PULMONOLOGY

Construct Validity of the ECCE: Latent Structure of a CEFR-Based High-Intermediate English Language Proficiency Test

Dr. Olivia Harris[1]*

[1]University of Melbourne, Language Testing and Assessment Unit, Melbourne, Australia

Abstract

The Common European Framework of Reference for Languages Companion Volume (CEFR CV; Council of Europe, 2018) emphasizes macro-functions of language (i.e., reception, production, interaction, and mediation). However, there seems to be little consensus on whether the macro-functions are commensurable with CEFR-based proficiency tests. This commentary focuses on the Examination for the Certificate of Competency in English (ECCE), which is based on the CEFR and assesses high-intermediate level English proficiency. The study explores the latent structure of the ECCE and its generalizability across groups (gender, age, and first language [L1]) to examine its construct validity, dimensionality of language proficiency, and commensurability with the CEFR macro-functions. The latent structure was examined through confirmatory factor analysis using performance scores from 9,700 test-takers. The results indicated that test-takers' performance on the ECCE could be best represented by a correlated three-factor model (i.e., reading/listening/lexico-grammar, writing, and speaking abilities). The correlated three-factor model also held irrespective of gender, age, and L1s (with the exception of vocabulary scores). Overall, the findings indicate that the correlated three-factor model is consistent with the constructs that the ECCE proposes to measure, is in line with the current

multi-componential view of language proficiency, and is partly commensurate with the CEFR macro-functions.

Keywords

Language proficiency; Construct validity; CEFR; ECCE; Confirmatory factor analysis

Introduction

The purpose of this commentary is to explore whether macro-functions (i.e., reception, production, interaction, and mediation) suggested by the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001) are comparable (or commensurable) with a standardized proficiency test that adopts the CEFR but measures the traditional four language skills (i.e., reading, listening, writing, and speaking).

Large-scale standardized language proficiency tests are often based on conceptual frameworks that define language proficiency (e.g., academic vs. general language use) and provide specifications related to the language skills and knowledge measured (e.g., vocabulary knowledge and reading comprehension), contexts (e.g., immersion vs. foreign language contexts), levels of proficiency (e.g., intermediate vs. advanced), and age of test-takers (e.g., young vs. adult). One of the conceptual frameworks widely used for developing language tests is the CEFR (Council of Europe, 2001). The main goal of the CEFR is to promote language teaching and learning as a means of communication. The CEFR is action-oriented, describes tasks that learners can do in a second language (L2), and provides a comprehensive framework for describing language proficiency with common reference levels.[1] The CEFR has had considerable influences not only in Europe but also in other parts of the world (Little, 2007, 2019; Alderson, 2007; Hulstijn, 2007; North, 2007). Some of the main contributions of the CEFR to the field of language assessment include its introduction of the six-level scales of language proficiency (A1 to C2); its use of common language tasks/activities that are argued to require "a comparable level of proficiency from language to language" (Little, 2007, p. 646); and its focus on communicative proficiency to support language leaning, teaching, and assessment

[1] In this article, 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.

across different languages. There are many criticisms of CEFR as well. For example, the CEFR's proficiency level descriptors were developed rather intuitively and were not validated using learner performances, which make the scientific foundations of the CEFR less rigorous. In addition, while the CEFR proficiency levels have been extensively adopted by language testing agencies, textbook publishers, and curriculum developers and have been widely used as gatekeepers by policymakers, the use of CEFR in such contexts have been rarely subject to thorough scrutiny (Alderson, 2007; Deygers, 2019). Recently, in response to some criticisms, the Council of Europe has published the CEFR Companion volume (CV; 2018).

The CEFR Companion Volume (CV; Council of Europe, 2018) explicitly defines proficiency as "the ability to perform communicative language activities… whilst drawing upon both general and communicative language competences" (Council of Europe, 2018, p. 32). The communicative language activities are categorized according to four macro-functions or modes of communication: reception, production, interaction, and mediation (i.e., using language to create meaning; e.g., mediating communication). The main difference between production and interaction is that the former involves sustained monologue with long turns and the latter involves both receptive and productive skills within conversational dialogs and short turns. Importantly, the four modes of communication "reflect more the way people actually use the language than do the four skills [reading, listening, writing, and speaking]" and "a move away from the matrix of four skills and three elements (grammatical structure, vocabulary, phonology/graphology; Council of Europe, 2018, p. 31)." While these four modes of communication were suggested in the 2001 CEFR (Council of Europe, 2001), to our knowledge, there is little consensus or discussion in the testing literature on whether the four macro-functions

are commensurable with proficiency tests that adopt the CEFR but measure the traditional four skills (reading, listening, writing, and speaking).

