Listening. Learning. Leading.®

Research Report
ETS RR–15-16

Examining the Internal Structure of the Test of English-for-Teaching (*TEFT*™)

Lin Gu

Sultan Turkan

Pablo Garcia Gomez

June 2015

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Examining the Internal Structure of the Test of English-for-Teaching (*TEFT*™)

Lin Gu, Sultan Turkan, & Pablo Garcia Gomez

Educational Testing Service, Princeton, NJ

ELTeach is an online professional development program developed by Educational Testing Service (ETS) in collaboration with National Geographic Learning. The ELTeach program consists of two courses: English-for-Teaching and Professional Knowledge for English Language Teaching (ELT). Each course includes a coordinated assessment leading to a score report and certificate for teachers of English as a foreign language (EFL). The Test of English-for-Teaching (*TEFT*™), the assessment component of the English-for-Teaching course, measures EFL teachers' command of English for teaching English in classroom settings, as presented in the course. In this study, we examined the internal structure of the TEFT assessment. Results of the analyses demonstrated the role of both skill and content in representing the test's internal structure. The final parcel model had a higher-order general factor and four first-order factors corresponding to reading, writing, listening, and speaking. The findings support the current score reporting practice, that is, to report a total scaled score along with score information on language skills and on language use in specific content areas.

Keywords English-for-teaching; internal structure; dimensionality; validity; multitrait – multimethod; confirmatory factor analysis

doi:10.1002/ets2.12060

Demand for English language teaching (ELT) has been increasing given the growing numbers of English language learners around the globe. One notable trend in recent years is to introduce English as a school subject earlier in the national school curricula in the English as a foreign language (EFL) context (Butler, 2004; Nunan, 2003). This expansion of ELT, especially at elementary and secondary school levels, has created a great need for training and professional development for EFL teachers of school-age English language learners. Teachers are expected to have the knowledge and skills needed to help learners achieve a good command of English. This expectation, in turn, has increased the need for assessments of ELT.

In responding to this global need, ETS collaborated with National Geographic Learning to develop the ELTeach program, an online professional development program consisting of two courses: English-for-Teaching and Professional Knowledge for ELT. Each course includes a coordinated assessment leading to a score report and certificate for individual teachers.

The assessment component of the English-for-Teaching course is the Test of English-for-Teaching (*TEFT*™). This test measures EFL teachers' control over functional English in classroom settings as presented in the English-for-Teaching course. Specifically, it assesses the essential English language skills a teacher needs to be able to prepare and enact the lesson in a standardized (usually national) curriculum in English in a way that is recognizable and understandable to another speaker of the English language (Young, Freeman, Hauck, Garcia Gomez, & Papageorgiou, 2014).

As the TEFT assessment is a new test, it is critical to investigate its internal structure to inform and validate inferences and scoring decisions made on the basis of test performance. A test's internal structure (or dimensionality) refers to the latent factor structure that underlies observed test performance. The internal structure of a test summarizes the patterns of responses by specifying the nature and number of underlying factors as well as the relationships among them.

Investigating a test's internal structure has been proposed as integral to building a validity argument with regard to score interpretation and use (Bachman, 2005; Chapelle, Enright, & Jamieson, 2008). A validity argument can be made based on the extent to which test takers' responses to test items are consistent with the way in which scores are reported. In this study, we examined the internal structure of the TEFT assessment to evaluate the extent to which the intended score-based interpretation and use for the test can be warranted. To be more specific, we investigated the extent to which the dimensionality of the TEFT assessment is compatible with the test's score reporting practice.

Corresponding author: L. Gu, E-mail: LGu001@ets.org

Table 1 Summary of Test Items by Skill and by Content Area

Skill	Managing the classroom	Understanding/communicating lesson content	Providing feedback	Total
Reading	0	21	12	33
Writing	2	10	9	21
Listening	4	20	6	30
Speaking	9	9	3	21
Total	15	60	30	105

Design of Study

Sample

Data from 1,307 test takers who took one form of the TEFT assessment during the pilot administration of the ELTeach program in the fall of 2012 were used in this study (for more details on the pilot testing, see Freeman, Katz, Le Dréan, Burns, & Hauck, 2013). Study participants were from 10 countries in Asia, Europe, and Latin America, with the two highest percentages from China (about 40%) and Italy (about 25%). Almost half of the participants (45%) indicated that they had studied English for an extensive period of time (9 years or more). About two-thirds of the participants reported that they were teaching either at a primary or a middle school grade at the time of testing. Among the participants who were teaching at the time of testing (about 74% of the total sample), the vast majority (about 83%) was public school teachers.

