ETS TOEFL

Quality Beyond Measure.

TOEFL® Research Report
TOEFL–RR-82
ETS RR–18-21

Evaluating Invariance in Test Performance for Adolescent Learners of English as a Foreign Language

Venessa Manna

Hanwook Yoo

Lora Monfils

December 2018

The *TOEFL*® test is the world's most widely respected English language assessment, used for admissions purposes in more than 130 countries including Australia, Canada, New Zealand, the United Kingdom, and the United States. Since its initial launch in 1964, the TOEFL test has undergone several major revisions motivated by advances in theories of language ability and changes in English teaching practices. The most recent revision, the *TOEFL iBT*® test, contains a number of innovative design features, including integrated tasks that engage multiple skills to simulate language use in academic settings and test materials that reflect the reading, listening, speaking, and writing demands of real-world academic environments. In addition to the TOEFL iBT, the TOEFL Family of Assessments has expanded to provide high-quality English proficiency assessments for a variety of academic uses and contexts. The TOEFL Young Students Series (YSS) features the *TOEFL*® *Primary*™ and *TOEFL Junior*® tests, designed to help teachers and learners of English in school settings. The *TOEFL ITP*® Assessment Series offers colleges, universities, and others an affordable test for placement and progress monitoring within English programs.

Since the 1970s, the TOEFL tests have had a rigorous, productive, and far-ranging research program. ETS has made the establishment of a strong research base a consistent feature of the development and evolution of the TOEFL tests, because only through a rigorous program of research can a testing company demonstrate its forward-looking vision and substantiate claims about what test takers know or can do based on their test scores. In addition to the 20-30 TOEFL-related research projects conducted by ETS Research & Development staff each year, the TOEFL Committee of Examiners (COE), composed of distinguished language-learning and testing experts from the academic community, funds an annual program of research supporting the TOEFL family of assessments, including projects carried out by external researchers from all over the world.

To date, hundreds of studies on the TOEFL tests have been published in refereed academic journals and books. In addition, more than 300 peer-reviewed reports about TOEFL research have been published by ETS. These publications have appeared in several different series historically: TOEFL Monographs, TOEFL Technical Reports, TOEFL iBT Research Reports, and TOEFL Junior Research Reports. It is the purpose of the current TOEFL Research Report Series to serve as the primary venue for all ETS publications on research conducted in relation to all members of the TOEFL Family of Assessments.

To obtain more information about the TOEFL programs and services, use one of the following:

E-mail: toefl@ets.org **Web site:** www.ets.org/toefl

ETS is an Equal Opportunity/Affirmative Action Employer.

As part of its educational and social mission and in fulfilling the organization's non-profit Charter and Bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

TOEFL Research Report Series and ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Evaluating Invariance in Test Performance for Adolescent Learners of English as a Foreign Language

Venessa Manna,[1] Hanwook Yoo,[1] & Lora Monfils[2]

1 Educational Testing Service, Princeton, NJ
2 Educational Testing Service, San Francisco, CA

In this study, we assessed the invariance in the factor structure underlying English-language proficiency for two groups of adolescent learners in Japan: students in middle school (ages 13 – 15 years) and students in high school (ages 16 – 18 years). Language proficiency was measured using the *TOEFL Junior*® Comprehensive test, an assessment designed to measure the English skills of adolescent English learners in non-English-speaking countries. The study results indicate that a correlated 4-factor model corresponding to the 4 language abilities of reading, listening, speaking, and writing best represents the nature of language proficiency in the 2 groups of adolescent English learners. Moreover, the factor structure is invariant across these 2 groups. However, there is a statistically significant difference in performance on the reading construct. The results are consistent across 2 random samples, thus providing confirmatory evidence of model invariance. This study provides empirical support for the current score-reporting practices for the TOEFL Junior Comprehensive test and suggests that the test scores have the same meaning across test takers from middle school and high school.

Keywords English as a foreign language; adolescent learners; CFA; measurement invariance; cross-validation

doi:10.1002/ets2.12208

Research into the development of first and second language (L1 and L2) skills acquisition, particularly the modalities of listening, speaking, reading, and writing, has been conducted by psychologists and linguists for a number of decades. Historically, the language modalities of reading and writing were thought to be secondary forms of language, dependent on the oral skills of listening and speaking (Berninger, 2000; Shanahan, 2006). This historical view of the four modalities was influenced by the development of language skills in children; that is, regardless of language, children's listening and speaking skills, which are developed prior to formal schooling, provide the foundation for learning reading and writing (Bozorgian, 2012).

More recently, researchers have noted that the four language modalities overlap in their development. For example, for both L1 and L2 acquisition, improvement in listening skills has been shown to positively affect the other skills (Bozorgian, 2012). Moreover, even though the development of reading and writing is affected by the oral modalities, reading and writing in turn can affect listening and speaking (Shanahan, 2006), in general and in situations where exposure to discourse (listening, speaking) in the target language is limited (Bozorgian, 2012). Research has also shown that, as in L1 learning, there is evidence to support theories of L2 learning that separate out comprehension and production. For example, for adult foreign language learners, the ability to understand the language (listening, reading) develops in advance of the ability to produce the language (speaking, writing; Osterhout, McLaughlin, Pitkänen, Frenck-Mestre, & Molinaro, 2006, as cited in Muñoz & Singleton, 2011).

Building on research in language development, researchers have been investigating two opposing theories of the relationship of the language modalities using factor analytic techniques: *divisible competence* and *unitary competence*. The divisible competence hypothesis suggests that language can be divided into components or modalities. For example, as discussed, components may comprise several modalities, such as oral versus written modalities (i.e., listening/speaking vs. reading/writing), receptive versus productive modalities (i.e., listening/reading vs. speaking/writing), or each modality as a different component from that of the other modalities — or even the possibility of several language components, such as phonology, syntax, and lexicon (Bachman, 2000; Lado, 1961; Oller Jr. & Hinofotis, 1980; Scholz, Hendricks, Spurling, Johnson, & Vandenburg, 1980). Under the divisible competence hypothesis, a factor analytic study would result in each component loading on a separate factor. For the unitary competence hypothesis, researchers have suggested that language

Corresponding author: H. Yoo, E-mail: hyoo@ets.org

proficiency is a cohesive skill, thus all four language modalities (or all language components) would load onto one factor (Oller Jr. & Hinofotis, 1980).

Early studies on language tests administered to English as a second language (ESL) and English as a foreign language (EFL) students in higher education provided empirical support for the unitary competence hypothesis (Oller, 1983; Oller Jr. & Hinofotis, 1980; Scholz et al., 1980). Several studies of English-language proficiency assessments, such as the *TOEFL*® test, the Foreign Service Institute Oral Interview, and the ESL Placement Examination from the University of California, Los Angeles, reported that the first factor accounted for the majority of the total variance, with very little residual variance remaining once the general factor had been taken into account. Additionally, multifactor models did not provide a clear pattern of loadings based on language modality. However, the results from these early research studies were questioned because of the statistical techniques used. For example, some studies used principal components analysis, whereas a more appropriate methodology would have been factor analysis (Joliffe, 1986; Widaman, 2007). Moreover, as Powers (1982) has noted, results of dimensionality studies may vary based on characteristics of the analysis sample, in general and as pertaining to English proficiency.

As researchers continued to investigate the construct of language proficiency, a new theory emerged that comprised both the divisible competence and unified competence hypotheses (Bachman, 1991). Oller Jr. and Hinofotis (1980) noted that "the choice between the unitary competence hypothesis and the possibility of separate skills is less clear" (p. 23). This notion is particularly evident when the first factor does not account for the majority of the variance. Researchers hypothesized that language proficiency has multiple components with one overall general competence and several smaller yet interrelated competencies (Bachman, 1991). Thus, as Vollmer and Sang (1983) argued, the two hypotheses are not "mutually exclusive."

Further studies on language tests administered to college students reported that ESL/EFL proficiency consists of one higher order factor and several correlated first-order ability factors that measure skills in reading, listening, speaking, and writing (Bachman, Davidson, Ryan, & Choi, 1995; Bachman & Palmer, 1996). In contrast, research by Hale, Rock, and Jirele (1989) on TOEFL found that the results supported the correlated two-factor model, in which one factor was defined by listening comprehension and the other factor was defined by nonlistening components consisting of word usage, written expression, vocabulary, and reading comprehension. These studies give further support to the combination of both language hypotheses.

