BRITISH COUNCIL

ENGLISH LANGUAGE
ASSESSMENT RESEARCH GROUP

FACTOR STRUCTURE AND FOUR-SKILL PROFILES OF THE APTIS TEST

AR-G/2021/2

Yo In'nami, Chuo University, Japan
Rie Koizumi, Juntendo University, Japan

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

ABSTRACT

This project examines the factor structure and examinees' skill profiles of the Aptis test to contribute to its validation agenda. Examining the factor structure of the Aptis test will allow us to examine whether, and to what degree, the skills intended to be measured by the Aptis team are indeed measured by the test. Additionally, research into Aptis examinees' skill profiles will help identify the characteristics of each skill, revealing examinees' strengths and weaknesses.

Concerning the factor structure of the Aptis test, when all seven countries were analysed as a single dataset, the results from confirmatory factor analyses showed that their factor structure was best explained by a bi-factor model, where a single, general L2 proficiency factor, as well as four skills of listening, reading, speaking and writing, influenced test scores directly. When the data were separately analysed for each country, six of those countries (Chile, Indonesia, Mexico, Poland, Spain, and Sri Lanka) had a factor structure that was best explained by a higher-order model. Meanwhile, Bangladesh revealed its structure as best explained by a bi-factor model. Further, multiple-sample analyses across countries showed that a higher-order model with no equality constraints was most consistent with the data.

The results showed that one of the following two models was supported: (a) the high factor loadings from the higher-order, L2 proficiency factor to each skill factor in the higher-order models; and (b) the high factor loadings from the single, general L2 proficiency factor to each subscore, along with the high correlations among skill factors in the bi-factor models. It was concluded that both factor structures supported reporting a total score along with separate scores for each skill, as provided for in Aptis score reports.

With regard to the examinees' skill profiles on the Aptis test, the results from latent profile analyses showed that examinees were classified into five profiles (i.e., groups). One of the groups demonstrated low performance in general, with worse performance on easier tasks and better performance on more difficult ones, especially in the listening, reading and speaking sections. Another group showed average performance in reading, listening and writing but extremely low performance in speaking. Recommendations for further improvement of the Aptis test are presented.

Authors

Yo In'nami is a Professor of English at Chuo University in Japan. He is interested in meta-analytic inquiry into the variability of effects and the longitudinal measurement of change in language proficiency. His publications have appeared in *International Journal of Testing*, *Language Assessment Quarterly*, *Language Learning*, *Language Testing*, *Language Testing in Asia*, *System*, and *TESOL Quarterly*.

Rie Koizumi is an Associate Professor of English at Juntendo University in Japan. Her research interests include assessing and modeling second language ability, performance and development. She has published her work in *Language Testing*, *Language Assessment Quarterly*, *System*, *Assessment in Education: Principles, Policy & Practice*, and other journals.

Acknowledgements

We would like to thank the British Council's enormous support for funding and guidance, including Karen Dunn for providing us with data files for this project, Karen, Jamie Dunlea, and Richard Spiby for answering questions and offering invaluable advice throughout the study, a reviewer for his/her comments on our progress report, and Mina Patel for keeping us on the track. We are also grateful to Saori Kubo and Eric H. Setoguchi for their statistical advice. All remaining errors are our own.

CONTENTS

List of figures

List of tables

Appendix B tables

Appendix C table

1. INTRODUCTION

As an innovative and newly developed English assessment tool from the British Council, the Aptis test has been gaining popularity in many parts of the world. It is used to measure the English language proficiency of students, teachers or employees in organisations and institutions. To ensure, and maintain, its high quality, the Aptis test was developed based on Weir's (2005) socio-cognitive framework, with validation studies conducted based on O'Sullivan (2011) and O'Sullivan and Weir (2011). While validity evidence for the Aptis test has been accumulated as reported in O'Sullivan (2012) and O'Sullivan and Dunlea (2015) and as validation studies are listed online (British Council, 2018a), there are two pertinent issues that need to be addressed. The first issue is to examine what the Aptis test actually measures, which may differ from that intended by the Aptis team, due, for example, to test formats and practical constraints despite the developer's concerted efforts to assess the intended construct. It is important to confirm that what is actually measured matches the intended test constructs. The finding indicates if, and to what degree, interpretation and use on the basis of Aptis test scores are appropriate.

Another issue to examine is the skill profiles of Aptis examinees. Although total test scores reflect overall proficiency and are more likely to be used in situations such as the initial screening of job applicants, skill profiles are more informative and valuable to examinees as they indicate strengths and weaknesses in their performance skills. Profiles may be flat (indicating that the four skills are at the same level) or uneven/jagged/non-flat (indicating that some skills are higher in proficiency than others). While skill profiles comprise an important piece of validity evidence, particularly for diagnostic purposes, empirical investigation into this area has been limited (Hulstijn, 2011, 2015).

Examining these two issues will provide one piece of validity evidence toward a better understanding of the construct measured by the Aptis test and the inferences we can draw from the test performance. The current project aims to achieve these goals by investigating:

1. the factor structure of the Aptis test in Study 1
2. the skill profiles of the Aptis test examinees in Study 2.

2. BACKGROUND

Language tests are used to make decisions for recruitment, admission, and passing/failure of a course, among others. A central theme in validity and validation is whether, and to what extent, the decisions made based on such test scores are valid. The importance of making valid decisions must be examined through validation studies for all language tests, including the Aptis test.

The Aptis test was developed based on two frameworks. One was Weir's (2005) socio-cognitive framework, where validation evidence consisted of context validity, theory-based validity, scoring validity, consequential validity, and criterion-related validity. The other framework was a reconceptualisation of Weir's (2005) socio-cognitive framework, where the types of validation evidence needed to make valid decisions were focused on relating to the test (i.e., task), the test-taker (e.g., cognitive processes), and the scoring, with more detailed classifications. Although both frameworks were used to develop the Aptis test, the latter, reconceptualised framework served as the basis for actual validation studies on the Aptis test, as shown in O'Sullivan (2011, 2012), O'Sullivan and Dunlea (2015), and O'Sullivan and Weir (2011). More validation studies are now listed online (British Council, 2018a), accumulating evidence to indicate the degree to which valid inferences can be drawn from Aptis test scores (for the historical background of the test, see Weir and O'Sullivan, 2017).

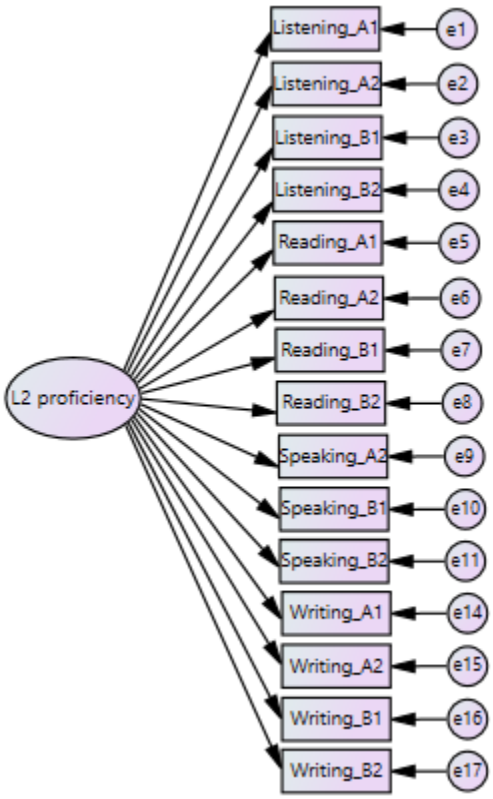
The current project was designed to contribute to this endeavor and accumulate validity evidence on the Aptis test by examining the factor structure of the test (Study 1) and investigating the skill profiles of Aptis test examinees (Study 2). The factor structure of a test seems to be situated at scoring validity in Weir's (2005) socio-cognitive framework and his reconceptualised framework, as Weir writes that scoring validity concerns the "internal reliability/validity of items" (p. 48). To the best of our knowledge, skill profiles are not explicitly described in these two frameworks and do not clearly exhibit how they are situated therein. However, information on learners' skill profiles describes how the relationship among constructs concurs with the underlying theory, and how score patterns, as often reported in score reports, demonstrate learners' strengths and weaknesses. Since both issues concern the meaning of test scores, researching skill profiles could likewise relate to scoring validity.

The factor structure of tests has been widely examined in existing literature. In a study on the factor structure of the Test of English for International Communication (TOEIC) designed to measure listening and reading, the developer's intended factor structure was found to concur with those operationalised in the test (In'nami & Koizumi, 2012). Similar findings were also reported for the Test of English as a Foreign Language Internet-based (TOEFL iBT®), showing that what was actually measured matched the intended test constructs (Sawaki & Sinharay, 2013, 2018; Sawaki, Stricker & Oranje, 2009).

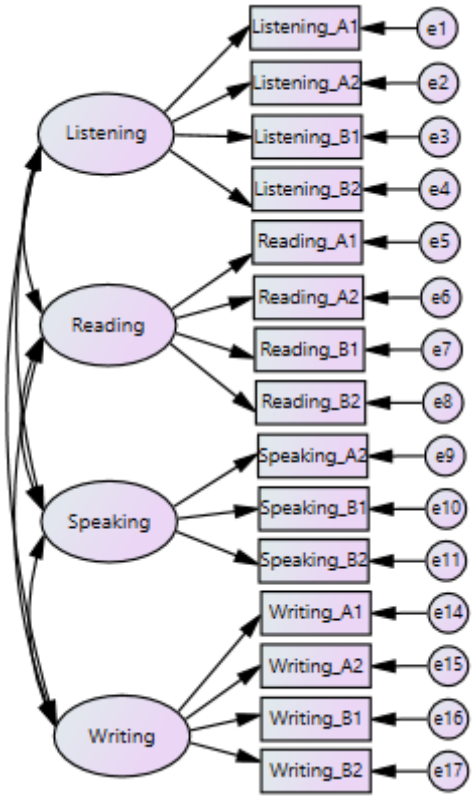
Behind this inquiry into the factor structure of tests lies a long-standing investigation into the factor structure of second language (L2) proficiency (e.g., Bachman & Palmer, 1982, 1989; In'nami, Koizumi & Nakamura, 2016; Llosa, 2007; Sawaki, 2007; Sawaki et al., 2009; Shin, 2005; Stricker & Rock, 2008). A structure comprising two or more factors is called multicomponential and has received more supporting evidence overall (e.g., Sawaki et al., 2009) as compared to a single-factor structure, where only one ability is hypothesised to explain score variance. Within multicomponential models, there have been different ways in which abilities have been structured: whether they are correlated (Model B or C in Figure 1), whether they are hierarchical with a general ability – L2 proficiency – to influence test performance through four-skill abilities (Model D), or whether such a general ability influences test performance directly (Model E). However, Harsch (2014) argues that "language proficiency can be conceptualised as single-factor and divisible [multicomponential], depending on the level of abstraction and the purpose of the assessment and score reporting" (p. 153). This suggests that in some cases, a single-factor (i.e., unitary) structure is possible.

We will examine which structure is supported by empirical data in the Aptis test: single-factor or multicomponential. This query is studied by investigating measurement invariance of data from the wide range of test-taker samples from different countries. For the applications of invariance testing in language testing, see Bae and Bachman (1998), Bae and Lee (2011), In'nami and Koizumi (2012), Purpura (1998), and Sawaki and Sinharay (2013, 2018).

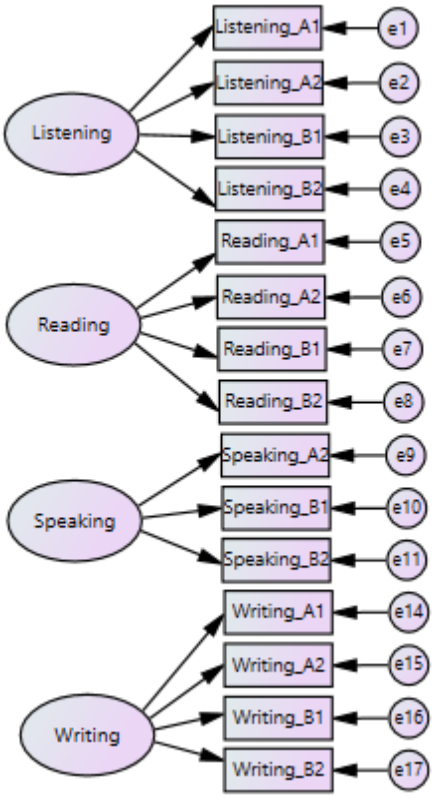
Figure 1: Five models. All factors are correlated in Model B



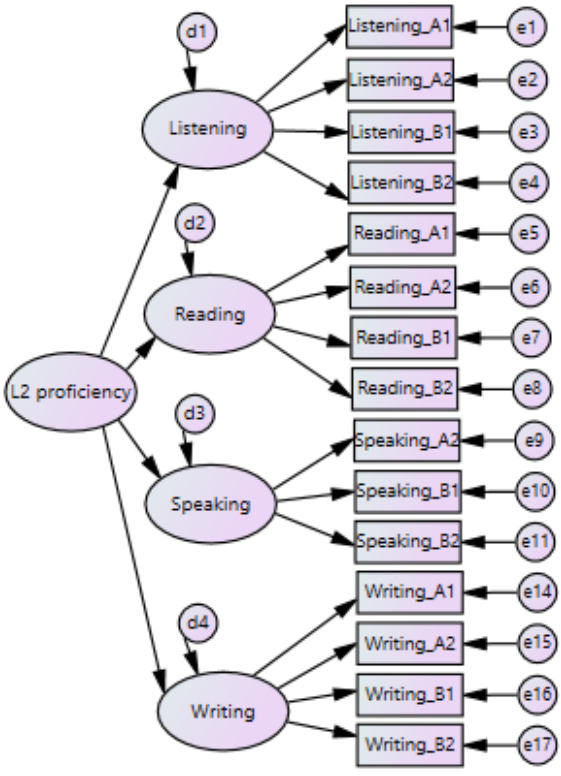
Model A: Single-factor model



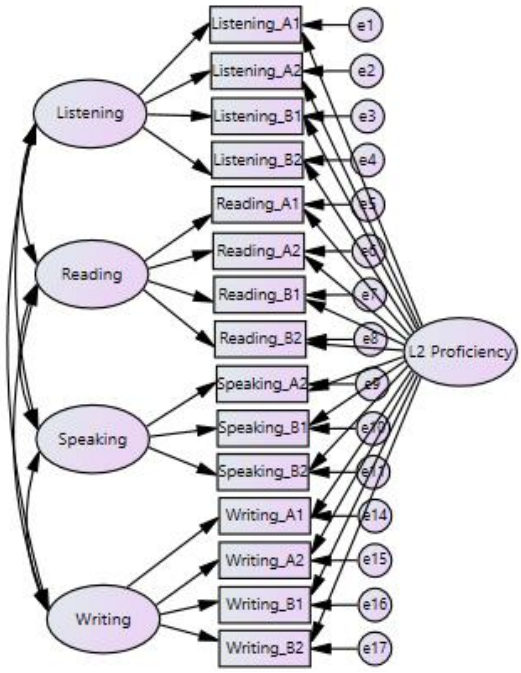
Model B: Correlated four-factor model



Model C: Uncorrelated four-factor model



Model D: Higher-order model



Model E: Bi-factor model

In contrast, learners' skill profiles and subskill profiles remain largely under-investigated; more research was recently called for by: Huhta, Alderson, Nieminen and Ullakonoja (2015); Hulstijn (2011, 2015); Hulstijn, Schoonen, De Jong, Steinel and Florijn (2012); and Harsch (2014). This is because skill profiles could be essential in providing diagnostic information on examinees' strengths and weaknesses. Such information can be used in many ways. For example, it would help learners plan how to improve their weak skills or further hone their strengths. Placement course directors would place students in courses that better meet their needs; for example, students with high reading ability but poor speaking ability would be placed in higher reading courses and lower speaking courses. In gate-keeping contexts, such as screening jobs or university applicants, committee members would select candidates whose skill profiles meet the requirements of their job description or admission criteria. As better decisions could be made based on sound skill profiles corroborated by empirical studies, research into skill profiles is a vital aspect of test validation when the test is intended to serve diagnostic and other purposes, as exemplified above. Thus, validation studies should cover whether profiles on score reports are useful and what typical and atypical profiles look like among test-takers. Even flat skill profiles should be useful to test users, as they can, for example, serve as a reference point against which they can compare themselves with other learners per skill.