The Test of English-for-Teaching (TEFT) Assessment

The development of the TEFT assessment followed a language for specific purposes (LSP) approach (Young et al., 2014). One of the main features of LSP tests is the specificity of the target language use (TLU) domain (Douglas, 2000). Identifying the characteristics of the TLU domain for a specific purpose lays the foundation on which language use can be contextualized. In the context of ELT, the TLU domain consists of language use situations that test takers will most likely encounter in a real-world context for the purpose of teaching English. To define the TLU domain, an LSP approach focuses on tasks that are typical of the TLU situations and on detailed analysis of language and skills needed to perform these tasks (Douglas, 2000; Dudley-Evans & St John, 1998; Hutchinson & Waters, 1987).

For the development of the TEFT assessment, a thorough list of tasks typically performed by ELT teachers in preparation for lessons and while teaching lessons was created on the basis of input from a global panel of experts as well as detailed analysis of curricula, textbooks, and classroom video recordings from various regions around the world. The design of the test organized the teacher tasks into three broad content areas based on the functional use of language in classroom settings: managing the classroom, understanding and communicating lesson content, and providing feedback. The tasks were also categorized into four groups by the primary language skill elicited, namely, reading, writing, listening, and speaking. This task analysis served as the foundation for both the English-for-Teaching course and the TEFT assessment. In the test, therefore, test takers are required to demonstrate their command of English for (a) engaging with students in simple, predictable classroom exchanges; (b) understanding content for students and tasks for the teacher as included in instructional materials and presenting lessons in class based on a defined curriculum and instructional materials; and (c) providing basic oral and written feedback to students (see Young et al., 2014, for a complete description of the TEFT design and development process).

The test form used in this study had a total of 105 teacher tasks as test items organized into two sections: Section A, Preparing for Lessons, and Section B, Teaching Lessons. Section A contained reading, writing, and listening items, and Section B had speaking, listening, and writing items. There were 33 reading tasks, 21 writing tasks, 30 listening tasks, and 21 speaking tasks. The listening and reading tasks were of selected response format (e.g., multiple choice) and were each scored dichotomously (0 or 1). The writing and speaking tasks required constructed responses and were each scored on a three-point scale. The Cronbach's α was 0.952 for the entire test, 0.830 for reading, 0.905 for writing, 0.752 for listening, and 0.907 for speaking. Each task was associated with a specific content area. The numbers of tasks focusing on each of the three content areas were 15, 60, and 30 respectively. Table 1 summarizes the number of test items by skill and by content area.

With regard to score reporting, the most prominent information on the score report is the total scaled score and the associated band and band descriptor. The total score, ranging from 400 to 700, is intended as an indicator of a test taker's overall command of the requisite English to teach English in English. In addition, scaled scores are reported to convey a test taker's ability to execute each of the four skills in the context of ELT, namely reading, writing, listening, and speaking. The scaled score for each skill ranges from 40 to 70. Also reported is the information on a test taker's command of English in the three content areas: managing the classroom, understanding and communicating lesson content, and providing feedback. For each content area, the number of score points earned out of the total available and the percentile information are reported.

Analyses

A multitrait–multimethod (MTMM) confirmatory factor analysis approach was taken to examine the influences of both skill and content on test performance. Proposed by Widaman (1985), the original purpose of such an approach is to investigate the influences of trait and test method (e.g., multiple-choice questions, short-answer questions, etc.) on test performance. We adopted this methodology to examine the roles of both skill and content factors in representing the internal structure of the test.

The MTMM approach consisted of two sequential steps. The first step was to establish a baseline model that has both skill and content factors. In this model, each item loads on one skill factor and one content factor. In light of the test's score reporting practice, two series of model testing were conducted to identify plausible baseline models.

