MICHIGAN LANGUAGE ASSESSMENT

# Latent Structure of the ECCE: Discovering Relationships among Reading, Listening, Writing, Speaking, and Lexico-Grammatical Ability

**Minkyung Kim**
**Scott A. Crossley**

SPAAN Research Grant
Working Paper

**2018-01**

Cambridge Assessment English
UNIVERSITY OF MICHIGAN
MichiganAssessment.org

# Table of Contents

## Index of Tables

## Abstract

The current study had two main purposes: (1) to examine the latent factor structure of the Examination for the Certificate of Competency in English (ECCE) and its generalizability across different groups (i.e., gender, age, and first language [L1]); and (2) to investigate the extent to which speaking and writing performances can be predicted by various individual differences (i.e., gender, L1, vocabulary knowledge, and listening and reading skills) and linguistic features as found in spoken and written responses. In the first analysis, the latent factor structure of the ECCE was examined through confirmatory factor analysis using performance scores from 9,700 test-takers. It was found that test-takers' performance on the ECCE could be best represented by a correlated three-factor model comprised of reading/listening/lexico-grammar, writing, and speaking abilities. Measurement invariance tests also reported that this three-factor model held equivalently across gender, age, and L1 (with the exception of vocabulary test scores). A second analysis using writing and speaking performances from 295 test-takers investigated the relationships among speaking and writing scores, individual differences, and linguistic features as found in spoken and written responses. A linear mixed effects modeling approach was used. The results indicated the speaking and writing models which combined individual differences and linguistic features explained 49.9% of the variance in speaking scores and 44.5% of the variance in writing scores. Overall, this study contributes to establishing the construct validity of the ECCE for measuring language competence and provides insights into links among speaking and writing scores, linguistic features, and individual differences.

## Introduction

The Examination for the Certificate of Competency in English (ECCE), developed by Michigan Language Assessment, is a test battery of standardized high-intermediate level English-as-a-foreign language (EFL) competency. Specifically, the ECCE aims at evaluating the B2 level of the Common European Framework of Reference (CEFR; Council of Europe, 2001). It assesses test-takers' English language proficiency in the skill areas of listening, reading, writing, speaking, vocabulary, and grammar. The ECCE is used for a variety of purposes, including educational program admissions, language course requirements, obtaining/improving employment, and personal interest (Michigan Language Assessment, 2017).

Given the widespread use of the ECCE, it is crucial to establish its construct validity (i.e., the degree to which a test measures the theoretical construct defined; Nunnally, 1978). A test with strong construct validity means that score interpretations are closely in line with the test's purposes and intentions. One of the approaches to examining construct validity is to examine the latent factor structure of the ECCE by analyzing the relationship between test scores and the constructs measured (e.g., Bae & Bachman, 1998; Bollen, 1989; Messick, 1996). In addition to establishing construct validity, it is equally important to examine the relationships among various skills assessed in the ECCE. Given that the ECCE includes language production (i.e., speaking and writing), it is worth examining how speaking and writing performances are associated with test-takers' individual differences (e.g., vocabulary knowledge, listening skills, and first languages [L1s]) and their ability in language use (e.g., lexical choices).

## Construct Validity and Latent Structures of Language Tests

Validity is defined as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (Messick, 1989, p. 11). That is, validity is a criterion for evaluating the extent to which a test measures what it is supposed to measure. Construct validity is a crucial measure of test validity that evaluates the extent to which a given test score can be indicative of the construct(s) that the test proposes to assess. To examine construct validity, previous research has explored the latent structure of factors (i.e., underlying, unobservable variables of multiple observed variables) of proficiency tests by examining the links among language abilities (e.g., ability to listen, speak, read, and write) as measured by various proficiency tests, such as the Test of English as a Foreign Language (TOEFL; Gu, 2014; Sawaki & Sinharay, 2013; Shin, 2005), the Michigan English Language Assessment Battery (MELAB; Wang, 2006), the Examination for the Certificate of Proficiency in English (ECPE; Wang, 2006), and the Test of English for International Communication (TOEIC; In'nami & Koizumi, 2011). That latent factors match the constructs which a test proposes to measure indicates that the test fulfills its purpose of measuring those constructs.

In establishing links between the latent construct of language competence and observed test performance, researchers have investigated whether second language (L2) competence (i.e., L2 ability) is unitary or consists of separable components.[1] While some research has found that language competence consists of a unitary component (e.g., Oller, 1976; Wang, 2006), other research has argued for language competence consisting of multiple components (e.g., Bachman & Palmer, 1982; Carroll, 1965, Gu, 2014, 2015; Harley, Allen, Cummins, & Swain, 1990; In'nami & Koizumi, 2011; Sawaki, Sinharay, & Oranje, 2009; Sawaki & Sinharay, 2013; Shin, 2005). The general consensus in L2 testing is that language competence consists of a general language component along with smaller, specific language components, such as vocabulary knowledge and writing skills (Sawaki, Sinharay, & Oranje, 2009).

[1] Here, an L2 is used as a broad term which is referred to as a language that is not a native language, including *n*th (e.g., second and third) languages and foreign languages.

Latent constructs of language competence vary depending on which language tests are used. Even in examining the same test, such as the TOEFL Internet-Based Test (iBT), different latent structures have been identified. For example, using a confirmatory factor analysis (CFA; a statistical procedure to test a relationship between observed variables and their underlying latent factors [or constructs] based on the latent factor structure pre-determined by researchers) approach, Gu (2014) found that the test was best represented by a two-factor model comprised of the ability to speak, and the ability to listen, read, and write for 370 test takers' performances. This study attributed the distinction between a speaking factor and a non-speaking factor to the potential effect of instruction on test-takers. While reading, listening, and writing skills have been traditionally emphasized as important language skills in TOEFL preparation and training, speaking skills have begun to be emphasized relatively recently since the TOEFL iBT (which included a mandatory speaking section) launched in 2005.

A different factorial representation of the TOEFL iBT was found by Sawaki and Sinharay (2013), who used a CFA approach with a larger sample ($n = 50{,}393$). Results indicated that test-takers' performance was best explained by a four-factor latent structure correlating to the four sub-skills (i.e., listening, speaking, reading, and writing). They also found the four factors were strongly correlated with each other ($.59 < r < .89$), although the speaking factor was relatively less strongly correlated with the other factors. In short, in examining construct representations of the TOEFL iBT, Gu (2014) and Sawaki and Sinharay (2013) suggested different factor models as best representing the test, while they had in common that speaking skills were distinguished from other language skills to some degree.

On the other hand, some studies have reported language competence consisting of a unitary general factor (e.g., Oller, 1976; Wang, 2006). For instance, in the investigation of the ECPE and the MELAB, using an exploratory factor analysis (EFA; a variable reduction procedure to decide the number of factors and the factor structure of a set of observed variables without a priori fixed factor model) approach, Wang (2006) found that in both tests a single general proficiency latent factor represented test-takers' observed performances. Specifically, results indicated that the latent structure of the ECPE ($n = 2{,}011$) was represented by a single proficiency factor, which represented general language proficiency and consisted of speaking, listening, grammar, cloze, vocabulary, and reading scores, while the latent structure of the MELAB ($n = 216$) was represented by another single general proficiency factor that consisted of writing, listening, grammar, cloze, vocabulary, and reading scores.[2] However, correlation analyses hinted at the separate nature of productive and non-productive skills. In the ECPE, speaking scores showed weak-to-moderate correlations with the other scores (listening, grammar, cloze, vocabulary, and reading; $.20 \leq r \leq .43$), whereas the other skills showed moderate-to-strong with each other ($.38 \leq r \leq .62$). Similarly, in the MELAB, composition scores showed weak correlations with the other scores (i.e. listening, grammar, cloze, vocabulary, and reading; $.14 < r < .20$), while the other scores showed strong correlations with each other ($.52 \leq r \leq .74$). Unlike Gu (2014) and Sawaki and Sinharay (2013), one caveat to interpreting Wang's (2006) study is that it used an EFA approach, and other alternative latent models were not tested.

Beyond examining construct validity, it is also important to test whether the structure of the test identified is generalizable across different groups (e.g., gender) to ensure that the test measures the same constructs for different groups. The generalizability of constructs identified in a test across different groups should not be taken for granted (Messick, 1989). Previous studies have examined whether the latent structure of language tests is generalizable across gender (Wang, 2006), target language contact (Gu, 2014), L1s (Sawaki & Sinharay, 2013), and randomly split samples of test-takers (In'nami & Koizumi, 2011). For instance, Gu (2014) found that a two-factor model of the TOEFL iBT (i.e., speaking and non-speaking factors) did not function differently between a study-abroad group (i.e., test takers who had been exposed to an English-speaking environment) and a group without study-abroad, which supported factorial invariance of the language competence measured by the TOEFL iBT. Sawaki and Sinharay (2013) indicated that a four-factor latent structure correlating to the four sub-skills (i.e., listening, speaking, reading, and writing) functioned equally for three different L1 groups (i.e., Arabic, Korean, and Spanish). Wang (2006) found that a single general proficiency latent factor of the MELAB and the ECPE, respectively, held equally across gender.

[2] Wang's (2006) study did not include ECPE writing or MELAB speaking data.

### Speaking and Writing Performance, Individual Differences, and Linguistic Features

In addition to establishing construct validity for language tests, it is also crucial to examine factors that have influences on assessing language production performances (i.e., speaking and writing), such as individual differences (characteristics of individuals; e.g., age, L1s, and vocabulary knowledge) and linguistic features found in test-takers' written and spoken responses. Many researchers have highlighted the importance of individual differences in assessing language competence and how these individual differences interact with their language performance in a test (e.g., Alderson & Banerjee, 2002; Bachman & Palmer, 1996; Harley, Cummins, Swain, & Allen, 1990). Other researchers also examined the links between test-takers' performances and linguistic features found in language samples, such as lexical, syntactic, and cohesive features (e.g., Biber, Gray, & Staples, 2014; Crossley & McNamara, 2012). Below, we briefly report previous findings about the relationships between individual differences and speaking/writing performance as well as between linguistic features found in spoken and writing responses and speaking/writing performance.

There are various individual differences that should be considered in language testing, including demographic/personal characteristics and language knowledge and skills (Alderson & Banerjee, 2002; Bachman & Palmer, 1996). Demographic characteristics include individual attributes which are not directly related to test-takers' language competence but which still may impact their language test performance (Bachman & Palmer, 1996). Among various demographic characteristics, three important ones are gender, age, and L1. Previous studies have demonstrated that females tend to show better performance on language tasks than males (e.g., Sunderland, 2000; Pavlenko & Piller, 2008). Age has also been considered crucial in language performance (e.g., Krashen, Long & Scarcella, 1979; MacWhinney, 2005; Nikolov & Djigunović, 2006). Particularly in EFL school contexts, it has been generally assumed that students in higher grades (e.g., 12th-grade students) would be more proficient English learners than those in lower grades (e.g., 3rd-grade students) not only because general school curriculum of an English subject expands with grade in many non-English speaking countries such as Brazil and China (Braine, 2005; Nikolov & Djigunović, 2006) but also because the human capacity of processing information (e.g., memorizing new words) increases with age until it peaks at around the age of 22 (Hulstijn, 2011; Salthouse, 2009). In addition, L1s have also been emphasized in terms of distances between L1s and L2s under the assumption that the more linguistically distant an L1 and an L2 are, the more difficult it is for individuals to learn the L2 (e.g., Van der Slik, 2010; Schepens, Van der Slik, & Van Hout, 2013).

In addition to demographic characteristics, test-takers' language skills and knowledge (e.g., vocabulary knowledge, and listening and reading skills) also influence speaking and writing performance. For instance, higher vocabulary knowledge in the L2 is linked to higher-rated speaking performance (e.g., de Jong et al., 2012; Koizumi & In'nami, 2015) and writing performance (e.g., Stæhr, 2008; Schoonen et al., 2003, 2011). Furthermore, it has been widely argued that listening and speaking skills share similar characteristics, such as the processing of oral language and the use of high-frequency lexical and grammatical features, while reading and writing share similar characteristics, such as the processing of written language and the use of low-frequency lexical and grammatical features (Bachman & Palmer, 1996; Hulstijn, 2011). Empirical studies have also reported moderate-to-strong correlations between listening and speaking skills (Liu & Costanzo, 2013; Sawaki & Sinharay, 2013; Wang, 2006), as well as moderate-to-strong correlations between reading and writing skills (Abu-Akel, 1997; Carson, Carrell, Silberstein, Kroll, & Kuehn, 1990; Sawaki & Sinharay, 2013).

Beyond individual differences, many researchers have also focused on how linguistic features found in L2 spoken and written responses relate to speaking scores (e.g., Kang, 2013; Laflair, Staples, & Egbert, 2015; Laflair & Staples, 2017) and writing scores (e.g., Crossley & McNamara, 2012; Leki, Cumming, & Silva, 2008) under the notion that L2 writers' linguistic production can impact raters' judgments of test-takers' speaking/writing performance. In writing contexts, higher-rated L2 essays tend to include greater lexical diversity, lower-frequency words, less familiar words, and more specific words (Crossley & McNamara, 2012; Crossley & McNamara, 2014; Guo, Crossley, & McNamara, 2013; Kyle & Crossley, 2016; Jarvis, 2002; Laufer & Nation, 1995); longer clauses and sentences with phrasal elaboration (Biber, Gray, & Poonpon, 2011; Crossley & McNamara, 2012; Lu, 2010; Yang, Lu, & Weigle, 2015); and fewer connectives and less word overlap between sentences (Crossley & McNamara, 2012). On the other hand, in speaking contexts, higher-rated L2 spoken responses tend to include a greater number of word types (i.e., unique words; Crossley & McNamara, 2013); fewer clausal *and* features (Biber et al., 2014); fewer first-person pronouns and fewer nouns (Kang, 2013; Laflair et al., 2015); fewer hesitation markers (Laflair et al., 2015); more likelihood adverbials (e.g., *maybe,* *possibly*; Laflair et al., 2015); greater causal cohesion (Crossley & McNamara, 2013); and more of the features found in oral narratives (Laflair & Staples, 2017).

## The Current Study

Previous studies examined latent structures of various tests including the TOEFL, the MELAB, and the TOEIC. However, no study on the latent structure of the ECCE has been conducted. In addition, while many studies have examined the relationship between individual differences and speaking/writing performances, and between linguistic features and speaking/writing performances, few studies have simultaneously examined links among individual differences, linguistic features, and speaking/writing performances.

To address the construct validity and the relationships among skills and knowledge tested in the ECCE, the current study conducts two main analyses. First, the latent factor structure of the ECCE will be examined through confirmatory factor analysis using scores from 9,700 test-takers. It will be further examined whether the latent structure is generalizable across different groups (i.e., gender, age, and L1) to ensure that the test assesses the same constructs. Investigating the latent structure of the ECCE and the generalizability of its structure would contribute to ensuring its construct validity by understanding the relationship between test scores and the constructs measured. Second, this study will investigate how speaking and writing performances can be predicted not only by various individual differences (i.e., gender, L1, age, vocabulary knowledge, and listening and reading skills) but also by linguistic features as found in spoken and written responses produced by test-takers. Linking speaking and writing scores with individual differences and linguistic features will help test administrators and teachers better understand how linguistic features and individual differences predict test-takers' speaking and writing performances. Thus, this study is guided by two main research questions (RQs) each with two corresponding sub-questions:

1. What is the relationship between test scores and the constructs measured in the ECCE?

    1.a. What is the latent structure of the ECCE that best represents test-takers' performances?