Currently the general view among researchers is that L2 proficiency is complex and multidimensional in nature (Carroll, 1965; Harsch, 2014; Oller, 1983). As noted earlier, this view has been supported by a number of factor analytic studies that found that language ability comprises one or more components representing the four language skills (reading, listening, speaking, and writing) that may be distinct, closely related, or hierarchically related to a global ability (Bachman et al., 1995; Sawaki & Sinharay, 2013; Sawaki, Stricker, & Oranje, 2009).

Research has also shown that the development of linguistic competence is complex and may be impacted by a number of test-taker characteristics, including cultural background, native language, cognitive ability, gender, and age (Bachman, 1990; Kunnan, 1998). Moreover, in studies that involved adult English learners, results indicate that the development of each of the four language skills may be differentially impacted by these background characteristics (Gradman & Hanania, 1991; Gu, 2014; Kunnan, 1995; Manna & Yoo, 2015; Wilson, 2000).

Although the dimensionality of language proficiency measured by ESL/EFL tests has been well documented, the majority of these studies have focused on adult learners. In contrast, research into the language ability assessed by tests designed for young learners of English is quite sparse. With the rise of English as a lingua franca and increasing implementation of teaching EFL to young learners, a greater understanding of English-language proficiency for young learners is needed to support best practice in both instruction and assessment (Edelenbos & Kubanek, 2009; Inbar-Lourie & Shohamy, 2009). As part of this, examination of invariance across young test-taker groups is needed to determine whether test scores reflect the same construct for various groups of test takers and thus inform score use within and across groups.

In a comprehensive review of the research on age and L2 acquisition, Muñoz and Singleton (2011) noted that although it is generally recognized that there is a relationship between age and L2 learning, the nature of that relationship has been a subject of controversy among researchers. In part, this is due to a historical focus on maturational factors (e.g., critical period hypothesis) with insufficient investigation into other factors that may impact L2 learning (Muñoz & Singleton, 2011), but it may also be due to historically divergent results that reflect methodological differences and challenges, particularly in the areas of "subject selection, data collection, and instrumentation" (DeKeyser, 2013, p. 52).

More recent studies have sought to examine the role of learning context and individual characteristics that may mediate age effects (Muñoz & Singleton, 2011). Results of studies conducted in instructional settings suggest that the effect of age on L2 learning covaries with learning environment (Lichtman, 2016; Muñoz, 2008). Moreover, the developmental relationships among language components may vary by learning conditions.

Among those studies that looked at L2 development and proficiency in young learners, the focus tended to be in the context of bilingual or ESL education. For example, several researchers have supported the distinction between social and academic language proficiency (Bailey, 2007; Carroll, 1983; Cummins, 1981). Hakuta, Butler, and Witt (2000) reported that in the context of English language learners (ELLs) in the United States, oral proficiency takes 3 – 5 years to develop, while academic English proficiency takes 4 – 7 years. In another study, Bae and Bachman (1998) looked at the impact of the background characteristics on Korean language proficiency of learners enrolled in a two-way Korean – English immersion program in a US elementary school setting. In this study, the researchers found that the correlated two-factor model best represented the structure of a Korean language test that included reading and listening tasks in both academic and general language settings. The study also found that Korean heritage learners showed less variability on listening, while nonnative speakers of Korean showed less variability on reading.

In the context of an EFL learning environment, a recent study by Gu (2015) examined the latent structure of language abilities of young ELLs using the *TOEFL Junior*® Comprehensive test (TJC). Initially developed to measure English proficiency for middle school ELLs in non-English-speaking countries, and currently administered to students ages 11+ years, this test assesses the academic and everyday English skills used in a school setting. Results indicate that language ability could be structurally represented by a higher order model, similar to that found in other factor analytic studies involving adult learners (Sawaki et al., 2009; Stricker & Rock, 2008). Specifically, the study found that test performance as measured by the TJC can be represented by a global English-language ability factor and four first-order factors corresponding to the four language skills: reading, listening, speaking, and writing. These results support the current score-reporting practice of providing an overall proficiency score as well as scores for each of the four skills.

The present study takes a look at the structure of English-language proficiency for learners of EFL and its invariance at a more granular level. As L2 acquisition is influenced by the learning environment, the study is limited to performance of ELLs in middle school and high school in Japan. Specifically, the study focuses on 13- to 15- and 16- to 18-year-olds and the possible impact of differences in age and corresponding school context (middle vs. high school) on how the language components are developmentally related to each other. The study also examines whether the two test-taker groups differ substantially on the means of the latent components representing English-language proficiency. In addition to investigating the factor structure of the abilities measured in the TJC across the two groups, this study uses cross-validation to examine the generalizability of the resulting factor structure across samples.

Building on the work of Gu (2015), the following research questions are addressed:

1. What is the latent structure of the TJC for learners of EFL in middle and high school?
2. To what extent is the latent structure invariant across these two groups of learners?
3. Are these results generalizable across samples?

The results of this study provide validity evidence to inform test score interpretation within and across the studied test-taker groups.

Method

Structure of the Measurement Instrument

To address the research questions of interest, this study used the computer-delivered TJC developed by Educational Testing Service. The test is not based on any specific curriculum and measures the academic and everyday English skills of students ages 11 years and older in non-English-speaking countries (Educational Testing Service, 2015). Intended to support teaching and learning by providing meaningful feedback on the proficiency level of students in the target population, the main emphasis of the test is the measurement of communicative competence.

The test consists of four sections: Reading, Listening, Speaking, and Writing. As noted by the design team, in practice, multiple language modalities may be required to complete a single language task. Therefore integrated tasks, which require multiple modalities (e.g., listening stimulus for speaking, listening, and reading stimulus for writing), are also included in the Speaking and Writing sections of the TJC.

Table 1 Test Design of the TOEFL Junior Comprehensive Test

Section	Total number of items	Scale score
Reading	28	140 – 160
Listening	28	140 – 160
Speaking	4	0 – 16
Writing	4	0 – 16
Total test	64	1 – 6

The Reading and Listening sections each contain 28 dichotomously scored selected response questions, with four options for each item. The Speaking and Writing sections each contain four constructed-response questions, each scored on a 0 – 4 scale. Separate scores are reported for each of the four sections in addition to the overall score level (see Table 1). The reported coefficient alpha reliability estimates (Cronbach, 1951) for the TJC were .90 for Reading, .90 for Listening, .87 for Speaking, and .83 for Writing. It should be noted that these estimates were obtained from the sample of test takers used to set the scale for the TJC (Wang, 2012).

Data

This study used data from an administration of the TJC in Japan in fall 2015. A total of 3,206 Japanese test takers took this test form. As the research questions focus on students in middle and high school, only those examinees ($n = 2{,}885$) who self-reported their ages to be 13 – 18 years, which corresponds to the Japanese ages for middle and high school, were included in the analyses (Japan Ministry of Education, Culture, Sport, Science, and Technology [MEXT], n.d.). As of 2011, fifth- and sixth-grade students in Japan are introduced to English (i.e., foreign language activity) as part of the primary school curriculum (MEXT, 2008). Thus sampled students in this study would have been studying English for approximately 4 – 5 years. The reliability estimates for this sample were .82 for Reading, .81 for Listening, .77 for Speaking, and .79 for Writing. These estimates were slightly lower than those reported by Wang (2012), but this is not surprising, because those estimates were based on a more diverse group of test takers.

Of the 2,885 middle and high school students, the majority were in the high school group (80%). Therefore, to address the disproportionate size of the two groups of study interest (middle school vs. high school), a random sample of 500 was taken from each group for a total analysis sample size of 1,000 (Sample A). A second sample was selected (Sample B) to conduct cross-validation analyses. Table 2 provides descriptive statistics for the two analysis samples as well as for all test takers for the fall 2015 administration. Also shown in Table 2 is the performance of the analysis samples by studied group (middle school vs. high school). The comparison of these summary statistics suggests that the performance of test takers who self-reported their ages as 13 – 18 years, from which our analysis samples were drawn, were similar to all test takers for the fall 2015 administration. Furthermore, Sample A and Sample B were similar to each other and to the total analysis sample from which they were drawn. Also, the performance of sampled high school test takers tended to be similar on average to those in middle school.

