understanding factual information, while higher-level candidates can perform higher-level cognitive processing like synthesising information from different texts. According to the design, one may expect to see a higher frequency of the "100" and "110" profiles rather than the "010" profile.

Overall, the findings suggest that the test can be used to effectively diagnose the strengths and weaknesses of B1, B2 and C1 candidiates. However, the observed skill progression does not completely match the theoretical design of the test.

6.4 RQ4: CDA classification reliability

The psychometric quality of CDA classification results must be evaluated before they are presented to stakeholders to aid decisions (Cui et al., 2012; Mirzaei et al., 2020). The classification accuracy at test level in this study was 0.836, above the rule of thumb of 0.7 suggested by previous research (Ravand, 2016). Such similar statistics were also observed in a previous CDA retrofitting study on TOEFL listening by Liu et al. (2018), which reported classification accuracy values of .71, .81, and .73 respectively for the easy, medium, and hard item banks.

Candidates' classification accuracy at the attribute level was also examined to see how candidates were correctly classified for each attribute. The classification accuracy estimates for the attributes in this study were all above 0.90, much higher than those reported in previous listening CDA retrofitting studies (e.g., Aryadoust, 2021), which ranged from 0.073 to 0.889. This suggests that attributes could be measured more sufficiently in tests structured for diagnostic uses than in existing proficiency tests to produce reliable results (Templin & Bradshaw, 2013).

Most importantly, instead of simply presenting classification accuracy at the test and attribute levels, the present study estimated classification accuracy at the pattern level, which can inform researchers and practitioners of the effectiveness of the test in classifying candidates into specific latent profiles (Iaconangelo, 2017). The results showed that the accuracy estimates for most profiles ranged from 0.629 to 0.926, with the exception of the profile "001" and "011", which had classification accuracy estimates of 0.569 and 0.499 respectively. The low accuracy of the two profiles can be explained from psychometric and psychological perspectives. Psychometrically, it may be attributed to the small number of candidates classified into the two profiles (i.e., 25 and 15 respectively), as classification accuracy generally decreases with sample size (Rupp, 2007). Psychologically, the response patterns of these students are somewhat aberrant. Their order of acquisition stands in contrast with the theoretical and empirical difficulty of the attributes. It was mentioned above that the latent profiles of "110" and "100" agree more with the test design. Therefore, caution needs to be exercised when interpreting the classification results for this specific group of candidates.

Overall, the findings indicate that the test provides accurate classifications of candidates at the test, attribute and pattern levels, with the exception of the less frequent subskill profiles "011" and "001".

7. LIMITATIONS

The study also suffers from several limitations. One major limitation of the study pertains to the generalisability of this study. The present study mainly attempted to validate the theorised Q-matrix on which the listening test of Aptis for Teens Advanced was designed, using a limited sample in a particular country. Further replication studies using other test forms and larger sample sizes from multiple regions are needed to examine whether the findings of the present study can be generalised to the Aptis for Teens Advanced listening test in general.

The second limitation is related to Q-matrix construction in this study. Previous CDA studies combined expert analyses and learners' verbal protocols to generate the Q-matrix and constantly modified the Q-matrix to reach optimal fit (e.g., Javidanmehr & Sarab, 2019; Mirzaei et al., 2020). The exploratory approach, however, was not adopted in this study, because the test in the study is made up of item clusters representing distinct attributes, which is functionally equivalent to specifying the Q-matrix prior to test construction. Therefore, the present study adopted a confirmationist approach to validation which primarily relied on quantitative evidence to validate the theoretical model and assessment design. Future research is warranted to triangulate our findings with learner verbal protocols to provide more robust evidence for the validity of the listening test as a diagnostic instrument.