    1.b. To what extent is the latent structure of the ECCE generalizable across gender, age, and L1?

2. What is the relationship among writing and speaking scores, test-takers' individual differences, and linguistic features found in speaking and writing samples?

2.a. To what extent are speaking scores predicted by test-takers' individual differences and linguistic features found in speaking samples?

2.b. To what extent are writing scores predicted by test-takers' individual differences and linguistic features found in writing samples?

## The Structure of the ECCE

To establish construct validity in a language test, interpretations of scores in the given test should be justified such that test scores indicate test-takers' language competence that the test intends to measure, and evidence that justifies those interpretations should be provided (Bachman & Palmer, 1996; Messick, 1989). Justifying the interpretations made based on test scores begins with how constructs are defined for a given test situation along with test purposes and design (Bachman & Palmer, 1996).

The main construct that the ECCE intends to measure is general English proficiency at the high-intermediate level. Specifically, the aim of the ECCE is to assess English proficiency at the B2 level of the Common European Framework of Reference (CEFR; Council of Europe, 2001). Language learners at this proficiency level:

- Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialization.
- Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party.
- Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.

(Council of Europe, 2001, p. 24)

Thus, the ECCE aims to measure three main elements of language competence: (a) understanding of complex input; (b) interacting fluently; and (c) producing clear text. Accordingly, the ECCE consists of four sections: Listening, Grammar/Vocabulary/Reading (GVR), Speaking, and Writing. The Listening and GVR sections relate to the ability to comprehend complex input, the Speaking section relates to the ability to speak in an interactive and fluent manner, and the Writing section relates to the ability to produce clear text.

The Listening section comprises two sub-sections: short conversations with 30 multiple-choice items (Part 1) and short talks with 20 multiple-choice items (Part 2). In Part 1, after listening to each short conversation, test-takers hear a question and are asked to select one of the three picture options that accurately answers the question. In Part 2, after listening to each short talk addressed by single speakers on different topics (e.g., a lecture about history and a talk delivered by a manager to his or her employees), test-takers are asked to answer four-to-six questions by selecting one of the four options that accurately answers the questions. The questions in both Parts assess test-takers' understanding of the given conversations and talks (e.g., understanding a main idea and details).

The GVR section consists of three sub-sections: Grammar with multiple-choice items, Vocabulary with multiple-choice items, and Reading with multiple-choice items. The Grammar and Vocabulary items ask test-takers to complete a sentence (e.g., "*It is better________ the job now rather than leave it for tomorrow.*") by selecting one of the four options that best completes the sentence (e.g., *finishes, to finish, finish*, and *finished*). The Reading section consists of two sub-sections: reading short passages (Part 1) and reading sets of four short texts related to each other by topic (Part 2). Each question in the Reading section has four options (i.e., one correct answer and three distractors), and assesses test-takers' literal and analytic understanding of the given passages. Test-takers are given 90 minutes to complete the entire GVR section.

In the Speaking section, test-takers participate in a structured multitask interview with one examiner. The Speaking section consists of four tasks. In Tasks 1–3, a hypothetical scenario is provided in which a test-taker is asked to solve a problem (e.g., deciding how to celebrate a town's 100th anniversary between two options). Task 1 requires the test-taker to figure out the problem by asking questions to the examiner, Task 2 to explain which option the test-taker thinks is best and why, and Task 3 to explain why the test-taker did not choose the other option. In Task 4, three elaboration questions related to the scenario are asked. Examples of the three elaboration questions are as follows:

- What is an important event that you remember? Why?;
- What are some ways people can remember special occasions in their community?; and
- Some people believe that public money should not be used for occasions like town anniversaries because such events do not directly benefit anyone. To what extent do you think this is true?

Task 1 is a warm-up activity for helping establishing rapport between the test-taker and the examiner, and thus it is unscored. Tasks 2 and 3 are scored together. Test-takers' performances on Tasks 2 and 3, and Task 4 are evaluated using the same analytic five-point rating scale with three criteria (i.e., *overall communicative effectiveness*, *language control/resources,* and *intelligibility/delivery*).[3] For Task 4, test-takers' performances on the three elaboration questions are separately evaluated for the criterion of *overall communicative effectiveness* (i.e., three different scores for each of the three questions), while being evaluated together for the criteria of *control/resources* and *intelligibility/delivery* (i.e., one score for all of the three questions).

The Writing section requires test-takers to read a short excerpt from a newspaper article about a situation or issue (e.g., increasing the cost of tickets for the city's professional soccer team) and then write a letter or essay giving an opinion about the situation or issue. Test-takers are provided 30 minutes to write the letter or essay. Each writing sample is rated separately by two expert raters using an analytic five-point rating scale with four criteria (i.e., *content and development, organization and connection of ideas, linguistic range and control,* and *communicative effect*).[4] Two ratings are summed. If two raters have nonadjacent scores for a writing sample, a third rater evaluates it.

## Analysis 1 Method

The purpose of this analysis was two-fold. First, it investigates the latent structure of the ECCE that best represented test-takers' performances (RQ1.a). Second it examines whether the latent structure of the ECCE could be generalizable across gender, L1, and age (RQ1.b).

### Data

We analyzed the response data of 9,700 ECCE test takers. The L1s of the test-takers included 14 different languages (see Table 1). The majority of test-takers were Greek-speaking (90.9%). Around 7

[3] The speaking rating rubrics are available on the Michigan Language Assessment website at http://michiganassessment.org/wp-content/uploads/2014/11/ECCE-Rating-Scale-Speaking-20140220.pdf.

[4] The writing rating rubrics are available on the Michigan Language Assessment website at http://michiganassessment.org/wp-content/uploads/2014/11/ECCE-Rating-Scale-Writing-20140220.pdf.

percent of the test-takers were Spanish-speaking. Table 2 presents the distribution of test-takers by gender. Among the 9,700 test-takers, 5,341 were female (55.1%) and 4,330 were male (44.6%). Gender was not reported for the remaining 29 test-takers (0.3%). Table 3 presents the distribution of test-takers by age.[5] The test-takers ranged in age from 10 to 61 with a mean of 15.91 (*SD* = 5.10). The test population primarily consisted of test-takers whose ages were between 13 and 16 (i.e., the first years of secondary school; 79.7%). These distributions by native languages, gender, and age were similar to those previously reported for the general test population (Michigan Language Assessment, 2017).

Table 1: ECCE Test-Takers by First Languages

| Native language | Number | Percentage |
|---|---|---|
| Greek | 8,814 | 90.9 |
| Spanish | 683 | 7.0 |
| Arabic | 73 | 0.8 |
| Portuguese | 69 | 0.7 |
| Albanian | 31 | 0.3 |
| Vietnamese | 14 | 0.1 |
| Other languages[a] | 16 | 0.1 |
| Total | 9,700 | 100.0 |

*Note.* [a] Other languages with fewer than 10 test-takers include Georgian, French, Romanian, Ukrainian, Armenian, Cambodian, German, and Macedonian.

Table 2: ECCE Test-Takers by Gender

| Native language | Number | Percentage |
|---|---|---|
| Female | 5,341 | 55.1 |
| Male | 4,330 | 44.6 |
| Unreported | 29 | 0.3 |
| Total | 9,700 | 100.0 |

[5] Age groups are based on the ECCE 2017 report.

Table 3: ECCE Test-Takers by Age

| Age | Number | Percentage |
|---|---|---|
| ≤ 12 | 291 | 3.0 |
| 13–16 | 7,728 | 79.7 |
| 17–19 | 606 | 6.2 |
| 20–22 | 330 | 3.4 |
| 23–25 | 244 | 2.5 |
| 26–29 | 185 | 1.9 |
| 30–39 | 174 | 1.8 |
| ≥ 40 | 118 | 1.1 |
| Missing data[a] | 24 | 0.2 |
| Total | 9,700 | 100 |

*Note.* [a] Missing data include those who did not report their age ($n$ = 5), those who likely erroneously reported their age as nine and below ($n$ = 8), and those who likely erroneously reported their age as 95 and above ($n$ = 11).

## Statistical Analysis

**Confirmatory factor analysis.** To examine the latent structure of the ECCE, we used Confirmatory factor analysis (CFA), which examines the relationships among observable variables (i.e., measurable variables or indicators) and their underlying latent variables (i.e., factors; Kline, 2011). CFA was used to test whether a latent variable model adequately represented the fit for the covariances (i.e., unstandardized correlations) between the observable variables (Kline, 2011). CFA was conducted using R (R Development Core Team, 2014) and *lavaan* packages (Rosseel, 2012). Multivariate normality was checked using Mardia's normalized estimate, with values below five considered to indicate multivariate normality (Byrne, 2006). Latent variables were represented by ovals, while observable variables were represented by squares. When evaluating the model, the latent variables were fixed at 1.0 such that factor loadings (i.e., measures of the influence a latent variable has on indicator variables) for each indicator variable were comparable.

**Hypothesized models.** In the CFA, five competing hypothesized models were constructed to determine which model would best represent the latent structure of the ECCE. Each model is briefly discussed below.

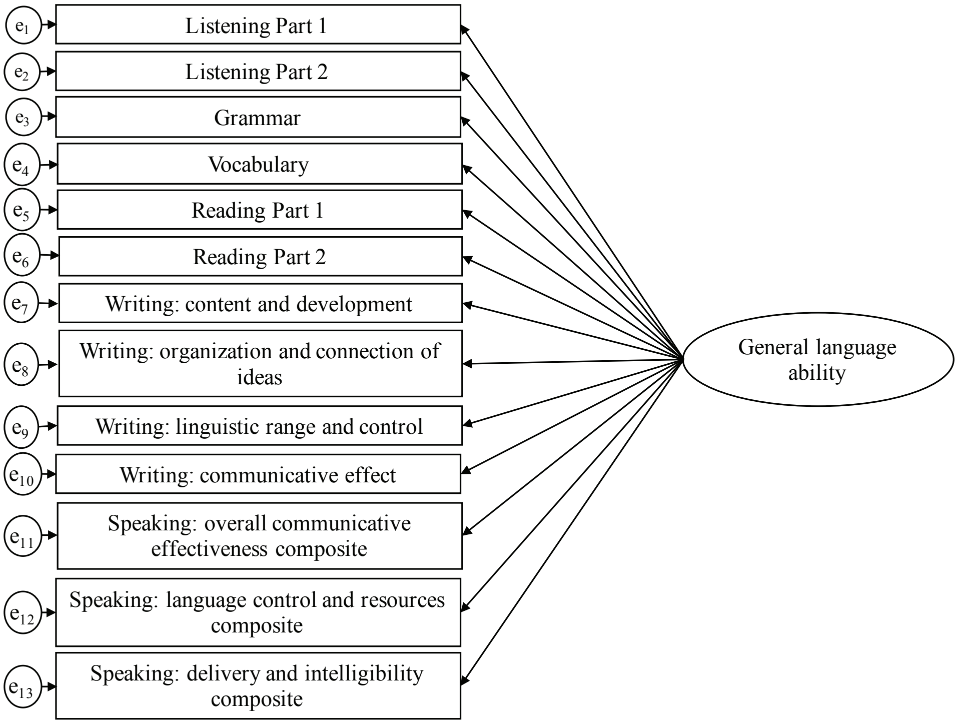
*Single-factor model (Figure 1).* In the single-factor model, five language abilities (i.e., reading, listening, writing, speaking, and lexico-grammatical abilities) load on the same factor (i.e., general language ability). As such, this model assumes that there is a general language proficiency informed by the five language skills which are not distinctive from each other at a latent level, suggesting the nature of language proficiency as a single unitary construct. This model was constructed based on previous research using the ECPE and the MELAB, which found that in both tests a single general proficiency latent factor underlay test-takers' observed performances (Wang, 2006).

*Correlated two-factor model (Figure 2).* In the correlated two-factor model, two distinct but correlated factors are specified: one for speaking and the other for listening, reading, lexico-grammar, and writing. This model is based on Gu (2014), who found that the two factors (i.e., speaking and listening/reading/writing) best represented the TOEFL iBT.

*Higher-order factor model (Figure 3).* In the higher-order factor model, a higher-order latent factor of general language proficiency is specified along with five first-order language ability latent factors (i.e., reading, listening, writing, speaking, and lexico-grammatical abilities). It is hypothesized that the positive correlations among the five language (first-order) factors are explained by a general (second- or higher-order) factor of language ability. This model was constructed in accordance with the scoring scheme of the ECCE, which reports each skill score along with a total score.

*Correlated five-factor model (Figure 4).* In the correlated five-factor model, five distinct but correlated factors are specified, each of which corresponds to reading, listening, writing, speaking, and lexico-grammatical abilities as measured by the ECCE.

*Correlated four-factor model (Figure 5).* The correlated four-factor model specifies four distinct but correlated factors that correspond to the four sections of the ECCE: listening, GVR, writing, and speaking. This model that corresponds to the test structure of the ECCE is in line with previous research on the TOEFL iBT (Sawaki & Sinharay, 2013) which found that test-takers' performance could be best explained by a four-factor latent structure correlating to the four language skills (i.e., listening, speaking, reading, and writing) as measured by the TOEFL iBT.



e1
Listening Part 1
e2
Listening Part 2
e3
Grammar
e4
Vocabulary
e5
Reading Part 1
e6
Reading Part 2
e7
Writing: content and development
e8
Writing: organization and connection of ideas
e9
Writing: linguistic range and control
e10
Writing: communicative effect
e11
Speaking: overall communicative effectiveness composite
e12
Speaking: language control and resources composite
e13
Speaking: delivery and intelligibility composite
General language ability

*Figure 1.* Single-factor model

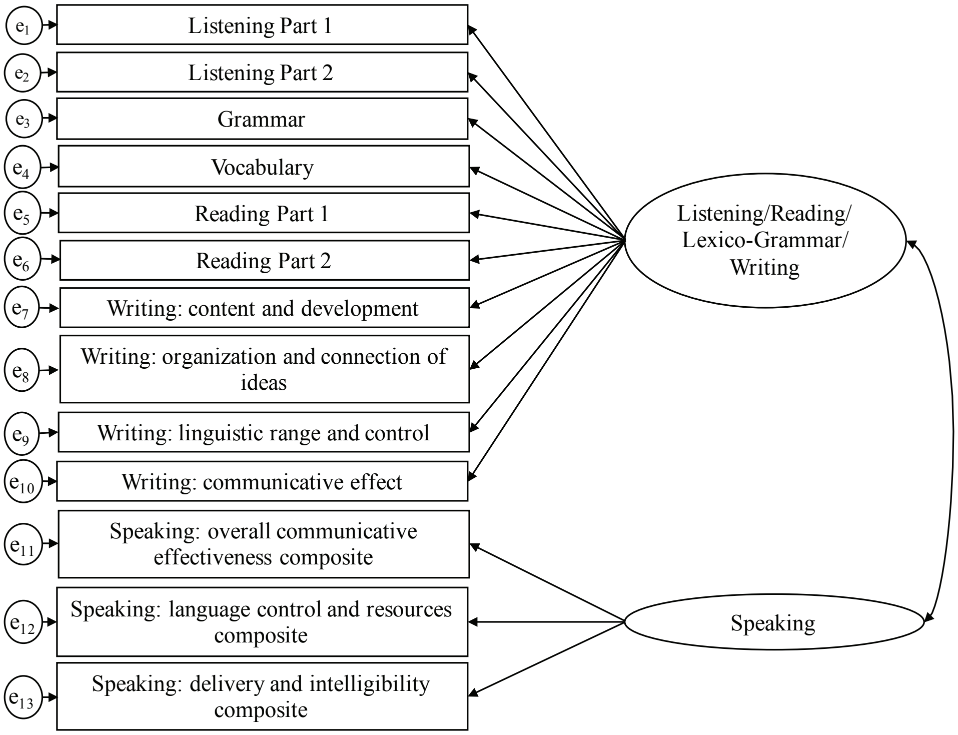

e1
Listening Part 1
e2
Listening Part 2
e3
Grammar
e4
Vocabulary
e5
Reading Part 1
e6
Reading Part 2
e7
Writing: content and development
e8
Writing: organization and connection of ideas
e9
Writing: linguistic range and control
e10
Writing: communicative effect
e11
Speaking: overall communicative effectiveness composite
e12
Speaking: language control and resources composite
e13
Speaking: delivery and intelligibility composite
Listening/Reading/ Lexico-Grammar/ Writing
Speaking