Table 2 Summary Statistics for Test-Taker Groups

		Reading		Listening		Speaking		Writing		Total test
Test-taker group	*N*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Median
All test takers	3,206	148.7	5.5	147.3	5.6	7.8	2.4	6.9	2.9	3
All 13 – 18[a]	2,885	148.2	5.4	146.9	5.4	7.6	2.3	6.7	2.8	3
Sample A	1,000	148.1	5.5	147.0	5.6	7.7	2.5	6.7	3.0	3
Middle school	500	147.5	5.4	147.0	5.7	7.7	2.7	6.7	3.2	3
High school	500	148.7	5.5	147.0	5.5	7.7	2.3	6.8	2.7	3
Sample B	1,000	147.9	5.3	147.0	5.5	7.7	2.4	6.7	2.9	3
Middle school	500	147.6	5.4	147.1	5.7	7.7	2.7	6.7	3.1	3
High school	500	148.3	5.2	146.8	5.3	7.6	2.2	6.8	2.6	3

[a]Test takers self-reporting age as 13 – 18 years, from which analysis Samples A and B were drawn.

Analyses

To evaluate the generalizability of results, each of the analyses described in the following pages was done separately for Sample A and Sample B, and results were compared across samples.

Confirmatory Factor Analysis

As a first step, an item-level confirmatory factor analysis (CFA) was conducted to determine whether the higher order factor structure found in Gu's (2015) study still held for the TJC given the more recent data collection. To this end, all three of the competing models (correlated four-factor, single-factor, and higher order factor) used in the Gu study were included for evaluation. The correlated four-factor model is motivated by the test design and hypothesizes the presence of four psychometrically distinct but correlated latent abilities defined by the items assessing each of the four language skills (Figure 1). The single-factor model hypothesizes that the four language skills are psychometrically indistinguishable from each other and that a single language ability underlies performance on the test (Figure 2). The higher order factor model assumes that there is a general ability factor that encompasses four latent factors corresponding to the four language skills (Figure 3).

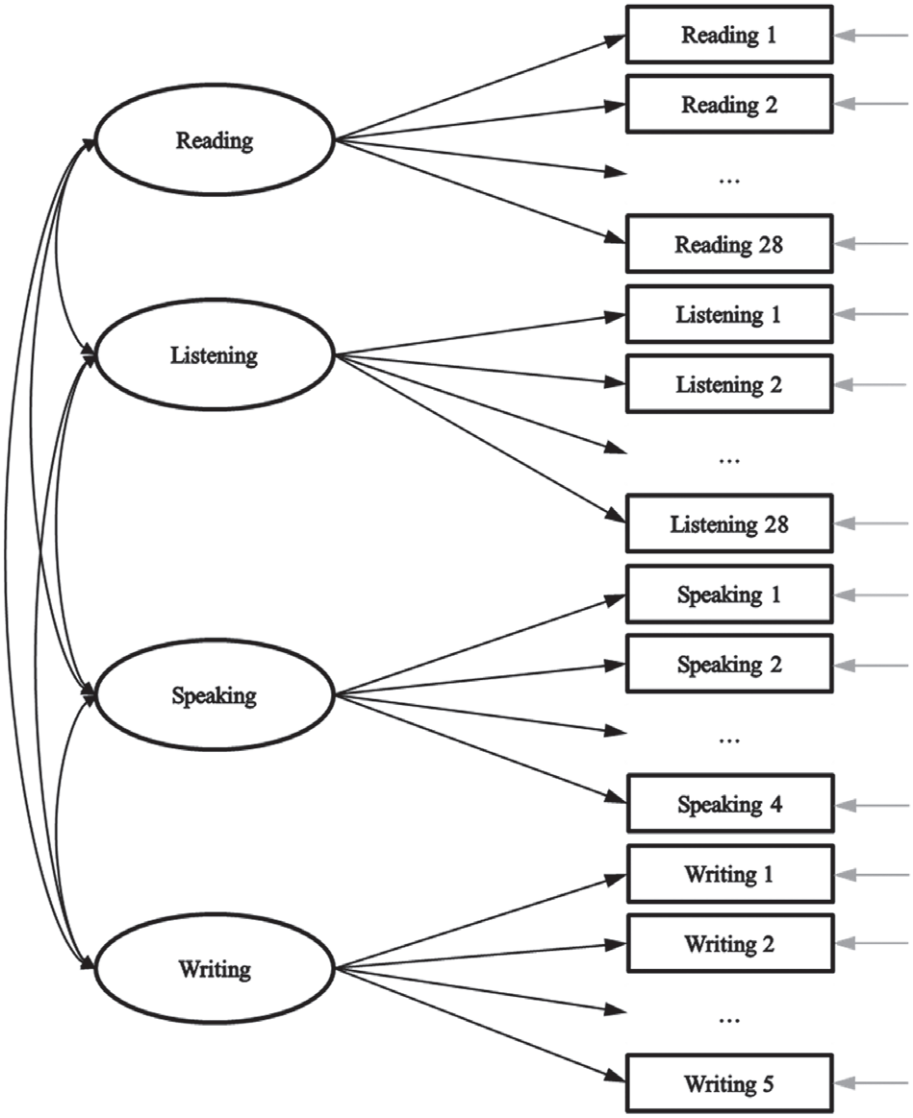

Figure 1 Schematic of correlated four-factor model.

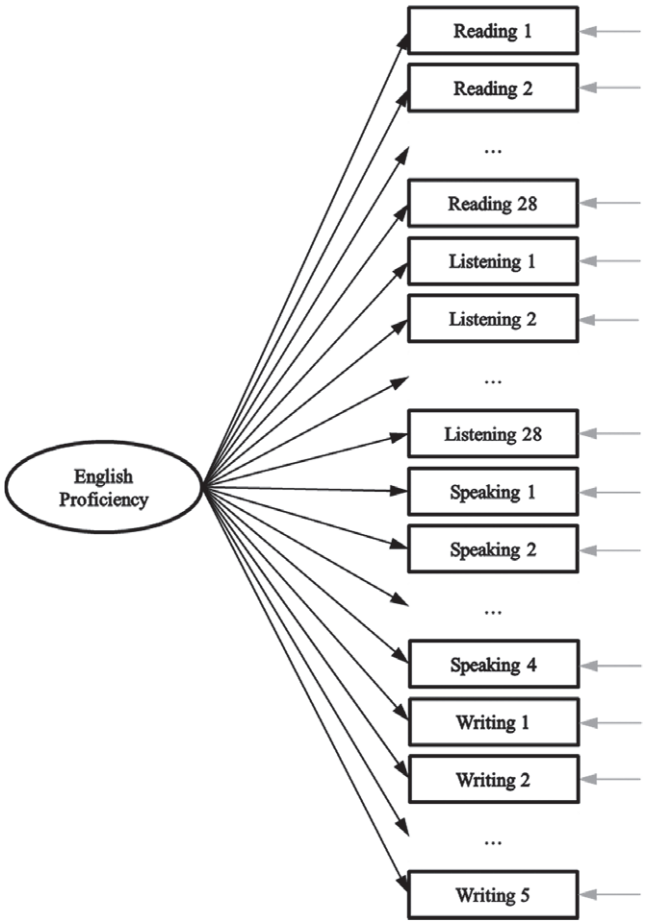

Figure 2 Schematic of single-factor model.

The model evaluation process involved the comparison of competing models (single-factor and higher order factor) against the least restrictive baseline model (correlated four-factor). Several goodness-of-fit indices were used to evaluate model-to-data fit. These included the comparative fit index (CFI) and the Tucker–Lewis index (TLI). Hu and Bentler (1999) recommended a cutoff value close to .95 for both of these indices. Two additional fit indices used were the root mean square error of approximation (RMSEA) and its 90% confidence interval (90% confidence interval), with a value of .05 or below indicating a good fit (Browne & Cudeck, 1993), and the Satorra–Bentler scaled chi-square difference test (SB$\Delta\chi^2$; Satorra & Bentler, 2001). In addition to comparison of goodness-of-fit indices, model parsimony and the reasonableness of individual parameter estimates (statistical significance, residuals, and modification indices) and correlations among the latent factors (with correlations greater than .90 considered extreme; Bagozzi & Yi, 1988; Sawaki et al., 2009) were considered in the evaluation of competing models.

All analyses were conducted with Mplus 7.4 (Muthén & Muthén, 2015). Parameter estimation was performed using a robust weighted least squares estimator with the diagonal weight matrix (WLSMV) method. The integrated speaking and writing items were loaded only on the target modality, for consistency with current score-reporting practices. For example, the integrated speaking items loaded only on the speaking factor and not on the listening or reading factor. The integrated writing items loaded only on the writing factor and not on the listening or reading factor. The first indicator of each factor was set to 1 for scale identification.