The TEFT assessment reports a separate score for each skill. We therefore tested three competing models that have latent skill factors (Figure 1) to examine the role of skills in explaining the test's internal structure. The first was a correlated four-factor model in which items in each skill section load on their respective skill factors, namely reading, writing, listening, and speaking, and the four skills are correlated with one another. The second was a higher-order model that consists of four first-order factors corresponding to the four skills and a higher-order general factor (G). In this model, items within each skill section load on their respective skill factors, and the four skills are independent conditional on a higher-order G factor; that is, the correlations among the skills are explained by G. The third was a bifactor model. In this model, each item loads on its respective skill factors and on G. This model imposes orthogonal relations between G and each of the skills while the skills are allowed to correlate with each other.

The test also reports score information for each content area. In light of that, we further hypothesized three competing models that have latent content factors (Figure 2) to examine the role of content in representing the test's internal structure. In the first model, the internal structure is summarized by three correlated content factors, namely, managing the classroom (M), understanding and communicating lesson content (U), and providing feedback (P). The second was a higher-order model in which three content-specific first-order factors are subsumed under a common underlying dimension, G. The third was a bifactor model in which each item loads on its respective content factors and on G. The content factors are correlated with each other and are independent of G.

On the basis of the results of the above model testing, plausible models that have both skill and content factors were identified and compared in search of the best-fitting model as the baseline model.

In the next step, the baseline model with both skill and content factors was compared to two models — one had only skill factors and the other only content factors — to investigate the unique influences of skill and content on test performance.

Analyses were conducted by using Mplus (Muthén & Muthén, 2010). The dataset contained both binary and ordinal variables at the item level. Finney and DiStefano (2013) suggested treating ordered categorical data with five categories or less as categorical instead of continuous and using robust diagonally weighted least squares (DWLS) estimators to adjust the parameter estimates, standard errors, and fit indices for the categorical nature of the data. We therefore treated all item-level variables as categorical and used the WLSMV estimator, a DWLS estimator provided by Mplus.

The adequacy and appropriateness of the models were evaluated based on three criteria: (a) values of selected global model fit indices, (b) individual parameter estimates, and (c) the principle of parsimony. The following DWLS-based fit indices were used for assessing model fit at the global level: chi-square (χ^2), comparative fit index (CFI), and root mean square error of approximation (RMSEA). A significant χ^2 value ($p < 0.01$) signals a poor model fit, although this value should be interpreted with caution because it is highly sensitive to sample size. A CFI value larger than 0.9 indicates an adequate model fit (Hu & Bentler, 1999). RMSEA values smaller than 0.05 can be interpreted as a sign of close model fit while values between 0.05 and 0.08 indicate adequate fit (Browne & Cudeck, 1993). Individual parameter estimates were