*Figure 2.* Correlated two-factor model

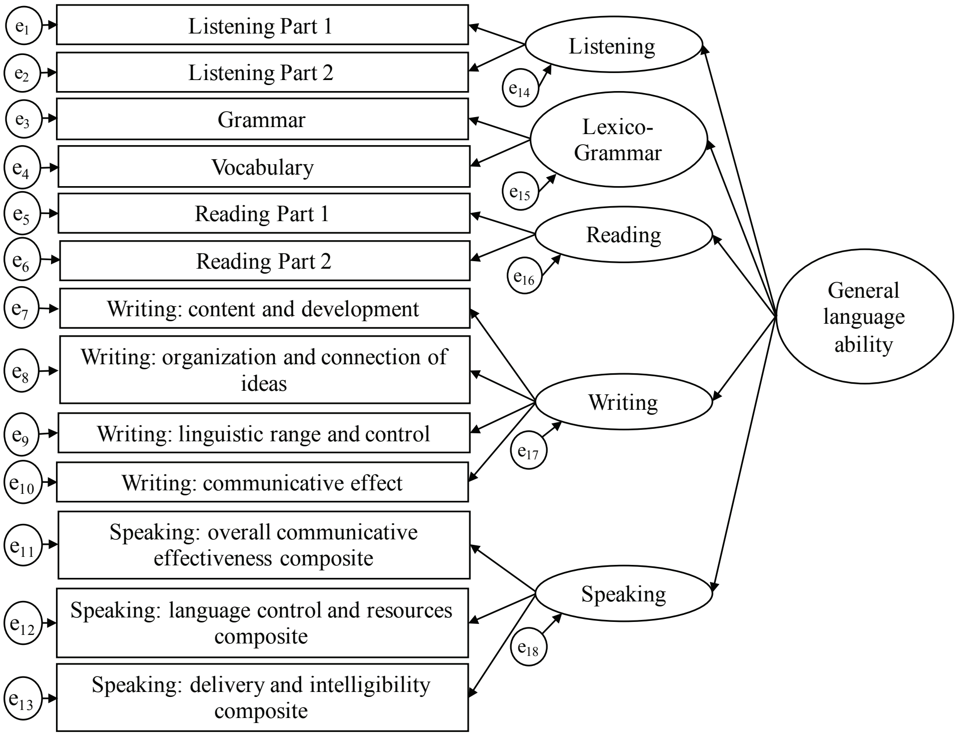

e1
Listening Part 1
e2
Listening Part 2
e3
Grammar
e4
Vocabulary
e5
Reading Part 1
e6
Reading Part 2
e7
Writing: content and development
e8
Writing: organization and connection of ideas
e9
Writing: linguistic range and control
e10
Writing: communicative effect
e11
Speaking: overall communicative effectiveness composite
e12
Speaking: language control and resources composite
e13
Speaking: delivery and intelligibility composite
Listening
e14
Lexico-Grammar
e15
Reading
e16
Writing
e17
Speaking
e18
General language ability

*Figure 3.* Higher-order factor model

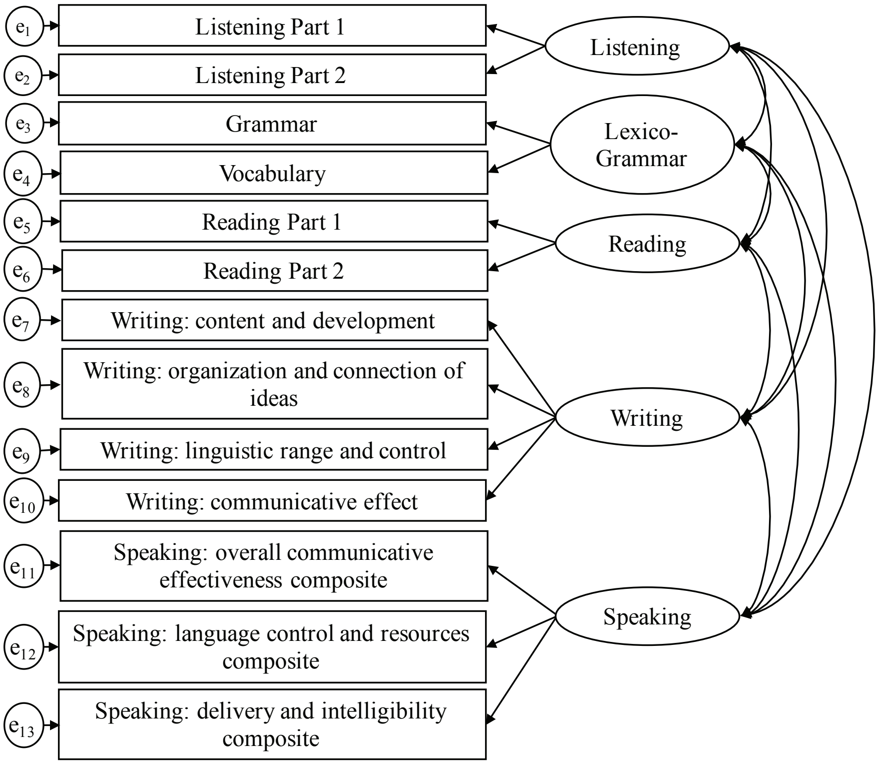

e1
Listening Part 1
e2
Listening Part 2
e3
Grammar
e4
Vocabulary
e5
Reading Part 1
e6
Reading Part 2
e7
Writing: content and development
e8
Writing: organization and connection of ideas
e9
Writing: linguistic range and control
e10
Writing: communicative effect
e11
Speaking: overall communicative effectiveness composite
e12
Speaking: language control and resources composite
e13
Speaking: delivery and intelligibility composite
Listening
Lexico-Grammar
Reading
Writing
Speaking

*Figure 4.* Correlated five-factor model

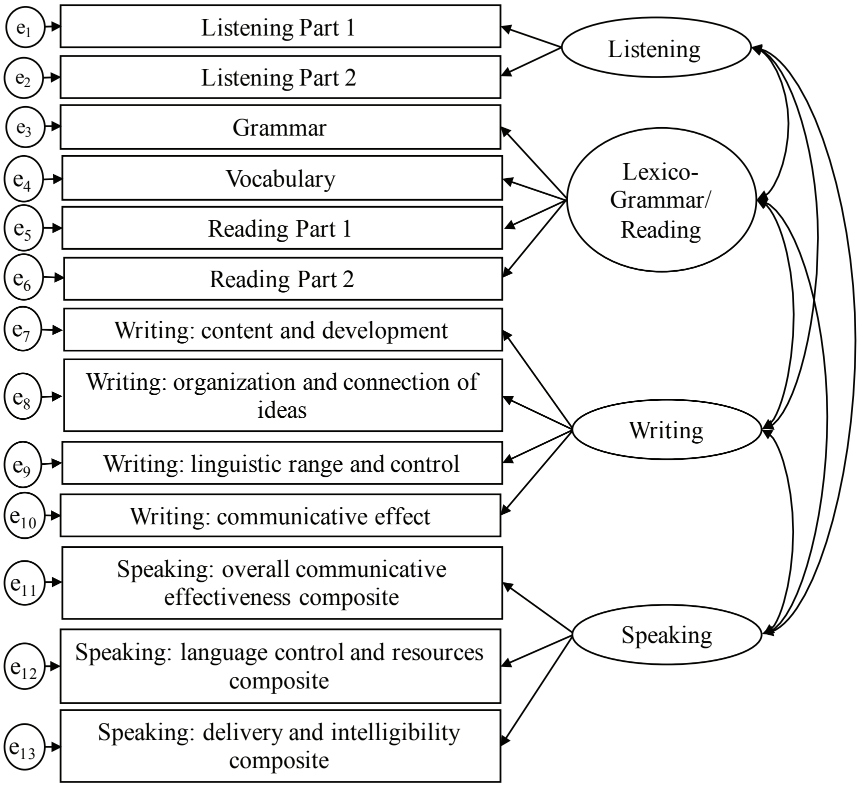

e1
Listening Part 1
e2
Listening Part 2
e3
Grammar
e4
Vocabulary
e5
Reading Part 1
e6
Reading Part 2
e7
Writing: content and development
e8
Writing: organization and connection of ideas
e9
Writing: linguistic range and control
e10
Writing: communicative effect
e11
Speaking: overall communicative effectiveness composite
e12
Speaking: language control and resources composite
e13
Speaking: delivery and intelligibility composite
Listening
Lexico-Grammar/ Reading
Writing
Speaking

*Figure 5.* Correlated four-factor model

**Goodness of fit of models.** To evaluate overall model fit, two criteria were used: goodness-of-fit measures and model parsimony. Six goodness-of-fit measures were used: the $\chi^2$ (Chi-square), comparative fit index (CFI), root mean square error of approximation (RMSEA), standardized root mean square residual (SRMR), Akaike's information criterion (AIC), and Bayesian information criterion (BIC). The $\chi^2$ measures absolute fit of the model to the data. When a latent model fit the data well, the $\chi^2$ statistic is statistically nonsignificant. However, $\chi^2$ is sensitive to reject the null hypothesis with large sample sizes such as found in the current data (Hair, Black, Babin, Anderson, & Tatham, 2006). Indicators of good model fit included CFI statistics greater than .95, RMSEA less than .06, and SRMR less than .05 (Hu & Bentler, 1999). AIC and BIC values were used to compare the relative fits of models. Smaller AIC or BIC values indicate better model fit to the data (Kass & Raftery, 1995).

With respect to model parsimony, when two models represented good fit to the data, a more parsimonious model was chosen. More parsimonious models contained fewer latent variables. In addition, multicollinearity between latent variables (defined as $r > .899$) was controlled for so as not to include latent variables that were not distinct enough (Sawaki et al., 2009). That is, higher correlations between two latent variables indicated that the two were similar enough to be considered as a single latent factor. Latent variables that showed multicollinearity with each other were combined to construct a single latent variable.

**Measurement invariance.** Invariance of measurement across groups was also tested for the final model. Measurement invariance examined the relationships between indicator variables and latent variables between groups (Beaujean, 2014). As such, holding measurement constant indicated that the latent model functions equivalently across different groups (i.e., the model is fair to different groups). Three different group variables were used: gender (i.e., male and female), L1s (Greek and Spanish groups which included more than 100 test-takers)[6], and age (i.e., young learners whose age was 12 or

[6] Generally, a minimum of 100 observations is recommended to construct a latent variable model (Loehlin, 1992). Thus, the other L1 groups which included less than 100 test-takers were not used for testing measurement invariance.

below, adolescent learners whose age was between 13 and 19, and adult learners whose age was 20 or above). When evaluating models for measurement invariance, the first indicator variable was fixed at 1.0 (Dimitrov, 2010).

The invariance measurement was evaluated with four sequential stages (Beaujean, 2014; Dimitrov, 2010): Configural invariance and three stages of measurement invariance (metric, scalar, and strict). Configural invariance (Model 0) was tested to examine whether the different groups simultaneously had the same number of latent variables which were formed by the same number of observed variables. Using metric/weak measurement invariance (Model 1), factor loadings on indicator variables were constrained to be equal across groups. Using scalar/strong measurement invariance (Model 2) for a given indicator, intercepts (i.e., means of indicator variables) were constrained to be equal across groups. Scalar invariance indicates that individuals at the same level of a given latent variable had the same value on the indicator variables regardless of group membership. Using strict measurement invariance (Model 3), for a given indicator, residual variances and covariances were constrained to be equal across groups. Strict invariance suggests that the indicator variables were measured with the same precision in each group.

When configural invariance across groups was met, measurement invariance tests were conducted for two nested models (Dimitrov, 2010): Model 1 vs. Model 0, Model 2 vs. Model 1, and Model 3 vs. Model 2. Invariance was interpreted in terms of the CFI difference ($\Delta\text{CFI} = \text{CFI}_{\text{constr.}} - \text{CFI}_{\text{unconstr.}}$).[7] When a ΔCFI value is lower than –.01, measurement invariance is not warranted (Dimitrov, 2010).

## Analysis 1 Results

### Descriptive Statistics

Table 4 summarizes means, standard deviations, ranges, skewness levels, and kurtosis levels for each test section. For the three analytic criteria for speaking performance, two or four separate scores for

[7] Due to the large sample size in the current study, differences in chi-square that are sensitive to sample sizes were not used in testing measurement invariance.

each criterion across different tasks were reported. However, for the CFA, in terms of the principle of parsimony, one score for each criterion for speaking performance was needed. To address this, under the assumption that the separate scores for each criterion would tap into the same aspect of speaking performance, factor analyses were conducted to reduce the separate scores that were based on the same criterion into single composite scores. The factor analyses confirmed that separate scores for each criterion loaded on each single factor: overall communicative effectiveness (OCE) with an eigenvalue of 3.109 that described 77.727% of the variance; language control and resources (LCR) with an eigenvalue of 1.808 that described 90.424% of the variance; and delivery and intelligibility (DI) with an eigenvalue of 1.816 that described 90.779% of the variance. After the unidimensionality of scores of each speaking criterion across different tasks was assured, composite scores for each of the three speaking criteria were calculated by averaging the separate scores.