Another limitation is that this study did not extend to item-level CDA analyses to examine the compensatory and/or noncompensatory nature of the listening subskills. Such investigation would require items measuring two or more attributes in different combinations; however, the items in this study were designed to measure either one attribute, or two attributes at the same time. Future diagnostic tests should consider adopting an assessment design where each attribute is measured in isolation as well as in combination with different attributes. When such a design is available, more research can be conducted to examine how the cognitive processes interact with one another to lead to successful EFL listening, which could give theorists and practitioners a better understanding of the mechanism of listening, a skill that is severely under-researched and the least understood in research and practice (Yi, 2017).

8. CONCLUSION

8.1 Summary of findings

The present study has adopted a CDA approach to the validation of the listening subtest of Aptis for Teens Advanced, which is built on a model of cognitive processing structurally compatible with diagnostic uses. The findings indicate that overall, the listening subtest of Aptis for Teens Advanced can function as a diagnostic assessment to provide meaningful diagnostic information to inform remedial learning and instruction. The main findings of each research question are summarised below.

RQ1: To what extent are the cognitive attributes measured by Aptis listening tasks distinguishable from each other?

It was found that the cognitive attributes measured by Aptis listening tasks could be distinguished from each other, as evidenced in: 1) the moderate degree of polychoric correlation coefficents among the three attributes; and 2) slightly higher yet not overwhelming number of candidates classified into flat profiles rather than jagged profiles. The distinctness of attributes in the present study is more evident than that in previous CDA research on existing listening proficiency tests, indicating that tests built with multidimensional attributes in test specifications serve as a more viable alternative for diagnostic uses than existing unidimensional tests. Nonetheless, further revisions need to be made in Aptis listening test design or item writing to better differentiate the attribute “factual information” from “meaning at discourse level”.

RQ2: To what extent are the empirical difficulty levels of the cognitive attributes in line with the test developers’ assumptions?

The analyses showed that factual information was the easiest attribute as expected; however, interpretive meaning at the utterance level was more difficult than meaning at discourse level for C1 candidates, contradicting the postulation of Field’s (2008, 2013) cognitive processing model that meaning at discourse level lies on the top of the hierarchy of subskills. As the test developers did not stipulate the relative difficulties of these two attributes, this finding is only marked for the attention of the test developers. In addition, the hierarchy of attribute difficulty varied slightly across candidates at different CEFR levels. The results showed that the empirical difficulty levels of the cognitive attributes are partially in line with the test developers’ assumptions.

RQ3: To what extent do the candidate profiles at different CEFR levels match the test developers' assumption?

The candidate profiles at different CEFR levels partially matched the test developers' assumption.

1. B1, B2 and C1 students displayed a wide range of subskill profiles, indicating that the test may function well in diagnosing the strengths and weaknesses of B1, B2 and C1 candidates. This is consistent with the test developers' intention that the listening test primarily targets B1, B2, and C1 students.
2. The mastery probabilities of three proficiency groups (i.e., B1, B2, C1) on the three attributes differed significantly from each other, indicating that the test could discriminate candidates from different proficiency levels on the three attributes.
3. No obvious strengths or weaknesses could be identified for candidates at the lower and higher end of the ability continuum (A0–A2, C2).

RQ4: To what extent does the test produce reliable diagnostic information about candidates' attribute mastery statuses?

The classification reliability at the test, attribute and pattern levels was satisfactory, with the exception of the pattern-level reliability for two less common profiles. The results indicate that overall the test could produce reliable diagnostic information about candidates' attribute mastery statuses. However, caution needs to be exercised when interpreting the classification results for candidates in the subskill profile "001" and "011", as these two patterns not only stand in contrast with the skill progression assumed in test design from a theoretical perspective, but also produce comparatively low classification accuracy from a psychometric perspective.

8.2 Recommendations concerning Aptis

This project has a number of recommendations for the Aptis test and the British Council.