Factorial Invariance Across Middle and High School Groups

Once the factor structure that best represented the internal structure of the test was identified, the factorial invariance of the CFA model across the middle and high school groups was evaluated. The first step in this process was to test for configural invariance, that is, to assess whether the best fitting factor structure is invariant across the two age groups.

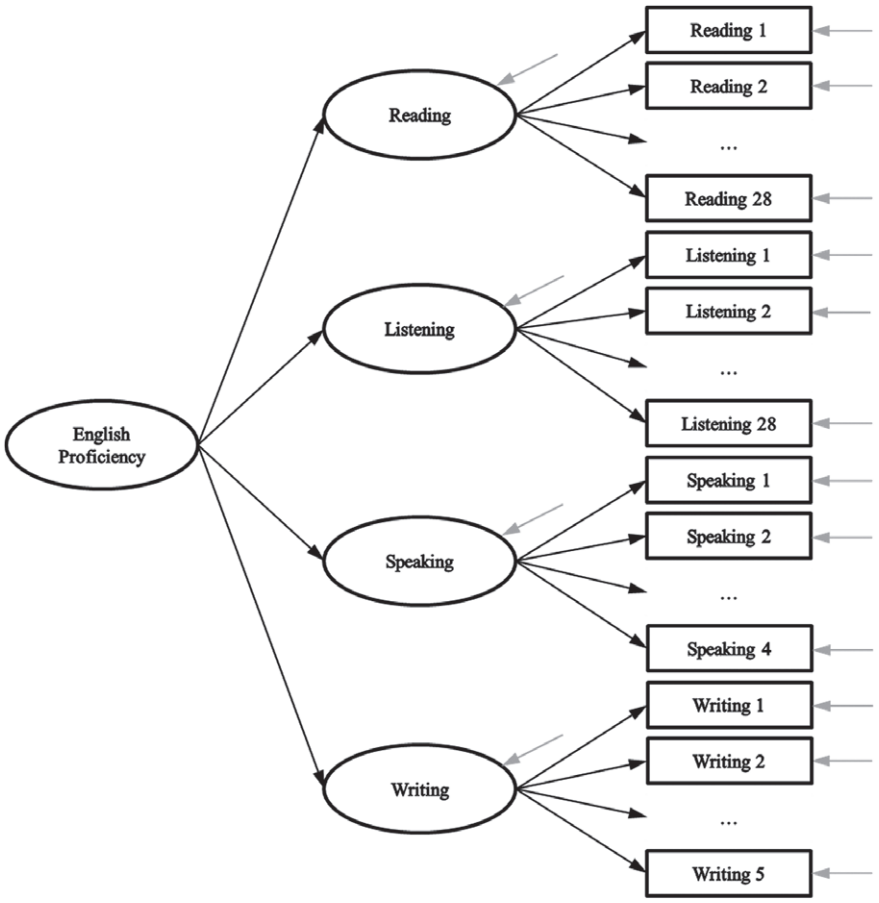

Figure 3 Schematic of higher order factor model.

Once configural invariance was established, measurement invariance (invariance of the factor loading, item thresholds, and error variances) was evaluated by testing a series of CFA models obtained by sequentially adding constraints to the configural invariance model, with the middle school group as the reference group. As a final step, structural invariance (invariance of the variances and covariances of the latent variables) was evaluated to compare the estimated mean of the latent factors. The models fitted and the comparisons they support are explained in more detail:

configural invariance (M0). In this model, the factorial structure was constrained to be the same across the middle school and high school groups. The factor variances were fixed to 1, and the factor means were fixed to 0 in each group for identification. The residual variances were not uniquely identified in the configural invariance model and therefore were constrained to 1 in both groups of test takers. This model supports the claim that the construct is operationalized in the same way in both groups.

weak measurement invariance (M1). The model was tested by constraining all factor loadings to be equal across groups. For model identification, the factor variance was fixed to 1 for the middle school group and freely estimated in the high school group, and the factor mean was fixed to 0 in both groups. All residual variances were constrained to 1 across groups. This model supports the claim that the relationship between the test items and the latent construct is equivalent across both groups of test takers.

strong measurement invariance (M2). This model was tested by constraining the factor loadings and thresholds to be the same across groups. In this model, the factor variances and means were fixed to 1 and 0, respectively, in the middle school group for identification, but both factor variances and means were freely estimated in the high school group. Again, the residual variances were constrained to be 1. Under strong measurement invariance, the latent means between groups can be compared.

strict measurement invariance (M4). This model was tested by constraining the factor loadings, thresholds, and error variances to be equal across groups. It is the same as M2 and is denoted as M4(=M2). This test for strict measurement invariance proceeds backward, such that a model with all residual variances freely estimated in the high school group (M3) was first fitted and then compared with a model in which all residual variances were fixed to 1 for both groups (M4). Strict measurement invariance provides support for the claim that the test items are measured with the same precision in each group.

structural invariance (M5). The establishment of strict measurement invariance facilitates a comparison of the variability and relationship among the constructs, that is, the test for structural invariance. In this study, the focus was on examining the differences of estimated factor means between two groups by constraining the factor variances and covariances. The factor means were fixed to 0 in the middle school group but freely estimated in the high school group.

The factorial invariance analyses were also conducted using Mplus 7.4. The weighted least squares means and variance (WLSMV) and the theta parameterization were used to estimate all models (Muthén & Muthén, 2015). As a result, the model fit statistics describe the fit of the item factor model to the polychoric correlation matrix among the items for each group. The model fit indices (CFI, TLI, and RMSEA) mentioned previously were also used to evaluate the global fit of the invariance models to the data. In addition, the SB$\Delta\chi^2$ test and changes in the CFI index were used in the sequence of invariance tests (M1 vs. M0; M2 vs. M1; M4 vs. M3; M5 vs. M4). In comparing two nested models, if a decrease in the value of CFI is greater than .01, then the more restrictive model should be rejected (Cheung & Rensvold, 2002).

Results

Confirmatory Factor Analysis

The overall goodness-of-fit statistics from the CFA of the three hypothesized models for Sample A and Sample B are presented in Table 3. The correlated four-factor model, which served as the baseline against which the two competing models were compared, had model fit indices that suggested that this model fit the data well (Sample A, RMSEA = .019, CFI = .976, TLI = .975; Sample B, RMSEA = .018, CFI = .968, TLI = .967). A comparison of the single-factor to the correlated four-factor model indicated that although the single-factor fit indices exceeded the criteria, they showed slightly worse fit (Sample A, RMSEA = .022, CFI = .968, TLI = .967, ΔCFI = −.008; Sample B, RMSEA = .021, CFI = .970, TLI = .969, ΔCFI = −.006). Similarly, the fit indices for the higher order model compared to the baseline model were acceptable (Sample A, RMSEA = .019, CFI = .976, TLI = .975, ΔCFI = .000; Sample B, RMSEA = .019, CFI = .975, TLI = .975, ΔCFI = −.001), with marginally worse fit for Sample B. Given the similarity of these fit statistics, the least restrictive four-factor model was selected as the best representation of the factor structure.

Across the two samples, the interfactor correlations for the four-factor model ranged from .810 to .932, with speaking and reading having the lowest correlation (see Tables 4 and 5). The factor loadings for the correlated four-factor model were also substantively interpretable and meaningful, with all estimates being statistically significant, $p < .001$. As shown in Appendixes A and B, standardized loadings for Reading vary in magnitude from .19 to .72 (Sample A) and from .25 to .72 (Sample B), for Listening from .14 to .77 (Sample A) and from .18 to .72 (Sample B), for Speaking from .74 to .81 (Sample A) and from .73 to .81 (Sample B), and for Writing from .63 to .81 (Sample A) and from .61 to .80 (Sample B). Also provided in Appendixes A and B are standard measurement errors that represent the percentage of variance in the indicators that is not explained by the target latent factor. Thus, given the model-to-data fit, the pattern of correlation among the latent factors and reasonableness of parameter estimates for both Sample A and Sample B, the correlated four-factor solution was chosen as the factor structure that best represented the performance of adolescent ELLs in Japan as measured by the TJC.