Table 4: Descriptive Statistics for Test Scores for Each Test Section (*n* = 9,700)

| Score | Mean | SD | Range | Maximum possible | Skewness | Kurtosis |
|---|---|---|---|---|---|---|
| Listening Part 1 | 16.37 | 3.967 | 2–22 | 22 | –.746 | –.082 |
| Listening Part 2 | 8.04 | 3.016 | 0–14 | 14 | –.033 | –.803 |
| Grammar | 16.85 | 5.152 | 2–26 | 26 | –.275 | –.645 |
| Vocabulary | 17.54 | 4.726 | 2–26 | 26 | –.364 | –.637 |
| Reading Part 1 | 3.28 | 1.368 | 0–5 | 5 | –.444 | –.677 |
| Reading Part 2 | 6.01 | 2.208 | 0–10 | 10 | –.099 | –.699 |
| Writing: content and development | 6.54 | 1.120 | 0–10 | 10 | .107 | 1.867 |
| Writing: organization and connection of ideas | 6.21 | .961 | 0–10 | 10 | .214 | 3.836 |
| Writing: linguistic range and control | 6.21 | 1.077 | 0–10 | 10 | .151 | 2.536 |
| Writing: communicative effect | 6.36 | 1.136 | 0–10 | 10 | .174 | 2.161 |
| Speaking OCE composite | 3.896 | .681 | 0–5 | 5 | –.633 | 1.974 |
| Speaking Tasks 2 and 3 OCE | 4.01 | .736 | 0–5 | 5 | –.631 | 1.386 |
| Speaking Task 4 Q1 OCE | 4 | .745 | 0–5 | 5 | –.610 | 1.295 |
| Speaking Task 4 Q2 OCE | 3.85 | .790 | 0–5 | 5 | –.488 | .820 |
| Speaking Task 4 Q3 OCE | 3.73 | .816 | 0–5 | 5 | –.309 | .424 |
| Speaking LCR composite | 3.68 | .713 | 0–5 | 5 | –.247 | .917 |
| Speaking Tasks 2 and 3 LCR | 3.74 | .746 | 0–5 | 5 | –.245 | .649 |
| Speaking Task 4 LCR | 3.62 | .753 | 0–5 | 5 | –.159 | .626 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Speaking DI composite | 4.016 | .723 | 0–5 | 5 | −.636 | 1.388 |
| Speaking Tasks 2 and 3 DI | 4.07 | .742 | 0–5 | 5 | −.634 | 1.206 |
| Speaking Task 4 DI | 3.96 | .775 | 0–5 | 5 | −.526 | .834 |

*Note*. OCE = overall communicative effectiveness; LCR = language control and resources; DI = delivery and intelligibility

The distributions for test scores were checked through skewness and kurtosis levels.[8] Three analytic scores for writing performance (i.e., organization and connection of ideas, linguistic range and control, and communicative effect) were not normally distributed: Their values for kurtosis were above 2 (i.e., distributions that are more clustered around the mean with higher peaks). Due to the non-normal distribution of these scores, the test results of multivariate normality (an assumption for conducting CFA) indicated non-normality of multivariate distribution. To address non-normality, estimator MLM (i.e., using standard maximum likelihood to estimate the model parameters with robust standard errors and a Satorra-Bentler scaled test statistic) was used. The MLM$\chi^2$ (i.e., Satorra–Bentler scaled chi-square; SB$\chi^2$) takes into account a scaling correction to estimate chi-square under non-normal conditions (Satorra & Bentler, 1994). For invariance model fit, SB$\chi^2$ was also used (Satorra & Bentler, 2001).

[8] The values for skewness and kurtosis between −2 and +2 were considered acceptable to indicate a shape close to normal distribution (George & Mallery, 2016, pp. 114–115).

## Confirmatory Factor Analysis

Correlation matrices among indicator variables (i.e., test scores) for CFA are shown in Table 5. All of the test scores showed moderate-to-strong correlations with each other with coefficients ranging from .307 to .817.

Table 5: Correlation Matrices for Indicator Variables (n = 9,700)

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 Listening P1 | 1 | | | | | | | | | | | |
| 2 Listening P2 | .649 | 1 | | | | | | | | | | |
| 3 Grammar | .675 | .662 | 1 | | | | | | | | | |
| 4 Vocabulary | .646 | .648 | .770 | 1 | | | | | | | | |
| 5 Reading P1 | .534 | .532 | .579 | .588 | 1 | | | | | | | |
| 6 Reading P2 | .504 | .561 | .592 | .609 | .516 | 1 | | | | | | |
| 7 Writing CD | .343 | .333 | .414 | .415 | .329 | .328 | 1 | | | | | |
| 8 Writing OCI | .333 | .325 | .397 | .389 | .314 | .311 | .765 | 1 | | | | |
| 9 Writing LRC | .403 | .390 | .502 | .480 | .368 | .371 | .719 | .765 | 1 | | | |
| 10 Writing CE | .372 | .358 | .447 | .442 | .348 | .352 | .817 | .779 | .788 | 1 | | |
| 11 Speaking OCE composite | .467 | .412 | .498 | .460 | .371 | .355 | .358 | .339 | .401 | .374 | 1 | |
| 12 Speaking LCR composite | .458 | .413 | .501 | .459 | .360 | .358 | .334 | .317 | .379 | .352 | .800 | 1 |
| 13 Speaking DI composite | .442 | .380 | .471 | .426 | .341 | .325 | .321 | .307 | .372 | .335 | .778 | .760 |

*Note*. All correlation coefficients are significant at $p < .001$. P = part; CD = content and development; OCI = organization and connection of ideas; LRC = linguistic range and control; CE = communicative effect; OCE = overall communicative effectiveness; LCR = language control and resources; DI = delivery and intelligibility

Using the CFA, statistics for evaluating overall model fit of the five hypothesized models were calculated (see Table 6 for fit statistics for each model). Due to the large sample size, SB$\chi^2$ values for all of the five models were significant. Thus, the significance of SB$\chi^2$ values was not counted as a goodness-of-fit criterion. As shown in Table 6, the results of the CFA indicated the fit for the single-factor model and the correlated two-factor model was poor, while the fit for the higher-order model, the correlated five-factor model, and the correlated four-factor model was excellent (see Appendix A for detailed CFA results for each of the five models along with parameter estimates). Among the three models of good fit, the correlated five- and four-factor models not only showed better fit in terms of CFI, RMSEA, and SRMR values than the higher-order model, but also were more parsimonious (i.e., fewer latent variables) than the higher-order model. Thus, the correlated five- and four-factor models were considered as better representations of the latent structure of the ECCE than the higher-order model.

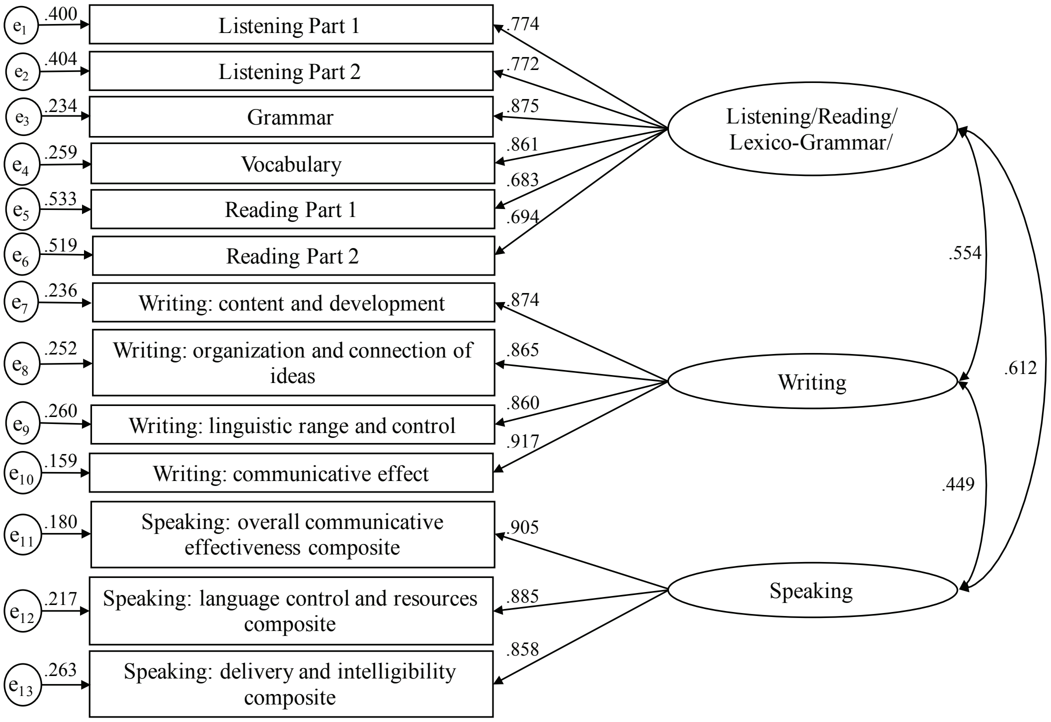
Table 6: Fit Statistics for the Five Models

| Model | SB$\chi^2$ | df | CFI | RMSEA | SRMR | AIC | BIC |
|---|---|---|---|---|---|---|---|
| Single-factor | 26286.217 | 65 | .568 | .204 | .124 | 416699.468 | 416979.483 |
| Correlated two-factor | 19887.414 | 64 | .673 | .179 | .102 | 403942.684 | 404229.879 |
| Higher-order | 1313.940 | 60 | .979 | .046 | .034 | 381982.165 | 382298.080 |
| Correlated five-factor | 1059.946 | 55 | .983 | .043 | .020 | 381714.260 | 382066.074 |
| Correlated four-factor | 1187.920 | 59 | .981 | .044 | .021 | 381842.085 | 382165.180 |

Between the correlated five- and four-factor models, the four-factor model was considered a better one than the five-factor model. The main reason was that in the five-factor model the latent variable of reading ability showed multicollinearity with the latent variable of lexico-grammar ability ($r$ = .938), which indicated that these two latent variables represented the same construct of reading/lexico-grammar ability as measured in the GVR section of the ECCE. Additionally, the four-factor model was more parsimonious (i.e., fewer latent variables) than the five-factor model.

In the four-factor model, the latent variable of reading/lexico-grammar ability also showed multicollinearity with the latent variable of listening ability ($r$ = .941), indicating that these two latent variables represented the same construct of receptive processing skills (i.e., understanding oral and written information). Thus, an additional correlated three-factor model was constructed with three separate but interacting latent variables: listening/reading/lexico-grammar, writing, and speaking abilities. The correlated three-factor model fit the data well: SB$\chi^2$(62) = 1452.438, CFI = .977, RMSEA = .048, SRMR = .023, AIC = 382115.795, and BIC = 382417.350. Although the three-factor model did not show better model fit than the four- or five-factor models, the three-factor model was chosen because it showed excellent fit along with fewer latent variables (i.e., more parsimonious), and no multicollinearity was found among the three latent variables. Figure 6 shows the correlated three-factor model along with parameter estimates. In the correlated three-factor model, the correlation between listening/reading/lexico-grammar and writing abilities demonstrated a strong effect size (*r* = .554) as did the correlation between listening/reading/lexico-grammar and speaking abilities (*r* = .612). The correlation between writing and speaking abilities demonstrated a moderate effect size (*r* = .449). However, in no cases was strong multicollinearity ($r > .899$) reported between the factors. Importantly, that these three latent variables showed moderate-to-strong correlations ($.449 \le r \le .612$) suggests that these abilities may tap into a general underlying language competence.

To sum up, the results of CFA indicated that the correlated three-factor model best represented the latent structure of the ECCE because it fit the data well, did not show multicollinearity among the three latent variables, and the most parsimonious. These results suggest that the ECCE measures three separate but correlated L2 abilities in a latent structure: listening/reading/lexico-grammar, writing, and speaking abilities.



e1
.400
Listening Part 1
.774
e2
.404
Listening Part 2
.772
e3
.234
Grammar
.875
e4
.259
Vocabulary
.861
e5
.533
Reading Part 1
.683
e6
.519
Reading Part 2
.694
Listening/Reading/
Lexico-Grammar/
.554
e7
.236
Writing: content and development
.874
e8
.252
Writing: organization and connection of ideas
.865
Writing
.612
e9
.260
Writing: linguistic range and control
.860
e10
.159
Writing: communicative effect
.917
.449
e11
.180
Speaking: overall communicative effectiveness composite
.905
e12
.217
Speaking: language control and resources composite
.885
Speaking
e13
.263
Speaking: delivery and intelligibility composite
.858

*Figure 6.* Correlated three-factor model

*Note.* Estimates are standardized and all significant ($p < .001$).

### Measurement Invariance

Using the correlated three-factor model as shown in Figure 6, measurement invariance was tested for three different group criteria: gender (i.e., male and female), L1s (i.e., Greek and Spanish), and age (i.e., young, adolescent, and adult learners). First, measurement invariance was tested across different gender (i.e., 5,341 female and 4,330 male test-takers) with the correlated three-factor model as a baseline model ($M_{Baseline}$). Table 7 shows the fit statistics for invariance assessment by gender. The goodness-of-fit indices showed that configural invariance across gender was supported (see fit statistics for Model 0 in Table 7). Given the evidence of configural invariance, metric measurement invariance was supported, such that the model constrained with metric measurement had good fit, and the ΔCFI value was greater than −.01 (see fit statistics for Model 1 in Table 7). Sequentially, scalar measurement invariance and strict measurement invariance were also supported (see fit statistics for Models 2 and 3 in Table 7). These results indicated that the indicator variables and the latent variables included in the correlated three-factor model were measured with the same level of precision across gender.

Table 7: Fit Statistics for Invariance Assessment Across Gender

| Model | $SB\chi^2$ | df | CFI | ΔCFI | RMSEA | SRMR |
|---|---|---|---|---|---|---|
| $Model_{baseline}$ | 1452.438 | 62 | .977 | - | .048 | .023 |
| Model 0 (Configural) | 1484.053 | 124 | .977 | - | .048 | .023 |
| Model 1 (Metric) | 1516.302 | 134 | .977 | .000 | .046 | .025 |
| Model 2 (Scalar) | 1580.441 | 144 | .976 | −.001 | .045 | .025 |
| Model 3 (Strict) | 1588.390 | 157 | .976 | .000 | .043 | .026 |

Measurement invariance was also tested across different age groups (i.e., young, adolescent, and adult learners). Among the 9,700 test-takers, 291 test-takers were 12 years old or below (i.e., young learners), 8,334 were between 13 years old and 19 years old (i.e., adolescent learners), and 1,051 were 20 years old or above (i.e., adult learners). Table 8 summarizes the measurement invariance results for test-takers grouped by age. Configural invariance (Model 0) across age was supported. The fit indices for subsequent models for measurement invariance (Models 1 to 3) across age were also met. For the pairwise comparisons of nested models (i.e., Model 1 vs. Model 0, Model 2 vs. Model 1, and Model 3 vs. Model 2), all of the ΔCFI values were greater than −.01, which supported strict measurement invariance.

Table 8: Fit Statistics for Invariance Assessment Across Age

| Model | SB$\chi^2$ | df | CFI | ΔCFI | RMSEA | SRMR |
|---|---|---|---|---|---|---|
| Model$_{\text{baseline}}$ | 1452.438 | 62 | .977 | - | .048 | .023 |
| Model 0 (Configural) | 1523.598 | 186 | .978 | - | .047 | .023 |
| Model 1 (Metric) | 1601.347 | 206 | .977 | −.001 | .047 | .026 |
| Model 2 (Scalar) | 2148.623 | 226 | .968 | −.009 | .051 | .028 |
| Model 3 (Strict) | 2261.375 | 252 | .967 | −.001 | .050 | .029 |

Next, measurement invariance was tested across different L1s (i.e., Greek and Spanish). Among the 9,700 test-takers in the current study, 8,814 were Greek-speaking, and 683 were Spanish-speaking. Table 9 summarizes the measurement invariance results for test-takers grouped by L1s. Configural invariance (Model 0) across L1s was supported. Metric measurement invariance was also supported (i.e., good fit and ΔCFI > −.01). However, scalar measurement invariance was not supported, such that the ΔCFI value was lower than −.01 (see fit statistics for Models 2 in Table 9). The examination of the modification indices revealed that the intercept of vocabulary test scores was substantially different across the two L1 groups. Thus, the constraints on the intercept of vocabulary scores was modified to be freed (i.e., non-invariant). The subsequent model supported scalar measurement invariance (i.e., good fit and ΔCFI > −.01; see fit statistics for Model 2$_{\text{partial}}$ in Table 9). Given the evidence of scalar measurement invariance, strict measurement invariance was supported (see fit statistics for Model 3 in Table 9). Regarding differences in vocabulary scores, a post hoc independent sample *t*-test was conducted to examine vocabulary score differences between Greek- and Spanish-speaking groups. The *t*-test result demonstrated that there was a significant difference in the vocabulary scores for the Greek-speaking group (Mean = 17.297, *SD* = 4.734) and the Spanish-speaking group (Mean = 20.191, *SD* = 3.54): *t*(9495) = −15.647, *p* < .001, Cohen's *d* = .692. This analysis indicated that on average, Spanish-speaking test-takers performed significantly better on the vocabulary test of the ECCE than Greek-speaking test-takers.