Despite these advantages of skill profiles, they have been under-utilised in practice and under-researched, with some exceptions described below. For example, Xi and Mollaun (2006) compared speaking performance across three criteria for analytic scoring (i.e., delivery, language use and topic development) as part of their investigation into the value of analytic scoring for the TOEFL Academic Speaking Test. The three scoring criteria with data obtained from 140 non-native speakers in the U.S. were highly or perfectly correlated (observed correlations for all tasks = .93 to .95; disattenuated correlations for all tasks = .98 to 1.00). This suggests little "profile variability" (p. 30) and limited use of analytic scoring for the TOEFL Academic Speaking Test.

Granfeldt and Ågren (2013) examined whether uneven profiles were present in written texts produced by 38 Swedish students of L3 French. The written data were analysed using ratings of the Common European Framework of Reference for Languages (CEFR) and Processing Theory ratings. The results showed an overall strong correlation between the CEFR ratings and Processing Theory (r_s = .86). Uneven profiles were observed, with communicative proficiency more advanced than morphosyntactic development and vice versa.

Huhta et al. (2015) attempted to create diagnostic profiles of foreign language readers and writers in their DIALUKI project (for details, see Alderson, Haapakangas, Huhta, Nieminen & Ullakonoja, 2015). Cognitive and linguistic tasks and motivation and background questionnaires were administered to Finnish-speaking learners of English as a foreign language and to Russian-speaking learners of Finnish as a second language. Both groups consisted of 8th graders and gymnasium students. Preliminary results with latent profile analysis indicated some variability in profile in foreign language reading and other skills. For example, both groups were weak in reading and strong in writing. They differed in that one group demonstrated lower ability to recognise word boundaries than the other group. This suggests even profiles between groups of similar traits and uneven profiles between skills.

Sawaki and Sinharay (2013) investigated the skill profiles of the TOEFL iBT test using standardised section scores. They reported three or four types of profiles, stating that the profiles were all flat, with four skills belonging to relatively the same level (e.g., low, medium and high levels). Their study demonstrated the difficulty of obtaining uneven profiles in the TOEFL iBT test, especially with a large number of examinees and a diverse population (n = 8,710 to 14,495, with Arabic, Korean, Spanish, and others as first languages).

Ginther and Yan (2018) reports on the relationship between the TOEFL iBT scores and the grade point average of first-year Chinese students at a university in the U.S. As part of this investigation, the participants' descriptive statistics for the TOEFL iBT were calculated. The results showed that, on average, reading scores were the highest (e.g., 24.02 out of 30 in 2011) and the speaking scores were the lowest (e.g., 19.75 out of 30 in 2011). The same results held across three academic years.

Finally, Koizumi and her associates have investigated skill profiles of Japanese learners of English using various four-skill standardised tests (e.g., Koizumi, 2015, using TOEIC®; Koizumi, Agawa & Asano, 2018, using the TOEFL iBT). Koizumi, Kashimada and Akimoto (2019) reported that skill profiles observed across five four-skill tests were similar: There was a limited proportion of learners with even profiles (9% to 29%) and learners with various types of uneven profiles; 20% to 40% of learners had speaking as their lowest skill, whereas 19% to 32% had reading as their highest skill.

Previous studies have shown that test results can reveal either even (or flat) or uneven profiles of skills and subskills. The aim of our second study is to expand this line of research (Study 2) by focusing on investigating the skill profiles of the Aptis test examinees. Accumulating more findings in addition to those obtained from previous studies will be useful in helping us better understand how various language features are related and thereby, how flat (e.g., Koizumi et al., 2019; Sawaki & Sinharay, 2013; Xi & Mollaun, 2006) or uneven (Ginther & Yan, 2018; Granfeldt & Ågren, 2013; Huhta et al., 2015) the profiles are.

3. STUDY 1: FACTOR STRUCTURE OF THE APTIS TEST

The purposes of Study 1 were two-fold: to investigate the factor structure of the Aptis test, and to examine whether the same structure holds across countries. Two research questions were examined.

1. **What is the factor structure of the Aptis test when all countries are analysed together, as well as individually?**
2. **To what extent does measurement invariance hold across countries?**

We examined the factor structure of the Aptis test by country and across countries. The rationale behind this was that widespread administration and use of the Aptis test in numerous countries could suggest that its scores might be a function of many test and examinee variables. As country is a variable that is related to scores (e.g., Sawaki & Sinharay, 2013, 2018; Sawaki, Stricker & Oranje, 2009), it would be of great interest and importance to examine the extent to which this would be the case for the Aptis test. This does not mean that other variables such as language, gender, age and language proficiency merit no research. They should be examined in future studies as well.

3.1 Method

3.1.1 Data

We analysed international comparison data collected by the British Council in 2015. The data included three versions of the Aptis test. Examinees took any one of the versions for each section: grammar and vocabulary, listening, reading, speaking, and writing skills. For example, some examinees took Version 1 for all skills, whereas others took Version 1 for grammar and vocabulary, Version 3 for listening, and Version 2 for reading, speaking, and writing. Some examinees only took grammar and vocabulary tests and did not undertake tests in the other four skills. We found 367 combinations for taking the tests among a total of 6,255 examinees. Of these, 1,493 examinees took Version 1 for all skills, which was by far the largest in number compared with the next largest number of examinees (n = 48) who took Version 1 for all skills excluding reading, which they did not take.

From these 1,493 examinees, we used the data of 1,270 examinees in seven countries: Bangladesh (403), Chile (117), Indonesia (95), Mexico (331), Poland (100), Spain (137), and Sri Lanka (87).[1]

[1]Note: The sample size was adequate for examining the five models for all countries combined and for each country, and for examining invariance across countries, based on the results for Monte Carlo simulation studies with Mplus (for simulation procedures, see for example In'nami & Koizumi, 2013; Muthén & Muthén, 2002). More specifically, based on Muthén and Muthén (2002), precision and power for each parameter in a model were estimated. Parameters were specified, and 10,000 samples (replications) were generated in each run. Results over the 10,000 replications were summarised.
Muthén and Muthén (2002), as summarised in In'nami and Koizumi (2013), is as follows:

> To determine if the sample size for a model is sufficient in terms of precision and power, precision of parameter estimates was examined using four criteria, and power using one criterion, following Muthén and Muthén (2002). First, parameter bias should not exceed |10%| for any parameter in the model. Second, this should also be the case for standard error bias. Third, the standard error bias for the parameter for which power is of particular interest should not exceed |5%|. Fourth, 95% coverage—the proportion of replications for which the 95% confidence interval covers the population parameter value—should fall between .91 and .98. One minus the coverage value equals the alpha level of .05. Coverage values should be close to the correct value of .95. Finally, power was evaluated as to whether it exceeded .80—a commonly accepted value for sufficient power (e.g., Cohen, 1988). Results based on these five criteria—parameter bias, standard error bias, standard error bias for the parameter of interest, 95% coverage, and power—were first calculated for each parameter for each model. Then the results for each criterion were averaged across models. (pp. 336–337)

Results showed that all the criterion above were overall satisfied, for example, with a small percentage of underpowered (i.e., less than .80) parameters found for all countries combined (12 [4.124%]) and for each country (2 [0.926%] for Bangladesh, 15 [6.944%] for Indonesia, 4 [1.852%] for Mexico, 1 [0.641%] for Poland, 8 [3.704%] for Spain, 6 [2.778%] for Sri Lanka).

First, in consultation with the British Council, we removed candidates with automatically generated reference numbers (e.g., AUTOee7d37b3c8fe4ba99b8e7f0e2169d465) because they are "more likely to be teachers/administrators checking the system, rather than…test-takers. Project participants should all have a manually allocated number" (British Council, personal communication, 15 May 2018), and kept those with numbers that had been manually allocated (e.g., ER16030, TT48, ICR0022). Second, we also arranged the data according to the number of examinees for each country and removed six countries that were judged to have too few examinees to enable analysis using confirmatory factor analysis. The country with the smallest number of examinees included in our analysis was Sri Lanka (n = 87). Algeria (n = 70) and the remaining five countries with smaller numbers of examinees were removed from the analysis. Finally, although this did not influence the data size, vocabulary and grammar tests were excluded from the analysis. This was because relationships between vocabulary and grammar, four skills, and L2 English proficiency were not clear enough to posit factor structures. As O'Sullivan and Dunlea (2015) commented, their score is reported separately from the total score and each skill score in a score report, given that the relationship between vocabulary and grammar and CEFR levels is not yet clear enough to judge the CEFR level of learners based on their performance in vocabulary and grammar (Council of Europe, 2001).

In the end, our data consisted of 1,270 examinees in seven countries. Note that the countries are the location where examinees took the Aptis test and may not necessarily indicate their nationality, but could be used as a proxy for nationality, as in this study. Generally, for each country, examinees were in an education or employment sector.

3.1.2 Test content and format

According to O'Sullivan and Dunlea (2015), the Aptis test is a non-adaptive, computer-based test. The listening and reading sections use selected-response formats, whereas the speaking and writing sections elicit examinees' open-ended performances.

The listening section had four parts: Part 1 corresponded to level A1 (5 items; 5 points in total); Part 2 corresponded to level A2 (6 items; 6 points in total); Part 3 corresponded to level B1 (7 items; 7 points in total); and Part 4 corresponded to level B2 (7 items; 7 points in total). The terms A1, A2, B1 and B2 refer to the language levels outlined in the CEFR. Part 1, lexical recognition, referred to the ability to understand specific information, such as numbers and names, in short monologues. Part 2, the identification of specific factual information, referred to the ability to understand specific information, such as concrete messages, in short monologues and conversations. Part 3 was the same as Part 2, the only difference being that it required information to be integrated from several parts of the text. Finally, Part 4, meaning representation and inference, referred to the ability to understand a speaker's attitude, opinion or intention by integrating information from several parts of the text. All items were presented in a four-option multiple choice format. Item-level data were available for the listening section, but, as explained below, they were aggregated to create Part (i.e., task)-level, composite scores.

The reading section had four parts: Part 1 corresponded to level A1 (1 task; 5 points in total); Part 2 corresponded to level A2 (1 task; 6 points in total); Part 3 corresponded to level B1 (1 task; 7 points in total); and Part 4 corresponded to level B2 (1 task; 7 points in total). Part 1, sentence level meaning, referred to careful, local-level reading in a three-option multiple-choice format. Part 2, inter-sentential cohesion, referred to careful, global-level reading, which required examinees to re-order jumbled sentences to complete a story. Part 3, text-level comprehension of short texts, referred to careful, global-level reading in which examinees filled each gap by selecting the most appropriate word among the options. Finally, Part 4, text-level comprehension of a long text, referred to careful and expeditious, global-level reading in which examinees filled each gap by selecting the most appropriate heading for each paragraph among the options. Part (i.e., task)-level, composite data were available for the reading section, with each examinee receiving a single total score for each focus area.

The speaking section used a semi-direct format of recording examinees' talk that raters evaluated afterwards. It had four parts: Part 1 corresponded to level A2 (1 task; 5 points in total); Parts 2 and 3 corresponded to level B1 (1 task; 5 points in total, for each); and Part 4 corresponded to level B2 (1 task; 6 points in total). Part 1 involved providing personal information, such as the examinee's name and some background information. Part 2 involved describing and expressing opinions, and providing reasons and explanations through being shown a photograph and describing and answering questions related to it. Part 3 involved describing, comparing, contrasting, as well as providing reasons and explanations: Examinees described, contrasted and compared two photographs, and answered relevant questions. Finally, Part 4 tested the examinee's ability to describe ideas in an integrated manner, while discussing opinions on an abstract topic. Again, Part (i.e., task)-level, composite data were available for the speaking section, with each examinee receiving a single total score for each focus area.

The writing section had four parts: Part 1 corresponded to level A1 (1 task; 5 points in total); Part 2 corresponded to level A2 (1 task; 5 points in total); Part 3 corresponded to level B1 (1 task; 5 points in total); and Part 4 corresponded to level B2 (1 task; 6 points in total). Part 1 involved providing word/phrase-level personal information, such as the examinee's name and some background information. Part 2 involved providing a sentence-level, short description of personal or concrete information. Part 3 required examinees to compose short, paragraph-level responses to a text, such as a question posted on an online message board. Finally, Part 4 tested the ability to write longer, paragraph-level responses in both informal and formal contexts. Again, Part (i.e., task)-level, composite data were available for the writing section, with each examinee receiving a single total score for each focus area.

The listening, reading, and writing sections consisted of A1 to B2 tasks, whereas the speaking section consisted of A2 to B2 tasks. Part-level, composite data were available for the reading, speaking, and writing tests. For the listening section data, dichotomous, item-level data were available, and the total scores in each part were summed into Part-level, composite data. The summation of scores is technically termed parceling; the product of parceling is called a parcel or a parcel score.

3.2 Analyses

To examine Research Question 1, confirmatory factor analyses were used to examine the factor structure of the Aptis test by testing five models that hypothesised relationships among variables:

- a single-factor model
- a correlated four-factor model
- an uncorrelated four-factor model
- a higher-order model
- a bi-factor model.

These models are presented in Figure 1. In each figure, the rectangles represent the observed variables, the ovals represent latent factors, and the circles represent measurement errors or residuals. The observed variables in this study were: (a) item-level dichotomous data for the listening section that were aggregated to create part-level, composite scores; and (b) part-level, composite data for the reading, speaking, and writing tests. The unidimensionality of each parcel for the listening section was examined and confirmed in order to aggregate the item-level score to form a parcel score (e.g., Little, Cunningham & Shahar, 2002; Little, Rhemtulla, Gibson & Schoemann, 2013; Meade & Kroustalils, 2006).

Among the five models, the higher-order model and bi-factor model may appear similar. They represent the structure of language ability that consists of L2 proficiency and four skills of listening, reading, speaking, and writing. This means that test scores are explained by L2 proficiency and the four skills. However, the models differ in two ways in which observed variables (or test scores in the current study) are explained (Dunn & McCray, 2020; Kline, 2016). First, in the higher-order model, L2 proficiency is assumed to influence test scores only through the four skills. In other words, there is no direct impact of L2 proficiency on test scores. In contrast, in the bi-factor model, such direct relationship is modeled between L2 proficiency and test scores. Second, in the higher-order model, L2 proficiency is modelled to influence test scores through the four skills. In this respect, L2 proficiency is directly related to the four skills. In the bi-factor model, L2 proficiency is modelled to be unrelated to the four skills. Thus, the relationships among L2 proficiency, the four skills, and test scores differ across the two models. The question of whether a higher-order or bi-factor model best represents the factor structure of the Aptis test had not been previously examined.

Adopting the same factor structure across countries is important when scores are compared across countries. However, in case of a score reporting format, adopting either the higher-order or bi-factor model seems to support the format of the Aptis test, which reports a single total score along with separate scores for each skill.

3.2.1 Preliminary analyses

Preliminary analyses on univariate and multivariate normality were conducted at the parcel (Part/Task) level for all countries combined and for each country separately. For all countries combined, the data were univariately normally distributed, as judged by the skewness and kurtosis values of |3.30| (the *z* score at $p < .01$; e.g., Tabachnick & Fidell, 2013) and histograms. The only exception was Reading Part 1, where the data had a peaked distribution (kurtosis = 5.00) with a high mean score (4.69 out of 5) and narrow standard deviation (0.65), suggesting that this part was very easy for examinees. The data were multivariate nonnormal as tested by Mardia's multivariate normality test available in an R package, MVN (Korkmaz, Goksuluk, & Zararsiz, 2016). There were no missing values. For each country, the data were univariately normal, except for Poland, where the data distribution often skewed and peaked. The data were multivariate nonnormal for all countries.

Given that some of the data were univariately and/or multivariately nonnormal, the MLMV estimator—a robust version of maximum likelihood estimation—was used to estimate model parameters with Mplus version 8 (Muthén & Muthén, 1998–2017a). One of the factor loadings from each factor was fixed to 1 for scale identification (see Appendix A for all the syntax used in the current study). The model fit was evaluated using the following: a comparative fit index (CFI) of .90 or higher; a standardised root mean square residual (SRMR) of .08 or lower; and root mean square error of approximation (RMSEA) values of 0.08 or lower (Browne & Cudeck, 1993). Chi-square difference tests were conducted to compare the five models for the Aptis test as they were nested (e.g., Brown, 2015). With statistical nonsignificance, a more parsimonious model, with fewer parameters to estimate (usually a model with a larger number of degrees of freedom), is selected; a statistically significant model with a larger chi-square value is not selected. Model fit and statistical criteria were used with substantive interpretability to evaluate each model.

3.2.2 Testing of the five models with aggregate data and data per country

After the examination of normality and missing cases, the aggregate data and data per country were analysed to select the best-fitting model among the single-factor, correlated four-factor, uncorrelated four-factor, higher-order, and bi-factor models. Once the best-fitting model had been selected for each country, multiple-sample (multiple-group/invariance) analysis was conducted to investigate the extent to which the same best-fitting model explained the data across countries.