Factorial Invariance Across Middle School and High School Groups

Table 6 shows the results of the test for configural invariance, that is, a test of whether the proposed correlated four-factor model fits the empirical data for each group. Results show excellent fit for the middle school group (Sample A, RMSEA = .016, CFI = .984, TLI = .983; Sample B, RMSEA = .017, CFI = .982, TLI = .981) as well as for the high school group (Sample A, RMSEA = .019, CFI = .968. TLI = .967; Sample B, RMSEA = .018, CFI = .970, TLI = .969), indicating

Table 3 Summary of Results for Confirmatory Factor Analysis

Model	*df*	SBχ^2	RSMEA (90% CI)	CFI	TLI	SB$\Delta\chi^2$	Δdf	*p*-Value	ΔCFI
Sample A									
Correlated four factor	2,009	2,729.6	.019 [.017, .021]	.976	.975				
Single factor	2,015	2,970.6	.022 [.020, .023]	.968	.967	153.7	6	.000	−.008
Higher order factor	2,011	2,739.5	.019 [.017, .021]	.976	.975	10.3	2	.006	.000
Sample B									
Correlated four factor	2,009	2,691.3	.018 [.017, .020]	.976	.975				
Single factor	2,015	2,882.1	.021 [.019, .022]	.970	.969	183.4	6	.000	−.006
Higher order factor	2,011	2,709.4	.019 [.017, .020]	.975	.975	18.5	2	.000	−.001

Note. CFI = comparative fit index; CI = confidence interval; RMSEA = root mean square error of approximation; SB = Satorra – Bentler; TLI = Tucker – Lewis index.

Table 4 Estimated Correlation Matrix of the Latent Variables for Confirmatory Factor Analysis, Sample A

	Correlated four factor				
	Reading	Listening	Speaking	Writing	
Correlated four factor					
Reading	1				
Listening	.874	1			
Speaking	.810	.871	1		
Writing	.872	.917	.921	1	
	Higher order factor				
	Reading	Listening	Speaking	Writing	Higher order factor
Higher order factor					
Reading	1				
Listening	.861	1			
Speaking	.829	.878	1		
Writing	.879	.931	.896	1	
Higher order	.901	.955	.919	.975	1

Table 5 Estimated Correlation Matrix of the Latent Variables for Confirmatory Factor Analysis, Sample B

	Correlated four factor				
	Reading	Listening	Speaking	Writing	
Correlated four factor					
Reading	1				
Listening	.895	1			
Speaking	.842	.858	1		
Writing	.885	.901	.932	1	
	Higher order factor				
	Reading	Listening	Speaking	Writing	Higher order factor
Higher order factor					
Reading	1				
Listening	.876	1			
Speaking	.860	.874	1		
Writing	.900	.914	.898	1	
Higher order	.929	.944	.927	.969	1

Table 6 Configural Invariance of Confirmatory Factor Analysis Across Groups

Group	*N*	*df*	SBχ^2	RSMEA (90% CI)	CFI	TLI
Sample A						
Middle school	500	2,009	2,258.5	.016 [.012, .019]	.984	.983
High school	500	2,009	2,381.8	.019 [.016, .022]	.968	.967
Sample B						
Middle school	500	2,009	2,286.5	.017 [.013, .020]	.982	.981
High school	500	2,009	2,317.3	.018 [.014, .021]	.970	.969

Note. CFI = comparative fit index; CI = confidence interval; RMSEA = root mean square error of approximation; SB = Satorra – Bentler; TLI = Tucker – Lewis index.

Table 7 Factorial Invariance of Confirmatory Factor Analysis Across Groups

Model	Comparison	*df*	SBχ^2	RSMEA (90% CI)	CFI	TLI	SB$\Delta\chi^2$	Δdf	*p*-Value	ΔCFI
Sample A										
M0		4,018	4,642.2	.018 [.015, .020]	.977	.976				
M1	M1 – M0	4,079	4,624.9	.016 [.014, .019]	.980	.980	70.1	61	.199	.003
M2	M2 – M1	4,167	4,717.3	.016 [.014, .019]	.980	.980	104.4	88	.111	.000
M3		4,102	4,716.9	.017 [.015, .020]	.977	.977				
M4 (=M2)	M4 – M3	4,167	4,717.3	.016 [.014, .019]	.980	.980	76.0	65	.166	.003
M5	M5 – M4	4,177	4,783.4	.017 [.015, .019]	.978	.978	24.0	10	.008	−.002
Sample B										
M0		4,018	4,604.2	.017 [.015, .019]	.977	.977				
M1	M1 – M0	4,079	4,532.0	.015 [.012, .017]	.983	.982	53.4	61	.746	.006
M2	M2 – M1	4,167	4,625.6	.015 [.012, .017]	.982	.982	106.7	88	.086	−.001
M3		4,102	4,697.9	.017 [.014, .019]	.977	.977				
M4 (=M2)	M4 – M3	4,167	4,625.6	.015 [.012, .017]	.982	.982	52.5	65	.868	.005
M5	M5 – M4	4,177	4,816.5	.017 [.015, .020]	.975	.976	35.5	10	.000	−.007

Note. CFI = comparative fit index; CI = confidence interval; M0 = configural invariance; M1 = invariant factor loadings; M2 = M1 + invariant thresholds; M2P = M1 + partially invariant thresholds; M3 = free residual variances; M4 (same as M2P) = M3 + invariant residual variances; M5 = M4 + invariant factor variances; RMSEA = root mean square error of approximation; SB = Satorra – Bentler; TLI = Tucker – Lewis index.

that the correlated four-factor model of English-language proficiency is supported in both groups. In both Sample A and Sample B, the standardized factor loadings with the baseline model for each group were statistically significant, $p < .001$. For the middle school group in Sample A, the values varied in magnitude from .16 to .74 for Reading, from .27 to .68 for Listening, from .76 to .85 for Speaking, and from .63 to .83 for Writing. For the high school group in Sample A, the values ranged from .23 to .68 for Reading, from .09 to .79 for Listening, from .70 to .75 for Speaking, and from .63 to .80 for Writing (see Appendix C). Similar values for the standardized factor loadings were also found with Sample B (see Appendix D).

As configural invariance was supported, the next step was to test for various aspects of measurement invariance by comparing nested models M1 through M5 using SB$\Delta\chi^2$ and ΔCFI (see Table 7). These models all had overall global fit indices that were acceptable across Sample A and Sample B (RMSEA < .05; CFI > .95; TLI > .95). The first comparison, a test for weak measurement invariance (M1 to M0; Sample A, SB$\Delta\chi^2$ nonsignificant, ΔCFI = .003; Sample B, SB$\Delta\chi^2$ nonsignificant, ΔCFI = .006), suggested that the factor loadings were the same across the middle school and high school groups, that is, that the test items were related to the latent factor equivalently across groups. In the second comparison (M2 to M1), a test for strong measurement invariance, the results suggested that in addition to the factor loadings, the threshold indicators were also the same across the two groups (Sample A, SB$\Delta\chi^2$ nonsignificant, ΔCFI = .000; Sample B, SB$\Delta\chi^2$ nonsignificant, ΔCFI = −.001), suggesting that the latent means can be compared across groups. The third comparison, a test for strict measurement (M4 – M3), indicated equal residual variances across the two groups (Sample A, SB$\Delta\chi^2$ nonsignificant, ΔCFI = .003; Sample B, SB$\Delta\chi^2$ nonsignificant, ΔCFI = .005); that is, the observed variables were invariant across groups, thus providing no evidence of measurement bias. Given these results and the consistency across the two samples, we concluded that strict measurement invariance, with invariance of all factor loadings, invariance of all thresholds, and invariance of all item uniqueness, was in place across the middle school and high school groups.

Table 8 Estimated Factor Mean Comparison From Measurement Invariance Model

	Sample A				Sample B			
Latent variable	Standardized estimate	*SE*	*p*-Value	Effect size[a]	Standardized estimate	*SE*	*p*-Value	Effect size[a]
Reading	.224	.069	.001	.23	.153	.070	.028	.15
Listening	.001	.069	.994	.00	−.040	.069	.559	.04
Speaking	.002	.071	.980	.00	−.060	.071	.396	.06
Writing	.065	.070	.354	.07	.035	.070	.614	.04

Note. The reference group is middle school (means are all zero), and the focal group is high school.
[a]Effect size was calculated as $\widehat{d} = \left|\widehat{k}_1 - \widehat{k}_2\right| / \widehat{\phi}^{1/2}$, where $\widehat{k}_1$ and $\widehat{k}_2$ are estimated means on the measured construct and $\widehat{\phi}$ is a within-group pooled variance estimate for scores on the measured construct (Hancock, 2001). These are similar to Cohen's (1988) *d*, where values of .2, .5, and .8 can be interpreted as small, medium, and large effect sizes, respectively.