- Since the listening test of Aptis for Teens Advanced is structurally compatible with diagnostic uses, it can be used to provide meaningful added information beyond the total score.
- Field's (2008, 2013) cognitive processing model may provide a theory of diagnosis for the design of diagnostic listening tests. Since all Aptis tests are built based on Field's model, the test developers may consider examining the potential of using other sections of the Aptis for Teens Advanced (e.g., reading) and/or other tests in Aptis test system (e.g., Aptis for Teens General) for diagnostic purposes.
- The listening test of Aptis for Teens Advanced could discriminate candidates from different proficiency levels (i.e., B1, B2, C1) on the three attributes (i.e., factual information, interpretive meaning at the utterance, meaning at discourse level). It functioned quite sastifactorily in diagnosing the strengths and weaknesses of B2 and C1 candidates.
- The listening test of Aptis for Teens Advanced can provide accurate classifications of candidates at the test and attribute levels. The classification reliability of the six common subskill profiles was satisfactory, yet caution needs to be taken when interpreting the two less frequent subskill profiles "011" and "001".

- More rigorous statistical analyses can be conducted to guarantee the psychometric quality of items in field testing. Although all Aptis listening items had been trialled on a sample of over 100 candidates and found to fit the Rasch model with no misfit, it did not guarantee discrimination in the live test. If the two items with negative discrimination estimates had been replaced with higher-quality items in operational testing, the reliability and diagnostic power of the test might have been enhanced.

- Aptis test developers need to reconsider the current practice of using the same items (Items 18–25) to measure multiple attributes (interpretive meaning at the utterance and meaning at discourse level), as attributes that are measured in isolation tend to yield higher classification power.

- Aptis test developers can conduct relevant analyses to see whether each item is a typical item targeting the attribute it purports to measure, if the test is to be used for diagnostic purposes in the real world.

REFERENCES

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, *52*(3), 317–332.

Alderson, C. (1990). Testing reading comprehension skills (Part One). *Reading in a Foreign Language*, *6*, 425–438.

Alderson, C. (2000). *Assessing reading*. Cambridge University Press.

Alderson, C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. Continuum.

Alderson, C. (2007). The challenge of (diagnostic) testing: Do we know what we are measuring? In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. Turner, & C. Doe (Eds.), *Language testing reconsidered* (pp. 21–39). University of Ottawa Press.

Alderson, C. (2010). Cognitive diagnosis and Q-matrices in language assessment: A commentary. *Language Assessment Quarterly*, *7*, 96–103. http://doi.org/10.1080/15434300903426748

Anderson, J. R. (2015). *Cognitive psychology and its implications* (8th ed.). New York, NY: Worth Publishers.

Aryadoust, V. (2021). A cognitive diagnostic assessment study of the listening test of the Singapore-Cambridge General Certificate of Education O-Level: Application of DINA, DINO, G-DINA, HO-DINA, and RRUM. *International Journal of Listening*, *35*(1), 29–52. http://doi.org/10.1080/10904018.2018.1500915

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford University Press.

Baker, F. B. (1985). *The basics of item response theory*. Portsmouth, NH: Heinemann.

Baker, F. B. (1987). Methodology review: Item parameter estimation under the one-, two-, and three-parameter logistic models. *Applied Psychological Measurement 12*, 111–141. http://doi.org/10.1177/014662168701100201

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

Bradshaw, L., Izsak, A., Templin, J., & Jacobson, E. (2014). Diagnosing teachers' understandings of rational numbers: Building a multidimensional test within the diagnostic classification framework. *Educational Measurement: Issues and Practice*, *33*(1), 2–14. http://doi.org/10.1111/emip.12020

Brindley, G. (1997). Investigating second language listening ability: Listening skills and item difficulty. In G. Brindley & G. Wigglesworth (Eds.), *Access: Issues in language test design and delivery* (pp. 65–85). National Centre for English Language Teaching and Research, Macquarie University.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Sage Publications.

Buck, G. (1990). *The testing of second language listening comprehension*. Unpublished doctoral dissertation, University of Lancaster.

Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing*, *15*(2), 119–157. http://doi.org/10.1177/026553229801500201

Cai, L., Thissen, D., & du Toit, S. (2011). *IRTPRO user's guide*. Scientific Software International.