3.2.3 Multiple-sample analysis

For Research Question 2, multiple-sample analyses were conducted to examine the equivalence of the factor structure of the Aptis test across countries. Based on Kline (2016), Brown (2015), and Vandenberg and Lance (2000), the best-fitting model was tested across countries for (1) no equality constraints (i.e., configural invariance [the same factor structure with no equal constraints on any parameters]) and (2) equal factor loadings (i.e., metric invariance). If (1) holds, it indicates that the same structure of factors underlies examinees' performance (i.e., the same number of factors and the same structure of factors, observed variables, and factor loading patterns [e.g., two factors loaded on three observed variables each, with each factor correlated]). However, this may differ across countries due to unequal factor loadings, unequal intercepts, or other unequal parameters in the factor structure. If (2) additionally holds, it also indicates that such a factor structure has equal factor loadings. This implies that the degree of relationship between the constructs and observed variables (i.e., Aptis Part scores) is equal across countries. The constructs are represented in the same way in each country. Note that a model is gradually restricted by adding constraints to examine the extent to which the model holds across countries.

As will be reported below, (a) Bangladesh and (b) the other remaining countries (i.e., Chile, Indonesia, Mexico, Poland, Spain and Sri Lanka) were analysed separately as each group of countries was best explained by different models of language ability structure. Further testing for intercepts, measurement error variances, factor variances, and factor covariances equality across countries was not pursued because multiple-sample analysis is sequential and the current results did not satisfy (2) equal factor loadings, as will be described below.

Measurement invariance was evaluated in two ways: first, with chi-square difference tests (with statistical nonsignificance suggesting support for further invariance testing) and second, with Chen's (2007) criteria for small sample size ($N < 300$). In Chen's (2007) criteria for testing factor loading invariance, a change of $\leqq$ -.005 in CFI, also supported by a change of $\leqq$ -0.010 in RMSEA or a change of $\leqq$ .025 in SRMR was required; for testing intercept invariance, the same criteria as for CFI and RMSEA and a change of $\leqq$ .005 in SRMR was required.

3.3. Results

3.3.1 Descriptive statistics

Descriptive statistics for the total and part scores are presented in Tables 1 and 2, respectively. As shown in Table 1, Poland performed the best and Mexico the worst. Further, as expected, average performance decreased as intended CEFR levels increased (see Table 2). For example, as the last row in Table 2 shows, average performance in the listening section was 81%, 70%, 55%, and 47% for A1-, A2-, B1-, and B2-level tasks, respectively. Finally, analyses often faced problems with not-positive-definite matrices.[2] In most cases, this involved the Reading A1 task. As partly explained in the *Preliminary analyses*, it was very easy for examinees and had a peaked distribution (kurtosis = 5.00) with a high mean score (4.69 out of 5) and narrow standard deviation (0.65). All the correlation matrices can be seen in Appendix B.

Table 1: Descriptive statistics for total scores

Group		Listening	Reading	Speaking	Writing
All	Mean	15.38	18.56	12.02	14.00
	SD	5.62	6.24	5.65	4.29
Bangladesh	Mean	14.18	17.73	11.75	13.11
	SD	5.18	6.32	5.46	4.40
Chile	Mean	18.85	21.86	13.09	15.50
	SD	4.24	3.84	4.73	2.57
Indonesia	Mean	15.62	19.60	12.07	15.06
	SD	3.95	4.78	4.07	3.01
Mexico	Mean	13.06	15.70	9.87	12.19
	SD	6.14	6.67	6.79	4.65
Poland	Mean	21.13	23.44	15.96	17.54
	SD	3.64	3.48	3.41	2.84
Spain	Mean	16.92	21.62	14.18	16.39
	SD	4.39	4.11	3.90	2.48
Sri Lanka	Mean	15.83	17.21	11.98	13.99
	SD	4.77	6.01	4.85	3.96
	k	25	4	4	4
	Total score	**25**	**25**	**21**	**21**
	Mean (%)	**62%**	**74%**	**57%**	**67%**

Note. *k* = Number of items/tasks.

[2] To investigate not-positive-definite matrices, we examined the outputs and found that this problem was often due to the negative variances of latent variables. To address this issue, we tried three ways, but none produced satisfactory results. Note the first and second methods were from Savalei and Kolenikov (2008), which we believe offers the most accessible and practical approaches to handling this issue. The third method was from the literature on Bayesian approaches (e.g., Brown, 2015; Kline, 2016). First, negative variances were tested for their statistical significance. If significant, it could indicate model misspecification. If not significant, the model is accepted. This allows for an "explicit examination of Heywood cases" and is a "more honest choice" (Savalei & Kolenikov, 2008, p. 166) than constraining a Heywood case to the closest admissible value or using SEM software's default settings. Almost all negative variances of latent variables were nonsignificant. Unfortunately, when we moved to multiple-sample analysis, we noticed that Mplus does not produce an output when the data include not-positive-definite matrices. Thus, this first method worked in a single-sample testing, but not in multiple-sample analysis. Second, the negative variances of latent variables were restricted to stay nonnegative (using the model constraint command in Mplus). The results had often convergent issues and they were not fixable even after increasing the number of iterations to 10,000. Thus, this second method failed as well. Third, the Bayes estimator was used by specifying the inverse Gamma distribution of those variances with mean = -1, 0, and 1 and variance = 0 and 0.01 in combination in the model priors command in Mplus. The results were still the same as those from the MLMV estimator and did not support the model. Thus, this third method failed as well. We failed to address the negative variances of latent variables across the three methods, and we had no choice but to fix those variances to zero. For those variances fixed to zero, see the rightmost column in Table 3.
Another potential cause of not-positive-definite matrices was the high correlations among variables. This was also addressed following the three methods above. First, the statistical significance of such correlations were tested and found that they were all statistically significant. This suggests the model was misspecified and led us to reject that model. The second and third methods were also implemented, resulting in the same results and providing no support for the model.

Table 2: Descriptive statistics for Part scores

		Listening				Reading				Speaking			Writing			
Group		**A1**	**A2**	**B1**	**B2**	**A1**	**A2**	**B1**	**B2**	**A2**	**B1**	**B2**	**A1**	**A2**	**B1**	**B2**
All	Mean	4.07	4.20	3.84	3.27	4.69	4.46	4.86	4.54	3.69	6.11	2.22	4.39	3.84	3.25	2.52
	SD	1.20	1.85	1.76	1.90	0.65	2.02	2.38	2.31	1.51	2.94	1.65	0.80	1.28	1.55	1.63
Bangladesh	Mean	4.00	3.82	3.57	2.79	4.71	4.17	4.85	4.00	3.66	6.11	1.98	4.22	3.46	3.06	2.37
	SD	1.24	1.88	1.70	1.62	0.60	2.14	2.34	2.41	1.60	2.82	1.49	0.86	1.42	1.52	1.72
Chile	Mean	4.66	5.23	4.67	4.30	4.94	5.41	5.61	5.91	3.91	6.44	2.74	4.57	4.27	3.54	3.12
	SD	0.77	1.29	1.45	1.86	0.24	1.28	1.99	1.36	1.29	2.44	1.46	0.58	0.85	1.23	1.18
Indonesia	Mean	4.21	4.55	3.60	3.26	4.89	4.48	4.96	5.26	3.63	6.11	2.34	4.62	4.11	3.66	2.67
	SD	0.97	1.38	1.29	1.49	0.37	2.06	1.94	1.86	1.13	2.02	1.44	0.62	1.22	1.21	1.24
Mexico	Mean	3.57	3.40	3.33	2.76	4.35	3.81	3.77	3.77	3.18	4.93	1.76	4.21	3.59	2.68	1.72
	SD	1.38	2.00	1.83	1.93	0.90	2.08	2.66	2.23	1.71	3.54	1.88	0.87	1.27	1.76	1.60
Poland	Mean	4.76	5.51	5.59	5.27	4.93	5.69	6.68	6.14	4.47	8.08	3.41	4.71	4.63	4.32	3.88
	SD	0.62	0.97	1.36	1.63	0.26	1.13	1.22	1.45	0.88	1.90	1.16	0.76	0.75	1.08	1.16
Spain	Mean	4.26	4.98	4.07	3.61	4.88	5.49	5.47	5.78	4.32	7.26	2.59	4.71	4.35	4.03	3.30
	SD	0.83	1.34	1.51	1.76	0.34	1.23	1.78	1.73	0.97	2.11	1.40	0.49	0.90	1.03	1.07
Sri Lanka	Mean	4.26	4.64	3.76	3.16	4.71	3.97	4.95	3.57	3.67	6.07	2.24	4.56	3.93	3.01	2.48
	SD	1.13	1.49	1.72	1.73	0.57	2.09	2.16	2.31	1.33	2.41	1.65	0.66	1.29	1.51	1.56
	k	**5**	**6**	**7**	**7**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
	Total score	**5**	**6**	**7**	**7**	**5**	**6**	**7**	**7**	**5**	**10**	**6**	**5**	**5**	**5**	**6**
	Mean (%)	**81%**	**70%**	**55%**	**47%**	**94%**	**74%**	**69%**	**65%**	**74%**	**61%**	**37%**	**88%**	**77%**	**65%**	**42%**

Note. *k* = Number of items/tasks

Table 3: Fit indices for the five models for each country

Group	Model	χ^2	*df*	CFI	RMSEA (90% CI)	SRMR	Fit ?	Best-fitting model[a]?	Note
All (*N* = 1,270)	Single factor	904.279*	90	.925	0.084 (0.079, 0.089)	.034			
	Correlated four factors	344.326*	84	.976	0.049 (0.044, 0.055)	.026	Yes		
	Uncorrelated four factors	4129.057*	90	.626	0.188 (0.183, 0.193)	.436			
	Higher-order	346.508*	86	.976	0.049 (0.044, 0.054)	.026	Yes		
	Bi-factor	115.279	69	.996	0.023 (0.015, 0.030)	.012	Yes	Yes	
Bangladesh (*n* = 403)	Single factor	353.662*	90	.912	0.085 (0.076, 0.095)	.039			
	Correlated four factors	189.113*	84	.965	0.056 (0.045, 0.066)	.040	Yes		
	Uncorrelated four factors	1199.640*	90	.628	0.175 (0.166, 0.184)	.408			
	Higher-order	189.176*	86	.965	0.055 (0.044, 0.065)	.040	Yes		
	Bi-factor	70.827	69	.999	0.008 (0.000, 0.031)	.017	Yes	Yes	
Chile (*n* = 117)	Single factor	198.207*	90	.722	0.101 (0.082, 0.120)	.085			
	Correlated four factors	Not-positive-definite (Listening-Reading factor correlation = 1.015, Reading-Writing factor correlation = 1.053)							
	Uncorrelated four factors	247.168*	90	.596	0.122 (0.104, 0.140)	.275			
	Higher-order	118.733*	88	.921	0.055 (0.025, 0.078)	.058	Yes	Yes	Reading factor variance negative (-0.001), fixed to 0; Writing factor variance 0.000, fixed to 0
	Bi-factor	No convergence[b]							
Indonesia (*n* = 95)	Single factor	132.311*	90	.832	0.070 (0.043, 0.095)	.075			
	Correlated four factors	105.853	84	.913	0.052 (0.000, 0.081)	.061	Yes		
	Uncorrelated four factors	200.329*	90	.562	0.114 (0.092, 0.135)	.266			
	Higher-order	106.982	86	.917	0.051 (0.000, 0.079)	.062	Yes	Yes	
	Bi-factor	Not-positive-definite							Writing factor variance tiny (0.005), fixed to 0
Mexico (*n* = 331)	Single factor	309.897*	90	.934	0.086 (0.076, 0.096)	.031			
	Correlated four factors	Not-positive-definite (Listening-Reading factor correlation = 1.004)							
	Uncorrelated four factors	1372.617*	90	.614	0.206 (0.197, 0.216)	.489			Speaking B1 variance negative (-0.036), fixed to 0
	Higher-order	160.429*	87	.978	0.050 (0.038, 0.063)	.027	Yes	Yes	Reading factor variance negative (-0.002), fixed to 0
	Bi-factor	Not-positive-definite (Listening-Reading factor correlation = 1.230, Listening-Writing factor correlation = 1.021)							

Group	Model	χ^2	*df*	CFI	RMSEA (90% CI)	SRMR	Fit ?	Best-fitting model[a]?	Note
Poland (*n* = 100)	Single factor	129.727*	90	.790	0.066 (0.039, 0.091)	.076			
	Correlated four factors	103.890	84	.895	0.049 (0.000, 0.077)	.064			
	Uncorrelated four factors	157.329*	90	.644	0.086 (0.064, 0.109)	.321			
	Higher-order	105.920	87	.900	0.047 (0.000, 0.075)	.064	Yes	Yes	Listening factor variance negative (-0.014), fixed to 0
	Bi-factor	No convergence[b]							
Spain (*n* = 137)	Single factor	146.255*	90	.854	0.068 (0.047, 0.087)	.063			
	Correlated four factors	111.512	84	.928	0.049 (0.019, 0.072)	.056	Yes		
	Uncorrelated four factors	249.322*	90	.585	0.114 (0.097, 0.131)	.290			
	Higher-order	112.705*	86	.930	0.048 (0.017, 0.070)	.056	Yes	Yes	
	Bi-factor	No convergence[b]							
Sri Lanka (*n* = 87)	Single factor	110.660	90	.955	0.051 (0.000, 0.081)	.049	Yes		
	Correlated four factors	Not-positive-definite							
	Uncorrelated four factors	282.037*	90	.584	0.157 (0.136, 0.177)	.395			
	Higher-order	95.181	87	.982	0.033 (0.000, 0.069)	.046	Yes	Yes	Listening factor variance negative (-0.010), fixed to 0
	Bi-factor	No convergence[b]							Speaking B2 variance negative (-0.491), fixed to 0; Speaking factor variance tiny (0.074), fixed to 0

Note. *df* = degrees of freedom. CFI = comparative fit index. RMSEA = root mean square error of approximation. CI = confidence interval. SRMR = standardised root mean square residual. *$p < .05$.

[a]The best-fitting model was determined by the chi-square difference test using the DIFFTEST option in Mplus.

[b]Convergence was not reached even after (a) "increasing the number of iterations" (Muthén & Muthén, 2017b, p. 524), (b) "using the preliminary parameter estimates as starting values" (Muthén & Muthén, 2017b, p. 524), and trying both (a) and (b) simultaneously.

Table 4: Parameter estimate of the bi-factor model for the all countries combined and Bangladesh

	ALL COUNTRIES COMBINED										BANGLADESH									
	Listening		Reading		Speaking		Writing		General		Listening		Reading		Speaking		Writing		General	
	Unst.	St.	Unst.	St.	Unst.	St.	Unst.	St.	Unst.	St.	Unst.	St.	Unst.	St.	Unst.	St.	Unst.	St.	Unst.	St.
L_A1	1	.654							1	.468	1	.682							1	.456
L_A2	1.360	.578							2.083	.633	1.067	.478							2.248	.674
L_B1	0.794	.354							2.116	.675	0.718	.356							1.871	.621
L_B2	0.626	.259							2.411	.715	0.383	.200							1.698	.593
R_A1			1	.685					0.389	.335			1	.643					0.352	.329
R_A2			2.053	.454					2.134	.593			2.425	.439					2.111	.556
R_B1			2.788	.524					2.925	.690			3.310	.548					2.671	.643
R_B2			2.100	.408					2.552	.622			2.598	.417					2.821	.659
S_A2					1	.761			1.160	.433					1	.815			1.191	.421
S_B1					1.941	.757			2.945	.564					1.697	.784			2.516	.504
S_B2					0.758	.528			2.001	.683					0.513	.447			1.986	.750
W_A1							1	.420	0.528	.372							1	.360	0.469	.308
W_A2							1.733	.464	1.128	.495							2.035	.443	1.293	.514
W_B1							2.654	.572	1.585	.573							2.548	.519	1.539	.572
W_B2							2.454	.504	1.933	.666							2.502	.450	2.259	.742
C[a]																				
R	0.321	.913									0.274	.837								
S	0.602	.669	0.374	.729							0.663	.604	0.365	.723						
W	0.224	.851	0.140	.933	0.276	.720					0.208	.799	0.114	.952	0.291	.723				

Note. Unst. = Unstandardised. St. = Standardised.

All unstandardised parameters except for those fixed to 1 for identification were statistically significant.

The standardised solution is STDYX in Mplus. C[a] = Correlations/covariances among first-order factors.