Table 9: Fit Statistics for Invariance Assessment Across L1s

| Model | $SB\chi^2$ | df | CFI | ΔCFI | RMSEA | SRMR |
|---|---|---|---|---|---|---|
| $\text{Model}_{\text{baseline}}$ | 1452.438 | 62 | .977 | - | .048 | .023 |
| Model 0 (Configural) | 1384.718 | 124 | .974 | - | .046 | .022 |
| Model 1 (Metric) | 1530.800 | 134 | .972 | –.002 | .047 | .026 |
| Model 2 (Scalar) | 2221.028 | 144 | .958 | –.014 | .055 | .028 |
| Model $2_{\text{partial}}$ (Scalar) | 1977.901 | 143 | .963 | –.009 | .052 | .027 |
| Model 3 (Strict) | 1925.126 | 156 | .964 | .001 | .049 | .028 |

In sum, the results of the measurement invariance analyses showed that the correlated three-factor model for the ECCE had equivalent latent representations with the same level of precise measurement across gender, L1s (with the exception of the vocabulary test scores), and age. Thus, the effect of the different group memberships (i.e., gender, age, and L1s) on establishing on the correlated three-factor model was minimal.

## Analysis 2 Method

The purpose of this analysis was to examine the relationships among speaking/writing scores, individual differences (i.e., personal/demographic variables, such as age, gender, L1s, as well as language skills and knowledge, such as reading, listening, writing, and vocabulary), and linguistic features found in test-takers' speaking and writing responses (RQ2). Specifically, two sub-analyses were conducted. First, speaking scores were predicted using individual differences and linguistic features found in speaking samples as explanatory variables (RQ2.a). Second, writing scores were predicted using individual differences and linguistic features found in writing samples as explanatory variables (RQ2.b). Reading, listening, writing, speaking and vocabulary skills and knowledge were based on test scores from the ECCE. Linguistic features found in speaking and writing responses were measured using natural language processing (NLP) tools.

## Data

We analyzed the response and performance data of 295 ECCE test takers.[9] Each test-taker completed the ECCE. Test-takers also provided their age, gender, and L1. Among the 295 test-takers, 181 (61.356%) were female and 114 (38.644%) were male. The test-takers ranged in age from 13 to 47 with a mean of 19.04 (*SD* = 5.717). The test population consisted of test-takers whose L1s were Spanish (*n* = 202; 68.475%) and Portuguese (*n* = 93; 31.525%).

The current data set included four different writing prompts and 20 different speaking prompts. Test-takers chose to produce either an essay or a letter based on a given prompt. Hand-written samples were scanned. The Speaking section includes four sequential tasks. Among the four speaking tasks, due to cost, we chose to analyze and transcribe Task 4 only, in which a test-taker is asked three independent questions (e.g., "*What is an important event that you remember? Why*?"). We selected Task 4 because it involves speaking ability only (i.e., test-takers' ability to answer the given questions) as compared to Tasks 1–3 which involve both listening ability (i.e., test-takers' understanding of the given scenario told by an examiner) and speaking ability. Audio files of speaking samples were trimmed to include Task 4 only, and then transcribed. Each transcript was divided into two files: one containing the test-taker's performance and the other containing the examiner's questions and responses. We used transcripts containing the test-takers' performances only. The test-taker transcripts then were cleaned to eliminate fillers (e.g., *um* and *er*) and interjections (e.g., *oh* and *ah*). We also deleted repetitions in false starts (e.g., "*the, the, the, teacher*" was modified to "*the teacher*") so as not to include these repeated items in word counts. The percentage of the nonlexical items (i.e., fillers, interjections, and repeated words in false starts) per student's transcript was calculated.

[9] Originally, we had a total of 300 test-taker performance data. Three test-takers' data were removed due to inaudibility of speaking samples. Also, two test-takers (one Cambodian and one Vietnamese) were removed to consider the L1 factor for further analysis by including Portuguese- and Spanish-speaking test-takers only.

### Linguistic features

We used three NLP tools to compute various linguistic features found in test-takers' writing and speaking performances. These linguistic features included lexical and phrasal features as measured by the Tool for the Automatic Analysis of Lexical Sophistication (TAALES; Kyle & Crossley, 2015), and cohesive features as measured by the Tool for the Automatic Analysis of Cohesion (TAACO; Crossley, Kyle, & McNamara, 2016). In addition, syntactic complexity was measured for writing data using the Syntactic Complexity Analyzer (Lu, 2010).[10] These NLP tools and computational indices are briefly discussed below. More detailed information of these tools is provided in Crossley, Kyle, and McNamara (2016), Kyle and Crossley (2015), and Lu (2010).

**Tool for the Automatic Analysis of Lexical Sophistication***.* TAALES (Kyle & Crossley, 2015) calculates approximately 400 lexical and phrasal indices, including lexical frequency (i.e., scores based on how often a word occurs in a reference corpus), lexical range (i.e., scores based on how many documents in a reference corpus include a word), psycholinguistic word information (e.g., familiarity, imageability, concreteness, and meaningfulness), word neighborhood (e.g., word neighborhood size and frequency indices for orthographic, phonographic, and phonological neighbors), word recognition norms (i.e., native English speakers' latencies and accuracies to lexical items during lexical decision and word naming tasks), semantic relations (e.g., hypernymy and polysemy), *n*-gram frequency (i.e., scores based on how often an *n*-gram occurs in a reference corpus), *n*-gram range (i.e., scores based on how many documents in a reference corpus contain an *n*-gram), *n*-gram association strength (i.e., how strongly a combination of words is attached to each other), and academic language (i.e., lexical and phrasal items that occur frequently in an academic corpus).

[10] We did not use SCA for speaking data because it was developed to measure syntactic complexity in writing data.

**Tool for the Automatic Analysis of Cohesion**. TAACO (Crossley et al., 2016) computes approximately 150 indices related to text cohesion, including the number of words, sentences[11], paragraphs[12], lemmas (i.e., the base form entered in the dictionary), content words, function words, *n*-grams, and part of speech tags (e.g., nouns, verbs, adjectives, and adverbs); type-token ratios[13] for all words, part of speech tags, content words, and function words; sentence overlap for all words, part of speech tags, content words, and function words; paragraph overlap for all words, part of speech tags, content words, and function words; and connectives such as logical connectives (e.g., *moreover*, *nevertheless*), and temporal connectives (e.g., *after*, *before*).

**Syntactic Complexity Analyzer**. SCA (Lu, 2010) computes 14 indices of syntactic complexity: three length of production units (i.e., mean length of clause, sentence, and T-unit); a sentence complexity ratio (i.e., clauses per sentence); four subordination indices (i.e., T-unit complexity ratio, complex T-unit ratio, dependent clauses per clauses, and dependent clauses per T-unit); three coordination indices (i.e., coordinate phrases per clause, coordinate phrases per T-unit, and sentence coordination ratio); and three particular structures (i.e., complex nominals per clause, complex nominals per T-unit, and verb phrases per T-unit).

### Statistical analysis

To predict speaking and writing scores, we used a linear mixed effects (LME) modeling approach, which considers both fixed effects (i.e., variables that potentially predict independent variables) and random effects (i.e., variables that are unrelated to independent variables and represent finite set levels

[11] In speaking data, an utterance was defined as a unit of speech bounded by pauses (Sato, 1988), and the end of each utterance was marked by a period. Thus, the number of sentences in speaking data indicated the number of utterances.

[12] In transcribing speaking samples, each turn for an interlocutor (either an examiner or a test-taker) was transcribed in a single paragraph in a text file. Thus, the number of paragraphs in speaking data indicated the number of turns produced by each test-taker.

[13] Types refer to the total number of unique, different words in a given text, while tokens refer to the total number of words of a given text.

of a factor or which only a random sample is available). In language testing contexts, LME models are useful because prompts are random effects (i.e., test-takers are randomly provided with one of thousands of potential prompts) that can be represented in LME models.

Before conducting LME analyses, in order to verify that analytic scores of speaking and writing performance loaded on the same factor and create a single independent score variable for each model, principal component analyses were conducted by entering the four analytic writing scores (i.e., content and development, organization and connection of ideas, linguistic range and control, communicative effect) and the five analytic speaking scores of Task 4 (i.e., overall communicative effectiveness for three sub-tasks, language control and resources, delivery and intelligibility), respectively. Composite speaking/writing scores were calculated by first multiplying each writing/speaking analytic criterion's factor loading with a test-taker's score for that criterion, and then summing these multiplies.

Using the composite speaking and writing scores, we computed three LME models for speaking and writing, respectively. The first model was constructed using individual difference variables including demographic/personal information (i.e., age, gender, and L1s) and other language scores (i.e., reading, listening, vocabulary, and grammar). Because the Reading and Listening sections have two sub-scores, respectively, composite reading and listening composite scores were also calculated in a manner similar to ones used to calculate composite speaking and writing scores. The second model was constructed using linguistic features found in speaking and writing samples. The final model was constructed using both individual difference variables and linguistic variables. To compare models, we used Akaike's information criterion (AIC) values and log-likelihood ratio tests. The model with a lower AIC index fits the data better (Maydeu-Olivares & Garcia-Forero, 2010), and log-likelihood ratio tests show which model is significantly better at a .05 significance level in terms of model fit.

In creating linguistic feature model, in order to verify that linguistic variables were meaningfully correlated with writing and speaking scores, we calculated correlations between composite speaking/writing scores and linguistic features as measured in speaking/writing samples. Linguistic variables that did not reach a correlation value of $|r| > .100$ with the composite scores (representing at least a small effect size, Cohen, 1988) were removed from further analyses. Then, the remaining linguistic variables were controlled for multicollinearity (defined as $r > .699$) in order not to include indices that measured similar linguistic features. Among variables that showed multicollinearity, the variable with the strongest correlation with the composite score was retained.

In constructing LME models, speaking/writing prompts were first added as random factors to ensure that effects of prompts were represented in the models (Baayen, Davidson & Bates, 2008). Second, with linguistic and individual difference variables (depending on the model of interest) added as fixed effects, models were developed by backward selection of the fixed effects using log-likelihood ratio tests following the convention of $t > 1.96$ at a .05 significance level (i.e., selecting the fixed effects that reached the significance level). Third, we tested interaction terms among the significant fixed effects by backward selection of the main and interaction effects. In addition, to fully test the random structure, we added a random slope adjustment for each significant fixed effect (i.e., the effect of prompts on linguistic features and gender) one-by-one (Barr, Levy, Scheepers, & Tily, 2013). We excluded the random slope terms that did not contribute to better goodness of fit of the model to prune irrelevant random effects. Finally, after adding relevant random slopes (if any), the optimal model was fitted by again backward fitting of the fixed effects.

To construct LME models, we used *R* (R Core Team, 2016) and the *lme4* package (Bates, Maechler, Bolker, & Walker, 2015). We also used the *LMERConvenienceFunctions* package (Tremblay & Ransijn, 2015) to perform backward selection of fixed effects and interaction effects, and test the significance of adding random slopes. In addition, we used the *lmerTest* package (Kuznetsova, Brockhoff, & Christensen, 2016) to calculate *p* values from the models. We also used the *MuMIn* package (Bartoń, 2017) to calculate two measures of variance explained from the models: a marginal *r*-squared that calculated the variance explained by the fixed effects only, and a conditional *r*-squared that calculated the variance explained by both the fixed and random effects.

## Analysis 2 Results

### Descriptive Statistics

Table 10 summarizes means, standard deviations, and ranges for each test section. The results of the principal component analyses confirmed that the four analytic writing scores loaded into the same factor (with an eigenvalue of 3.368 that accounted for 84.218 of the variance); the three analytic speaking scores loaded into the same factor (with an eigenvalue of 3.88 that accounted for 77.609 of the variance); the two reading scores loaded into the same factor (with an eigenvalue of 1.504 that accounted for 75.185 of the variance); and the two listening scores loaded into the same factor (an eigenvalue of 1.746 that accounted for 1.746 that accounted for 87.297 of the variance). Thus, composite writing, speaking, reading, and listening scores were calculated, respectively.[14] Correlations among these scores are displayed in Table 11. Because grammar scores showed multicollinearity with both of vocabulary scores ($r$ = .735) and listening scores ($r$ = .745), grammar scores were excluded from further consideration.

Table 10: Descriptive Statistics for Test Scores ($n$ = 295)

| Score | Mean | SD | Range | Maximum possible |
|---|---|---|---|---|
| Listening Part 1 | 14.610 | 4.086 | 6–22 | 22 |
| Listening Part 2 | 9.428 | 3.168 | 2–14 | 14 |
| Listening composite | 24.130 | 6.346 | 9.340–33.624 | 33.624 |
| Grammar | 18.420 | 4.935 | 4–26 | 26 |
| Vocabulary | 21.030 | 3.437 | 9–26 | 26 |
| Reading Part 1 | 3.906 | 1.216 | 0–5 | 5 |
| Reading Part 2 | 7.219 | 1.907 | 1–10 | 10 |
| Reading composite | 9.645 | 2.367 | 1.734–13.005 | 13.005 |
| Writing: content and development | 7.199 | 1.481 | 4–10 | 10 |
| Writing: organization and connection of ideas | 6.785 | 1.321 | 4–10 | 10 |
| Writing: linguistic range and control | 6.788 | 1.365 | 2–10 | 10 |

[14] A formula for calculating a composite speaking score was: [(score of overall communicative effectiveness for Question 1) × .879 + (score of overall communicative effectiveness for Question 2) × .888 + (score of overall communicative effectiveness for Question 3) × .878 + (score of language control and resources) × .850 + (score of delivery and intelligibility) × .908]. A formula for calculating a composite writing score was: [(score of content and development) × .909 + (score of organization and connection of ideas) × .913 + (score of language range and control) × .911 + (score of communicative effect) × .938]. A formula for calculating a composite reading score was: [(score of Reading Part 1) × .867 + (score of Reading Part 2) × .867]. A formula for calculating a composite listening score was: [(score of Listening Part 1) × .934+ (score of Listening Part 2) × .934].

| | | | | |
|---|---|---|---|---|
| Writing: communicative effect | 7.040 | 1.594 | 2–10 | 10 |
| Writing composite | 25.530 | 4.860 | 10.990–36.710 | 36.710 |
| Speaking Task 4 Question 1 OCE | 3.919 | .854 | 2–5 | 5 |
| Speaking Task 4 Question 2 OCE | 3.892 | .886 | 2–5 | 5 |
| Speaking Task 4 Question 3 OCE | 3.818 | .923 | 2–5 | 5 |
| Speaking Task 4 LCR | 3.603 | .872 | 1–5 | 5 |
| Speaking Task 4 DI | 3.859 | .941 | 1–5 | 5 |
| Speaking composite | 16.820 | 3.477 | 7.048–22.015 | 22.015 |

*Note*. OCE = overall communicative effectiveness; LCR = language control and resources; DI = delivery and intelligibility

Table 11: Correlations Among Grammar, Vocabulary, Reading, Listening, Writing, And Speaking Scores

| | Grammar | Vocabulary | Reading | Listening | Writing |
|---|---|---|---|---|---|
| Vocabulary | .735 | 1 | | | |
| Reading | .672 | .669 | 1 | | |
| Listening | .745 | .639 | .663 | 1 | |
| Writing | .389 | .364 | .414 | .347 | 1 |
| Speaking | .564 | .477 | .464 | .554 | .318 |

*Note*. Reading, listening, writing, and speaking scores are composite scores; All correlation coefficients are significant at $p < .001$.