The establishment of measurement invariance allowed testing for structural invariance (M5 – M4). Although the SB$\Delta\chi^2$ was significant, the negligible change in ΔCFI (Sample A, −.002; Sample B, −.007) supported structural invariance across the middle school and high school groups. Moreover, on the basis of the CFI results, because the most restricted tested model (M5) fit the data well across the two groups, and its fit indices were only slightly smaller in comparison to those for M4, this model was adopted; that is, the correlated four-factor model functioned well across the two groups, and all the factor loadings, measurement error variances, factor variances, and factor covariances were equal across the two groups.

Estimated Factor Mean Difference

The establishment of structural invariance across the two groups of test takers indicates that (a) factor loading, (b) indicator threshold, and (c) factor variances and covariances did not vary across test-taker groups. Therefore the factor means from this model can be compared across two groups using those test takers in the middle school group as the reference group. These results are shown in Table 8 and indicate that there are no statistically significant differences in performance between the middle school and high school groups on the listening, speaking, and writing latent constructs. In contrast, for the reading latent construct, the high school group scored higher on average than the middle group (Sample A, standardized estimate = .224, effect size = .23; Sample B, standardized estimate = .153, effect size = .15). Overall, both Sample A and Sample B showed the same pattern of estimated factor mean differences.

Discussion

As noted in the introduction, this study sought to extend the work of Gu (2015) by investigating the invariance in the factor structure underlying English-language proficiency in two groups of adolescent learners in Japan, students in middle school (aged 13 – 15 years) and students in high school (aged 16 – 18 years), as measured by the TJC. Specifically, this study was designed to investigate the latent structure of the TJC, its invariance across the middle school and high school groups, and the generalizability across multiple samples. Although both studies used the same test (TJC), there were differences in study sample characteristics shown to influence L2 acquisition, such as L1 and academic context (Lichtman, 2016; Muñoz, 2008; Muñoz & Singleton, 2011; Powers, 1982). Specifically, there were differences in the age spans (11 – 15 vs. 13 – 18 years) and sample sizes (one sample of 436 vs. two samples of 1,000 test takers). In addition, this study used a Japanese-only sample to control for L1 and national academic context, whereas the sample in Gu's (2015) study included students from 15 countries, with slightly more than half native Korean speakers.

Results of the present study indicate that the latent structure of English-language proficiency as measured by the TJC was best represented by the correlated four-factor model corresponding to the four communicative language skills of reading, listening, speaking, and writing. Moreover, there was structural invariance across the two groups of learners of EFL. This result was consistent across two random samples (Samples A and B), thus providing confirmatory evidence of model invariance and consistency of score meaning across and within the studied middle and high school test-taker groups.

Although the selected latent factor model differed from Gu's (2015) higher order factor model (represented by a global English-language ability factor and four first-order factors corresponding to the four language skills: reading, listening,

speaking, and writing), the implications for score reporting are consistent. Specifically, the high correlations across the four latent factors found in this study were consistent with Gu in providing support for the current practice of reporting scores for each of the four modalities as well as total test. The study results also provide support for curriculum development and classroom learning strategies with a focus on all four modalities of communicative language proficiency to foster language development for the young EFL learners rather than focusing on individual modalities in isolation.

The study's findings also support the factorial invariance of the language ability constructs across the middle school and high school groups as measured by the TJC; that is, the correlated four-factor model was representative of the factor structure across these two groups of test takers. Moreover, the results show not only strong measurement invariance (i.e., equal factor loadings and equal indicator thresholds) but also strict measurement invariance (i.e., equal factor loading, equal indicator thresholds, and equal item uniqueness) across the two groups of test takers. This is an indication that there is no substantial differential functioning of the test items across the test-taker groups and that the test items are measured with the same precision in each group. In addition, structural invariance was upheld; that is, the variability and correlational relationships among the factors were the same across both groups. These results provide validity evidence in support of the consistency of the meaning of scores across the two groups.

Interestingly, the correlated four-factor model has been found to represent English-language proficiency of adult learners using the *TOEFL iBT*® test (Manna & Yoo, 2015; Sawaki & Sinharay, 2013), despite the differences in target age groups and assessment instruments. However, unlike adult learners, for whom differential performance across the four latent constructs is not unexpected due to greater variation in background characteristics, the levels of proficiency across the latent constructs in the two groups of adolescent learners studied tended to be similar. One exception was the reading construct, for which the high school students had a higher proficiency level (small effect size). This may be due to older test takers having greater proficiency in L1 reading skills that transfer to L2 reading proficiency (Cummins, 1979; Hasselgreen, 2012; Norris, Davis, & Timpe-Laughlin, 2017). However, these similarities and differences may also be due to how English is taught in Japan at middle school and high school and to other factors, such as test-taker characteristics and sociocultural norms.

Many of the research studies into the nature of L2 proficiency have been based on adult learners. Furthermore, the studies that looked at young learners tended not to include a comprehensive assessment of the four language abilities of reading, listening, speaking, and writing. This study allows us to extend our view of L2 proficiency and the relationship among the four language abilities from adult learners of EFL to adolescent learners of EFL in middle school and high school. However, this study is limited in that test-taker background characteristics that can impact test performance were not included in the analyses. In particular, information on L1 proficiency; years studying English; time in additional study outside of the classroom; and exposure to the target language, such as time in English-speaking countries (Hasselgreen, 2012; Muñoz, 2008; Muñoz & Singleton, 2011), were not available for this study. An additional limitation stems from the proximity of the study sample, drawn from the fall 2015 test administration, to the initial formal introduction of English instruction to fifth- and sixth-grade students in Japan (MEXT, 2008). This limited the full cycle of implementation for the older students and may have contributed to the generally low scores of the full study sample.

As noted earlier, this study was limited to students in Japan by design. Future research is needed to examine the consistency of results for adolescent EFL learners in other countries. In addition, research is needed to investigate the relationship of other English learner background characteristics and performance on the TJC as well as other tests designed for young and adolescent learners. These characteristics include ability level, L1, and instructional context variables including type, intensity, and duration of instruction. Longitudinal studies may also be of interest to examine the stability of results across developmental trajectories and to inform the efficacy of curriculum implementation.

References

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

Bachman, L. F. (1991). What does language testing have to offer? *TESOL Quarterly, 25*, 671–704. https://doi.org/10.2307/3587082

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing, 17*, 1–42. https://doi.org/10.1177/026553220001700101

Bachman, L. F., Davidson, F., Ryan, K., & Choi, I.-C. (1995). *An investigation into the comparability of two tests of English as a foreign language: The Cambridge–TOEFL comparability study*. New York, NY: Cambridge University Press.

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

Bae, J., & Bachman, L. (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. https://doi.org/10.1177/026553229801500304

Bagozzi, R. P., & Yi, Y. (1988). On the evaluation of structural equation models. *Journal of the Academy of Marketing Science*, *16*, 74–94. https://doi.org/10.1007/BF02723327

Bailey, A. L. (2007). Introduction: Teaching and assessing students learning English in school. In A. L. Bailey (Ed.), *The language demands of school: Putting academic English to the test* (pp. 1–26). New Haven, CT: Yale University Press.

Berninger, V. (2000). Development of language by hand and its connections with language by ear, mouth, and eye. *Topics in Language Disorders, 20*, 65–84. https://doi.org/10.1097/00011363-200020040-00007

Bozorgian, H. (2012). Listening skill requires a further look into second/foreign language learning. *International Scholarly Research Notices (ISRN) Education, 2012*, Article 810129.

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

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, NY: 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.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*, 233–255. https://doi.org/10.1207/S15328007SEM0902_5

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

Cronbach, L. J. (1951). Coefficient alpha and internal structure of tests. *Psychometrika, 16*, 297–334. https://doi.org/10.1007/BF02310555

Cummins, J. (1979). Cognitive/academic language proficiency, linguistic interdependence, the optimum age question and some other matters. *Working Papers on Bilingualism*, *19*, 121–129.

Cummins, J. (1981). Age on arrival and immigrant second language learning in Canada: A reassessment. *Applied Linguistics, 11*, 132–149. https://doi.org/10.1093/applin/II.2.132

DeKeyser, R. M. (2013). Age effects in second language learning: Stepping stones toward better understanding. *Language Learning, 63*, 52–67. https://doi.org/10.1111/j.1467-9922.2012.00737.x

Edelenbos, P., & Kubanek, A. (2009). Early foreign language learning: Published research, good practice and main principles. In M. Nikolov (Ed.), *The age factor and early language learning* (pp. 39–58). Berlin, Germany: Mouton de Gruyter.