In addition, while proficiency levels of the CEFR have been adopted by a number of standardized tests, models of language proficiency as measured by CEFR-based tests have not been widely tested. Examining how components of language proficiency measured by standardized tests interact within a conceptual framework is important for the establishment of construct validity (i.e., the degree to which a test measures the theoretical construct defined). That is, for standardized tests that typically assess multiple language skills (e.g., listening, speaking, reading, and writing; Carroll, 1965), it is important to assess the latent configuration of the tests, examine whether the separately measured skills adequately comprise the underlying construct of language proficiency, and analyze whether the latent configuration reflects the constructs that the tests are meant to measure and the conceptual framework that the tests adopt (e.g. Bae & Bachman, 1998; Messick, 1996).

Our study was motivated by an interest in the latent structure of a language proficiency test based on the CEFR. It is also motivated by the lack of previous research that links CEFR-based proficiency tests to the macro-functions of communication (i.e., reception, production, interaction, and mediation) that have been emphasized in the CEFR CV. The test of interest is the Examination for the Certificate of Competency in English (ECCE), which is based on the CEFR. The ECCE, developed by Michigan Language Assessment, is a standardized test battery of high-intermediate level English-as-a-foreign language (EFL) competency at the B2 level of the CEFR. The ECCE is used for a variety of purposes, including educational program admissions, language course requirement, obtaining/improving employment, and personal

interest (Michigan Language Assessment, 2017)[2]. The main purpose of our study is to explore the latent structure of the ECCE that can best represent test-takers' performances to help validate a latent structure in a CEFR based test and examine the dimensionality of language proficiency as well as the commensurability between the latent structure and the CEFR macro-functions. We also examine the generalizability of the latent structure of the ECCE across gender, first languages (L1s), and age. Investigating the model of language proficiency measured by the ECCE will help not only to establish its construct validity, but also to provide information about the nature of language proficiency models based on the CEFR and derived mainly from intermediate-level EFL adolescent learners. As well, the study will provide information about the potential links of the language skills and knowledge measured by ECCE to the macro-functions of communication as presented in in the CEFR CV.[3]

In the following, we will first provide a review of previous research on construct validity and the dimensionality of language proficiency. We then describe the materials and methods used in this study. We then report the results and discuss them in terms of the construct validity, the dimensionality of language proficiency, and the commensurability with the CEFR macro-functions. We conclude with implications of our findings for language assessment.

Construct Validity and Dimensionality of Language Proficiency

Construct validity has been a main concept of interest in the field of assessment since Cronbach and Meehl's (1955) seminal work. Cronbach and Meehl suggested construct validity be defined theoretically and highlight the relationship between the test and the proposed interpretations. Influenced by Cronbach and Meehl (1955), Messick (1989) defined construct

[2] The majority of test-takers of the ECCE are adolescent EFL learners. However, preteens and adults also take the ECCE for various reasons.
[3] In this study, we do not cover the mode of mediation because the ECCE, as in other many tests, does not include mediation as a testing component.

validity 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" (p. 11). While many general frameworks of construct validation have been developed (e.g., Cronbach, 1971; Loevinger, 1957; Messick 1989), one of the most influential frameworks related to construct validation is Kane's argument-based approach to validation (Kane, 2013). According to Kane (2013), "[t]he process outlined within the argument-based approach is basically quite simple. First, state the claims that are being made in a proposed interpretation or use (the IUA [interpretation/use argument]), and second, evaluate these claims (the validity argument)" (p. 9). That is, the argument-based approach involves (1) specifying what is being claimed in the interpretation and use of the test (i.e., IUA), which are formulated mainly during the development stage; and (2) evaluating the plausibility of the proposed interpretation and use (i.e., the validity argument), which is primarily examined during the appraisal stage. Both of the specification (the IUA) and the evaluation (the validity argument) are necessary for validation processes (Kane, 2013, p. 16). The evidence that supports certain claims can be provided by conducting various different analyses, such as content analyses, confirmatory factor analyses, generalizability analyses, interrater reliability analyses, and item-response-theory-based analyses (Kane, 2013).

Given that language tests are often developed based on language proficiency models (e.g., Bachman & Palmer, 1982; Canale & Swain, 1980), one way to establishing construct validity for language tests is to provide evidence that test scores reflect the theoretical construct(s) of language proficiency that the test proposes to assess. To do so, previous research has examined the dimensionality of language proficiency and the latent structure of proficiency tests by examining potential models of language proficiency and selecting the one with the best

fit using confirmatory factor analysis (CFA). This approach can be further subcategorized into two approaches: theory-driven (e.g., using theoretical models of language competence) and skill-based (e.g., reading, listening, writing, and speaking). Pioneer studies (Bachman & Palmer, 1982; Harley et al., 1990) tested theoretically-based models of language proficiency. Bachman and Palmer (1982) examined models of communicative competence, which presumably consisted of grammatical, pragmatic, and sociolinguistic competences, using CFA. Findings indicated that the model with the best fit was composed of a higher-order, general factor along with two first-order, specific factors (i.e., grammatical/pragmatic competence and sociolinguistic competence). Adopting the communicative competence framework proposed by Canale and Swain (1980) and using CFA, Harley et al. (1990) found two factors: a "general language proficiency factor" and a "written method factor" (p. 15).