Cai, Y., Tu, D., & Ding, S. (2018). Theorems and methods of a complete Q-matrix with attribute hierarchies under restricted Q-matrix design. *Frontiers in Psychology*, *9*, 1413. http://doi.org/10.3389/fpsyg.2018.01413

Chen, J. (2017). A residual-based approach to validate Q-matrix specifications. *Applied Psychological Measurement*, *41*, 277–293. http://doi.org/10.1177/0146621616686021

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement*, *50*(2), 123–140. http://doi.org/10.1111/j.1745-3984.2012.00185.x

Chiu, C. Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement*, *37*, 598–618. http://doi.org/10.1177/0146621613488436

Cui, Y., Gierl, M. J., & Chang, H.-H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *Journal of Educational Measurement*, *49*(1), 19–38. http://doi.org/10.1111/j.1745-3984.2011.00158.x

de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement*, *45*(4), 343–362. http://doi.org/10.1111/j.1745-3984.2008.00069.x

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*(2), 179–199. http://doi.org/10.1007/s11336-011-9207-7

de la Torre, J., & Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika, 81*, 253–273. http://doi.org/10.1007/s11336-015-9467-8

DiBello, L., Roussos, L., & Stout, W. (2007). Cognitive diagnosis Part I. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics (Vol. 26): Psychometrics* (pp. 979–1030). Elsevier.

du Toit, M. (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT [Computer manual].* Scientific Software International.

Fairbairn, J., Spiby, R., & Dunlea, J. (2017). *China Teens Higher test and task design for listening, reading and writing components*. British Council.

Fernández, E. M., & Cairns, H. S. (2018). Overview. In E. M. Fernández & H. S. Cairns (Eds.), *The handbook of psycholinguistics* (pp. 185–192). John Wiley & Sons.

Field, J. (2008). *Listening in the language classroom*. Cambridge University Press.

Field, J. (2013). Cognitive validity. In L. Taylor & A. Geranpayeh (Eds.), *Examining listening* (pp. 77–151). Cambridge University Press.

Freedle, R., & Kostin, I. (1996). *The prediction of TOEFL listening comprehension item difficulty for mini-talk passages: Implications for construct validity*. TOEFL Research Report RR 96-29. Educational Testing Service.

Gierl, M. J., & Cui, Y. (2008). Defining characteristics of diagnostic classification models and the problem of retrofitting in cognitive diagnostic assessment. *Measurement: Interdisciplinary Research and Perspectives, 6*(4), 263–268. http://doi.org/10.1080/15366360802497762

Gierl, M. J., Cui, Y., & Zhou, J. (2009). Reliability and attribute-based scoring in cognitive diagnostic assessment. *Journal of Educational Measurement*, *46*, 293–313. http://doi.org/10.1111/jedm.2009.46.issue-3

Haberman, S. J., & von Davier, M. (2006). Some notes on models for cognitively based skills diagnosis. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics (Vol. 26): Psychometrics* (pp. 1031–1038). Elsevier. http://doi.org/10.1016/S0169-7161(06)26040-1

Harding, L., Alderson, C., & Brunfaut, T. (2015). Diagnostic assessment of reading and listening in a second or foreign language: Elaborating on diagnostic principles. *Language Testing*, *32*(3), 317–336. http://doi.org/10.1177/0265532214564505

Henning, G. (1992). Dimensionality and construct validity of language tests. *Language Testing*, *9*(1), 1–11. http://doi.org/10.1177/026553229200900102

Iaconangelo, C. (2017). *Uses of classification error probabilities in the three-step approach to estimating cognitive diagnosis models*. (Unpublished doctoral dissertation). Rutgers University. https://rucore.libraries.rutgers.edu/rutgers-lib/55495/PDF/1/play/

Im, S., & Corter, J. E. (2011). Statistical consequences of attribute misspecification in the rule space method. *Educational and Psychological Measurement*, *71*(4), 712–731. http://doi.org/10.1177/0013164410384855

Jang, E. E. (2010). Demystifying a Q-matrix for making diagnostic inferences about L2 reading skills: The author responds. *Language Assessment Quarterly*, *7*, 116–117. http://doi.org/10.1080/15434300903559225

Jang, E. E., Kim, H., Vincett, M., Barron, C., and Russell, B. (2019). Improving IELTS reading test score interpretations and utilisation through cognitive diagnosis model-based skill profiling. *IELTS Research Reports Online Series, No. 2.* British Council, Cambridge Assessment English and IDP: IELTS Australia.