## Constructing speaking models

**Speaking prompts as a random intercept.** As a basis LME model, a random intercept model was created by including the speaking prompt factor as a random intercept, and explained 3.440% of the variance in speaking composite scores.

**Speaking individual difference model.** An LME model predicting speaking composite scores was created using three demographic variables (i.e., L1s, gender, and age) and four test scores (i.e., reading, listening, vocabulary, and writing scores) as fixed effects, and prompts as a random effect. This model reported significant main effects for L1s, vocabulary, and listening scores (see Table 12). Neither significant interaction nor random slope effects were reported. The results indicated that higher speaking scores were predicted by higher listening scores ($t = 7.441$, $p < .001$) and higher vocabulary scores ($t$ = 3.455, $p$ < .001). In addition, Spanish-speaking test-takers received higher speaking scores than Portuguese-speaking test-takers ($t$ = 3.244, $p$ = .001). This model reported a marginal $R^2$ of .350 and a conditional $R^2$ of .362.

Table 12: LME Model Predicting Speaking Scores Using Individual Difference Variables

| Fixed effect | Estimate | Standard error | $t$ | $p$ |
|---|---|---|---|---|
| (Intercept) | 5.194 | 1.144 | 4.540 | <.001 |
| Listening | .258 | .035 | 7.441 | <.001 |
| Vocabulary | .215 | .062 | 3.455 | .001 |
| L1 (Portuguese baseline) | 1.220 | .376 | 3.244 | .001 |

**Speaking linguistic model**. An LME model predicting speaking composite scores was created using linguistic features (as measured by TAALES and TAACO) as fixed effects, and prompts as a random effect. Results indicated that higher-rated speaking samples included more function word types ($t$ = 4.666, $p$ < .001); greater function word overlap across immediately adjacent turns ($t$ =3.308, $p$ = .001); bigrams with stronger associations (as measured by Mutual Information scores[15]; $t$ =3.278, $p$ = .001); greater adverb overlap in the next two utterances ($t$ =2.792, $p$ = .006); fewer non-lexical items (e.g., fillers and false starts; $t$ =–2.723, $p$ = .007); trigrams with stronger associations (as measured by Mutual Information scores*;* $t$ =2.658, $p$ = .008); more adverb types ($t$ =2.444, $p$ = .015); and nouns with more polysemous meanings ($t$ =2.259, $p$ = .025; see Table 13). Neither significant interaction or random slope effects were reported. This model reported a marginal $R^2$ of .425 and a conditional $R^2$ of .432.

[15] *N*-grams with higher Mutual Information are the ones made up of strongly associated low-frequency words (e.g., *exultant triumph*; Evert, 2008).

Table 13: LME Model Predicting Speaking Scores Using Linguistic Features

| | Estimate | Standard error | *t* | *p* |
|---|---|---|---|---|
| (Intercept) | −4.717 | 2.389 | −1.974 | 0.049 |
| Number of function word types | .155 | .033 | 4.666 | <.001 |
| Function word overlap in the next turn (binary) | 3.341 | 1.010 | 3.308 | .001 |
| Bigram mutual information (COCA spoken corpus) | 4.128 | 1.259 | 3.278 | .001 |
| Adverb overlap in the next two utterances | 2.633 | .943 | 2.792 | .006 |
| Percentage of nonlexical items | −.056 | .021 | −2.723 | .007 |
| Trigram mutual information (COCA magazine corpus) | 2.145 | .807 | 2.658 | .008 |
| Number of adverb types | .128 | .052 | 2.444 | .015 |
| Polysemous nouns | .425 | .188 | 2.259 | .025 |

**Speaking individual difference and linguistic model.** An LME model predicting speaking composite scores was created using linguistic features (as measured by TAALES and TAACO) and individual differences (demographic variables and other test scores) as fixed effects, and prompts as a random effect. Results indicated that higher listening scores predicted higher speaking scores ($t = 7.908$, $p < .001$). Results also indicated that higher-rated speaking samples included more function word types ($t = 6.471$, $p < .001$); greater function word overlap across immediately adjacent turns ($t = 4.030$, $p < .001$); bigrams with stronger associations ($t = 2.831$, $p = .005$); greater adverb overlap in the next two utterances ($t = 2.790$, $p = .006$); and trigrams with stronger associations ($t = 2.408$, $p = .017$; see Table 14). Neither significant interaction nor random slope effects were reported. This model reported a marginal $R^2$ of .496 and a conditional $R^2$ of .499.

Table 14: LME Model Predicting Speaking Scores Using Linguistic Features And Individual Difference Variables

| | Estimate | Standard error | *t* | *p* |
|---|---|---|---|---|
| (Intercept) | −5.738 | 1.839 | −3.120 | .002 |
| Listening scores | .200 | .025 | 7.908 | <.001 |
| Number of function word types | .172 | .027 | 6.471 | <.001 |
| Function word overlap across the next utterance (binary) | 3.706 | .920 | 4.030 | <.001 |

| | | | | |
|---|---|---|---|---|
| Bigram mutual information (COCA spoken corpus) | 3.313 | 1.170 | 2.831 | .005 |
| Adverb overlap across the next two sentences | 2.421 | .868 | 2.790 | .006 |
| Trigram mutual information (COCA magazine corpus) | 1.790 | .743 | 2.408 | .017 |

**Speaking model comparisons.** We compared the three speaking models in terms of AIC values and the variance explained from the models (see Table 15). The combined model of individual differences and linguistic features had the lowest AIC value (1385.0), which indicated this model fit the data better than the individual difference model (with an AIC value of 1454.8) and the linguistic model (with an AIC value of 1427.9). In addition, larger variance in the composite speaking scores was explained by the combined model (49.9%) than the individual difference model (36.2%) and the linguistic model (43.2%). Additionally, the results of log-likelihood ratio tests indicated that the linguistic model was significantly better than the individual difference model ($\chi^2(5) = 36.869$, $p < .001$); the combined model was significantly better than the individual difference model ($\chi^2(3) = 75.782$, $p < .001$); but no difference was reported between the linguistic model and the combined model ($\chi^2(2) = 0$, $p = 1$). Additionally, the variance explained by the random prompt factor ranged from .3% to 1.2% across the three speaking models, indicating that the random effects of speaking prompts were negligible.

Table 15: Speaking Model Comparisons

| Speaking model | AIC | Marginal $R^2$ | Conditional $R^2$ |
|---|---|---|---|
| Individual difference model | 1454.8 | .350 | .362 |
| Linguistic model | 1427.9 | .425 | .432 |
| Individual difference and linguistic model | 1385.0 | .496 | .499 |

### Constructing writing models

**Writing prompts as a random intercept.** The writing prompt factor had eight levels with four prompts by two genres (essay vs. letter). With the writing prompt factor as a random intercept, the random intercept model explained 3.304% of the variance in writing composite scores.

**Writing individual difference model.** An LME model predicting writing composite scores was created using three demographic variables (i.e., L1s, gender, and age) and four test scores (i.e., reading, listening, vocabulary, and speaking scores) as fixed effects, and prompts as a random effect. This model reported significant main effects for reading, listening, L1s, gender, and age (see Table 16). Neither significant interaction nor random slope effects were reported. The results indicated that higher writing scores were predicted by higher reading scores ($t = 4.525$, $p < .001$) and higher listening scores ($t = 3.532$, $p < .001$). In addition, Spanish-speaking test-takers received higher writing scores than Portuguese-speaking test-takers ($t = 3.388$, $p = .001$); male test-takers received lower writing scores than female test-takers ($t = -3.366$, $p < .001$); and older test-takers received higher writing scores than younger test-takers ($t = .111$, $p = .017$). This model reported a marginal $R^2$ of .264 and a conditional $R^2$ of .290.

Table 16: LME Model Predicting Writing Scores Using Individual Difference Variables

| | Estimate | Standard error | *t* | *p* |
|---|---|---|---|---|
| (Intercept) | 11.827 | 1.740 | 6.799 | < .001 |
| Reading | .621 | .137 | 4.525 | < .001 |
| Listening | .196 | .056 | 3.532 | < .001 |
| L1 (Portuguese baseline) | 2.323 | .686 | 3.388 | .001 |
| Gender (female baseline) | −1.698 | .504 | −3.366 | < .001 |
| Age | .111 | .046 | 2.399 | .017 |

**Writing linguistic model**. An LME model predicting writing composite scores was created using linguistic features (as measured by TAALES, TAACO, and SCA) as fixed effects, and prompts as a random effect. Results indicated that higher-rated writing samples included more lemma types ($t = 10.250$, $p < .001$); more academic words from the Academic Word List 1[16] (Coxhead, 2000; $t = 3.075$, $p = .002$) as well as from the Academic Word List 8 ($t = 2.406$, $p = .017$); trigrams with stronger associations (as measured by Mutual Information scores; $t = 2.950$, $p = .004$); lower verb type-token ratios (i.e., more repetitions of the same verbs; $t = -2.720$, $p = .007$); and trigrams whose directional

[16] The sub-lists of the Academic Word List are available at: https://www.victoria.ac.nz/lals/resources/academicwordlist/publications/awlsublists1.pdf

associations were stronger (as measured by Delta P[17]; $t = 2.591$, $p = .010$; Table 17). This model reported a marginal $R^2$ of .391 and a conditional $R^2$ of .400.

Table 17: LME Model Predicting Writing Scores Using Linguistic Features

| | Estimate | Standard error | *t* | *p* |
|---|---|---|---|---|
| (Intercept) | 7.891 | 2.683 | 2.941 | .003 |
| Number of lemma types | .124 | .012 | 10.250 | < .001 |
| Academic Word List 1 (normed) | 72.485 | 23.576 | 3.075 | .002 |
| Trigram mutual information (COCA newspaper corpus) | 2.498 | .847 | 2.950 | .004 |
| Verb type-token ratio | −6.288 | 2.312 | −2.720 | .007 |
| Trigram Delta P (COCA fiction corpus) | 242.358 | 93.543 | 2.591 | .010 |
| Academic Word 8 (normed) | 282.733 | 117.498 | 2.406 | .017 |

**Writing individual difference and linguistic model**. An LME model predicting writing composite scores was created using linguistic features (as measured by TAALES, TAACO, and SCA) and individual differences (demographic variables and other test scores) as fixed effects, and prompts as a random effect. Results indicated that higher writing scores were predicted by higher reading scores ($t = 3.424$, $p = .001$) and higher vocabulary scores ($t = 2.179$, $p = .030$). In addition, male test-takers received lower writing scores than female test-takers ($t = -2.868$, $p = .004$). Results also indicated that higher-rated writing samples included more lemma types ($t = 8.339$, $p < .001$); lower verb type-token ratios (i.e., more repetitions of the same verbs; $t = -3.179$, $p = .002$); more academic words from the Academic Word List 8 (Coxhead, 2000; $t = 2.793$, $p = .030$); and trigrams whose directional

[17] Delta P considers directionality of *n*-grams because association strengths of *n*-grams are not symmetrical (Gries, 2013). For example, a bigram, *artificial intelligence*, has a higher Delta P score than another bigram, *intelligence artificial*, because *intelligence* is combined with other words, such as *agencies* and *sources*, more frequently than *artificial*.

associations were stronger ($t = 1.985$, $p = .048$; see Table 18). This model reported a marginal $R^2$ of .437 and a conditional $R^2$ of .445.

Table 18: LME Model Predicting Writing Scores Using Linguistic Features And Individual Difference Variables

| | Estimate | Standard error | *t* | *p* |
|---|---|---|---|---|
| (Intercept) | 10.706 | 1.924 | 5.564 | < .001 |
| Number of lemma types | .101 | .012 | 8.339 | < .001 |
| Reading | .419 | .122 | 3.424 | .001 |
| Verb type-token ratio | −7.118 | 2.239 | −3.179 | .002 |
| Gender (female baseline) | −1.279 | .446 | −2.868 | .004 |
| Academic Word List 8 (normed) | 315.850 | 113.093 | 2.793 | .006 |
| Vocabulary | .189 | .087 | 2.179 | .030 |
| Trigram Delta P (COCA fiction corpus) | 179.055 | 90.199 | 1.985 | .048 |

**Writing model comparisons.** We compared the three writing models in terms of AIC values and the variance explained from the models (see Table 19). The combined writing model of individual differences and linguistic features had the lowest AIC value (1614.8), which indicated this model fit the data better than the individual difference model (with an AIC value of 1689.5) and the linguistic model (with an AIC value of 1640.0). Furthermore, more composite writing score variance was explained by the combined model (44.5%) than the individual difference model (29.5%) and the linguistic model (40.0%). Additionally, the results of log-likelihood ratio tests indicated that the combined model was significantly better than the individual difference model ($\chi^2(1) = 76.691$, $p < .001$) and the linguistic model ($\chi^2(1) = 27.181$, $p < .001$); and the linguistic model was significantly better than the individual difference model ($\chi^2(0) = 49.51$, $p < .001$). Finally, the variance explained by the random prompt factor ranged from .8% to 2.4% across the three models, indicating that the random effects of writing prompts were negligible.

Table 19: Writing Model Comparisons

| Writing model | AIC | Marginal $R^2$ | Conditional $R^2$ |
|---|---|---|---|
| Individual difference model | 1689.5 | .271 | .295 |
| Linguistic model | 1640.0 | .391 | .400 |
| Individual difference and linguistic model | 1614.8 | .437 | .445 |

## Discussion

The two main purposes of this study were to investigate the relationship between test scores and the constructs measured in the ECCE (RQ1), and examine the relationship among writing and speaking scores, test-takers' individual differences, and linguistic features found in speaking and writing samples (RQ2). Discussion related to each RQ is presented below.

### RQ1: Relationship between test scores and the constructs measured in the ECCE

This study examined the latent structure of the ECCE that best represented test-takers' performances, using 9,700 test-takers' performance data (RQ1.a). The results of CFA indicated that among various plausible latent models, a correlated three-factor model that consisted of listening/reading/lexico-grammar, writing, and speaking abilities was considered the best model because it had excellent fit without multicollinearity among the three factors.

The correlated three-factor model is consistent with the constructs that the ECCE proposes to measure. The ECCE intends to measure three main constructs of language competence: (a) understanding of complex input; (b) interacting fluently; and (c) producing clear text. These three constructs correspond to each of the three latent factors, such that the Listening/Reading/Lexico-Grammar factor relates to understanding of input, the Speaking factor to interacting fluently, and the Writing factor to producing clear text. Thus, our findings provide evidence that the construct validity is established in the ECCE because its latent structure matches the construct which the ECCE is supposed to measure. The correlated three-factor model is also in line with the current multi-componential view of language competence in the language testing literature (Bachman & Palmer, 1982; Carroll, 1983; Gu, 2014; Sawaki & Sinharay, 2013), such that language competence consists of both of divisible language factors (i.e., listening/reading/lexico-grammar, writing, and speaking) and a general underlying language competence (based on moderate-to-strong correlations among the three latent variables).