3.3.2 Testing of the five models with the aggregate data

Row 2 in Table 3 summarises the fit indices for the five models tested for all the countries combined. Of the five models, the correlated four-factor model fit the data well (CFI = .976, RMSEA = 0.049 [90% confidence interval: 0.044, 0.055], and SRMR = .026), as did the higher-order model (CFI = .976, RMSEA = 0.049 [0.044, 0.054], and SRMR = .026) and the bi-factor model (CFI = .996, RMSEA = 0.023 [0.015, 0.030], and SRMR = .012). A comparison of these three models revealed the bi-factor model to be the best model to represent the structure of the abilities measured in the Aptis test for the entire body of the examinees in the current data. The standardised parameter estimates in Table 4 show that each skill factor was overall well explained by tasks (.259 for the Listening B2 task to .761 for the Speaking A2 task). Each skill was also explained by the general proficiency factor (.335 for the Reading A1 task to .715 for the Listening B2 task). Thus, both individual skills (i.e., listening, reading, speaking, and writing) and general proficiency are considered to explain the performance measured in the Aptis test.

3.3.3 Testing of the five models with each country

In Table 3, rows 3 and below show the overall goodness of fit of the five models for each country. For example, the Bangladesh data were well explained by the correlated four-factor model (CFI = .965, RMSEA = 0.056 [0.045, 0.066], and SRMR = .040), the higher-order model (CFI = .965, RMSEA = 0.055 [0.044, 0.065], and SRMR = .040), and the bi-factor model (CFI = .999, RMSEA = 0.008 [0.000, 0.031], and SRMR = .017). In contrast, the single-factor model and uncorrelated four-factor model did not fit the data well. These results suggest the multi-componentiality of the abilities measured in the Aptis test for the Bangladesh examinees. When the three-fitting models were compared, the bi-factor model was found to better explain the data. Thus, the bi-factor model was judged to best represent the structure of the abilities measured in the Aptis test for the Bangladesh examinees. This process of testing and comparing models was repeated for each country to select the best-fitting model.

Across the countries, the results were summarised as follows.

- First, the single-factor (i.e., unitary) model fit the data only for Sri Lanka.
- Second, the correlated four-factor model fit the data for Bangladesh, Indonesia and Spain.
- Third, the uncorrelated four-factor model did not fit the data for any countries.
- Fourth, the higher-order model fit the data in all countries.
- Fifth, the bi-factor model fit the data only for Bangladesh.
- Sixth, when these well-fitting models were compared, the higher-order model best represented the structure of the abilities measured in the Aptis test for all countries, except for Bangladesh. The best-fitting model for Bangladesh was the bi-factor model.
- Seventh, the bi-factor model often had convergence issues, perhaps because it was too complex to analyse given the current sample size. As the best-fitting models differed between Bangladesh (the bi-factor model) and the other countries (the higher-order model), and Bangladesh was the only country against which no other countries could be meaningfully compared in terms of the best-fitting model, Bangladesh was excluded from the following multiple-sample analysis.

3.3.4 Multiple-sample analysis

The higher-order model was tested across all countries, except for Bangladesh, for (1) no equality constraints and (2) equal factor loadings. As Table 5 shows, the higher-order model with no equality constraints (Model 1) was simultaneously examined across the six countries. The result was satisfactory (CFI = .954, RMSEA = 0.046 [0.036, 0.056], and SRMR = .048) and suggests that the higher-order model explained the data well across the countries. The test of equal factor loadings (Model 2) also suggested the good fit of the model.

Table 5: Fit indices for the tests of measurement invariance of the higher-order model across Chile, Indonesia, Mexico, Poland, Spain and Sri Lanka

	χ^2	*df*	CFI	RMSEA (90% Confidence Interval)	SRMR	Fit?	χ^2 difference test	ΔCFI	ΔRMSEA	ΔSRMR	Retained?
Model 1: No equality constraints	683.533*	521	.954	0.046 (0.036, 0.056)	.048	Yes	--	--	--	--	Yes
Model 2: Equal factor loadings	802.078*	576	.936	0.052 (0.043, 0.061)	.106	Yes	χ^2 = 126.821, *df* = 55, *p* = .000	.018	.006	.058	No

Note. *df* = degrees of freedom. CFI = comparative fit index. RMSEA = root mean square error of approximation. CI = confidence interval. SRMR = standardised root mean square residual. *p < .05. ΔCFI, ΔRMSEA, and ΔSRMR = change in CFI, RMSEA, and SRMR values, respectively

.

Table 6: Parameter estimates in Model 1 of the no equality constraints, higher-order model across Chile, Indonesia, Mexico, Poland, Spain and Sri Lanka (continued over)

	CHILE										INDONESIA									
	Listening		Reading		Speaking		Writing		Higher-order		Listening		Reading		Speaking		Writing		Higher-order	
	Unst.	St.	Unst.	St.	Unst.	St.	Unst.	St.	Unst.	St.	Unst.	St.	Unst.	St.	Unst.	St.	Unst.	St.	Unst.	St.
L_A1	1	.600									1	.738								
L_A2	2.058	.736									1.552	.806								
L_B1	2.210	.702									1.151	.639								
L_B2	2.928	.726									1.030	.494								
R_A1			1	.374									1	.480						
R_A2			10.496	.731									7.341	.628						
R_B1			19.182	.859									8.392	.764						
R_B2			7.779	.511									6.024	.573						
S_A2					1	.769									1	.830				
S_B1					2.359	.957									1.850	.856				
S_B2					1.229	.834									1.131	.735				
W_A1							1	.425									1	.333		
W_A2							1.715	.492									3.091	.527		
W_B1							2.567	.513									3.392	.582		
W_B2							2.465	.514									4.132	.690		
H[a]																				
L									1	.992									1	.778
R									0.195	1									0.304	.958
S									1.225	.566									1.357	.806
W									0.537	1									0.350	.950

Note. Unst. = Unstandardised. St. = Standardised. All unstandardised parameters except for those fixed to 1 for identification were statistically significant.
The standardised solution is STDYX in Mplus. H[a] = Higher-order factor loadings.

Table 6: Parameter estimates in Model 1 of the no equality constraints, higher-order model across Chile, Indonesia, Mexico, Poland, Spain and Sri Lanka (continued)

	MEXICO										POLAND									
	Listening		Reading		Speaking		Writing		Higher-order		Listening		Reading		Speaking		Writing		Higher-order	
	Unst.	St.	Unst.	St.	Unst.	St.	Unst.	St.	Unst.	St.	Unst.	St.	Unst.	St.	Unst.	St.	Unst.	St.	Unst.	St.
L_A1	1	.769									1	.733								
L_A2	1.649	.874									1.614	.759								
L_B1	1.343	.779									1.703	.572								
L_B2	1.403	.773									2.550	.715								
R_A1			1	.738									1	.425						
R_A2			2.442	.775									8.973	.863						
R_B1			3.610	.896									10.077	.899						
R_B2			2.290	.680									9.880	.743						
S_A2					1	.867									1	.742				
S_B1					2.329	.977									2.330	.802				
S_B2					1.140	.902									1.367	.768				
W_A1							1	.643									1	.639		
W_A2							1.643	.724									1.045	.676		
W_B1							2.702	.861									1.543	.689		
W_B2							2.418	.844									1.496	.625		
H[a]																				
L									1	.999									1	1
R									0.624	1									0.208	.870
S									1.243	.887									1.011	.705
W									0.518	.982									0.967	.912

Note. Unst. = Unstandardised. St. = Standardised. All unstandardised parameters except for those fixed to 1 for identification were statistically significant.
The standardised solution is STDYX in Mplus. H[a] = Higher-order factor loadings.

Table 6: Parameter estimates in Model 1 of the no equality constraints, higher-order model across Chile, Indonesia, Mexico, Poland, Spain and Sri Lanka (continued)

	SPAIN										SRI LANKA									
	Listening		Reading		Speaking		Writing		Higher-order		Listening		Reading		Speaking		Writing		Higher-order	
	Unst.	St.	Unst.	St.	Unst.	St.	Unst.	St.	Unst.	St.	Unst.	St.	Unst.	St.	Unst.	St.	Unst.	St.	Unst.	St.
L_A1	1	.642									1	.689								
L_A2	1.993	.792									1.424	.742								
L_B1	1.949	.687									1.437	.652								
L_B2	2.425	.736									1.600	.721								
R_A1			1	.445									1	.589						
R_A2			5.207	.645									4.147	.660						
R_B1			9.962	.857									5.771	.889						
R_B2			7.497	.661									5.407	.778						
S_A2					1	.751									1	.655				
S_B1					2.634	.906									2.604	.939				
S_B2					1.311	.682									1.655	.870				
W_A1							1	.284									1	.496		
W_A2							3.213	.497									2.027	.513		
W_B1							5.132	.691									3.896	.842		
W_B2							5.577	.725									3.993	.837		
H[a]																				
L									1	.907									1	1
R									0.289	.917									0.406	.949
S									1.183	.789									1.045	.935
W									0.257	.902									0.391	.936

When a model with equal factor loadings (Model 2) was compared with another model with no equality constraints (Model 1), Model 2 was not supported, as shown by statistically significant (χ^2 = 126.821, *df* = 55, *p* = .000) and more-than-minimum or small changes in fit indices (ΔCFI = .018, ΔRMSEA = 0.006, ΔSRMR = .058). The results indicate that Model 2 was more statistically degraded than Model 1, and Model 2 was rejected. This suggests that Model 1 is the final model and that the same number of factors and factor loading patterns holds across countries, but the size of the factor loadings differ. In other words, the basic factor structure is the same across countries: The number of the constructs measured and their indication by observed variables (i.e., Aptis Part scores) hold across countries. However, the exact strength of such indications/relationships among the constructs and observed variables differs in each country. This means that the constructs are represented in different manners across the countries (equal factor loadings not retained). Also note that (3) equal factor loadings and equal intercepts were not tested because (2) equal factor loadings were not supported.

Table 6 shows the parameter estimates for the final model (Model 1) that had the same factor structure but differing sizes of factor loadings: Factor loadings were not constrained and were allowed to vary across countries. For example, the standardised factor loadings from L_A2 (the Listening A2 task), L_B1 (the Listening B1 task), and L_B2 (the Listening B2 task) were .736, .702, and .726 for Chile, .806, .639, and .494 for Indonesia, .874, .779, and .773 for Mexico, .759, .572, and .715 for Poland, .792, .687, and .736 for Spain, and .742, .652, and .721 for Sri Lanka. Overall, factor loadings were medium to high (.374 to .992 for Chile; .333 to .958 for Indonesia; .643 to .999 for Mexico; 425 to .912 for Poland; .284 to .917 for Spain; .496 to .949 for Sri Lanka) and suggest the appropriateness of the higher-order model across the countries.

3.4 Discussion and conclusion

Two research questions were examined in relation to the factor structure of the Aptis test, referred to as Research Question 1 (all data combined and analysed versus data analysed separately for each country) and Research Question 2 (measurement invariance across countries).

Research Question 1 was: *What is the factor structure of the Aptis test when all countries are analysed together as well as individually?* First, the results showed that when all countries were combined and analysed, their ability structure was best explained by a bi-factor model, where a single, general L2 proficiency factor as well as the four skills of listening, reading, speaking, and writing influenced test scores directly. Second, when analysed per country, the best-fitting model was a bi-factor model for Bangladesh and a higher-order model for all other countries (i.e., Chile, Indonesia, Mexico, Poland, Spain and Sri Lanka). Thus, the bi-factor model was adopted when all data were combined and analysed and when Bangladesh was analysed. For the remaining countries, the higher-order model best explained the data.

The final model differed between (a) all data combined and Bangladesh (i.e., a bi-factor model) AND (b) the other countries (i.e., a higher-order model). As seen in Table 3, regarding (a), both the bi-factor and higher-order models fit the data: when compared, the bi-factor model fit the data better than did the higher-order one. Bangladesh had the largest number of test-takers (*n* = 403) and may have reflected the tendency of all data combined (*N* = 1,270). By contrast, regarding (b), the higher-order model fit the data, whereas the bi-factor model did not converge or had a not-positive-definite matrix. These issues with the bi-factor model were somewhat consistent with those reported in earlier studies, where bi-factor models with good fit to the data had occasional problems with interpretability and unstable parameter estimates (e.g., Sawaki & Sinharay, 2018).

The finding that the final adopted models varied, although they were both theoretically plausible, was in line with Sawaki et al. (2009) and Sawaki and Sinharay (2018). These two studies investigated a factor structure of the TOEFL iBT using the same test specifications and different datasets and adopted different models: a higher-order model in Sawaki et al. (2009) and a correlated four-factor model in Sawaki and Sinharay (2018). As an explanation for the mixed results across the studies, Sawaki and Sinharay (2013, the original internal report used as a basis for the 2018 article) wrote the following:

> A possible explanation may be the somewhat increased homogeneity of the data for the samples analysed in this study, presumably reflecting differential levels of motivation and familiarity with the TOEFL iBT format between the field study participants and the operational test-takers involved in this study. (p. 89)

To the extent that the factor structure of a test changes according to differences in samples, it is unsurprising that in the current study the factor structure of the Aptis test was a function of a sample of examinees analysed. This could have been particularly true given the wide range of samples spanning seven countries in our dataset.

The support for the bi-factor or higher-order model indicates the multicomponentiality of L2 proficiency, in concurrence with previous studies (Bachman & Palmer, 1982, 1989; In'nami et al., 2012; In'nami et al., 2016; Llosa, 2007; Sawaki, 2007; Sawaki & Sinharay, 2013, 2018; Sawaki et al., 2009; Shin, 2005; Stricker, & Rock, 2008). The bi-factor model had not been previously tested except for Sawaki et al. (2009) and Sawaki and Sinharay (2013), perhaps because a bi-factor model has only been more widely used in the recent past (see examples in Brown, 2015; Kline, 2016). In Sawaki et al. (2009), the correlated four-factor model explained performance better than the bi-factor model. The higher-order model was tested in numerous previous studies and was often found to be the best-fitting model. Altogether, the findings from previous and current studies suggest that L2 proficiency is multicomponential, with an L2 proficiency factor, and that reporting a single total score along with separate scores for each skill, as practiced in the Aptis test, is justified.

Research Question 2 asked: *To what extent does measurement invariance hold across countries?* As the best-fitting models differed between Bangladesh (the bi-factor model) and all other countries (the higher-order model), and Bangladesh was the only country against which no other countries could be meaningfully compared in terms of the best-fitting model, Bangladesh was excluded from investigation into measurement invariance using multiple-sample analysis. In other words, whether the same higher-order model from Research Question 1 was supported across countries was more closely examined. The results showed that the higher-order model with no equality constraints was most consistent with the data. This suggests that the same number of factors and factor loading patterns hold across the countries, but that the size of the factor loadings differ. Thus, the constructs are represented in different manners across countries (equal factor loadings not retained). This provides the weakest evidence for invariance.

Moreover, the results suggest that the same factor structure was not retained across countries (i.e., a bi-factor model for all the data combined and for Bangladesh; a higher-order model for the remaining six countries). This evidence is troublesome as it indicates that the Aptis test measures listening, reading, speaking, and writing with different structures and with a different degree of precision across countries. This would not be a problem if test scores were reported for each country, but would be an issue if scores were compared across countries. However, the adoption of two statistically different but theoretically similar models may not be a serious concern in terms of score reporting, because the bi-factor and higher-order models both support the format of the Aptis test, which reports a single total score along with separate scores for each skill.

Further, although the factor loadings in the final models were overall medium to high, W_A1 (the Writing A1 task) loaded rather low on the writing ability factor, except for Mexico and Poland (standardised parameter estimate = .643 and .639 for Mexico and Poland; .425, .333, .284, and .496 for Chile, Indonesia, Spain, and Sri Lanka, respectively). This suggests that the Writing A1 task did not contribute much to the measurement of writing ability, probably because this is an A1-level task and was very easy for examinees. Table 2 shows that this task had a mean score of 4.21 to 4.71 (possible total score = 5; 88% correct overall) and had less discriminability, as indicated by small standard deviations (0.49 to 0.86). This may not be a problem because the model was overall consistent with the data. Neither is it a problem in terms of content relevance and representativeness because the Aptis test is designed to include A1-level tasks. Furthermore, it is expected that when a nonadaptive test aims to assess a wide range of proficiency, it needs to present test-takers with tasks at diverse difficulty levels. As a result, some tasks may be too easy or difficult for examinees and may not have the intended level of discrimination.

Nevertheless, when test specifications are revised, test developers might want to consider the following: The Writing A1 task could be revised, for example, by increasing the number of tasks or levels in one task (currently 1 task; 5 points in total, which could be increased to 2 tasks or 8 points in total for 1 task). They could better discriminate between examinees and lead to a higher factor loading from the Writing A1 task to the writing ability factor.

These results also held for the Reading A1 task and Listening B2 task, which was either very easy or difficult and did not discriminate the examinees well, loading rather low on the reading ability factor. This could also be addressed in the same manner.