Javidanmehr, Z., & Sarab, A. M. R. (2019). Retrofitting non-diagnostic reading comprehension assessment: Application of the G-DINA model to a high stakes reading comprehension test. *Language Assessment Quarterly*, *16*(3), 294–311. http://doi.org/10.1080/15434303.2019.1654479

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with non-parametric item response theory. *Applied Psychological Measurement*, *25*(3), 258–272. http://doi.org/10.1177/01466210122032064

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement*, *49*(1), 59–81. http://doi.org/10.1111/j.1745-3984.2011.00160.x

Lee, Y., & Sawaki, Y. (2009a). Application of three cognitive diagnosis models to ESL reading and listening assessments. *Language Assessment Quarterly*, *6*(3), 239–263. http://doi.org/10.1080/15434300903079562

Lee, Y., & Sawaki, Y. (2009b). Cognitive diagnosis approaches to language assessment: An overview. *Language Assessment Quarterly*, *6*(3), 172–189. http://doi.org/10.1080/15434300902985108

Lei, P.-W., & Li, H. (2014). *Fit indices' performance in choosing cognitive diagnostic models and Q-matrices*. Paper presented at the Annual Meeting of the National Council on Measurement in Education (NCME), Philadelphia, PA.

Lei, P.-W., & Li, H. (2016). Performance of fit indices in choosing correct cognitive diagnostic models and Q-matrices. *Applied Psychological Measurement, 40*(6), 405–417. http://doi.org/10.1177/0146621616647954

Li, H., & Suen, H. K. (2013). Constructing and validating a Q-matrix for cognitive diagnostic analyses of a reading test. *Educational Assessment*, *18*(1), 1–25. http://doi.org/10.1080/10627197.2013.761522

Li, H., Hunter, C. V., & Lei, P-W. (2016). The selection of cognitive diagnostic models for a reading comprehension test. *Language Testing*, *33*(3), 391–409. http://doi.org/10.1177/0265532215590848

Liu, H., You, X., Wang, W., Ding, S., & Chang, H.-H. (2014). Large-scale implementation of computerized adaptive testing with cognitive diagnosis in China. In Y. Cheng & H.-H. Chang (Eds.), *Advancing methodologies to support both summative and formative assessments* (pp. 245–261). Information Age Publishing, Inc.

Liu, J. C. (2016). On the consistency of Q-matrix estimation: A commentary. *Psychometrika*, *82*, 523–527. http://doi.org/10.1007/s11336-015-9487-4

Liu, R. (2018). Misspecification of attribute structure in diagnostic measurement. *Educational and Psychological Measurement*, *78*(4), 605–634. http://doi.org/10.1177/0013164417702458

Liu, R., Huggins-Manley, A. C., & Bulut, O. (2018). Retrofitting diagnostic classification models to responses from IRT-based assessment forms. *Educational and Psychological Measurement*, *78*(3), 357–383. http://doi.org/10.1177/0013164416685599

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates, Inc.

Lumley, T. J. N. (1993). Reading comprehension sub-skills: teachers' perceptions of content in an EAP test. *Melbourne Papers in Applied Linguistics*, *2*, 25–55.