Regarding the three factors identified in the ECCE, it should be noted that although the two separate sections (i.e., GVR and Listening sections) were combined into one latent factor, these results do not necessarily devalue the current score reports of the two separate section scores. This is because while listening and reading have in common that comprehension of both oral and written language involves constructing a coherent mental representation of input (Kintsch, 1998), they differ not only in the mode of input (i.e., audio vs. visual) but also in test takers' involvement in control of input (i.e., test takers can take control over processing written input but not spoken input). Thus, listening and reading are similar, but still separable (de Bot, Paribackht, & Wesche, 1997; In'nami & Koizumi, 2011). In addition, the combined nature of the listening/reading/lexico-grammar latent factor indicates that the processing of lexical and grammatical information at the sentential level closely relates to that of longer input at the discourse level, and that knowing lexical meanings and grammatical structures likely helps test-takers to understand longer stretches of written and oral messages appropriately.

This study also examined the generalizability of the latent structure of the ECCE that best represented test-takers' performances across genre, age, and L1 (RQ1.b). Overall, measurement invariance tests reported that the correlated three-factor model was generalizable across gender, L1s (with the exception of vocabulary test scores), and age. Specifically, our findings that the three-factor model of the ECCE was fully generalizable across gender and age indicate that the model invariantly measured the indicator variables and the latent variables across male and female test-takers and across different age groups. That is, the ECCE measured the same constructs (i.e., the three latent factors) for different groups of gender and age. On the other hand, the three-factor model was partially generalizable across different L1s such that the results of the measurement invariance tests supported strict measurement invariance for the model with the exception of the intercepts of vocabulary scores. The notion that the intercepts of vocabulary scores were not equally measured across L1s indicates that the vocabulary test elicited different responses from Greek-speaking and Spanish-speaking test-takers. In addition, the result that Spanish-speaking test-takers performed significantly better on the vocabulary test of the ECCE than Greek-speaking test-takers may indicate the effects of cognates (i.e., words that share similar meaning and form across languages; van Hell & De Groot, 1998) in vocabulary tests. That is, because the Spanish language is linguistically closer to the English language than the Greek language is (Miller & Chiswick, 2005; Van der Slik, 2010), Spanish-speaking test-takers might be more advantaged in completing vocabulary tests than Greek-speaking test-takers. Thus, more attention drawn to considering cognate effects in vocabulary tests would merit consideration.

### RQ2: Examining the relationships among individual differences, linguistic features, and speaking and writing scores

This study examined the extent to which individual differences (i.e., gender, L1, age, vocabulary knowledge, and listening and reading skills) and linguistic features as found in spoken and written responses predicted speaking performances (RQ2.a) and writing performances (RQ2.b). Overall, the results indicated the models which combined individual differences and linguistic features explained larger variances in speaking scores (49.9%) and writing scores (44.5%) than models based only on individual differences and models based on linguistic features alone. The models are discussed below.

#### Individual difference models of speaking and writing

The individual difference models for speaking and writing explained 36.2% of the variance in speaking scores and 29.5% of the variance in writing scores, respectively. Similarities between these speaking and writing models were found. Both individual difference models indicated that higher listening scores were predictive of higher speaking/writing scores, suggesting that the ability to comprehend oral input is important not only for oral production but also written production. In addition, both individual difference models indicated that L1s were important demographic characteristics of test takers, such that Spanish-speaking test-takers performed better than Portuguese-speaking test-takers. However, this result may not be attributable to L1-L2 language differences because both Spanish and Portuguese belong to the Indo-European language family as English does, and the linguistic distance between Spanish and English is similar to that between Portuguese and English (Chiswick & Miller, 2005). Instead, better performance on the part of Spanish-speaking test-takers may relate to other factors such as test-takers' nationalities, school curriculum, and motivation to learn English.

Differences between the speaking and writing individual difference model also merit discussion. The speaking model included higher vocabulary scores as a predictor of higher speaking scores, supporting the importance of vocabulary knowledge in L2 speaking performance (de Jong et al., 2012; Koizumi & In'nami, 2015). On the other hand, the writing model included higher reading scores as a predictor of higher writing scores, supporting the close relationship between reading and writing in the L2 (Abu-Akel, 1997; Carson et al., 1990; Sawaki & Sinharay 2013). The writing model also indicated that female test-takers performed better than male test-takers, which is in line with previous research

(Sunderland, 2000; Pavlenko & Piller, 2008). Furthermore, the writing model suggested that older test-takers performed better than younger test-takers (Nikolov & Djigunović, 2006). Interestingly, that the age factor was included in the writing model but not in speaking model may reflect different levels of intellectual skills involved across speaking and writing tasks (Hulstijn, 2011). In the ECCE, writing mainly involves higher level information processing because it requires test-takers to present their opinion about a situation or issue, while speaking largely involves lower level information processing because it asks test-takers to interact with the examiner with non-academic, casual topics. Thus, older test-takers who generally have better intellectual skills (Salthouse, 2009) are likely to be also better writers than younger test-takers, but not necessarily better speakers.

## Linguistic models of speaking and writing

The linguistic models for speaking and writing explained 43.2% of the variance in speaking scores and 40.0% of the variance in writing scores, respectively. A similarity existed between the speaking and writing linguistic models, such that both higher-rated speaking and writing samples tended to include *n*-grams whose associations were stronger (as measured by Mutual Information and Delta *P* scores). This trend tallies with some previous studies which have found that more proficient L2 learners tend to use *n*-grams whose associations are stronger (Kim, Crossley, & Kyle, 2018; Kyle, Crossley, & Berger, in press). The findings also support the importance of using multiword units (that frequently co-occur) for proficient L2 speaking and writing (Bestgen & Granger, 2014; Durrant & Schmitt, 2009). However, it should be noted that register variation across spoken and written discourse was also found, such that in the speaking model, *n*-gram association scores were based on the spoken and informal reference corpora (i.e., the COCA spoken and magazine), and in the writing model, *n*-gram association scores were based on the written reference corpora (i.e., the COCA newspaper and fiction). This difference in reference corpora may be in line with the characteristic of interpersonal, informal spoken registers in the interview-based L2 speaking test context, the characteristic of formal written registers in the opinion-based L2 writing test context (Biber et al., 2011, 2016).

Many differences between the speaking and writing models were also found. First, while both of the models included the number of word types (i.e., the number of unique lexical items, rather than the total number of words itself) as predictors of higher L2 speaking and writing scores, there were differences in specific types of words included in each model. The speaking linguistic model included the numbers of function word types and adverb types as predictors of speaking scores, such that higher-rated L2 speaking samples contained more function word types and adverb types. This finding may indicate that raters of L2 speaking performance tend to be influenced by test-takers' use of various function word types (e.g., use of determiners, pronouns, and conjunctions), which in turn may help raters build links among entities mentioned during speaking performance. For instance, test-takers' use of some function words that refer back to previously mentioned entities (e.g., *she* and *they*) may facilitate raters' referential processing. Our finding also indicates that speaking raters also tend to respond positively to test-takers' use of various adverb types (e.g., adverbs that express relation of manner, degree, level of certainty). This finding generally supports the positive relationships between the use of adverbs and speaking scores (Laflair et al., 2015). In contrast to the speaking model, the writing model included the number of lemma types as a predictor of writing scores, such that higher-rated L2 writing samples contained more lemma types. This finding is in line with previous findings that the use of more diverse lexical items is linked to higher L2 writing performance, particularly in time-limited testing settings (Crossley & McNamara, 2012; Treffers-Daller, 2013).

Another difference between the linguistic speaking and writing models is related to the use of words. The speaking model indicated that the use of more polysemous words (i.e., words with multiple meanings) was predictive of higher speaking scores. This finding is in line with previous research which has found the positive relationship between the use of polysemous words and speaking development and scores (Crossley, Salsbury, & McNamara, 2010; Saito, Webb, Trofimovich, & Isaacs, 2016). On the other hand, the writing model demonstrated that the greater use of academic words was predictive of higher writing scores. This finding corroborates previous studies which have shown that the use of academic words is indicative of higher-rated academic writing (Douglas, 2013).

A third difference between the linguistic speaking and writing models is associated with cohesion. The speaking model showed that greater overlap of function words and adverbs was predictive of higher speaking scores. This finding indicates that greater overlap of function words across adjacent turns may facilitate raters' cohesive understanding of test-takers' speech. In addition, greater overlap of adverbs across adjacent utterances may be related to test-takers' repeated expressions of stance (e.g., *certainly, actually*) and repeated use of amplifiers (e.g., *really, totally*), which in turn may have a positive influence on raters' evaluation of test-takers' performances. On the other hand, the writing model indicated that greater repetitions of verbs in L2 writing samples (as measured by lower verb type-token ratios) was predictive of higher writing scores. This finding suggests that cohesion built through greater verb repetitions may be related to higher-rated L2 writing.

Finally, the linguistic model of speaking indicated that higher-rated speaking performances tended to include fewer non-lexical words (e.g., fillers, and repeated words in false starts). This finding is in line with previous research that has shown that fluent speech (e.g., fewer hesitation markers) is predictive of higher speaking scores (Iwashita, Brown, McNamara, & O'Hagan, 2008; Laflair et al., 2015).

### The combined models of speaking and writing

The combined linguistic and individual differences models for speaking and writing explained 49.9% of the variance in speaking scores and 44.5% of the variance in writing scores, respectively. These models performed better than the individual differences models and the linguistic models in terms of the goodness of fit and the variance explained. The combined speaking model included one individual difference variable (i.e., listening scores) and five linguistic features (i.e., function word types, function word overlap across adjacent utterances, bigram mutual information, adverb overlap across adjacent turns, and trigram mutual information), all of which were included either in the individual difference model and the linguistic model. The combined writing model included three individual difference variables (i.e., reading scores, vocabulary scores, and gender) and four linguistic features (i.e., lemma types, verb type-token ratio, academic words, trigram Delta *P*), all of which except for vocabulary scores were also included in the individual difference model and the linguistic model.

These findings are generally similar to those reported in the individual and linguistic models. One finding different from the individual and linguistic models is that while higher listening scores were predictive of higher speaking scores, higher reading and vocabulary scores were predictive of higher writing scores, supporting a distinction between oral and written language (Hulstijn, 2011). More specifically, according to Hulstijn (2011), language proficiency is developed along with two different paths: basic language cognition (involving oral language, such as listening and speaking, and high-frequency lexical and grammatical features) and higher language cognition (involving written language and literacy skills, such as reading and writing, and low-frequency lexical and grammatical features). In this respect, the close relationship between listening and speaking may be attributed to the notion that both listening and speaking tests measure test-takers' ability to process language in the basic language cognition domain, while the close relationship between reading and writing may be attributed to the notion that both reading and writing tests measure test-taker's ability to process language in higher language cognition domain (Hulstijn, 2011).

## Conclusion

We examined the latent factor structure of the ECCE and its generalizability across different groups (i.e., gender, age, and L1) and investigated the extent to which speaking and writing performances were predicted by various individual differences (i.e., gender, L1, age, vocabulary knowledge, and listening and reading skills) and linguistic features as found in spoken and written samples. Findings indicated that the latent structure of the ECCE could be best represented by a correlated three-factor model comprised of reading/listening/lexico-grammar, writing, and speaking abilities for the entire sample as well as for different groups across gender, L1s (with exception of vocabulary test scores), and age. Findings also showed that the speaking and writing models which combined individual differences and linguistic features explained 49.9% of the variance in speaking scores and 44.5% of the variance in writing scores.

This study has three important implications in language assessments in general and in the context of the ECCE in particular. First, the study contributed to examining construct validity of the ECCE for measuring language competence. The findings support the notion that the ECCE measures the constructs (i.e., understanding complex input, interacting fluently, and producing clear text) proposed. Second, the study showed evidence for generalizability of the latent structure of the ECCE across gender, L1 (with exception of vocabulary test scores), and age, which further enhances the construct validity of the ECCE. That is, measurement invariance of the ECCE in assessing the three latent factors (i.e. reading/listening/lexico-grammar, writing, and speaking) is less likely to be influenced by gender, L1, and age. Lastly, this study provides a clear sketch of the links among speaking and writing scores, linguistic features found in speaking and writing samples, and individual differences. We found that some individual difference variables (e.g., L1s) were predictive of both speaking and writing scores, while some other individual differences (e.g., listening scores and age) were predictive of either speaking or writing scores. We also found that some linguistic features (e.g., association strengths of *n*-grams) were predictive of both speaking and writing scores, while some other linguistic features (e.g., the use of polysemous words and academic words) were predictive of either speaking or writing scores. Overall, these findings may provide evidence to support a distinction between oral language proficiency (using the basic language cognition domain) and written language proficiency (using the higher language cognition domain).

This study also has limitations. For instance, in examining the latent structure of the ECCE across different L1, only two language groups (i.e., Greek- and Spanish-speaking) were included. Including more diverse L1 groups merits consideration. Also, we considered a limited range of demographic/personal characteristics of test-takers (i.e., age, L1, and gender). Future studies could include other characteristics, such as test-takers' nationalities and length of studying English.

## References

Abu-Akel, A. (1997). On reading-writing relationships in first and foreign languages. *JALT Journal, 19*(2), 198–216.

Alderson, J. C., & Banerjee, J. (2002). Language testing & assessment (Part 2). *Language Teaching, 35*, 79–113.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed- effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412.

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

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

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*, 380–414.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278.

Bartoń, K. (2017). *MuMIn: Multi-Model Inference*. R package version 1.40.0. https://CRAN.R-project.org/package=MuMIn

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.1-9, https://CRAN.R-project.org/package=lme4.

Beaujean, A. A. (2014) Latent variable modeling using R: A step-by-step guide. NY: Routledge.

Bestgen, Y., & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. Jo*urnal of Second Language Writing, 26*, 28–41.

Biber, D., & Gray, B. (2013). Discourse characteristics of writing and speaking task types on the TOEFL iBT test: A lexico-grammatical analysis. (Research Report No. RR-13-05). Princeton, NJ: Educational Testing Service.

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly, 45*, 5–35.

Biber, D., Gray, B., & Staples, S. (2016). Predicting patterns of grammatical complexity across language exam task types and proficiency levels. *Applied Linguistics*, *37*(5), 639–668.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: John Wiley and Sons.

Braine, G. (Ed.). (2005). Teaching English to the world: History, curriculum, and practice. Mahwah, NJ: Laurence Erlbaum.

Byrne, B. M. (2006). Structural equation modeling with EQS: Basic concepts, applications, and programming (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Carroll, J. B. (1965). Fundamental consideration in testing for English language proficiency of foreign students. In H. B. Allen (Ed.), *Teaching English as a second language: A book of readings* (pp. 364–372). New York: McGraw-Hill.

Carroll, J. B. (1983). Psychometric theory and language testing. In J. W. Oller, Jr. (Ed.), *Issues in language testing research* (pp. 80–107). Rowley, MA: Newbury House.

Carson, J. E., Carrell, P. L., Silberstein, S., Kroll, B., & Kuehn, P. (1990). Reading-writing relationships in first and second language. *TESOL Quarterly, 24*(2), 245–266.

Celce-Murcia, M., & Larsen-Freeman, D. (1999). *The grammar book*. Boston: Heinle & Heinle.

Cohen, J. (1988). Statistical power analysis for the behavioural sciences. Hillsdale, NJ: Erlbaum.

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

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The role of cohesion, readability, and lexical difficulty. *Journal of Research in Reading, 35*, 115–135.

Crossley, S. A., & McNamara, D. S. (2013). Applications of Text Analysis Tools for Spoken Response Grading. *Language Learning & Technology, 17* (2), 171–192.