Moreover, the support for the bi-factor model for Bangladesh and the higher-order model for the six countries justifies reporting a single total score along with separate scores for each skill. This is positive evidence for the Aptis test, which provides a score for each skill and, for examinees who went through the four-skill tests, an overall score. In fact, we observed the mostly high correlations between four skills in the bi-factor models for all countries combined and Bangladesh (r = .720 to .952; except for .604 to .669) and the mostly high higher-order loadings on the first-order, four-skill factors (.700 to 1.00; except for .566 for Chile). This suggests that the four skills were highly correlated, although not high enough to be considered a single construct. These results indicate the psychometric distinctiveness of, and yet close relationship among, each skill section, and reinforce support for the score report format of the Aptis test, which is consistent with our finding from Research Question 1.

4. STUDY 2: APTIS EXAMINEES' FOUR-SKILL PROFILES

The purpose of Study 2 was to investigate the four-skill profiles of the Aptis test examinees. One research question was examined.

Research Question 3: Are there different language profiles across learner groups? If so, what different language profiles are present?

4.1 Method

4.1.1 Data

The same data as in Study 1 were used: (a) item-level dichotomous data for the listening section that were aggregated to create part-level, composite scores; and (b) part-level, composite data for the reading, speaking, and writing tests. The data consisted of 1,270 examinees across seven countries: Bangladesh (403), Chile (117), Indonesia (95), Mexico (331), Poland (100), Spain (137), and Sri Lanka (87). It should be noted that we could only obtain scores in each part for each skill, not the total scores for each skill, which is shown on the score report. As a reviewer suggested, it would be interesting to see what four-skill profiles look like using such total skill scores. This is a topic for future investigation.

4.1.2 Test content and format

These are the same as those in Study 1. The listening, reading, speaking, and writing Part scores were used.

4.2 Analyses

To examine Research Question 3, latent profile analyses were used to identify score patterns in the Aptis test. Results from latent profile analysis were statistically compared and the best model selected. Latent profile analyses differ from cluster analysis, where results are visually (and not statistically) compared. For further details, see, for example, Flaherty and Kiff (2012), and Muthén and Muthén (2017b).

The number of profiles was examined in four aspects. The first two aspects are based on Nylund, Asparouhov and Muthén (2007), and the other two aspects are based on general practice in the use of latent profile analysis. First, models were estimated by successively increasing the number of profiles (i.e., 2, 3, 4 … profiles) and by evaluating the improvement in the results with statistical significance using bootstrapped likelihood ratio tests (BLRT). Obtaining statistical significance suggests further improvement in the number of profiles, while nonsignificance suggests no such improvement in the number of profiles.[3] The penultimate number of profiles that produced nonsignificant improvement in profiles (e.g., K-1), compared with the subsequent number of profiles (e.g., K), was considered the best number of profiles. For example, if a 3-profile model was nonsignificantly improved over a 2-profile model, the 2-profile model was adopted. Second, models were evaluated with the Bayesian information criterion (BIC; Schwartz, 1978). Models with smaller values are considered better models.

[3] Statistical nonsignificance is preferred in chi-square difference tests in Study 1 because it supports further invariance across countries. Statistical significance is preferred in BLRT in Study 2 because it suggests further improvement in the number of profiles. Thus, the direction of statistical (non)significance in relation to positive results is reversed between Studies 1 and 2.

Third, models were evaluated in terms of the quality of their profiles. For this purpose, entropy was used, ranging from 0 to 1, with higher values suggesting a better profile result. An entropy of .80 or above indicates a good result (Clark & Muthén, n.d.). Fourth, models were also evaluated in terms of the quality of the profiles with average latent profile probabilities for most likely latent profile membership. These probabilities refer to the probability of individuals being classified into particular profiles. A probability of .70 or above indicates a good result (Nagin, 2005). Results from these four aspects were considered with substantive interpretability to determine the number of profiles. All analyses were conducted with the MLR estimator—a robust version of maximum likelihood estimation and the default estimator for latent profile analysis—in Mplus version 8 (Muthén & Muthén, 1998–2017a).

Unlike in Study 1, the data were combined for analysis and were not separated per country. This was because at the per country level, the number of parameters was greater than the sample size (i.e., the model was too complex to estimate relative to the number of examinees), as indicated by warning messages in Mplus. Finally, before the latent profile analyses, the data were standardised separately for each part, with means = 0 and standard deviations = 1, to facilitate interpretation across the tasks that consisted of different number of items/tasks and different total scores.

4.3 Results

4.3.1 The number of profiles

Increasing the number of profiles up to five consistently produced statistically significant results, suggesting the improvement in the number of profiles. However, when the number of profiles was specified as 6, the matrix was not positive definite due to the intercept of Reading_A1 (the Reading A1 task) for Profile 2. Removing this task did not solve the problem as it led to the discovery of another not-positive-definite matrix. Instead, we closely inspected the five-profile model and decided to adopt it as the final model. As Table 7 shows, this model was significantly better than the four-profile model ($p < .001$), had a lower BIC value (40459.518), and had a satisfactory entropy (.934, which exceeded the .80 criterion). The classification accuracy was high (.926 to .980, both of which exceeded the .70 criterion), suggesting that 92.6% to 98.0% of examinees were classified into a particular group, with a tiny percentage found to belong to more than one group. Based on these results, the five-profile model was adopted as the final model.

Table 7: Fit indices of the latent profile models

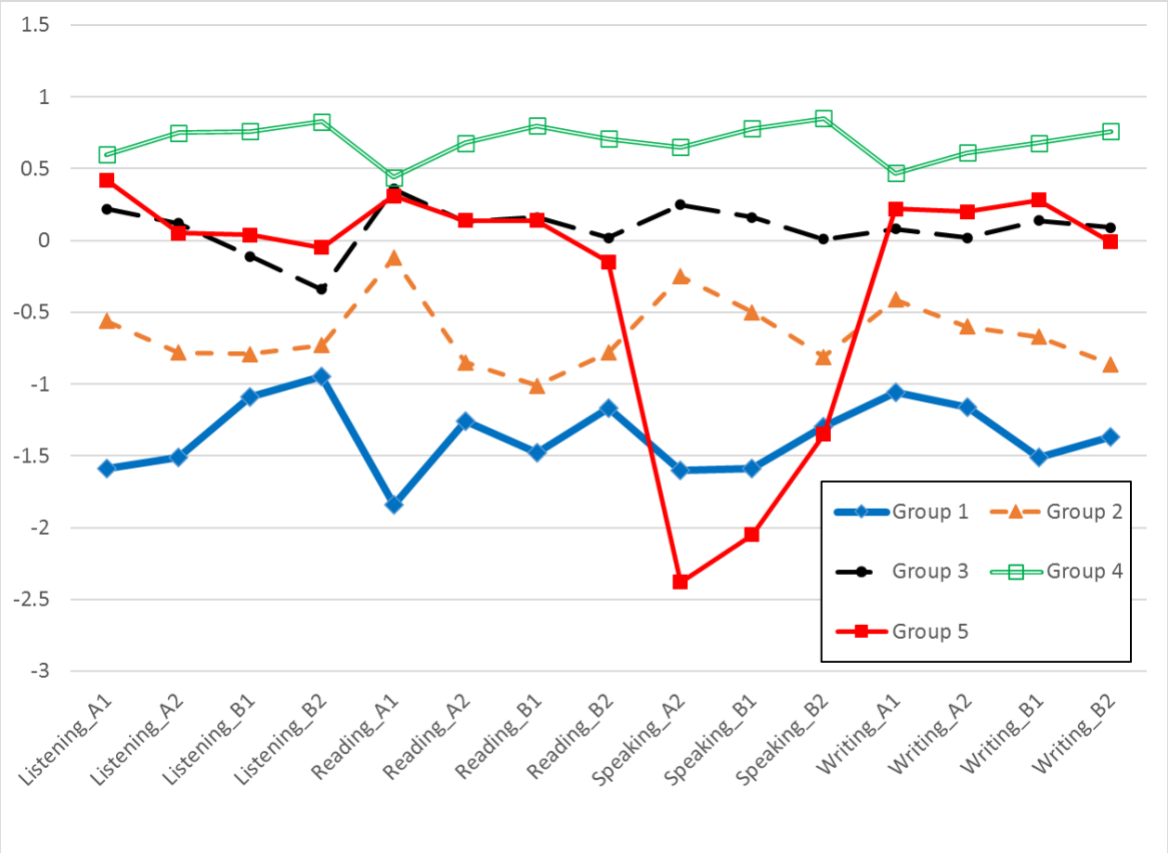
	BLRT[a]	BIC[b]	Entropy	Average classification probabilities					
					1	2			
2 profiles	$p < .001$	44309.378	.970	1	.986	.014			
				2	.005	.995			
					1	2	3		
3 profiles	$p < .001$	42004.968	.939	1	.979	.021	.000		
				2	.012	.950	.038		
				3	.000	.017	.983		
					1	2	3	4	
4 profiles	$p < .001$	41138.174	.914	1	.966	.000	.000	.034	
				2	.000	.940	.029	.031	
				3	.000	.025	.975	.000	
				4	.053	.026	.000	.922	
					1	2	3	4	5
5 profiles	$p < .001$	40459.518	.934	1	.980	.018	.000	.000	.002
				2	.014	.948	.033	.000	.005
				3	.000	.022	.926	.051	.001
				4	.000	.000	.030	.970	.000
				5	.019	.002	.000	.000	.979

Note. [a]bootstrapped likelihood ratio test. [b]Bayesian information criterion.

4.3.2 Interpretation of the profiles

The five-profiles are presented in Figure 2 (see Appendix C for details). Examinees in Profile 1 (i.e., Group 1) comprised 14.3% of the data (*n* = 182 out of 1,270; see Table 8) and was termed the "Low group with better performance in more difficult tasks" because this group generally represented examinees with low proficiency, performing worse on easier tasks and better on more difficult ones, especially in the listening, reading, and speaking sections. Group 2 comprised 16.0% of the data (*n* = 203) and was termed the "Low group (with the expected pattern)". Group 3 composed 26.0% of the data (*n* = 333) and was termed the "Intermediate group" because it represented examinees with average proficiency. Group 4 comprised 41.0% of the data (*n* = 519) and represented examinees with relatively high levels of proficiency across tasks; it was termed the "Advanced group." Group 5 comprised 2.7% of the data (*n* = 33) and represented examinees with average proficiency, except in speaking. This group was termed the "Intermediate group with a low speaking skill." In general, these five profiles classify examinees according to proficiency.

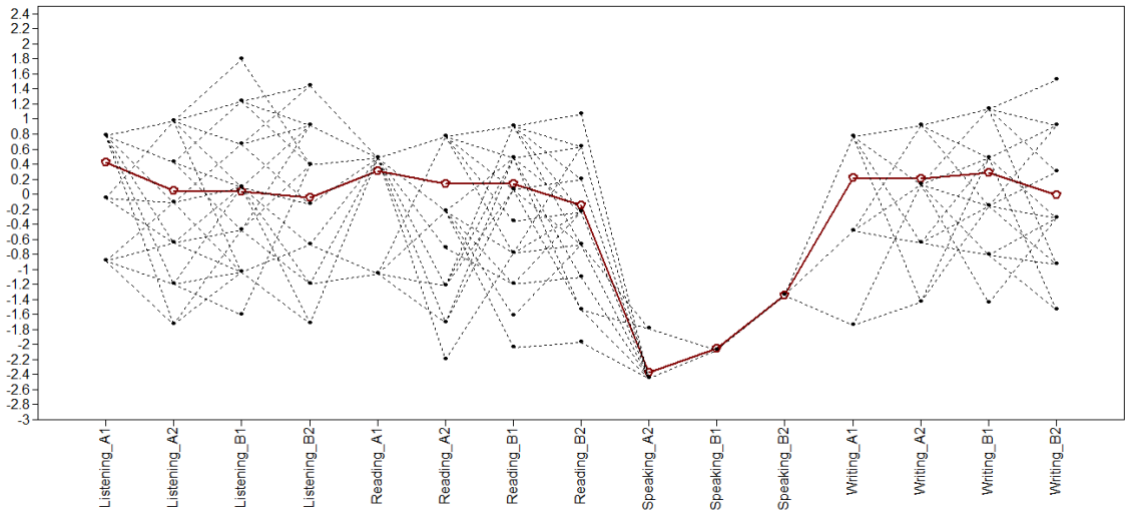
Figure 2: Latent profiles of the five groups



Note. The x axis shows Parts, and the y axis shows standardised raw scores with means = 0 and standard deviations = 1. Group 1 (14.3%), Group 2 (16.0%), Group 3 (26.0%), Group 4 (41.0%), and Group 5 (2.7%).

Next, the low speaking scores of individuals in Group 5 were inspected. As Figure 3 shows, performance varied among examinees in terms of listening, reading and writing. In contrast, speaking performance varied little, indicating the examinees' consistently low performance across speaking tasks.

Figure 3: Plots of each examinee's performance pattern in latent profiles of Group 5



Note. The x axis shows Parts, and the y axis shows standardised raw scores with means = 0 and standard deviations = 1. The solid, brown line shows the estimated means of Group 5. The dotted lines show observed values. This figure was drawn in Mplus using the following steps: "Plot" → "View plots" → "estimated means and observed individual values" → "Show one class per window," "Use most likely class," "All individuals," and "Individual data."

A detailed breakdown of each group by country is presented in Figure 4 and Table 8 (seen in vertically presented percentages in round brackets). For example, 13% of Bangladesh's examinees were in Group 1, 19% in Group 2, 32% in Group 3, 32% in Group 4, and 3% in Group 5. In the all-country data, the largest group was Group 4 (41.0%), followed by Group 3 (26.0%). The finding that Groups 3 and 4 were the main groups was observed in Bangladesh, Chile, Indonesia, and Spain. In Mexico, Group 1 was dominant. In Poland, Group 4 was dominant. In Sri Lanka, Groups 2 3, and 4 were dominant. Three points should be mentioned here: First, results with large percentages in Groups 2 (low group), 3 (intermediate group), and 4 (advanced group) generally reflected the types of test-takers who took this test. Second, Mexico had a large group of participants in Group 1 (35%), as did Bangladesh (13%). Third, the percentages of Group 5 test-takers were small across countries, with the highest being Chile's 5%.

When the results were compared horizontally (see Table 8), either Bangladesh or Mexico was predominant, because they accounted for a large proportion of all examinees (32% and 26%, respectively). For example, in Group 1, Mexico and Bangladesh accounted for 63% and 30% of test-takers, respectively.