Ma, W., de la Torre, J., Sorrel, M., & Jiang, Z. (2020). *Package 'GDINA'*. Retrieved from https://cran.r-project.org/web/packages/GDINA/GDINA.pdf

Madison, M. J., & Bradshaw, L. P. (2015). The effects of Q-matrix design on classification accuracy in the log-linear cognitive diagnosis model. *Educational and Psychological Measurement*, *75*(3), 491–511. http://doi.org/10.1177/0013164414539162

Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research*, *49*(4), 305–328. http://doi.org/10.1080/00273171.2014.911075

Min, S., & He, L. (2022). Developing individualized feedback for listening assessment: Combining standard setting and cognitive diagnostic assessment approaches. *Language Testing*, 39(1), 90–116. http://doi.org/10.1177/026553222199547

Mirzaei, A., Vincheh, M. H., & Hashemian, M. (2020). Retrofitting the IELTS reading section with a general cognitive diagnostic model in an Iranian EAP context. *Studies in Educational Evaluation*, *64*, 1–10. http://doi.org/10.1016/j.stueduc.2019.100817

Nájera, P., Sorrel, M. A., & Abad, F. J. (2019). Reconsidering cutoff points in the general method of empirical Q-matrix validation. *Educational and Psychological Measurement, 79*(4), 727–753. http://doi.org/10.1177/0013164418822700

Neyman, J., & Pearson, E. S. (1992). On the problem of the most efficient tests of statistical hypotheses. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in statistics* (pp. 73–108). Springer.

Oller, J. W., Jr. (1983). Evidence for a general proficiency factor: An expectancy grammar. In J. W. Oller Jr. (Ed.), *Issues in language testing research* (pp. 3–10). Newbury House Publishers.

O'Sullivan, B., Dunlea, J., Spiby, R., Westbrook, C., & Dunn, K. (2020). *Aptis General Technical Manual Version 2.2* (Technical Report TR/2020/001). British Council. https://www.britishcouncil.org/sites/default/files/aptis_technical_manual_v_2.2_final.pdf

Pae, H. K., & Greenberg, D. (2014). The relationship between receptive and expressive subskills of academic L2 proficiency in nonnative speakers of English: A multigroup approach. *Reading Psychology*, *35*(3), 221–259. http://doi.org/10.1080/02702711.2012.684425

Ranjbaran, F., & Alavi, S. M. (2017). Developing a reading comprehension test for cognitive diagnostic assessment: A RUM analysis. *Studies in Educational Evaluation*, *55*, 167–179. http://doi.org/10.1016/j.stueduc.2017.10.007

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* The University of Chicago Press.

Ravand, H. (2016). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. *Journal of Psychoeducational Assessment*, *34*(8), 782 –799. http://doi.org/10.1177/0734282915623053

Ravand, H., & Robitzsch, A. (2018). Cognitive diagnostic model of best choice: a study of reading comprehension, *Educational Psychology*, *38*(10), 1255–1277. http://doi.org/10.1080/01443410.2018.1489524

Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement*, *47*(3), 361–372. http://doi.org/10.1111/j.1745-3984.2010.00118.x

Romero, S. J., Ordonez, X. G., Ponsoda, V., & Revuelta, J. (2014). Detection of Q-matrix misspecification using two criteria for validation of cognitive structures under the least squares distance model. *Psicologica: International Journal of Methodology and Experimental Psychology*, *35*(1), 149–169.

Rost, M. (2016). *Teaching and researching listening* (3rd ed.). Longman.

Roussos, L. A., DiBello, L. V., Stout, W., Hartz, S. M., Henson, R. A., & Templin, J. L. (2007). The fusion model skills diagnosis system. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 275–318). Cambridge University Press.

Rupp, A. A. (2007). The answer is in the question: A guide for describing and investigating the conceptual foundations and statistical properties of cognitive psychometric models. *International Journal of Testing*, *7*, 95–125. http://doi.org/10.1080/15305050701193454

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford Press.