Recent studies have adopted a four-skills approach under the assumption that four skills (i.e., listening, reading, speaking, and writing) can be distinguished empirically and comprise language proficiency (Carroll, 1965; Gu, 2014, 2015; In'nami & Koizumi, 2012; Sawaki & Sinharay, 2013; Shin, 2005). Research has reported that models of language proficiency vary depending on which language tests are analyzed. Even in examining the same test, such as the Test of English as a Foreign Language Internet-based test (TOEFL iBT), previous findings differently identified its model of language proficiency. For example, using a CFA 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' performance, while Sawaki and Sinharay (2013) using a larger sample (N = 50,393) found that test-takers' performance was best explained by a four-factor model correlating to the four sub-skills (i.e., listening, speaking, reading, and writing).

Beyond examining models of language proficiency, it is also important to test whether the latent model identified is generalizable across different groups (e.g., gender and L1s) to ensure that the test measures the same constructs for different groups because the generalizability of constructs identified in the model across different groups is not granted (Messick, 1989). Previous studies have examined whether models of language proficiency are generalizable across gender (Wang, 2006), target language contact (Gu, 2014), L1s (Sawaki & Sinharay, 2013), and different samples (In'nami & Koizumi, 2012). For instance, Sawaki and Sinharay (2013) indicated that a four-factor language proficiency model consisting of the four sub-skills (i.e., listening, speaking, reading, and writing) functioned equally for three different L1 groups (i.e., Arabic, Korean and Spanish).

Current Study

To summarize, previous studies on the latent configuration of proficiency tests have generally supported the notion that language proficiency is multi-dimensional, such that it consists of a general component and smaller, specific components (Bachman, Davidson, Ryan, & Choi, 1995; Bachman & Palmer, 1982; Fouly, Bachman, & Cziko, 1990; Gu, 2014, 2015; Harley, Allen, Cummins, & Swain, 1990; Sawaki, Sinharay, & Oranje, 2009). While previous studies have examined models of language proficiency by either testing theoretical models or using various proficiency tests, to our knowledge, little research has been conducted on componential models of language proficiency as measured by tests based on the CEFR or tests aiming at intermediate-level of language proficiency primarily in EFL adolescent learners. In addition, given that macro-functions of language (reception, production, interaction, and mediation) emphasized in the CEFR CV (Council of Europe, 2018), it is important to examine the potential links between these macro-functions and a language test based on the CEFR. To

address these gaps, test-takers performances on the ECCE were analyzed in this study because the ECCE is based on the CEFR, assesses intermediate-level English proficiency, and is primarily used to assess EFL adolescent learners.[4] Specifically, we examine the construct validity of the ECCE by establishing the latent structure of language proficiency within the test. By doing so, we investigate the appropriateness of the ECCE scores as indicators of constructs based on the CEFR, the multi-dimensional nature of language proficiency in consideration of specific language knowledge and skills, and the commensurability between the latent configuration of the ECCE and the macro-functions of the CEFR. In addition, we examine whether the latent structure of language proficiency identified for the ECCE is generalizable across different groups (e.g., gender and age) in order to ensure that the model of language proficiency identified for the ECCE assesses the same constructs for different groups. Thus, our study is guided by two main questions:

1. What latent structure of the ECCE best represents test-takers' performances?
2. To what extent is the latent structure of the ECCE generalizable across gender, age, and first languages (L1s)?

Materials and Method

The Structure of the ECCE

The main construct that the ECCE intends to measure is general English proficiency at the high-intermediate level (i.e., the B2 level of the CEFR; Council of Europe, 2001). The ECCE was developed based on the CEFR descriptors of the B2 level.[5] According to the CEFR,

[4] A majority of ECCE test-takers are Greek-speaking. An increasing number of EFL adolescent learners from Greece take the ECCE because in Greece, English became the primary foreign language throughout formal schooling after Greece joined the European Union, and the Greece government prioritizes improving English proficiency for promoting economic competitiveness (British Council, 2018).

[5] Michigan Language Assessment submitted test specifications that support the claim that the ECCE assesses the CEFR B2 level to the Council of Europe (Michigan Language Assessment, personal communication, 2019).

language learners at the B2 proficiency level can "understand the main ideas of complex text on both concrete and abstract topics"; "interact with a degree of fluency and spontaneity"; and "produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue" (Council of Europe, 2001, p. 24). Thus, the ECCE aims to measure three constructs of language competence: understanding of complex input, interacting fluently, and producing clear text (Michigan Language Assessment, 2017). Accordingly, the ECCE consists of four sections: Listening, Grammar/Vocabulary/ Reading (GVR), Speaking, and Writing sections. The Listening and Reading 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.