Figure 4: Percentages of latent profiles for each country

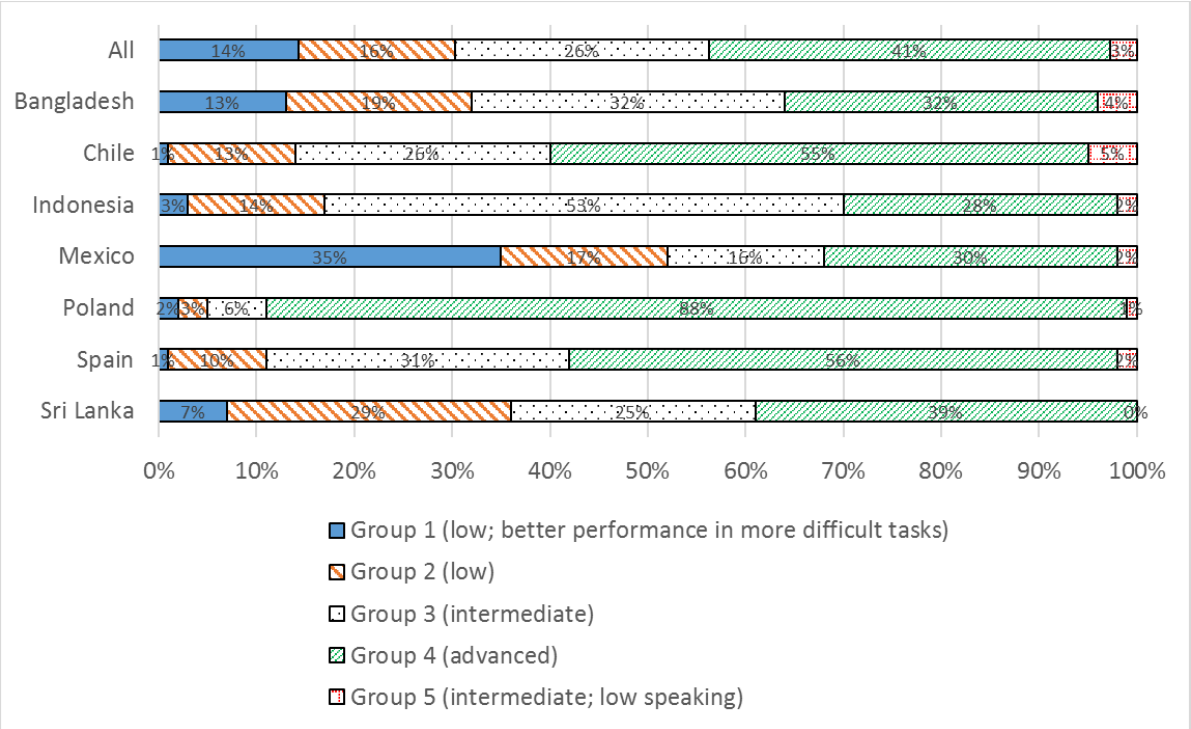


Table 8: Breakdown (number of examinees) of the groups by country

	All	Bangladesh	Chile	Indonesia	Mexico	Poland	Spain	Sri Lanka
Group 1 (low; better performance in more difficult tasks)	182 (14.3%) [100%]	54 (13%) [30%]	1 (0%) [0%]	3 (3%) [2%]	115 (35%) [63%]	2 (2%) [1%]	1 (0%) [0%]	6 (7%) [3%]
Group 2 (low)	203 (16.0%) [100%]	78 (19%) [38%]	15 (13%) [7%]	13 (14%) [6%]	55 (17%) [27%]	3 (3%) [1%]	14 (10%) [7%]	25 (29%) [12%]
Group 3 (intermediate)	333 (26.0%) [100%]	128 (32%) [38%]	31 (26%) [9%]	50 (53%) [15%]	53 (16%) [16%]	6 (6%) [2%]	43 (31%) [13%]	22 (25%) [7%]
Group 4 (advanced)	519 (41.0%) [100%]	129 (32%) [25%]	64 (55%) [12%]	27 (28%) [5%]	100 (30%) [19%]	88 (88%) [17%]	77 (56%) [15%]	34 (39%) [7%]
Group 5 (intermediate; low speaking)	33 (2.7%) [100%]	14 (3%) [42%]	6 (5%) [18%]	2 (2%) [6%]	8 (2%) [24%]	1 (1%) [3%]	2 (1%) [6%]	0 (0) [0%]
Total	**1,270 (100%) [100%]**	**403 (100%) [32%]**	**117 (100%) [9%]**	**95 (100%) [7%]**	**331 (100%) [26%]**	**100 (100%) [8%]**	**137 (100%) [11%]**	**87 (100%) [7%]**

Note. ( ) = Percentage when numbers in one country are compared (vertically) within country. [ ] = Percentage when numbers in one group are compared (horizontally) across countries.

4.4. Discussion and conclusion

To examine the linguistic profiling of Aptis test examinees, one research question was examined. Research Question 3 asked: *What different language profiles are present across groups?* The results showed that examinees were classified into five profile groups. Groups 1 and 2 were the low groups, with Group 1 scoring lower on A1-level tasks than B2-level tasks in listening, reading, and writing. Group 2 showed an opposite, normal trend of scoring higher on A1-level tasks than B2-level tasks in listening, reading, and writing. Groups 3 and 4 were the Intermediate and Advanced groups, respectively. Group 5 was also an Intermediate group, but demonstrated extremely low performance across all speaking tasks.

In contrast to Groups 2 (low group), 3 (intermediate group), and 4 (advanced group), whose profile appeared normal considering that the purpose of a proficiency test such as the Aptis test is to spread participants, Groups 1 and 5 were characterised by unexpected profiles. In Group 1, the average score patterns were unusual, as examinees performed worse on easier tasks and better on more difficult tasks, except for writing tasks. There is a possibility that examinees encountered mechanical problems in solving easy questions. Another plausible reason may be an order effect, but this is unlikely in reading and listening because in the Aptis test, tasks are always ordered in increasing difficulty (i.e., A1, A2, B1, and B2, in this order). For example, in reading, Figure 2 shows that examinees in Group 1 scored the highest on B2 tasks, the second highest on A2 tasks, the third highest on B1 tasks, and this order was not the same as the one in which the tasks were presented. In contrast, in speaking, test-takers had lower scores on easier tasks, probably because they were not used to recording their speech; once they learned how to do it, they may have been able to demonstrate their ability better and thus obtained higher scores for more difficult tasks. Thus, less familiarity with the semi-direct speaking format of the test might explain some of Group 1's results. To investigate this speculation, future analysis of the test task responses and test-taking processes of the examinees in this group is needed. When we consider the implications of this phenomenon, to reduce the number of examinees in Group 1, test developers and administrators may need to routinely take a record of possible technical and test-takers' behavioural problems and consider ways to respond to them, for example, by providing more practice tasks for getting used to speaking test formats.

The reason for the extremely low speaking performance of Group 5, which was observed across all speaking tasks, was not very clear. Additionally, Group 5 performed better in more difficult tasks (e.g., better performance in the Speaking B1 task than the Speaking A2 task). This was difficult to understand because Group 5 demonstrated average performance in the Al tasks of listening, reading, and writing skills. The average performance in writing and yet the extremely low performance in speaking were also difficult to understand because these production skills are generally closely related and tend not to produce such a huge gap between the skills. The two explanations we hypothesised are as follows. First, there may indeed be some learners with such profile, in which speaking skill is much lower than listening, reading, and writing. These types of learners have been consistently reported in Koizumi et al. (2018) among Japanese learners of English when scores in five four-skill tests were analysed. Second, there might have been mechanical problems in administering the speaking section and/or recording examinees' voices. Whether the same five profiles are observed and whether the same extremely low speaking performance of Group 5 is observed require further research with different data. Replicating the results may rule out sampling variation and indicate the generalisability of the results and the possible need to provide examinees in Group 5 with specific feedback on speaking.

Taken together, the reasons for the unexpected profiles in Groups 1 and 5 do not seem to be related to problems with the test. Thus, the results can provide some diagnostic information to test-takers and developers, plus evidence of positive validity to users of the Aptis test for diagnostic purposes. However, there is a caveat: Some may state that when the results are considered across skills, almost all examinees (i.e., Groups 1 through 4, which in total consisted of 97.3% of the examinees) seem to have relatively flat profiles across skills, but the current results do not suggest such profile information.

This is because raw scores in each part for each skill were separately standardised, and we were not able to rigorously compare scores across parts and skills. The results suggest score patterns when the means are fixed to zero. Future analysis using comparable scores with the same full marks would provide information on flat or uneven profiles across parts and skills.

5. RECOMMENDATIONS FROM STUDIES 1 AND 2

Five recommendations are made based on the findings of the current project. First, positive evidence has been provided for using the Aptis score report format with a total score along with separate scores for each skill, because a bi-factor or higher-order factor structure was shown in Research Questions 1 and 2 in Study 1.

Second, it is acceptable to score performance and report test results in a score report for each country, but it is not highly recommended that such performance be compared across countries. This is based on our finding for Research Question 2 in Study 1: Different factor structures were found between Bangladesh and the other seven countries, and the requirement of having equal factor loadings across the seven countries for score comparisons was not met. One could compare performance across countries, but should remember that differences in performance could be spurious due to item(s) or section(s) functioning differently across countries. However, this limitation may not be serious in case of the Aptis test, which does not seem to place high priority on international comparisons.

Third, while the current test is acceptable, in the future, test developers might consider revising the test specifications of the Writing A1 task (for example, by including more items) to better discriminate between examinees and make it more relevant to the writing ability. The Writing A1 task did not load well on the writing construct in all countries, except for Mexico and Poland, likely due to its relatively low level of difficulty. The same goes for the Reading A1 task and Listening B2 task, which was either very easy or difficult for examinees, as shown by a high or low mean score and a narrow standard deviation. Considerations might be given to the revision of the test specifications to add more items so that more variance and better discrimination could make these tasks load heavily on the writing, reading, and listening construct. This recommendation is mainly based on our finding for Research Questions 1 and 2 in Study 1.

Fourth, test developers might want to consider routinely analysing examinees' skill profiles. In addressing Research Question 3 in Study 2, we identified five groups with different skill profiles, with two groups performing in a manner unexpected from test developers. Routine analysis could be simple or detailed using item-level scores, section-level scores, scaled skill scores, or CEFR levels for each skill. Such an analysis would reveal the examinees' skill characteristics and provide evidence on validity for diagnostic purposes.

Finally, all the findings and suggestions above would be further strengthened by examining whether the same results are obtained for different versions of the Aptis test. The current data were from Version 1 of the Aptis test, because Versions 2 and 3 had smaller sample size and were not included into the current study. If more data are added to Versions 2 and 3 and the same results as in the current analysis are obtained, this suggests that our finding is not unique to Version 1 and generalises to different versions, indicating the essential nature of the constructs measured in the Aptis test. If the results are not replicated in Versions 2, 3, or both, this warrants close scrutiny because particular items or tasks that only appear in either version of the test may be responsible for the divergent results. Such further analysis can be augmented by including information on test format (paper-based or computer-based) or examinees' background.

REFERENCES

Alderson, J. C., Haapakangas, E.-L., Huhta, A., Nieminen, L., & Ullakonoja, R. (2015). *The diagnosis of reading in a second or foreign language*. New York, NY: Routledge.

Bachman, L. F., & Palmer, A. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly*, *16*, pp. 449–465. doi:10.2307/3586464

Bachman, L. F., & Palmer, A. (1989). The construct validation of self-ratings of communicative language ability. *Language Testing*, *6*, pp. 14–29. doi:10.1177/026553228900600104

Bae, J., & Bachman, L. F. (1998). A latent variable approach to listening and reading: testing factorial invariance across two groups of children in the Korean/English Two-Way Immersion Program. *Language Testing*, *15*, pp. 380–414. doi:10.1177/026553229801500304

Bae, J., & Lee, Y.-S. (2011). The validation of parallel test forms: 'Mountain' and 'beach' picture series for assessment of language skills. *Language Testing*, *28*, pp. 155–177. doi:10.1177/0265532210382446

British Council. (2018a). *Research*. Retrieved from http://www.britishcouncil.org/aptis/research

British Council. (2018b). *What is Aptis?* Retrieved from https://www.britishcouncil.org/exam/aptis/what

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd edition). New York, NY: Guilford.

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). Newbury Park, CA: Sage.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*, pp. 456–466. doi:10.1037/0033-2909.105.3.456

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, *14*, pp. 464–504. doi:10.1080/10705510701301834

Clark, S. L., & Muthén, B. (n.d.). *Relating latent class analysis results to variables not included in the analysis*. Retrieved from https://www.statmodel.com/download/relatinglca.pdf

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment.* Cambridge University Press. Retrieved from http://rm.coe.int/1680459f97

Council of Europe. (2009). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR): A manual.* Strasbourg, France: Language Policy Division. Retrieved from http://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680667a2d

Dunn, K. J., & McCray, G. (2020). The place of the bifactor model in confirmatory factor analysis investigations into construct dimensionality in language testing. *Frontiers in Psychology*, *11*(1357), pp. 1–16. doi:10.3389/fpsyg.2020.01357

Flaherty, B. P., & Kiff, C. J. (2012). Latent class and latent profile models. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf & K. J. Sher (Eds.), *APA handbook of research methods in psychology: Vol. 3: Data analysis and research publication* (pp. 391–404). Washington, DC: American Psychological Association.

Ginther, A., & Yan, X. (2018). Interpreting the relationships between TOEFL iBT scores and GPA: Language proficiency, policy, and profiles. *Language Testing*, *35*, pp. 271–295. doi:10.1177/0265532217704010

Granfeldt, J., & Ågren, M. (2013). Stages of processability and levels of proficiency in the Common European Framework of Reference for Languages: The case of L3 French. In A. F. Mattsson & C. Norrby (Eds.), *Language acquisition and use in multilingual contexts: Theory and practice* (pp. 28–38). Lund, Sweden: Lund University. Retrieved from http://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=4072080&fileOId=4075031

Harsch, C. (2014). General language proficiency revisited: Current and future issues. *Language Assessment Quarterly*, *11*, pp. 152–169. doi:10.1080/15434303.2014.902059

Huhta, A., Alderson, J. C., Nieminen, L., & Ullakonoja, R. (2015). *Diagnostic profiling of foreign language readers and writers: Exploring the usefulness of latent profile analysis in diagnostic assessment research.* Paper presented at the 37th Language Testing Research Colloquium, Eaton Chelsea. Toronto, ON, Canada.

Hulstijn, J. H. (2011). Language proficiency in native and non-native speakers: An agenda for research and suggestions for second-language assessment. *Language Assessment Quarterly*, *8*, pp. 229–249. doi:10.1080/15434303.2011.565844

Hulstijn, J. H. (2015). *Language proficiency in native and non-native speakers: Theory and research.* Amsterdam, The Netherlands: John Benjamins.

Hulstijn, J. H., Schoonen, R., De Jong, N. H., Steinel, M. P., & Florijn, A. (2012). Linguistic competences of learners of Dutch as a second language at the B1 and B2 levels of speaking proficiency of the Common European Framework of Reference for Languages (CEFR). *Language Testing*, *29*, pp. 203–221. doi:10.1177/0265532211419826

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

In'nami, Y., & Koizumi, R. (2013). Review of sample size for structural equation models in second language testing and learning research: A Monte Carlo approach. *International Journal of Testing*, *13*, pp. 329–353. doi:10.1080/15305058.2013.806925

In'nami, Y., Koizumi, R., & Nakamura, K. (2016). Factor structure of the Test of English for Academic Purposes (TEAP®) test in relation to the TOEFL iBT® test. *Language Testing in Asia*, *6*(1), pp. 1–23. doi:10.1186/s40468-016-0025-9

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York, NY: Guilford.

Koizumi, R. (2015). Factor structure and four-skill profiles of the TOEIC® tests among Japanese university learners of English. *Annual Review of English Language Education in Japan (ARELE)*, *26*, pp. 109–124. Retrieved from https://www.jstage.jst.go.jp/article/arele/26/0/26_KJ00010090532/_article

Koizumi, R., Agawa, T., & Asano, K. (2018, March). *Deriving useful information on skill imbalance from TOEFL iBT® scores*. Paper presented at the American Association for Applied Linguistics 2018 Conference, Chicago, Illinois, U.S.

Koizumi, R., Kashimada, Y., & Akimoto, T. (2019, August). *Nihonjin eigo gakushusha no yongino reberu no zure no tokucho: GTEC CBT taipu no baai* [Characteristics of four-skill profiles of Japanese learners of English]. Presented at the 45th Japan Society of English Language Education (JASELE) Annual Conference, Hirosaki University, Aomori, Japan.

Korkmaz S, Goksuluk D, & Zararsiz G. (2016). *MVN*. Retrieved from: https://cran.r-project.org/web/packages/MVN/MVN.pdf

Little, T. D., Cunningham, W. A., & Shahar, G. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling*, *9*, pp. 151–173. doi:10.1207/S15328007SEM0902_1

Little, T. D., Rhemtulla, M., Gibson, K., & Schoemann, A. M. (2013). Why the items versus parcels controversy needn't be one. *Psychological Methods*, *18*, pp. 285–300. doi:10.1037/a0033266

Llosa, L. (2007). Validating a standards-based classroom assessment of English proficiency: A multitrait-multimethod approach. *Language Testing*, *24*, pp. 489–515. doi:10.1177/0265532207080770

Meade, A. W., & Kroustalils, C. W. (2006). Problems with item parceling for confirmatory factor analytic tests of measurement invariance. *Organizational Research Methods*, *9*, pp. 369–403. doi:10.1177/1094428105283384

Muthén, L. K., & Muthén, B. O. (1998–2017a). Mplus (Version 8) [Computer software]. Los Angeles, CA: Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, *9*, pp. 599–620. doi:10.1207/S15328007SEM0904_8

Muthén, L. K., & Muthén, B. O. (2017b). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.

Nagin, D. S. (2005). *Group-based modeling of development*. Cambridge, MA: Harvard University Press.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*, *14*, pp. 535–569. doi:10.1080/10705510701575396

O'Sullivan, B. (2011). Language testing. In J. Simpson (Ed.), *Routledge Handbook of Applied Linguistics* (pp. 259–273). Oxford, U.K.: Routledge.

O'Sullivan, B. (2012). *Aptis test development approach: Aptis technical report* (ATR-1). London: British Council. Retrieved from: http://www.britishcouncil.org/sites/britishcouncil.uk2/files/aptis-test-dev-approach-report.pdf

O'Sullivan, B., & Dunlea, J. (2015). *Aptis General technical manual* (Version 1.0). TR/2015/005. London: British Council. Retrieved from: https://www.britishcouncil.org/sites/default/files/aptis_general_technical_manual_v-1.0.pdf

O'Sullivan, B., & Weir, C. (2011). Language testing and validation. In B. O'Sullivan (Ed.), *Language Testing: Theory and Practice* (pp. 13–32). Oxford, UK: Palgrave.