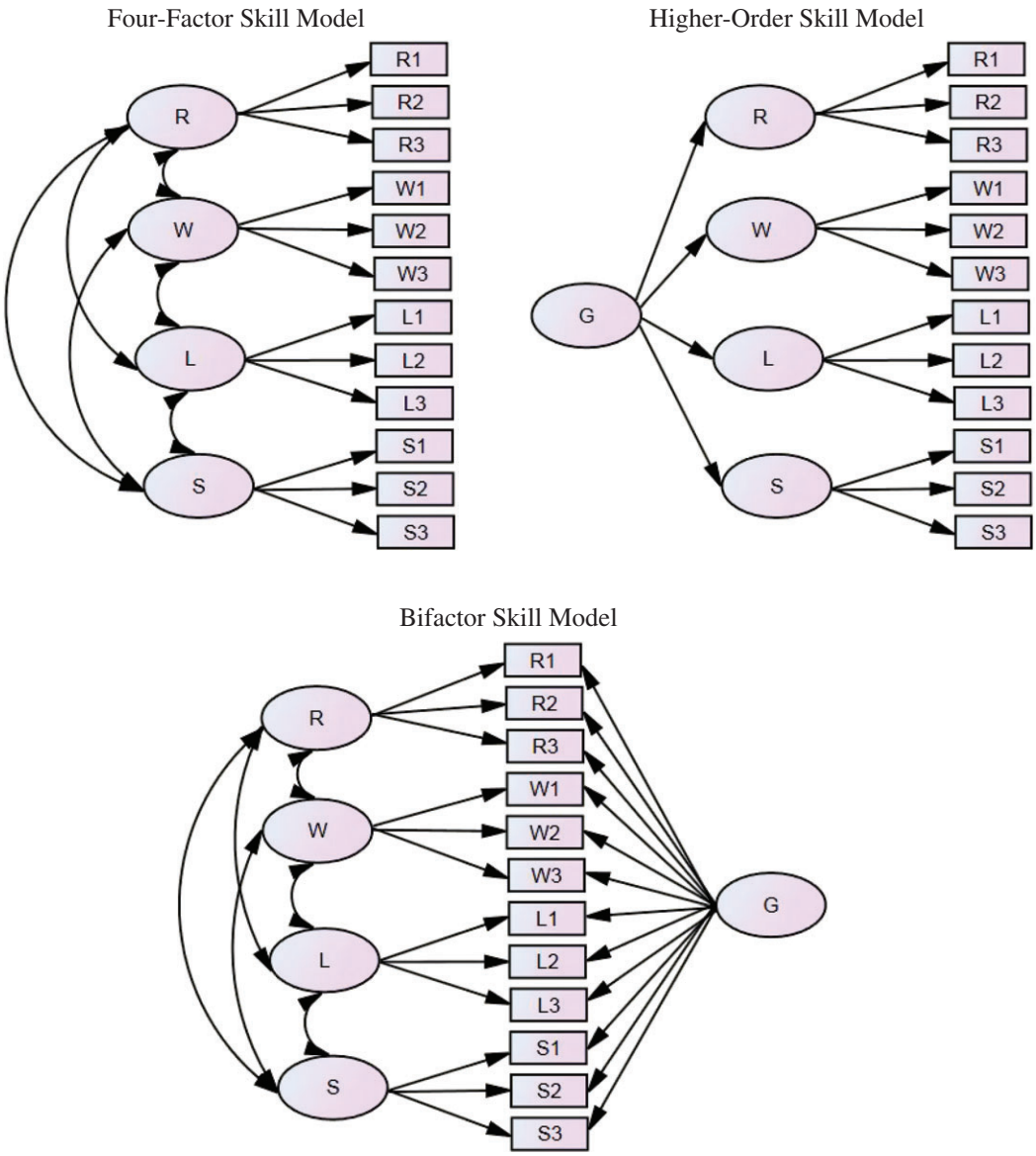

Figure 1 Schematic representations of the skill models. R = reading; W = writing; L = listening; S = speaking; G = general factor.

also examined for appropriateness and significance. Previous researchers (Sawaki, Stricker, & Oranje, 2009; Stricker & Rock, 2008) considered a correlation of 0.9 or greater to indicate extremely high interfactor correlations. This criterion was adopted to screen out models with high inter-factor dependency. The principle of parsimony, which favors a simpler model over a more complicated one when two models fit equivalently, was invoked when choosing between competing models with similar fits. When comparing models, we evaluated the significance of chi-square difference ($\Delta\chi^2$) and change in CFI (ΔCFI) to determine which model provided the best fit to the data, taking into consideration model parsimony. A nonsignificant $\Delta\chi^2$ test result suggests the equivalence of model fit. Model equivalence is also indicated by a ΔCFI less than or equal to 0.01 (Cheung & Rensvold, 2002). If two models fit equivalently, the simpler model should be chosen based on the principle of parsimony.

Results

The Multitrait–Multimethod (MTMM) Analysis

The results of testing the three competing skill models and the three competing content models are summarized in Table 2.