Crossley, S. A., & McNamara, D. S. (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing, 26*, 66–79.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods, 48*(4), 1227–1237.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2010). The development of polysemy and frequency use in English second language speakers. La*nguage Learning, 60*, 573–605.

de Bot, K., Paribakht, T. S., & Wesche, M. B. (1997). Toward a lexical processing model for the study of second language vocabulary acquisition: Evidence from ESL reading. *Studies in Second Language Acquisition, 19*, 309–329.

de Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition*, *34*, 5–34.

Dimitrov, D. M. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development, 43*, 121–149.

Douglas, R. D. (2013). The lexical breadth of undergraduate novice level writing competency. *Canadian Journal of Applied Linguistics, 16*, 152–170.

Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics in Language Teaching, 47*, 157–177.

Evert, S. (2008). Corpora and collocations. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics. An international handbook* (pp. 1212–1248). Berlin/New York: Mouton de Gruyter.

George, D., & Mallery, M. (2016). *IBM SPSS statistics 23 step by step: A simple guide and reference* (14th Ed.). New York: Routledge.

Gries, S. T. (2013). 50-something years of work on collocations: What is or should be next ... In*ternational Journal of Corpus Linguistics, 18*, 1, 37–166.

Gu, L. (2014). At the interface between language testing and second language acquisition: Language ability and context of learning. *Language Testing, 31*(1), 111–133.

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing, 18*(3), 218–238.

Hair, J. F, Jr., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate analysis.* NJ: Pearson Prentice-Hall, Englewood Cliffs.

Harley, B., Cummins, J., Swain, M., & Allen, P. (1990). The nature of language proficiency. In B. Harley, P. Allen, J. Cummins, & M. Swain (Eds.), *The development of second language proficiency* (pp. 7–25). New York: Cambridge University Press.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55.

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

In'nami, Y., & Koizumi, R. (2011). Factor structure of the revised TOEIC® test: A multiple-sample analysis. *Language Testing, 29*(1), 131–152.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics, 29*(1), 24–49.

Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing, 19*(1), 57–84.

Kang, O. (2013). Linguistic analysis of speaking features distinguishing general English exams at CEFR levels. *Cambridge English: Research Notes, 52*, 40–48.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*, 773–795.

Kim, M., Crossley, S. A., & Kyle, K. (2018). Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development, and writing quality. *The Modern Language Journal, 102(1),* 120–141.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition.* Cambridge, England: Cambridge University Press.

Kline, R. B. (2011). Principles and practice of structural equation modeling (3rd ed.). New York: Guilford.

Koizumi, R., & In'nami, Y. (2015). Vocabulary knowledge and speaking proficiency among second language learners from novice to intermediate levels. *Journal of Language Teaching and Research, 4*(5), 900–913.

Krashen, S. D., Long, M. A., & Scarcella, R. C. (1979). Age, rate and eventual attainment in second language acquisition. *TESOL quarterly*, 573–582.

Kunnan, A. J. (1998). An introduction to structural equation modeling for language assessment research. *Language Testing, 15*(3), 295–332

Kuznetsova, A, Brockhoff, P. B., & Christensen, R. H. B. (2015). *lmerTest: Tests in linear mixed effects models*. R package version 2.0-29, http://CRAN.R-project.org/package=lmerTest.

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly, 49*(4), 757–786.

Kyle, K., & Crossley, S. A. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing, 34,* 12–24.

Kyle, K., Crossley, S. A., & Berger, C. (in press). The tool for the analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Method.*

Laflair, G. T., & Staples, S. (2017). Using corpus linguistics to examine the extrapolation inference: A case study of a high-stakes speaking assessment. *Language Testing, 34*(4), 451–475.

LaFlair, G. T., Staples, S., & Egbert, J. (2015). Variability in the MELAB speaking task: Investigating linguistic characteristics of test-taker performance in relation to rater severity and score. *CaMLA Working Papers (2015–04).*

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics, 16*(3), 307–322.

Liu, J., & Costanzo, K. (2013). The relationship among TOEIC listening, reading, speaking, and writing skills. In D.E. Powers (Ed.), *The research foundation for the TOEIC tests: A compendium of studies* (Vol. II, 2.1–2.25). Princeton, NJ: Educational Testing Service.

Loehlin, J. C. (1992). Latent variable models: An introduction to factor, path. and structural analysis (2nd ed.). Hillsdale, NJ: Erlbaum

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics, 15*(4), 474–496.

MacWhinney, B. (2005). A unified model of language development. In J. F. Kroll & A. M. B. De Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 49–67). Oxford: Oxford University Press.

Maydeu-Olivares, A., & Garcia-Forero, C. (2010). Goodness-of-fit testing. *International encyclopedia of education*, *7*(1), 190–196.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13–103). New York: American Council on Education and Macmillan.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*, 241–256.

Michigan Language Assessment (2017). *ECCE 2016 Report*. Ann Arbor, MI: Michigan Language Assessment

Miller, P. W., & Chiswick, B. R. (2005). Linguistic distance: A quantitative measure of the distance between English and other languages. *Journal of Multilingual and Multicultural Development, 26*(1), 1–11.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution, 4*(2), 133–142.

Nikolov, M., & Djigunović, J. M. (2006). Recent research on age, second language acquisition, and early foreign language learning. *Annual review of applied linguistics*, *26*, 234-260.

Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.

Oller, J.W., Jr. (1976). Evidence of a general language proficiency factor: An expectancy grammar. *Die Neuren Sprachen, 76,* 165–174

Pavlenko, A., & Piller, I. (2008). Language education and gender. In S. May & N. H. Hornberger (Eds.), *Encyclopedia of language and education: Vol. 1. Language policy and political issues in education* (pp. 57–69). New York: Springer.

R Core Team. (2016). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36.

Saito, K., Webb, S., Tro movich, P., & Isaacs, T. (2016). Lexical problems of comprehensible second language speech. *Studies in Second Language Acquisition, 38*, 677–701.

Salthouse, T. A. (2009). When does age-related cognitive decline begin? *Neurobiology of Aging*, *30*, 507–514.

Sato, C. (1988). Origins of complex syntax in interlanguage development. *Studies in Second Language Acquisition, 10*, 371–395.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399–419). Thousands Oaks, CA: Sage.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66*, 507–514.

Sawaki, Y., & Sinharay, S. (2013). *Investigating the value of section scores for the TOEFL iBT test* (TOEFL iBT Research Report No. TOEFLiBT-21). Princeton, NJ: Educational Testing Service.

Sawaki, Y., Stricker, L., & Oranje, A. (2009). Factor structure of the TOEFL Internet-based test. *Language Testing, 26*, 5–30.

Schepens, J., Van der Slik, F., & Van Hout, R. (2013). The effect of linguistic distance across Indo-European mother tongues on learning Dutch as a second language. In L. Borin & A. Saxena (Eds.), *Approaches to measuring linguistic differences* (pp. 199–230). Berlin, Germany: De Gruyter.

Schoonen, R., Van Gelderen, A., De Glopper, K., Hulstijn, J., Simis, A., Snellings, P. & Stevenson, M. (2003). First language and second language writing: The role of linguistic fluency, linguistic knowledge and metacognitive knowledge. *Language Learning, 53*(1), 165–202.

Schoonen, R., Van Gelderen, A., Stoel, R. D., Hulstijn, J., & De Glopper, K. (2011). Modeling the development of L1 and EFL writing proficiency of secondary school students. *Language Learning, 61*(1), 31–79.

Shin, S.-K. (2005). Did they take the same test? Examinee language proficiency and the structure of language tests. *Language Testing, 22*(1), 31–57.

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal, 36*(2), 139–152.

Sunderland, J. (2000). Issues of language and gender in second and foreign language education. *Language Teaching, 33*(4), 203–223.

Treffers-Daller, J. (2013). Measuring lexical diversity among L2 learners of French: An explora- tion of the validity of D, MTLD, and HD-D as measures of language ability. In S. Jarvis & M. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures* (pp. 79–103). Amsterdam: Benjamins.

Tremblay, A., & Ransijn, J. (2015). *LMERConvenienceFunctions: Model selection and post-hoc analysis for (G)LMER models*. R package version 2.10. https://CRAN.R-project.org/package=LMERConvenienceFunctions.

Van der Slik, F. W. (2010). Acquisition of Dutch as a second language: The explanative power of cognate and genetic linguistic distance measures for 11 West European first languages. *Studies in Second Language Acquisition, 32*(3), 401–432.

van Hell, J. G., & De Groot, A. (1998). Conceptual representation in bilingual memory: Effects of concreteness and cognate status in word association. *Bilingualism: Language and Cognition, 1*(3), 193–211.

Wang, S. (2006). Validation and invariance of factor structure of the ECPE and MELAB across gender. *Spaan Fellow Working Papers in Foreign Language Assessment, 4*(1), 41-56.

Yang, W., Lu, X., & Weigle, S. C. (2015). Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. *Journal of Second Language Writing, 28*, 53–67.

## Appendix A: Confirmatory factor analysis (CFA) results for the five models

CFA Results for the Single-Factor Model

| Factor/indicator | Estimate | Z[a] |
|---|---|---|
| General language ability | | |
| Reading Part 1 | .650 | 79.615 |
| Reading Part 2 | .654 | 80.621 |
| Listening Part 1 | .737 | 87.077 |
| Listening Part 2 | .720 | 96.411 |
| Grammar | .828 | 115.807 |
| Vocabulary | .808 | 112.618 |
| Speaking: overall communicative effectiveness | .656 | 64.536 |
| Speaking: language control and resources | .646 | 67.041 |
| Speaking: delivery and intelligibility | .617 | 61.858 |
| Writing: content and development | .624 | 50.322 |
| Writing: organization and connection of ideas | .608 | 44.474 |
| Writing: linguistic range and control | .686 | 55.300 |
| Writing: communicative effect | .658 | 53.088 |

*Note*. [a] Based on robust standard errors. Estimates are standardized and all significant at $p < .001$.

CFA Results for the Correlated Two-Factor Model

| Factor/indicator | Estimate | Z[a] |
|---|---|---|
| Listening/Reading/Lexico-Grammar/Writing | | |
| Listening Part 1 | .738 | 87.104 |
| Listening Part 2 | .731 | 98.628 |
| Reading Part 1 | .661 | 81.400 |
| Reading Part 2 | .668 | 82.957 |
| Grammar | .836 | 118.320 |
| Vocabulary | .823 | 116.777 |
| Writing: content and development | .633 | 51.474 |
| Writing: organization and connection of ideas | .617 | 45.551 |
| Writing: linguistic range and control | .693 | 56.152 |
| Writing: communicative effect | .667 | 54.288 |
| Speaking | | |
| Speaking: overall communicative effectiveness | .906 | 79.389 |
| Speaking: language control and resources | .884 | 88.145 |

| | | |
|---|---|---|
| Speaking: delivery and intelligibility | .858 | 77.824 |
| Covariance | | |
| Listening/Reading/Lexico-Grammar/Writing <–> Speaking | .632 | 70.466 |

*Note.* [a] Based on robust standard errors. Estimates are standardized and all significant at $p < .001$.

CFA Results for the Higher-Order Model

| Factor/indicator | Estimate | Z[a] |
|---|---|---|
| Listening | | |
| Listening Part 1 | .808 | 22.485 |
| Listening Part 2 | .803 | 22.140 |
| Reading | | |
| Reading Part 1 | .713 | 14.547 |
| Reading Part 2 | .724 | 14.438 |
| Lexico-Grammar | | |
| Grammar | .886 | 4.565 |
| Vocabulary | .869 | 4.107 |
| Speaking | | |
| Speaking: overall communicative effectiveness | .905 | 57.087 |
| Speaking: language control and resources | .885 | 61.388 |
| Speaking: delivery and intelligibility | .858 | 55.760 |
| Writing | | |
| Writing: content and development | .874 | 79.722 |
| Writing: organization and connection of ideas | .865 | 65.751 |
| Writing: linguistic range and control | .860 | 75.587 |
| Writing: communicative effect | .917 | 86.748 |
| General language ability | | |
| Listening | .950 | 20.955 |
| Reading | .955 | 14.094 |
| Lexico-Grammar | .981 | 11.358 |
| Speaking | .625 | 40.565 |
| Writing | .568 | 50.977 |

*Note.* [a] Based on robust standard errors. Estimates are standardized and all significant at $p < .001$.

CFA Results for the Correlated Five-Factor Model

| Factor/indicator | Estimate | Z[a] |
|---|---|---|
| Listening | | |

| | | |
|---|---|---|
| Listening Part 1 | .807 | 97.601 |
| Listening Part 2 | .803 | 107.591 |
| Reading | | |
| Reading Part 1 | .713 | 82.658 |
| Reading Part 2 | .724 | 83.370 |
| Lexico-Grammar | | |
| Grammar | .886 | 125.705 |
| Vocabulary | .869 | 124.787 |
| Speaking | | |
| Speaking: overall communicative effectiveness | .905 | 79.360 |
| Speaking: language control and resources | .885 | 88.148 |
| Speaking: delivery and intelligibility | .858 | 77.847 |
| Writing | | |
| Writing: content and development | .874 | 76.419 |
| Writing: organization and connection of ideas | .864 | 62.832 |
| Writing: linguistic range and control | .861 | 69.614 |
| Writing: communicative effect | .917 | 82.018 |
| Covariances | | |
| Listening <–> Reading | .920 | 123.059 |
| Listening <–> Lexico-Grammar | .931 | 195.097 |
| Listening <–> Speaking | .604 | 58.223 |
| Listening <–> Writing | .502 | 53.666 |
| Reading <–> Lexico-Grammar | .938 | 148.621 |
| Reading <–> Speaking | .555 | 51.735 |
| Reading <–> Writing | .537 | 54.948 |
| Lexico-Grammar <–> Speaking | .607 | 58.770 |
| Lexico-Grammar <–> Writing | .562 | 73.428 |
| Speaking <–> Writing | .449 | 43.020 |

*Note*. [a] Based on robust standard errors. Estimates are standardized and all significant at $p < .001$.

CFA Results for the Correlated Four-Factor Model

| Factor/indicator | Estimate | Z[a] |
|---|---|---|
| Listening | | |
| Listening Part 1 | .808 | 97.711 |
| Listening Part 2 | .802 | 107.335 |
| Reading/Lexico-Grammar | | |
| Reading Part 1 | .683 | 83.780 |

| | | |
|---|---|---|
| Reading Part 2 | .695 | 86.659 |
| Grammar | .879 | 124.919 |
| Vocabulary | .866 | 124.767 |
| Speaking | | |
| Speaking: overall communicative effectiveness | .906 | 79.378 |
| Speaking: language control and resources | .885 | 88.151 |
| Speaking: delivery and intelligibility | .858 | 77.829 |
| Writing | | |
| Writing: content and development | .874 | 76.419 |
| Writing: organization and connection of ideas | .864 | 62.832 |
| Writing: linguistic range and control | .861 | 69.614 |
| Writing: communicative effect | .917 | 82.018 |
| Covariances | | |
| Listening <–> Reading/Lexico-Grammar | .941 | 215.366 |
| Listening <–> Speaking | .604 | 58.246 |
| Listening <–> Writing | .502 | 53.669 |
| Reading/Lexico-Grammar <–> Speaking | .604 | 61.783 |
| Reading/Lexico-Grammar <–> Writing | .564 | 75.190 |
| Speaking <–> Writing | .449 | 43.021 |

*Note.* [a] Based on robust standard errors. Estimates are standardized and all significant at $p < .001$.