Educational Testing Service. (2015). *Handbook for the TOEFL Junior Comprehensive test*. Princeton, NJ: Educational Testing Service.

Gradman, H., & Hanania, E. (1991). Language learning background factors and ESL proficiency. *Modern Language Journal, 75*, 39–51. https://doi.org/10.1111/j.1540-4781.1991.tb01081.x

Gu, L. (2014). At the interface between language testing and second language acquisition: Language ability and context of learning. *Language Testing*, *31*, 111–133. https://doi.org/10.1177/0265532212469177

Gu, L. (2015). Language ability of young English language learners: Definition, configuration, and implications. *Language Testing, 32*, 21–38. https://doi.org/10.1177/0265532214542670

Hakuta, K., Butler, Y. G., & Witt, D. (2000). *How long does it take English learners to attain proficiency*? (Policy Report No. 2000-1). Berkeley, CA: University of California Linguistic Minority Research Institute.

Hale, G. A., Rock, D. A., & Jirele, T. (1989). *Confirmatory factor analysis of the Test of English as a Foreign Language* (TOEFL Research Report No. 32). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1982.tb01327.x

Hancock, G. R. (2001). Effect size, power, and sample size determination for structured means modeling and MIMIC approaches to between-groups hypothesis testing of means on a single latent construct. *Psychometrika, 66*, 373–388. https://doi.org/10.1007/BF02294440

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

Hasselgreen, A. (2012). Adapting the CEFR for the classroom assessment of young learners' writing. *Canadian Modern Language Review, 69*, 415–435. https://doi.org/10.3138/cmlr.1705.415

Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55. https://doi.org/10.1080/10705519909540118

Inbar-Lourie, O., & Shohamy, E. (2009). Assessing young language learners: What is the construct? In M. Nikolov (Ed.), *The age factor and early language learning* (pp. 83–96). Berlin, Germany: Mouton de Gruyter.

Japan Ministry of Education, Culture, Sport, Science, and Technology. (n.d.). *Principles guide Japan*. Retrieved from http://www.mext.go.jp/en/policy/education/overview/index.htm

Japan Ministry of Education, Culture, Sports, Science, and Technology. (2008). *Shougakou gakushu shidou yoryo* [Course of study for elementary school education – Chapter 4 Foreign language activities]. Retrieved from http://www.mext.go.jp/component/english/__icsFiles/afieldfile/2011/03/17/1303755_011.pdf

Joliffe, I. T. (1986). Principal component analysis and factor analysis. In I. T. Joliffe (Ed.), *Principal component analysis* (pp. 115 – 128). New York, NY: Springer.

Kunnan, A. J. (1995). *Test taker characteristics and test performance: A structural equation modeling approach*. Cambridge, England: Cambridge University Press.

Kunnan, A. J. (1998). Approach to validation in language assessment. In A. J. Kunnan (Ed.), *Validation in language assessment* (pp. 1 – 16). Mahwah, NJ: Lawrence Erlbaum Associates.

Lado, R. (1961). *Language testing: The construction and use of foreign language tests*. New York, NY: McGraw-Hill.

Lichtman, K. (2016). Age and learning environment: Are children implicit second language learners? *Journal of Child Language, 43*, 707 – 730. https://doi.org/10.1017/S0305000915000598

Manna, V., & Yoo, H. (2015). *Investigating the relationship between test-taker background characteristics and test performance in a heterogeneous English-as-a-second-language test population: A factor analytic approach* (Research Report No. RR-15-25). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12072

Muñoz, C. (2008). Age-related differences in foreign language learning: Revisiting the empirical evidence. *International Review of Applied Linguistics, 46*, 197 – 220.

Muñoz, C., & Singleton, D. (2011). A critical review of age-related research on L2 ultimate attainment. *Language Teaching*, *44*, 1 – 35. https://doi.org/10.1017/S0261444810000327

Muthén, L., & Muthén, B. (2015). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén and Muthén.

Norris, M. J., Davis, J., & Timpe-Laughlin, V. (2017). *Second language educational experiences for adult learners*. New York, NY: Routledge.

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

Oller, J. W., Jr., & Hinofotis, F. B. (1980). Two mutually exclusive hypotheses about second language ability: Indivisible or partially divisible competence. In J. W. Oller Jr. & K. Perkins (Eds.), *Research in language testing* (pp. 13 – 23). Rowley, MA: Newbury House.

Osterhout, L., McLaughlin, J., Pitkänen, I., Frenck-Mestre, C., & Molinaro, N. (2006). Novice learners, longitudinal designs, and event-related potentials: A means for exploring the neurocognition of second language processing. *Language Learning, 56*, 199 – 230. https://doi.org/10.1111/j.1467-9922.2006.00361.x

Powers, D. E. (1982). Selecting samples for testing the hypothesis of divisible versus unitary completence in language proficiency. *Language Learning, 32*, 331 – 335. https://doi.org/10.1111/j.1467-1770.1982.tb00975.x

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*, 507 – 514. https://doi.org/10.1007/BF02296192

Sawaki, Y., & Sinharay, S. (2013). *Investigating the value of section scores for the TOEFL iBT test* (TOEFL iBT Research Report No. 21). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2013.tb02342.x

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

Scholz, G., Hendricks, D., Spurling, R., Johnson, M., & Vandenburg, L. (1980). Is language ability divisible or unitary? A factor analysis of twenty-two English proficiency tests. In J. W. Oller Jr. & K. Perkins (Eds.), *Research in language testing* (pp. 24 – 33). Rowley, MA: Newbury House.

Shanahan, T. (2006). Relations among oral language, reading and writing development. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 171 – 183). New York, NY: Guildford Press.

Stricker, L. J., & Rock, D. A. (2008). *Factor structure of the TOEFL Internet-based test across subgroups* (TOEFL iBT Research Report No. 07). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2008.tb02152.x

Vollmer, H. J., & Sang, F. (1983). Competing hypotheses about second language ability: A plea for caution. In J. W. Oller Jr. (Ed.), *Issues in language testing research* (pp. 29 – 79). Rowley, MA: Newbury House.

Wang, L. (2012). *TOEFL Junior Comprehensive test pilot study* (Unpublished manuscript).

Widaman, K. F. (2007). Common factors versus components: Principals and principles, errors and misconceptions. In R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at 100* (pp. 177 – 203). Mahwah, NJ: Lawrence Erlbaum Associates.

Wilson, K. (2000). *An exploratory dimensionality assessment of the TOEIC test* (Research Report No. RR-00-14). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2000.tb01837.x

Appendix A

Standardized Parameter Estimates for the Correlated Four-Factor Model (Sample A)

Item	Reading estimate (error)	Listening estimate (error)	Speaking estimate (error)	Writing estimate (error)
1	.50 (.75)	.60 (.64)	.74 (.45)	.71 (.50)
2	.38 (.86)	.30 (.91)	.75 (.44)	.63 (.60)
3	.55 (.70)	.54 (.71)	.75 (.44)	.73 (.47)
4	.54 (.71)	.48 (.77)	.81 (.34)	.79 (.38)
5	.23 (.95)	.52 (.73)		.81 (.35)
6	.39 (.85)	.43 (.81)		
7	.57 (.67)	.77 (.40)		
8	.56 (.69)	.45 (.79)		
9	.49 (.76)	.46 (.78)		
10	.64 (.59)	.31 (.91)		
11	.58 (.66)	.14 (.98)		
12	.66 (.56)	.63 (.61)		
13	.72 (.48)	.47 (.78)		
14	.41 (.83)	.55 (.70)		
15	.19 (.96)	.41 (.83)		
16	.39 (.85)	.41 (.83)		
17	.35 (.88)	.48 (.77)		
18	.46 (.79)	.57 (.67)		
19	.58 (.66)	.42 (.83)		
20	.60 (.64)	.63 (.61)		
21	.53 (.72)	.61 (.63)		
22	.47 (.78)	.42 (.83)		
23	.58 (.67)	.65 (.58)		
24	.61 (.63)	.53 (.72)		
25	.56 (.69)	.67 (.55)		
26	.64 (.59)	.56 (.69)		
27	.47 (.78)	.25 (.94)		
28	.33 (.89)	.48 (.77)		

Note. Error value tabled is the standard measurement error representing the percentage of variance in the indicator not explained by the target latent factor.