In relation to the CEFR, the Listening and Reading sections in ECCE are related to the reception mode. The Speaking section relates to the spoken interaction mode (rather than the spoken production mode) because it is interview-based, involving both receptive (i.e., the test-taker's listening to the interviewer) and productive (i.e., the test-taker's responses) skills. The Writing section relates to the written production mode (rather than the written interaction mode) because the writing task is to produce a letter or an essay, mainly involving sustained monologue and not including receptive skills except with understanding the prompt. While Michigan Language Assessment does not outline the links between Grammar and Vocabulary sections and the CEFR in an explicit manner, these sections relate to linguistic competence as part of the communicative language competences outlined in the CEFR (Council of Europe, 2001, p. 13). However, it should be mentioned that descriptions of the macro-functions of language in the

However, it should be noted that the CEFR is not a standardized scale, and, thus, no institution monitors or coordinates its use (Council of Europe, 2018, p. 26).

CEFR are based on professional opinion and consensus rather than evidence (Cumming, 2009).

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. The questions in Part 1 assess test-takers' understanding of the given conversation. In Part 2, after listening to each short talk (e.g., a lecture about history and a talk between a manager and 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 Part 2 assess test-takers' literal and analytic understanding of the given talk (e.g., understanding a main idea and details of the talk).

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 virtual 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 (e.g., *What is an important event that you remember? Why?*). Task 1 is a warm-up activity for helping establishing rapport between the test-taker and the examiner, and thus unscored. For Tasks 2, 3, and 4, the examiner evaluates the test-taker's performance using an analytic five-point rating scale with three criteria (i.e., *overall communicative effectiveness, language control/resources*, and *intelligibility/delivery*).[6] As of the criterion of overall communicative effectiveness for Task 4, test-taker's performance on the three elaboration questions is separately evaluated (i.e., three different scores for each of the three questions). As of the criterion of *control/resources* and *intelligibility/delivery*, respectively, for Task 4, test-taker's performance on the three elaboration questions is evaluated together (i.e., one score for all of the three questions). The inter-rater reliability for the total score of the Speaking section was .909 in the form of Kendall's coefficient of concordance (Michigan Language Assessment, personal communication, September 4, 2019). It was measured as part of routine rater monitoring, using test data that were scored by both the examiner and internal raters.

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. 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*

[6] 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.

and control, and *communicative effect*).[7] Two ratings are summed. If two raters have nonadjacent scores for a writing sample, a third rater evaluates it. Test-takers are provided 30 minutes to write the letter or essay.[8] The inter-rater for the total score of the Writing section between raters was .817 in the form of Kendall's coefficient of concordance (Michigan Language Assessment, personal communication, September 4, 2019).

Data

We analyzed 9,700 test-taker response data on the ECCE. Test-takers were learners of English as an L2. The L1s of the test-takers included 14 different languages (see Table 1). The majority of test-takers were Greek-speaking (90.9%). Around seven percent of the test-takers were Spanish-speaking. 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 (.3%). 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 (Michigan Language Assessment, 2017).

[INSERT TABLE 1 NEAR HERE]

Statistical Analysis

Confirmatory factor analysis. To examine the latent structure of the ECCE, we used confirmatory factor analysis (CFA), which investigates the relationships among measurable variables (i.e., observable variables or indicators) and their latent variables (i.e., factors; Kline, 2011). We chose to use CFA for two main reasons. First, because the ECCE was developed

[7] 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.

along with specification for its interpretation and use, one of the next steps to examine the plausibility of the proposed interpretation and use (i.e., the validity argument) is to examine construct validity by linking test scores with underlying factors. By examining the underlying structure of the ECCE, evidence or counter-evidence for a claim that the ECCE measures general English proficiency in three modes (i.e., reception, written production, and spoken interaction) can be provided. Second, by using CFA, which models an underlying structure using latent factors, convergent and discriminant validity evidence can be provided (Kane, 2006). For example, observed scores that share similarities in the underlying structure are expected to load on the appropriate same latent factor, which would provide convergent evidence for the test. On the other hand, observed scores that reflect distinct characteristics are expected to load on different latent factors, which would provide for the discriminability of the factors.

To conduct CFA, R (R Development Core Team, 2018) and *lavaan* packages (Rosseel, 2012) were used. Multivariate normality was checked using Mardia's normalized estimate, with values below five considered to indicate multivariate normality (Byrne, 2006). In a latent model, latent variables were shown in ovals, while observable variables were shown in squares. When evaluating the model, the latent variables were fixed at 1.0, such that factor loadings for each indicator variable (i.e., estimates of the impact a latent variable has on indicator variables) 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 proficiency). As such, this model assumes that there is a general language proficiency informed by the five language abilities which are not distinctive from each other at a latent level, suggesting the nature of language proficiency as a single unitary construct.

[INSERT FIGURE 1 NEAR HERE]

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) which found that the two factors (i.e., speaking and listening/reading/writing) best represented the TOEFL iBT.

[INSERT FIGURE 2 NEAR HERE]

Higher-order factor model (Figure 3). In the higher-order factor model, a general, higher-order language proficiency latent factor is specified along with five first-order language latent factors (i.e., reading, listening, writing, speaking, and lexico-grammatical abilities. This model was constructed in accordance with the scoring scheme of the ECCE, which reports each skill score along with a total score.

[INSERT FIGURE 3 NEAR HERE]

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.