Purpura, J. E. (1998). Investigating the effects of strategy use and second language test performance with high-and low-ability test takers: A structural equation modelling approach. *Language Testing*, *15*, pp. 333–379. doi:10.1177/026553229801500303

Savalei, V., & Kolenikov, S. (2008). Constrained versus unconstrained estimation in structural equation modeling. *Psychological Methods*, *13*, pp. 150–170. doi:10.1037/1082-989X.13.2.150

Sawaki, Y. (2007). Construct validation of analytic rating scales in a speaking assessment: Reporting a score profile and a composite. *Language Testing*, *24*, pp. 355–390. doi:10.1177/026553220707720

Sawaki, Y., & Sinharay, S. (2013). *Investigating the value of section scores for the TOEFL iBT® test.* (TOEFL iBT® Research Report No. TOEFL-iBT–21). Retrieved from: http://origin-www.ets.org/Media/Research/pdf/RR-13-35.pdf

Sawaki, Y., & Sinharay, S. (2018). Do the TOEFL iBT® section scores provide value-added information to stakeholders? *Language Testing*, *35*, pp. 529–556. doi:10.1177/0265532217716731

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

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

Shin, S.-K. (2005). Did they take the same test? Examinee language proficiency and the structure of language tests. *Language Testing*, *22*, pp. 31–57. doi:10.1191/0265532205lt296oa

Stricker, L. J., & Rock, D. A. (2008). *Factor structure of the TOEFL® Internet-based test across subgroups*. (TOEFL iBT Research Report No. TOEFL-iBT-07). Retrieved from: http://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2008.tb02152.x/pdf

Tabachnick, B., G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Harlow, Essex, UK: Pearson.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, pp. 4–70. doi:10.1177/109442810031002

Weir, C. J. (2005). *Language Testing and Evaluation: An evidence-based approach*. New York, NY: Palgrave Macmillan.

Weir, C. J., & O'Sullivan, B. (2017). *Assessing English on the global stage: The British Council and English language testing 1941–2016*. South Yorkshire, UK: Equinox.

Xi, X., & Mollaun, P. (2006). *Investigating the utility of analytic scoring for the TOEFL Academic Speaking Test (TAST).* (TOEFL iBT Research Report, RR-06-07). Retrieved from: http://www.ets.org/Media/Research/pdf/RR-06-07.pdf

APPENDIX A: MPLUS SYNTAX

(1) Mplus syntax for confirmatory factor analysis (Study 1: Research Question 1)
(1.1) Single factor model
(1.2) Correlated four factor model
(1.3) Uncorrelated four factor model
(1.4) Higher-order model
(1.5) Bi-factor model
(2) Mplus syntax for multiple-sample analysis (Study 1: Research Question 2)
(2.1) No equality constraints (i.e., configural invariance [the same factor structure with no equal constraints on any parameters])
(2.2) Mplus syntax for multiple-sample analysis: Equal factor loadings (i.e., metric invariance)
(3) Mplus syntax for latent profile analysis (Study 2: Research Question 3)

(1) Mplus syntax for confirmatory factor analysis (Study 1: Research Question 1)
Analysis using the aggregate data is presented here. For analyses on each country, the same procedure was repeated by replacing the data file.
(1.1) Single factor model

```
title: Aptis

data: file is Parcel_allcountries.txt;
  LISTWISE=ON;

variable:
  names are l1-l4 r1-r4 s2-s4 w1-w4 country; ! l1-l4 refer to the Listening A1, A2, B1, and B2 tasks, respectively. r1-r4 refer to the Reading A1, A2, B1, and B2
        tasks, respectively. s2-s4 refer to the Speaking A2, B1, and B2 tasks, respectively. w1-w4 refer to the Writing A1, A2, B1, and B2 tasks, respectively.
  usevariables are l1-l4 r1-r4 s2-s4 w1-w4;
  missing = .;

analysis:
  estimator = mlmv;

model:
  f1 by l1-l4 r1-r4 s2-s4 w1-w4;

savedata: difftest = parcel_allcountries_unitary;
  estimates = parcel_allcountries_unitary.dat;!for power analysis!

output: sampstat stand tech4;
```

(1.2) Correlated four-factor model

```
data: file is Parcel_allcountries.txt;
  LISTWISE=ON;

variable:
  names are l1-l4 r1-r4 s2-s4 w1-w4 country; ! l1-l4 refer to the Listening A1, A2, B1, and B2 tasks, respectively. r1-r4 refer to the Reading A1, A2, B1, and B2
        tasks, respectively. s2-s4 refer to the Speaking A2, B1, and B2 tasks, respectively. w1-w4 refer to the Writing A1, A2, B1, and B2 tasks, respectively.
  usevariables are l1-l4 r1-r4 s2-s4 w1-w4;
  missing = .;

analysis:
  estimator = mlmv;
  difftest = Parcel_allcountries_bifactor;

model:
  f2 by l1-l4;
  f3 by r1-r4;
  f4 by s2-s4;
  f5 by w1-w4;

savedata: difftest = parcel_allcountries_correlated;
  estimates = parcel_allcountries_correlated.dat;!for power analysis!

output: sampstat stand tech4;
```

(1.3) Uncorrelated four-factor model

```
title: Aptis

data: file is Parcel_allcountries.txt;
  LISTWISE=ON;

variable:
  names are l1-l4 r1-r4 s2-s4 w1-w4 country; ! l1-l4 refer to the Listening A1, A2, B1, and B2 tasks, respectively. r1-r4 refer to the Reading A1, A2, B1, and B2
        tasks, respectively. s2-s4 refer to the Speaking A2, B1, and B2 tasks, respectively. w1-w4 refer to the Writing A1, A2, B1, and B2 tasks, respectively.
  usevariables are l1-l4 r1-r4 s2-s4 w1-w4;
  missing = .;

analysis:
```

```
  estimator = mlmv;

model:
  f2 by l1-l4;
  f3 by r1-r4;
  f4 by s2-s4;
  f5 by w1-w4;
  f2 f3 f4 f5 with f2@0 f3@0 f4@0 f5@0;

savedata: difftest = parcel_allcountries_uncorrelated;
  estimates = parcel_allcountries_uncorrelated.dat;!for power analysis!

output: sampstat stand tech4;
```

(1.4) Higher-order model

```
title: Aptis

data: file is Parcel_allcountries.txt;
  LISTWISE=ON;

variable:
  names are l1-l4 r1-r4 s2-s4 w1-w4 country; ! l1-l4 refer to the Listening A1, A2, B1, and B2 tasks, respectively. r1-r4 refer to the Reading A1, A2, B1, and B2
        tasks, respectively. s2-s4 refer to the Speaking A2, B1, and B2 tasks, respectively. w1-w4 refer to the Writing A1, A2, B1, and B2 tasks, respectively.
  usevariables are l1-l4 r1-r4 s2-s4 w1-w4;
  missing = .;

analysis:
  estimator = mlmv;
  difftest = parcel_allcountries_correlated;

model:
  f2 by l1-l4;
  f3 by r1-r4;
  f4 by s2-s4;
  f5 by w1-w4;
  f2 f3 f4 f5 with f2@0 f3@0 f4@0 f5@0;
  f6 by f2 f3 f4 f5;

savedata:
```

```
  difftest = parcel_allcountries_higherorder;
  estimates = parcel_allcountries_higherorder.dat;!for power analysis!

output: sampstat stand tech4;
```

(1.5) Bi-factor model

```
title: Aptis

data: file is Parcel_allcountries.txt;
  LISTWISE=ON;

variable:
  names are l1-l4 r1-r4 s2-s4 w1-w4 country; ! l1-l4 refer to the Listening A1, A2, B1, and B2 tasks, respectively. r1-r4 refer to the Reading A1, A2, B1, and B2
        tasks, respectively. s2-s4 refer to the Speaking A2, B1, and B2 tasks, respectively. w1-w4 refer to the Writing A1, A2, B1, and B2 tasks, respectively.
  usevariables are l1-l4 r1-r4 s2-s4 w1-w4;
  missing = .;

analysis:
  estimator = mlmv;
  difftest = Parcel_allcountries_higherorder;

model:
  f2 by l1-l4;
  f3 by r1-r4;
  f4 by s2-s4;
  f5 by w1-w4;
  f6 by l1-l4 r1-r4 s2-s4 w1-w4;
  f6 with f2@0 f3@0 f4@0 f5@0;

savedata: difftest = parcel_allcountries_bifactor;
  estimates = Parcel_allcountries_bifactor.dat;!for power analysis!

output: sampstat stand tech4;
```

(2) Mplus syntax for multiple-sample analysis (Study 1: Research Question 2)

(2.1) No equality constraints (i.e., configural invariance [the same factor structure with no equal constraints on any parameters])

```
title: Aptis

data: file (Poland) = parcel_Poland.txt;
```

```
    file (Chile) = parcel_Chile.txt;
    file (Indonesia) = parcel_Indonesia.txt;
    file (Spain) = parcel_Spain.txt;
    file (SriLanka) = parcel_SriLanka.txt;
    file (Mexico) = parcel_Mexico.txt;
    listwise=on;

variable:
  names = l1-l4 r1-r4 s2-s4 w1-w4; ! l1-l4 refer to the Listening A1, A2, B1, and B2 tasks, respectively. r1-r4 refer to the Reading A1, A2, B1, and B2 tasks,
          respectively. s2-s4 refer to the Speaking A2, B1, and B2 tasks, respectively. w1-w4 refer to the Writing A1, A2, B1, and B2 tasks, respectively.
  usevariables = l1-l4 r1-r4 s2-s4 w1-w4;
  missing = .;

analysis:
  type =general;
  estimator = mlmv;
  model = nomeanstructure;
  information = expected;
  iteration = 10000;

model:
  f2 by l1-l4;
  f3 by r1-r4;
  f4 by s2-s4;
  f5 by w1-w4;
  f2 f3 f4 f5 with f2@0 f3@0 f4@0 f5@0;
  f6 by f2 f3 f4 f5;

  [f2@0 f3@0 f4@0 f5@0 f6@0];!factor means constrained to zero

  model Indonesia:
  f2 by l2-l4;
  f3 by r2-r4;
  f4 by s3-s4;
  f5 by w2-w4;
  f2 f3 f4 f5 with f2@0 f3@0 f4@0 f5@0;
  f6 by f3 f4 f5;
  [l1-l4 r1-r4 s2-s4 w1-w4]; !intercepts not constrained equal
  model Spain:
  f2 by l2-l4;
```

```
f3 by r2-r4;
f4 by s3-s4;
f5 by w2-w4;
f2 f3 f4 f5 with f2@0 f3@0 f4@0 f5@0;
f6 by f3 f4 f5;
[l1-l4 r1-r4 s2-s4 w1-w4]; !intercepts not constrained equal
model SriLanka:
f2 by l2-l4;
f3 by r2-r4;
f4 by s3-s4;
f5 by w2-w4;
f2 f3 f4 f5 with f2@0 f3@0 f4@0 f5@0;
f6 by f3 f4 f5;
      f2@0;!fix f2 variance to avoid negative residual
[l1-l4 r1-r4 s2-s4 w1-w4]; !intercepts not constrained equal
model Poland:
f2 by l2-l4;
f3 by r2-r4;
f4 by s3-s4;
f5 by w2-w4;
f2 f3 f4 f5 with f2@0 f3@0 f4@0 f5@0;
f6 by f3 f4 f5;
      f2@0;!fix f2 variance to avoid negative residual
[l1-l4 r1-r4 s2-s4 w1-w4]; !intercepts not constrained equal
model Chile:
f2 by l2-l4;
f3 by r2-r4;
f4 by s3-s4;
f5 by w2-w4;
f2 f3 f4 f5 with f2@0 f3@0 f4@0 f5@0;
f6 by f3 f4 f5;
      f3@0;!fix f3 variance to avoid negative residual
      f5@0;!fix f5 variance to avoid negative residual
[l1-l4 r1-r4 s2-s4 w1-w4]; !intercepts not constrained equal
model Mexico:
f2 by l2-l4;
f3 by r2-r4;
f4 by s3-s4;
f5 by w2-w4;
```

```
  f2 f3 f4 f5 with f2@0 f3@0 f4@0 f5@0;
  f6 by f3 f4 f5;
       f3@0;!fix f3 variance to avoid negative residual
  [l1-l4 r1-r4 s2-s4 w1-w4]; !intercepts not constrained equal

  output: sampstat stand tech1 tech4 modindices;

  SAVEDATA: DIFFTEST IS MGCFA_configural_noBangladesh.dat;
  estimates = MGCFA_configural_noBangladesh_power.dat;
```

(2.2) Mplus syntax for multiple-sample analysis: Equal factor loadings (i.e., metric invariance)

```
title: Aptis

data: file (Poland) = parcel_Poland.txt;
     file (Chile) = parcel_Chile.txt;
     file (Indonesia) = parcel_Indonesia.txt;
     file (Spain) = parcel_Spain.txt;
     file (SriLanka) = parcel_SriLanka.txt;
     file (Mexico) = parcel_Mexico.txt;
     VARIANCES=NOCHECK; !all Polish examinees scored correctly. Set this or otherwise will result in error
     listwise=on;

variable:
  names = l1-l4 r1-r4 s2-s4 w1-w4; ! l1-l4 refer to the Listening A1, A2, B1, and B2 tasks, respectively. r1-r4 refer to the Reading A1, A2, B1, and B2 tasks,
          respectively. s2-s4 refer to the Speaking A2, B1, and B2 tasks, respectively. w1-w4 refer to the Writing A1, A2, B1, and B2 tasks, respectively.
  usevariables are l1-l4 r1-r4 s2-s4 w1-w4;
  missing = .;

analysis:
  type =general;
  estimator = mlmv;
  model = nomeanstructure;
  information = expected;
  iteration = 10000;
  DIFFTEST IS MGCFA_configural_noBangladesh.dat;

model:
  f2 by l1-l4;
  f3 by r1-r4;
  f4 by s2-s4;
```

```
f5 by w1-w4;
f2 f3 f4 f5 with f2@0 f3@0 f4@0 f5@0;
f6 by f2 f3 f4 f5;

[f2@0 f3@0 f4@0 f5@0 f6@0];!factor means constrained to zero

model Indonesia:
!f2 by l2-l4;
!f3 by r2-r4;
!f4 by s3-s4;
!f5 by w2-w4;
f2 f3 f4 f5 with f2@0 f3@0 f4@0 f5@0;
!f6 by f3 f4 f5;
[l1-l4 r1-r4 s2-s4 w1-w4]; !intercepts not constrained equal
model Spain:
!f2 by l2-l4;
!f3 by r2-r4;
!f4 by s3-s4;
!f5 by w2-w4;
f2 f3 f4 f5 with f2@0 f3@0 f4@0 f5@0;
!f6 by f3 f4 f5;
[l1-l4 r1-r4 s2-s4 w1-w4]; !intercepts not constrained equal
model SriLanka:
!f2 by l2-l4;
!f3 by r2-r4;
!f4 by s3-s4;
!f5 by w2-w4;
f2 f3 f4 f5 with f2@0 f3@0 f4@0 f5@0;
!f6 by f3 f4 f5;
[l1-l4 r1-r4 s2-s4 w1-w4]; !intercepts not constrained equal
     f2@0;!fix f2 variance to avoid negative residual
model Poland:
!f2 by l2-l4;
!f3 by r2-r4;
!f4 by s3-s4;
!f5 by w2-w4;
f2 f3 f4 f5 with f2@0 f3@0 f4@0 f5@0;
!f6 by f3 f4 f5;
     f2@0;!fix f2 variance to avoid negative residual
```

```
  [l1-l4 r1-r4 s2-s4 w1-w4]; !intercepts not constrained equal
  model Chile:
  !f2 by l2-l4;
  !f3 by r2-r4;
  !f4 by s3-s4;
  !f5 by w2-w4;
  f2 f3 f4 f5 with f2@0 f3@0 f4@0 f5@0;
  !f6 by f3 f4 f5;
       f3@0;!fix f3 variance to avoid negative residual
       f5@0;!fix f5 variance to avoid negative residual
  [l1-l4 r1-r4 s2-s4 w1-w4]; !intercepts not constrained equal
  model Mexico:
  !f2 by l2-l4;
  !f3 by r2-r4;
  !f4 by s3-s4;
  !f5 by w2-w4;
  f2 f3 f4 f5 with f2@0 f3@0 f4@0 f5@0;
  !f6 by f3 f4 f5;
       f3@0;!fix f3 variance to avoid negative residual
  [l1-l4 r1-r4 s2-s4 w1-w4]; !intercepts not constrained equal

output: sampstat stand tech1 tech4 modindices (all 4);

savedata:  DIFFTEST IS MGCFA_configural_metric_noBangladesh.dat;
        estimates = MGCFA_configural_metric_noBangladesh_power.dat;
```