Appendix B

Standardized Parameter Estimates for the Correlated Four-Factor Model (Sample B)

Item	Reading estimate (error)	Listening estimate (error)	Speaking estimate (error)	Writing estimate (error)
1	.47 (.78)	.56 (.69)	.73 (.47)	.68 (.54)
2	.48 (.77)	.32 (.90)	.76 (.42)	.61 (.62)
3	.48 (.77)	.49 (.77)	.73 (.47)	.75 (.43)
4	.55 (.70)	.46 (.79)	.81 (.34)	.78 (.39)
5	.25 (.94)	.50 (.75)		.80 (.36)
6	.44 (.81)	.41 (.83)		
7	.58 (.66)	.72 (.49)		
8	.53 (.72)	.50 (.75)		
9	.52 (.73)	.47 (.78)		
10	.63 (.60)	.28 (.92)		
11	.57 (.67)	.18 (.97)		
12	.64 (.59)	.66 (.56)		
13	.72 (.48)	.45 (.80)		
14	.33 (.89)	.57 (.68)		
15	.25 (.94)	.46 (.79)		
16	.37 (.87)	.48 (.77)		
17	.36 (.87)	.52 (.73)		
18	.39 (.85)	.62 (.62)		
19	.58 (.67)	.39 (.85)		
20	.56 (.69)	.60 (.64)		
21	.47 (.78)	.56 (.68)		
22	.42 (.82)	.39 (.85)		
23	.56 (.68)	.67 (.56)		
24	.56 (.69)	.50 (.76)		
25	.53 (.72)	.66 (.57)		
26	.60 (.65)	.55 (.70)		
27	.43 (.82)	.23 (.95)		
28	.30 (.91)	.45 (.80)		

Note. Error value tabled is the standard measurement error representing the percentage of variance in the indicator not explained by the target latent factor.

Appendix C

Standardized Parameter Estimates for the Correlated Four-Factor Model From Middle and High School Groups (Sample A)

	Middle school				High school			
Item	Reading estimate (error)	Listening estimate (error)	Speaking estimate (error)	Writing estimate (error)	Reading estimate (error)	Listening estimate (error)	Speaking estimate (error)	Writing estimate (error)
1	.53 (.72)	.58 (.67)	.76 (.42)	.76 (.43)	.47 (.78)	.64 (.59)	.72 (.48)	.66 (.57)
2	.48 (.77)	.27 (.93)	.78 (.39)	.63 (.61)	.32 (.90)	.35 (.88)	.70 (.51)	.63 (.60)
3	.50 (.75)	.56 (.69)	.78 (.40)	.79 (.38)	.61 (.63)	.52 (.73)	.72 (.49)	.66 (.57)
4	.52 (.73)	.44 (.81)	.85 (.28)	.83 (.31)	.57 (.68)	.55 (.70)	.75 (.44)	.74 (.45)
5	.20 (.96)	.53 (.72)		.82 (.34)	.27 (.93)	.51 (.74)		.80 (.37)
6	.43 (.81)	.47 (.78)			.34 (.89)	.40 (.84)		
7	.59 (.65)	.76 (.42)			.55 (.70)	.79 (.38)		
8	.48 (.77)	.52 (.73)			.64 (.59)	.38 (.86)		
9	.49 (.76)	.47 (.78)			.48 (.77)	.46 (.79)		
10	.63 (.60)	.36 (.87)			.65 (.58)	.25 (.94)		
11	.53 (.72)	.20 (.96)			.65 (.58)	.09 (.99)		
12	.65 (.58)	.64 (.60)			.68 (.54)	.63 (.61)		
13	.74 (.46)	.48 (.77)			.70 (.51)	.48 (.77)		
14	.42 (.82)	.52 (.73)			.39 (.85)	.60 (.65)		
15	.16 (.98)	.48 (.77)			.23 (.95)	.34 (.89)		
16	.41 (.83)	.45 (.80)			.36 (.87)	.37 (.86)		
17	.36 (.87)	.51 (.74)			.33 (.89)	.45 (.80)		
18	.47 (.78)	.55 (.70)			.45 (.80)	.60 (.64)		
19	.56 (.69)	.42 (.83)			.60 (.64)	.43 (.82)		
20	.58 (.66)	.59 (.65)			.62 (.61)	.68 (.54)		
21	.48 (.78)	.61 (.63)			.59 (.65)	.62 (.62)		
22	.44 (.81)	.44 (.81)			.50 (.75)	.39 (.84)		
23	.58 (.67)	.68 (.54)			.58 (.67)	.61 (.63)		
24	.59 (.66)	.51 (.74)			.62 (.61)	.56 (.69)		
25	.60 (.64)	.68 (.53)			.51 (.75)	.66 (.57)		
26	.66 (.57)	.58 (.67)			.61 (.63)	.53 (.72)		
27	.43 (.81)	.29 (.92)			.51 (.74)	.21 (.95)		
28	.32 (.90)	.49 (.76)			.34 (.88)	.46 (.78)		

Note. Error value tabled is the standard measurement error representing the percentage of variance in the indicator not explained by the target latent factor.

Appendix D

Standardized Parameter Estimates for the Correlated Four-Factor Model From Middle and High School Groups (Sample B)

	Middle school				High school			
Item	Reading estimate (error)	Listening estimate (error)	Speaking estimate (error)	Writing estimate (error)	Reading estimate (error)	Listening estimate (error)	Speaking estimate (error)	Writing estimate (error)
1	.48 (.77)	.58 (.66)	.76 (.43)	.74 (.46)	.46 (.79)	.55 (.70)	.71 (.50)	.61 (.63)
2	.47 (.78)	.31 (.90)	.78 (.39)	.65 (.57)	.51 (.74)	.34 (.89)	.72 (.48)	.57 (.68)
3	.48 (.77)	.55 (.70)	.77 (.40)	.78 (.39)	.47 (.78)	.41 (.83)	.66 (.56)	.72 (.49)
4	.55 (.70)	.43 (.81)	.84 (.30)	.81 (.34)	.56 (.69)	.50 (.75)	.78 (.39)	.74 (.45)
5	.26 (.93)	.58 (.67)		.83 (.31)	.24 (.94)	.41 (.83)		.77 (.41)
6	.47 (.78)	.47 (.78)			.41 (.84)	.34 (.88)		
7	.60 (.64)	.77 (.41)			.57 (.68)	.66 (.56)		
8	.51 (.74)	.54 (.70)			.55 (.70)	.45 (.80)		
9	.52 (.73)	.49 (.76)			.52 (.73)	.44 (.80)		
10	.67 (.55)	.36 (.87)			.60 (.65)	.18 (.97)		
11	.57 (.67)	.18 (.97)			.57 (.67)	.19 (.97)		
12	.67 (.56)	.65 (.58)			.62 (.62)	.68 (.54)		
13	.73 (.47)	.47 (.78)			.70 (.50)	.43 (.82)		
14	.39 (.85)	.55 (.70)			.26 (.93)	.60 (.65)		
15	.19 (.96)	.48 (.77)			.32 (.90)	.43 (.81)		
16	.42 (.83)	.43 (.82)			.31 (.90)	.55 (.70)		
17	.35 (.88)	.56 (.69)			.38 (.86)	.48 (.77)		
18	.42 (.82)	.60 (.64)			.34 (.88)	.64 (.60)		
19	.57 (.68)	.40 (.84)			.59 (.66)	.39 (.85)		
20	.55 (.70)	.60 (.64)			.56 (.69)	.60 (.64)		
21	.46 (.79)	.59 (.66)			.48 (.77)	.55 (.70)		
22	.45 (.80)	.38 (.85)			.40 (.84)	.40 (.84)		
23	.57 (.68)	.67 (.55)			.56 (.68)	.67 (.55)		
24	.56 (.68)	.48 (.77)			.55 (.70)	.52 (.73)		
25	.56 (.69)	.67 (.55)			.50 (.75)	.65 (.58)		
26	.65 (.57)	.59 (.66)			.53 (.72)	.51 (.74)		
27	.42 (.82)	.29 (.92)			.44 (.81)	.16 (.98)		
28	.35 (.88)	.47 (.78)			.26 (.93)	.42 (.82)		

Note. Error value tabled is the standard measurement error representing the percentage of variance in the indicator not explained by the target latent factor.

Suggested citation:

Manna, V., Yoo, H., & Monfils, L. (2018). *Evaluating invariance in test performance for adolescent learners of English as a foreign language* (Research Report No. RR-18-21). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12208

Action Editor: Donald Powers

Reviewers: Jaime Cid and Maria Elena Oliveri