[INSERT FIGURE 4 NEAR HERE]

Correlated four-factor model (Figure 5). The correlated four-factor model specifies four distinct but correlated factors that correspond to the four different sections of the ECCE: listening, GVR, writing, and speaking sections.

[INSERT FIGURE 5 NEAR HERE]

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 χ^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). When a latent model fit the data well, the chi-square statistic is statistically nonsignificant. However, χ^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 .950, RMSEA less than .060, and SRMR less than .050 (Hu & Bentler, 1999). Indicators of acceptable model fit included CFI statistics greater than .900, RMSEA less than .080, and SRMR less than .080 (Hu & Bentler, 1999). 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 is chosen, such that 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). Latent variables that showed multicollinearity with each other were combined to construct a single latent variable.

Measurement invariance. Invariance of measurement across different groups is also tested for the final model. Measurement invariance assesses the relationships between indicator and latent variables between groups (Beaujean, 2014). Holding measurement invariant indicates that the latent model is equivalently constructed across different groups. Three different group variables are used: gender (i.e., male and female), L1s (Greek and Spanish groups which

included more than 100 test-takers)[9], and age (i.e., young learners whose age was 12 or below, adolescent learners whose age was between 13 and 19, and adult learners whose age was 20 or above). These three age groups represent pre-teens, teens, and adults (Michigan Language Assessment, personal communication, 2019). A detailed procedure for measurement invariance is provided in Supplementary Material A.

Results

Descriptive Statistics

Table 2 summarizes means, standard deviations, ranges, skewness levels, and kurtosis levels for each test section. For three analytic criteria for speaking performance, two or four separate scores for each criterion across different tasks were reported. However, for the CFA, we needed one score for each criterion for speaking performance. 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.

[9] 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.

[TABLE 2 NEAR HERE]

The distributions for test scores were checked through skewness and kurtosis levels. The values for skewness and kurtosis between –2 and +2 were considered acceptable to indicate a shape close to normal distribution (George & Mallery, 2010). 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 MLMχ^2 (i.e., Satorra–Bentler scaled chi-square; SBχ^2) takes into account a scaling correction to estimate chi-square under non-normal conditions (Satorra & Bentler, 1994). For invariance model fit, SBχ^2 was also used (Satorra & Bentler, 2001).

Results of Confirmatory Factor Analysis

Correlation matrices among indicator variables (i.e., test scores) for CFA are shown in Table 3. All of the test scores showed moderate-to-strong correlations with each other with coefficients ranging from .307 to .817.

[INSERT TABLE 3 NEAR HERE]

Using the CFA, statistics for evaluating overall model fit of the five hypothesized models were calculated (see Table 4 for fit statistics for each model). Due to the large sample size, SBχ^2 values for all of the five models were significant. Thus, the significance of SBχ^2 values was not counted as a goodness-of-fit criterion. As shown in Table 4, the results of the CFA indicated the

single-factor model and the correlated two-factor model showed poor fit, while the higher-order model, the correlated five-factor model, and the correlated four-factor model showed excellent fit.[10] 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. Between the correlated five- and four-factor models, the four-factor model was considered a better one than the five-factor model. This was primarily because 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.

[INSERT TABLE 4 NEAR HERE]

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: SBχ^2(62) = 1452.438, CFI = .977, RMSEA = .048, SRMR = .023, AIC = 382115.795, and BIC =

[10] We tested another model based on a reviewer's suggestion which consisted of two factors: one factor included language use and control (i.e., speaking, writing, grammar, and vocabulary), and a second factor which included receptive language skills (i.e., listening and reading). The model fit was bad (SBχ^2[64] = 27059.509, CFI = .652, RMSEA = .209, SRMR = .161, AIC = 413255.246, and BIC = 413449.103), and, thus, this model was not considered further.

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 without no multicollinearity among the three latent variables. In this model, the correlation between listening/reading/lexico-grammar and writing abilities demonstrated a strong effect size (r = .554; Cohen, 1988) 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. That these three latent variables showed moderate-to-strong correlations suggests that these abilities may tap into a general underlying language proficiency.[11] Figure 6 shows the correlated three-factor model along with parameter estimates.

To further examine whether there may be a higher-order factor that links the correlated three factors in an underlying structure, another model that added a higher-order factor onto the three correlated model was created and tested. This higher-order model with three first-order factors showed excellent fit: SBχ^2(62) = 1452.438, CFI = .977, RMSEA = .048, SRMR = .025, AIC = 382089.795, and BIC = 382298.012. While the higher-order model with three first-order factors showed a model fit as good as the three correlated model, the correlated three-factor model was more parsimonious and showed a slightly better fit than the higher-order model. Thus, we chose the correlated three-factor model as the best model.