(3) Mplus syntax for latent profile analysis (Study 2: Research Question 3)

```
title: Aptis

data: file = parcel_allcountries.txt;
  listwise = on;

variable:
  names = l1-l4 r1-r4 s2-s4 w1-w4; ! l1-l4 refer to the Listening A1, A2, B1, and B2 tasks, respectively. r1-r4 refer to the Reading A1, A2, B1, and B2 tasks,
          respectively. s2-s4 refer to the Speaking A2, B1, and B2 tasks, respectively. w1-w4 refer to the Writing A1, A2, B1, and B2 tasks, respectively.

  usevariables = l1-l4 r1-r4 s2-s4 w1-w4;
  classes = c(5); !This is an example of 5 classes. Change the value to 1 through 4 in turn to estimate models with these different numbers of classes.
  missing = .;
```

```
define: standardize l1-l4 r1-r4 s2-s4 w1-w4;

analysis:
   type = mixture;
   estimator = mlr;
   starts = 1000 250;
   stiterations = 50;
   lrtbootstrap = 50;
   lrtstarts = 0 0 100 20;

output: sampstat tech1 tech4 tech14;

plot:
   type = plot3;
   series = l1 (1) l2 (2) l3 (3) l4 (4) r1 (5) r2 (6) r3 (7)
        r4 (8) s2 (9) s3 (10) s4 (11) w1 (12) w2 (13) w3 (14) w4 (15);

savedata:
   file = latentclass_parcel_allcountries_5class;
   save = cprob;
   format = free;
```

APPENDIX B: CORRELATION MATRICES

Table B1: Correlation matrix for all countries combined (N = 1,270)

	L_A1	L_A2	L_B1	L_B2	R_A1	R_A2	R_B1	R_B2	S_A2	S_B1	S_B2	W_A1	W_A2	W_B1	W_B2
L_A1	1														
L_A2	0.667	1													
L_B1	0.540	0.638	1												
L_B2	0.520	0.599	0.606	1											
R_A1	0.582	0.596	0.471	0.383	1										
R_A2	0.543	0.629	0.532	0.510	0.497	1									
R_B1	0.634	0.701	0.641	0.610	0.590	0.661	1								
R_B2	0.521	0.606	0.506	0.527	0.466	0.576	0.643	1							
S_A2	0.551	0.561	0.467	0.444	0.517	0.507	0.598	0.502	1						
S_B1	0.595	0.639	0.551	0.544	0.563	0.572	0.682	0.567	0.822	1					
S_B2	0.571	0.648	0.585	0.586	0.512	0.580	0.680	0.587	0.689	0.786	1				
W_A1	0.427	0.461	0.362	0.383	0.406	0.387	0.466	0.401	0.388	0.415	0.416	1			
W_A2	0.490	0.528	0.467	0.433	0.452	0.529	0.567	0.517	0.468	0.530	0.520	0.405	1		
W_B1	0.587	0.649	0.543	0.530	0.554	0.567	0.669	0.588	0.562	0.634	0.621	0.450	0.544	1	
W_B2	0.594	0.676	0.604	0.569	0.533	0.595	0.706	0.632	0.572	0.654	0.651	0.438	0.555	0.683	1

Table B2: Correlation matrix for Bangladesh (N = 403)

	L_A1	L_A2	L_B1	L_B2	R_A1	R_A2	R_B1	R_B2	S_A2	S_B1	S_B2	W_A1	W_A2	W_B1	W_B2
L_A1	1														
L_A2	0.626	1													
L_B1	0.518	0.616	1												
L_B2	0.407	0.508	0.437	1											
R_A1	0.539	0.518	0.418	0.28	1										
R_A2	0.522	0.587	0.443	0.381	0.452	1									
R_B1	0.588	0.627	0.561	0.471	0.566	0.598	1								
R_B2	0.549	0.578	0.504	0.479	0.455	0.552	0.679	1							
S_A2	0.538	0.514	0.436	0.359	0.529	0.480	0.593	0.534	1						
S_B1	0.556	0.556	0.445	0.395	0.512	0.535	0.633	0.566	0.852	1					
S_B2	0.546	0.649	0.560	0.502	0.477	0.573	0.654	0.636	0.679	0.728	1				
W_A1	0.357	0.353	0.288	0.283	0.372	0.310	0.408	0.377	0.315	0.336	0.334	1			
W_A2	0.461	0.477	0.432	0.366	0.455	0.518	0.584	0.503	0.450	0.495	0.523	0.274	1		
W_B1	0.574	0.605	0.485	0.402	0.499	0.538	0.620	0.578	0.533	0.600	0.604	0.340	0.526	1	
W_B2	0.579	0.661	0.603	0.497	0.500	0.571	0.719	0.672	0.587	0.646	0.706	0.384	0.593	0.667	1

Table B3: Correlation matrix for Chile (N = 117)

	L_A1	L_A2	L_B1	L_B2	R_A1	R_A2	R_B1	R_B2	S_A2	S_B1	S_B2	W_A1	W_A2	W_B1	W_B2
L_A1	1														
L_A2	0.656	1													
L_B1	0.439	0.466	1												
L_B2	0.381	0.414	0.578	1											
R_A1	0.264	0.382	0.366	0.197	1										
R_A2	0.389	0.600	0.436	0.564	0.251	1									
R_B1	0.481	0.593	0.605	0.656	0.314	0.622	1								
R_B2	0.325	0.446	0.334	0.319	0.089	0.399	0.404	1							
S_A2	0.282	0.232	0.242	0.318	0.010	0.280	0.426	0.232	1						
S_B1	0.271	0.307	0.359	0.484	0.150	0.391	0.472	0.31	0.742	1					
S_B2	0.283	0.298	0.432	0.490	0.154	0.376	0.455	0.297	0.622	0.797	1				
W_A1	0.115	0.319	0.343	0.402	0.314	0.274	0.355	0.223	0.143	0.198	0.227	1			
W_A2	0.257	0.335	0.294	0.378	0.121	0.451	0.396	0.267	0.211	0.357	0.366	0.162	1		
W_B1	0.234	0.384	0.421	0.363	0.111	0.308	0.457	0.232	0.125	0.274	0.275	0.242	0.339	1	
W_B2	0.304	0.431	0.332	0.244	0.149	0.379	0.510	0.363	0.179	0.138	0.229	0.165	0.183	0.289	1

Table B4: Correlation matrix for Indonesia (N = 95)

	L_A1	L_A2	L_B1	L_B2	R_A1	R_A2	R_B1	R_B2	S_A2	S_B1	S_B2	W_A1	W_A2	W_B1	W_B2
L_A1	1														
L_A2	0.611	1													
L_B1	0.452	0.505	1												
L_B2	0.389	0.338	0.408	1											
R_A1	0.472	0.257	0.353	0.050	1										
R_A2	0.404	0.323	0.328	0.312	0.369	1									
R_B1	0.394	0.478	0.341	0.304	0.360	0.545	1								
R_B2	0.283	0.463	0.162	0.224	0.163	0.291	0.412	1							
S_A2	0.376	0.431	0.323	0.175	0.310	0.403	0.456	0.325	1						
S_B1	0.340	0.477	0.312	0.244	0.295	0.375	0.432	0.423	0.738	1					
S_B2	0.286	0.334	0.308	0.197	0.303	0.447	0.513	0.458	0.569	0.617	1				
W_A1	0.098	0.156	0.205	-0.006	0.240	0.028	0.258	0.159	0.161	0.090	0.201	1			
W_A2	0.223	0.268	0.209	0.177	0.261	0.282	0.382	0.422	0.428	0.338	0.425	0.278	1		
W_B1	0.241	0.434	0.333	0.272	0.181	0.256	0.399	0.476	0.303	0.260	0.374	0.324	0.249	1	
W_B2	0.364	0.407	0.327	0.303	0.271	0.350	0.467	0.321	0.463	0.533	0.509	0.224	0.298	0.429	1

Table B5: Correlation matrix for Mexico (N = 331)

	L_A1	L_A2	L_B1	L_B2	R_A1	R_A2	R_B1	R_B2	S_A2	S_B1	S_B2	W_A1	W_A2	W_B1	W_B2
L_A1	1														
L_A2	0.656	1													
L_B1	0.559	0.674	1												
L_B2	0.585	0.678	0.688	1											
R_A1	0.623	0.689	0.554	0.497	1										
R_A2	0.567	0.677	0.609	0.594	0.556	1									
R_B1	0.690	0.773	0.727	0.715	0.629	0.716	1								
R_B2	0.500	0.579	0.524	0.520	0.504	0.514	0.624	1							
S_A2	0.587	0.657	0.543	0.531	0.568	0.546	0.623	0.500	1						
S_B1	0.676	0.767	0.665	0.668	0.636	0.643	0.751	0.580	0.860	1					
S_B2	0.656	0.752	0.670	0.672	0.608	0.658	0.765	0.591	0.752	0.877	1				
W_A1	0.524	0.562	0.459	0.468	0.465	0.501	0.555	0.414	0.507	0.542	0.534	1			
W_A2	0.574	0.606	0.529	0.499	0.527	0.588	0.629	0.535	0.576	0.648	0.603	0.537	1		
W_B1	0.661	0.734	0.614	0.616	0.669	0.647	0.751	0.595	0.667	0.729	0.722	0.547	0.618	1	
W_B2	0.633	0.720	0.652	0.645	0.596	0.653	0.754	0.555	0.597	0.710	0.723	0.517	0.573	0.745	1

Table B6: Correlation matrix for Poland (N = 100)

	L_A1	L_A2	L_B1	L_B2	R_A1	R_A2	R_B1	R_B2	S_A2	S_B1	S_B2	W_A1	W_A2	W_B1	W_B2
L_A1	1														
L_A2	0.557	1													
L_B1	0.265	0.491	1												
L_B2	0.485	0.508	0.564	1											
R_A1	0.210	0.267	0.178	0.191	1										
R_A2	0.581	0.605	0.461	0.495	0.272	1									
R_B1	0.590	0.591	0.438	0.491	0.476	0.795	1								
R_B2	0.453	0.545	0.517	0.580	0.353	0.592	0.653	1							
S_A2	0.521	0.426	0.264	0.291	0.102	0.481	0.385	0.391	1						
S_B1	0.384	0.395	0.287	0.428	0.136	0.424	0.342	0.392	0.605	1					
S_B2	0.361	0.448	0.338	0.427	0.232	0.441	0.420	0.450	0.539	0.629	1				
W_A1	0.538	0.424	0.149	0.377	0.051	0.377	0.325	0.369	0.313	0.284	0.297	1			
W_A2	0.568	0.430	0.227	0.416	0.074	0.519	0.411	0.496	0.435	0.384	0.292	0.559	1		
W_B1	0.461	0.469	0.366	0.519	0.264	0.485	0.575	0.461	0.275	0.223	0.296	0.473	0.385	1	
W_B2	0.381	0.406	0.425	0.522	0.176	0.402	0.387	0.540	0.313	0.404	0.382	0.352	0.392	0.435	1

Table B7: Correlation matrix for Spain (N = 137)

	L_A1	L_A2	L_B1	L_B2	R_A1	R_A2	R_B1	R_B2	S_A2	S_B1	S_B2	W_A1	W_A2	W_B1	W_B2
L_A1	1														
L_A2	0.542	1													
L_B1	0.386	0.537	1												
L_B2	0.476	0.569	0.536	1											
R_A1	0.364	0.375	0.312	0.251	1										
R_A2	0.324	0.492	0.392	0.308	0.325	1									
R_B1	0.426	0.580	0.528	0.546	0.389	0.529	1								
R_B2	0.293	0.481	0.193	0.389	0.202	0.518	0.562	1							
S_A2	0.359	0.337	0.274	0.384	0.333	0.193	0.427	0.330	1						
S_B1	0.435	0.450	0.499	0.471	0.345	0.399	0.542	0.423	0.699	1					
S_B2	0.395	0.431	0.422	0.467	0.281	0.371	0.524	0.352	0.514	0.591	1				
W_A1	0.244	0.248	0.198	0.199	0.146	0.202	0.319	0.141	0.184	0.182	0.222	1			
W_A2	0.240	0.311	0.360	0.218	0.086	0.249	0.422	0.215	0.090	0.327	0.197	0.270	1		
W_B1	0.394	0.398	0.385	0.424	0.113	0.324	0.510	0.390	0.396	0.506	0.422	0.076	0.292	1	
W_B2	0.349	0.465	0.400	0.442	0.196	0.384	0.456	0.476	0.354	0.482	0.309	0.127	0.397	0.535	1

Table B8: Correlation matrix for Sri Lanka (N = 87)

	L_A1	L_A2	L_B1	L_B2	R_A1	R_A2	R_B1	R_B2	S_A2	S_B1	S_B2	W_A1	W_A2	W_B1	W_B2
L_A1	1														
L_A2	0.594	1													
L_B1	0.448	0.487	1												
L_B2	0.508	0.472	0.413	1											
R_A1	0.479	0.505	0.403	0.330	1										
R_A2	0.408	0.493	0.507	0.488	0.380	1									
R_B1	0.538	0.690	0.627	0.606	0.507	0.600	1								
R_B2	0.461	0.509	0.455	0.572	0.390	0.572	0.691	1							
S_A2	0.534	0.439	0.414	0.519	0.378	0.451	0.508	0.426	1						
S_B1	0.576	0.642	0.563	0.618	0.520	0.500	0.747	0.719	0.567	1					
S_B2	0.532	0.556	0.446	0.614	0.431	0.408	0.663	0.589	0.633	0.823	1				
W_A1	0.373	0.334	0.244	0.356	0.344	0.257	0.416	0.310	0.295	0.383	0.362	1			
W_A2	0.377	0.311	0.394	0.243	0.495	0.311	0.405	0.311	0.249	0.428	0.343	0.169	1		
W_B1	0.560	0.575	0.487	0.601	0.463	0.413	0.637	0.507	0.514	0.700	0.669	0.459	0.488	1	
W_B2	0.537	0.550	0.523	0.618	0.459	0.405	0.620	0.617	0.524	0.747	0.644	0.399	0.403	0.696	1

APPENDIX C: DESCRIPTIVE STATISTICS OF LATENT PROFILE GROUPS

Table C1: Descriptive statistics for Part scores for each latent profile group

		Listening				Reading				Speaking			Writing			
Group		**A1**	**A2**	**B1**	**B2**	**A1**	**A2**	**B1**	**B2**	**A2**	**B1**	**B2**	**A1**	**A2**	**B1**	**B2**
Group 1	Mean	-1.59	-1.51	-1.09	-0.95	-1.84	-1.26	-1.48	-1.17	-1.60	-1.59	-1.30	-1.06	-1.16	-1.51	-1.37
	SD	0.65	0.58	0.70	0.69	0.63	0.68	0.51	0.71	0.53	0.48	0.55	0.85	0.77	0.64	0.61
Group 2	Mean	-0.56	-0.78	-0.79	-0.73	-0.12	-0.85	-1.01	-0.78	-0.25	-0.50	-0.81	-0.41	-0.60	-0.67	-0.86
	SD	0.65	0.58	0.70	0.69	0.63	0.68	0.51	0.71	0.53	0.48	0.55	0.85	0.77	0.64	0.61
Group 3	Mean	0.22	0.12	-0.11	-0.34	0.36	0.13	0.16	0.02	0.25	0.16	0.01	0.08	0.02	0.14	0.09
	SD	0.65	0.58	0.70	0.69	0.63	0.68	0.51	0.71	0.53	0.48	0.55	0.85	0.77	0.64	0.61
Group 4	Mean	0.60	0.75	0.76	0.83	0.44	0.68	0.80	0.71	0.65	0.78	0.85	0.47	0.61	0.68	0.76
	SD	0.65	0.58	0.70	0.69	0.63	0.68	0.51	0.71	0.53	0.48	0.55	0.85	0.77	0.64	0.61
Group 5	Mean	0.42	0.05	0.04	-0.05	0.31	0.14	0.14	-0.15	-2.38	-2.05	-1.35	0.22	0.20	0.28	-0.01
	SD	0.65	0.58	0.70	0.69	0.63	0.68	0.51	0.71	0.53	0.48	0.55	0.85	0.77	0.64	0.61

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.

FACTOR STRUCTURE AND FOUR-SKILL PROFILES OF THE APTIS TEST

Yo In'nami,
Chuo University, Japan
Rie Koizumi,
Juntendo University, Japan

AR-G/2021/2

ARAGs RESEARCH REPORTS ONLINE

ISSN 2057-5203