Sawaki, Y., Kim, H. J., & Gentile, C. (2009). Q-Matrix construction: Defining the link between constructs and test items in large-scale reading and listening comprehension assessments. *Language Assessment Quarterly*, *6*(3), 190–209. http://doi.org/10.1080/15434300902801917

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

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464. http://doi.org/10.1214/aos/1176344136

Sessoms, J. & Henson, R. A. (2018). Applications of diagnostic classification models: A literature review and critical commentary. *Measurement: Interdisciplinary Research and Perspectives*, *16*(1), 1–17. http://doi.org/10.1080/15366367.2018.1435104

Sinharay, S., Puhan, G., & Haberman, S. J. (2010). Reporting diagnostic scores in educational testing: Temptations, pitfalls, and some solutions. *Multivariate Behavioral Research*, *45*(3), 553–573. http://doi.org/10.1080/00273171.2010.483382

Sinharay, S., Puhan, G., Haberman, S. J., & Hambleton R. K. (2019). Subscores: when to communicate them, what are their alternatives, and some recommendations. In D. Zapata-Rivera (Ed.), *Score reporting research and applications* (pp. 80–107). Routledge Taylor & Francis Group.

Song, M.-Y. (2008). Do divisible subskills exist in second language (L2) comprehension? A structural equation modeling approach. *Language Testing*, *25*(4), 435–464. http://doi.org/10.1177/0265532208094272

Stout, W., Henson, R., DiBello, L, & Shear, B. (2019). The reparameterized unified model system: A diagnostic assessment modeling approach. In: M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models* (pp. 47–79). Springer.

Swaminathan, H., & Gifford, J. A. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika*, *50*, 349–364. http://doi.org/10.1007/BF02294110

Templin, J., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification*, *30*(2), 251–275. http://doi.org/10.1007/s00357-013-9129-4

Templin, J., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*(3), 287–305. http://doi.org/10.1037/1082-989X.11.3.287

Tengberg, M. (2018). Validation of sub-constructs in reading comprehension tests using teachers' classification of cognitive targets. *Language Assessment Quarterly*, *15*(2), 169–182. http://doi.org/10.1080/15434303.2018.1448820

Toprak, E., Aryadoust, V., & Goh, C. (2019). The log-linear cognitive diagnosis modeling (LCDM) in second language listening assessment. In V. Aryadoust & M. Raquel (Eds.), *Quantitative data analysis for language assessment Volume II: Advanced methods* (pp. 56–78). Routledge.

Vandergrift, L. (2007). Recent developments in second and foreign language listening comprehension research. *Language Teaching*, *40*, 191–210. http://doi.org/10.1017/S0261444807004338

Wang, W., Song, L., Chen, P., Meng, Y., & Ding, S. (2015). Attribute-level and pattern-level classification consistency and accuracy indices for diagnostic assessment. *Journal of Educational Measurement*, *52*(4), 457–476. http://doi.org/10.1111/jedm.12096

Weir, C. J., Yang, H., & Jin, Y. (2000). *An empirical investigation of the componentiality of L2 reading in English for academic purposes*. Cambridge University Press.

Yi, Y.-S. (2017). Probing the relative importance of different attributes in L2 reading and listening comprehension items: an application of cognitive diagnostic models. *Language Testing*, *34*(3), 1–19. http://doi.org/10.1177/0265532216646141

Yu, X., & Cheng, Y. (2019). Data-driven Q-matrix validation using a residual-based statistic in cognitive diagnostic assessment. *British Journal of Mathematical and Statistical Psychology*, *73*, 145–179. http://doi.org/10.1111/bmsp.12191

British Council Assessment Research Awards and Grants

If you're involved or work in research into assessment, then the British Council Assessment Research Awards and Grants might interest you.

These awards recognise achievement and innovation within the field of language assessment and form part of the British Council's extensive support of research activities across the world.

A CASE FOR THE DIAGNOSTIC USE OF THE LISTENING TEST OF APTIS FOR TEENS ADVANCED: A COGNITIVE DIAGNOSTIC ASSESSMENT STUDY

AR-G/2024/1

Shangchao Min
Hongwen Cai

ARAGs RESEARCH REPORTS ONLINE

ISSN 2057-5203

The British Council is the United Kingdom's international organisation for cultural relations and educational opportunities**.**

ISSN 2057-5203

9 772057 520008 >

www.britishcouncil.org/aptis/research