[INSERT FIQURE 6 NEAR HERE]

[11] To complement the results of confirmatory factor analysis, exploratory factor analysis was also conducted. For the exploratory factor analysis, a Promax rotation method (i.e., correlated solution) was used. Results indicated that the first three factors showed eigenvalues greater than 1 and explained 76.469% of the shared variance in the ECCE data. In addition, the scree plot showed noticeable declines in the eigenvalues until the third component. Thus, a three- factor solution was considered appropriate. These first three factors were the same as those of the correlated three-factor model resulting from confirmatory factor analysis (i.e., listening/reading/lexico-grammar, writing, and speaking). Thus, the results of the exploratory factor analysis support the results of the confirmatory factor analysis.

To sum up, the results of the CFA indicated that the correlated three-factor model best represented the latent structure of the ECCE because it fit the data well without multicollinearity among the three latent variables. This model also hints at the existence of general underlying language proficiency based on moderate-to-strong correlations among the latent factors. These results suggest that for test-takers of the ECCE included in the current study, most of whom are Greek-speaking and Spanish-speaking EFL learners in secondary school, the ECCE likely measures three corelated L2 abilities (i.e., listening/reading/lexico-grammar, writing, and speaking abilities) in a latent structure.

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 as a baseline model ($M_{Baseline}$). Full results of measurement invariance are provided in Supplementary Material A. Results indicated that measurement invariance was fully supported for gender and age groups. On the other hand, measurement invariance was partially supported for L1 groups, such that the intercepts of vocabulary test scores were substantially different across the two L1 groups. A post hoc *t*-test showed that Spanish-speaking test-takers performed significantly better on the vocabulary test of the ECCE than Greek-speaking test-takers. To summarize, the results of the measurement invariance analyses showed that the correlated three-factor 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.

Discussion

Best-Fitting Latent Model of the ECCE

The purpose of the current study was to investigate the latent structure of the ECCE using 9,700 test-taker performance data, and to examine whether the latent structure best representing the ECCE was generalizable across gender, age, and L1s. The first research question examined the latent structure of the ECCE. The results of the 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 the best model because it had excellent fit as well and did not demonstrate multicollinearity among the three factors.

The correlated three-factor model that consisted of listening/reading/lexico-grammar, writing, and speaking abilities can be discussed from at least three perspectives. First, in terms of construct validity, this correlated three-factor model is consistent with the constructs that the ECCE proposes to measure. Specifically, the ECCE intends to measure three main constructs of language proficiency at the B2 level of language proficiency: understanding of complex input, interacting fluently, and 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 (i.e., reception), the Speaking factor to interacting fluently (i.e., spoken interaction), and the Writing factor to producing clear text (i.e., written production). Thus, there is evidence that the ECCE measures those constructs, providing construct validity for ECCE. In relation to Kane's argument-based approach to validation (Kane, 2013), this study provides evidence of a claim associated with construct validity of the ECCE, such that the three correlated factors in the latent structure of the ECCE can be indicative the three constructs (i.e., reception,

written production, and spoken interaction) linked to the CEFR. In addition, the correlated three factors support convergent validity, such that activities related to reception (and processing written vocabulary and structures), written production, and spoken interaction within the scheme of the CEFR loaded on respective relevant latent factors. The correlated three factors also support discriminant validity, such that different modes (i.e., reception, written production, and spoken interaction) are distinctively revealed in the underlying structure.

Second, the correlated three-factor model also supports the current multi-componential view of language proficiency in the language testing literature (Bachman & Palmer, 1982; Carroll, 1983; Gu, 2014; Sawaki et al., 2009; Sawaki & Sinharay, 2013), such that language proficiency consists of divisible language factors (i.e., listening/reading/lexico-grammar, writing, and speaking). Additionally, that the three factors showed moderate-to-strong correlations with each other hints at a general underlying language proficiency that may yield these correlations.

Lastly, the correlated three-factor model can be discussed in terms of commensurability between the constructs measured by the ECCE and the macro-functions suggested in the CEFR (Council of Europe, 2018). The ECCE writing and speaking sections seem commensurate with written production and spoken interaction of the macro-functions, respectively. However, the ECCE listening and Grammar/Vocabulary/Reading (GVR) sections, which were combined into one factor in the latent configuration, do not seem commensurate with macro-functions. For example, listening and reading items likely correspond to reception, but grammar and vocabulary are more related to linguistic competence which learners may draw on to perform the macro-functions of language. This may be partly because test-takers' processing of lexical and grammatical information at the sentential level (related to linguistic competence) closely relates to that of longer input at the discourse level, and that knowing lexical meanings and grammatical

structures helps test-takers to perform reception-related activities (i.e., understanding longer stretches of written and oral messages appropriately) than production-related activities. That is, it seems that the CEFR's macro-function of reception is closely related to linguistic competence as measured by written lexical and grammatical forms, possible because both require processing linguistic input regardless of differences in modes (oral vs. written) and types of processing (grammar/vocabulary vs. comprehension). However, it should be noted that the listening/reading/lexico-grammar latent factor may also to some extent reflect the effect of the test format. That is, the latent factor may reflect the shared characteristics of multiple-choice test items (i.e., selected responses), being distinctive from the open-ended speaking and writing items (i.e., constructed responses).[12]

Generalizability of the Latent Model of the ECCE

The second research question examined the generalizability of the correlated three-factor model of the ECCE. Our findings that the correlated three-factor 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 for different groups of gender and age. On the other hand, the correlated 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. These findings indicate that measurement invariance of the ECCE in assessing the correlated three-

[12] It is possible that the test format effect could be the primary reason for the combined listening/reading/lexico-grammar factor. However, the effect of the test format cannot be examined in this study because each language skill/knowledge was measured by one response type only, not allowing for an analysis on the effect of different test formats. In addition, previous studies on the latent structure of language tests have reported that multiple-choice items loaded on different factors (e.g. Sawaki & Sinharay, 2013). Thus, using multiple choice items does not always lead to their loading on the same factor.

factor model is unlikely to be influenced by gender, L1s (with the exception of the intercepts of vocabulary scores), or age.

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 because of shared lexical items across the languages.

Limitations

Prior to presenting implications of this study, it is important to discuss its limitations. First, given the limited nature of the data used in this study, findings of this study are not necessarily applicable to other CEFR-based language tests or different learner populations. In addition, to examine the dimensionality of the ECCE, we adopted a confirmatory approach using a total score for each language skill and knowledge measured. To examine the latent structure of a test, it would be also be important to adopt an exploratory approach using item-based analyses (e.g., multidimensional item response theory-based approach). Also, in examining the latent structure of the ECCE across different L1s, only two language groups (i.e., Greek- and Spanish-speaking) were included. Including more diverse L1 groups merits consideration.

Other limitations are related to the test analyzed in this study. In the ECCE, reading, listening, and lexico-grammar skills were assessed in multiple-choice format only. Considering that test formats affect test-takers' performances (In'nami & Koizumi, 2009), it is possible that the results may have biased as a result this formatting. Thus, future studies should examine whether reading, listening, and lexico-grammar skills are also closely related when assessed in test formats other than multiple-choice items, such as open-ended items. In addition, grammar and vocabulary were measured in an isolated fashion (i.e., filling in the blanks in the given sentences), and thus not perfectly fitting into the CEFR that emphasizes communicative activities. Furthermore, while the CEFR argues for mediation as one of the four main modes of communicative activities, many tests based on the CEFR, including the ECCE, do not include the evaluation of test-takers' mediation activities. In future studies, it will be important to examine whether assessing mediation would be relevant to the context of language assessment, particularly to standardized testing contexts, and if so, how mediation can be measured (Deygers, 2019).

Implications for Language Testing

Findings of this study have multiple implications for language testing. Firstly, our findings provide evidence that the different skills and knowledge measured by the ECCE constitute multi-dimensional underlying factors (i.e., reception, written production, and spoken interaction). This finding demonstrates that it is important to evaluate test-developers' claims that are made during proposed interpretations or the appraisal stage (Kane, 2013). In addition, the multi-dimensional nature of performances on the ECCE of EFL learners (primarily Greek-speaking and Spanish-speaking secondary school students) builds on previous research on multi-dimensional frameworks of language proficiency (Gu, 2014, 2015; In'nami & Koizumi, 2012;

Sawaki et al., 2009, Sawaki & Sinharay, 2013).

Secondly, the intertwined nature of the listening/reading/lexico-grammar skills that share characteristics of understanding input empirically supports part of the CEFR descriptive scheme that categorizes reading and listening into one broader mode, reception. This finding is also in line with the increasing focus on the importance of a more integrated framework (rather than an isolated four-skills framework) in the language assessment literature, such that language should be considered as a system of integrated skills for real-life language use and communication (e.g., Faulkner-Bond, Wolf, Wells, & Sireci, 2018). In light of the combined nature of listening, reading, and lexico-grammar skills, we can further think of the next generation of CEFR-based language tests in which reading and listening (optionally along with grammar and vocabulary) are assessed in an integrated manner (e.g., answering questions after both reading and listening). By doing so, imbalanced time allocations in many language tests in which receptive modes receive much more time than productive modes (e.g., ECCE, TOEFL iBT, and IELTS) may be alleviated by integrating reading and listening sections and reducing the time allocated for receptive modes.

Thirdly, findings suggest that assessing linguistic competence (e.g., lexical and grammar knowledge) may be less necessary than thought especially if assessed in multiple-choice format that is accompanied by reading and listening test items based on the CEFR. Because the CEFR focuses on linguistic competence within communicative language activities, using multiple-choice items in an isolated manner to assess test-takers' linguistic competence is less likely to fit into the CEFR scales. In addition, given that reception (i.e., listening and reading) and linguistic competence (i.e., lexico-grammar) loaded on the same factor in the latent structure of the ECCE and that reception is emphasized as one of the main communicative modes in the CEFR

(European Council, 2001), assessing reading and listening skills may be sufficient. Furthermore, given that the assessment of grammar and vocabulary in isolated sentences and multiple-choice items is still not uncommon in standardized language tests (e.g., TOEIC and TOEFL IPT), developers of these tests should reconsider the value of having isolated grammar and vocabulary test items.

Fourthly, given that the correlated three-factor model was only partially generalizable across different L1s due to differences in the intercepts of vocabulary scores (i.e., Spanish-speaking test-takers receiving higher vocabulary scores than Greek-speaking test-takers), test developers should carefully consider cognate effects in vocabulary tests if test-takers come from various L1 backgrounds.

Fifthly, this study makes links to a proficiency test based on the CEFR and macro-functions suggested in the CEFR CV (Council of Europe, 2018). The findings hint at the separate operations of the macro-functions of reception, written production, and spoken interaction in the underlying structure of the ECCE. However, other macro-functions including written interaction and spoken production are not included in the test, and thus were not examinable in our study. In a future study, to support arguments on the four macro-functions of language suggested in the CEFR CV, it would be important to design or examine a test that contains all of the macro-functions suggested by the CEFR CV, and then explore the latent structure of the test, the dimensionality and operations of the macro-functions, and the commensurability between the test and the macro-functions. By doing so, such studies could reveal whether the four macro-functions assessed by a CEFR-based language test are psychometrically separable, and, if so, how they constitute distinct constructs of language proficiency.

Lastly, while the ECCE is developed to measure the B2 level of English proficiency based on the CEFR, it is less clear to what extent the ECCE measures this level or how it may differ from other proficiency tests aimed at different levels of proficiency, such as the TOEFL and the Michigan English Language Assessment Battery (MELAB). Future studies comparing language tests which assess different levels of proficiency (e.g., intermediate vs. advanced) could provide detailed specifications on how knowledge and skills in an L2 differ between given levels of language proficiency.

As a concluding remark, as Little (2019) states, it seems that "language education without [the CEFR] has become unimaginable" (p. 571) as the impact of the CEFR has spread around the globe. However, it should be noted that one aspect of the CEFR that lags behind its widespread impact is its unclear status in terms of scientific foundations (Alderson 2007; Deygers, 2019; Hulstijn, 2007; Little, 2019). As such, one of the contributions that researchers and practitioners can make in the field of language assessment is to seek empirical evidence that shows whether or not the CEFR's claims can be empirically validated using test-taker performance. Our study focused on one of the CEFR's claims (i.e., four macro-functions of language use) and examined the commensurability between components of a CEFR-based language test and the CEFR's macro-functions. Beyond the macro-functions of language use, more research is needed to examine whether CEFR-based tests are in line with the claims that CEFR reflects broader discussions of language teaching, learning, and assessment (Little, 2019). Lastly, we hope that the findings of this study, to some degree, can encourage language test developers, who make use of the CEFR as a basis for test design, including developers of the ECCE, to critically reconsider their tests in terms of how well they reflect the main goal of the CEFR (i.e., helping learners perform communicative tasks in the real world and exercise their agency in the target

language), and actively engage in reflection, communication, and discussion in the domains of language learning, teaching, and assessment.

Acknowledgements

This paper reports on research funded through Michigan Language Assessment's Spaan Research Grant Program, 2017. We are grateful to Michigan Language Assessment for their support and feedback. We are also greatly indebted to the anonymous reviewers and the editor Dr. Constant Leung for their helpful comments and suggestions.

References

Alderson, J. C. (2007). The CEFR and the need for more research. *The Modern Language Journal*, *91*(4), 659–663.

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.

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

British Council. (2018). *EU 2025: The future demand for English language in the European Union*. British Council.

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

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1,* 1–47.

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.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ:

Erlbaum.

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

Council of Europe. (2018). *Common European Framework of Reference for Languages: learning, teaching, assessment companion volume with new descriptors*. Council of Europe.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement*, 2nd ed. (pp. 443–507). Washington, DC: American Council on Education.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.

Cumming, A. (2009). Language assessment in education: Tests, curricula and teaching. *Annual Review of Applied Linguistics, 29*, 90–100.

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.

Deygers, B. (2019). The CEFR companion volume: Between research-based policy and policy-based research. *Applied Linguistics.*

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

George, D., & Mallery, M. (2010). *SPSS for windows step by step: A simple guide and reference, 17.0 update* (10th Ed.). Boston: Pearson.

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.

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.

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.

In'nami, Y., & Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing*, *26*(2), 219–244.

In'nami, Y., & Koizumi, R. (2012). Factor structure of the revised TOEIC® test: A multiple-sample analysis. *Language Testing, 29*(1), 131–152.

Kane, M. (1992). An argument-based approach to validation. *Psychological Bulletin, 112*, 527–535.

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement*, 4th ed. (pp. 17–64), Westport, CT: American Council on Education and Praeger.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73.

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

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.

Little, D. (2019). Proficiency guidelines and frameworks. In J. Schwieter & A. Benati (Eds.), *The Cambridge Handbook of Language Learning* (pp. 550–574). Cambridge University

Press.

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

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, Monograph Supplement, 3*, 635–694.

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: Cambridge Michigan Language Assessments.

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.

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

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.

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.

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.

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.