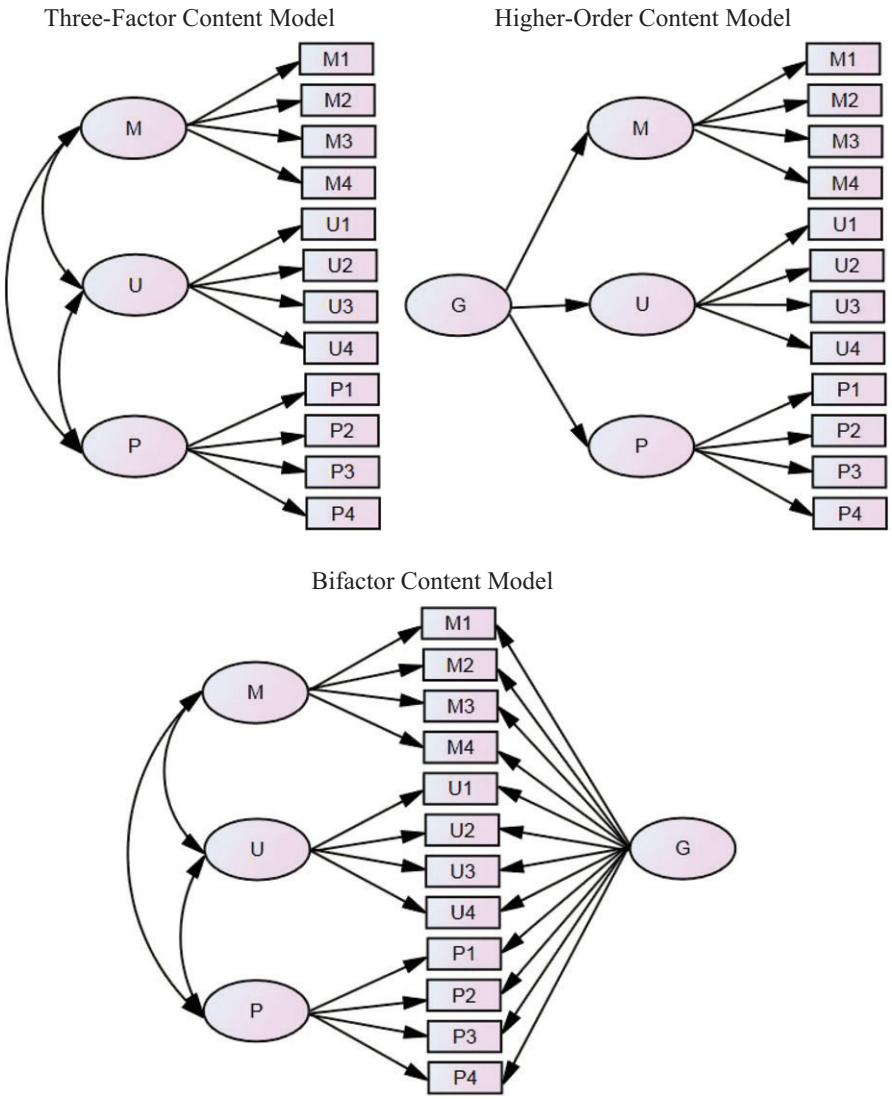

Figure 2 Schematic representations of the content models. M = managing the classroom; U = understanding and communicating lesson content; P = providing feedback; G = general factor.

With regard to the skill models, the bifactor model did not converge and was, consequently, discarded. The higher-order model and the four-factor model both had adequate fit at the global level. The correlations among the skills in the four-factor model were all moderate (0.752 – 0.800), indicating that the skills were distinct. Both the four-factor model and the higher-order model provided plausible factorial solutions in capturing the relationships among the skills.

The results of testing the content models showed that, at the global level, all three models exhibited adequate fit to the data. We further inspected the individual parameter estimates. A negative residual variance was found with the higher-order model, which rendered the model inadmissible. In the bifactor model, with the presence of a general factor, most of the loadings on the content-specific factors were insignificant. This model was therefore discarded. We also found that in the three-factor model the correlation between two content factors, M and U, reached 0.953, casting doubt on the distinctness of these two factors.

We subsequently decided to test an additional model, a correlated two-factor model, in which a common dimension underlies the items previously associated with the M and U factors, and this factor is correlated with a distinct P factor. The results (see Table 2) showed that this model fitted the data well, with the two factors correlated at 0.828.

Although the three-factor content model had a factor correlation larger than the cut-off value, we decided to keep this model in the following analysis because this model corresponds to the design of the test tasks and the practice of

Table 2 Results of Testing Skill and Content Models

Model	χ^2	df	Parameters	CFI	RMSEA
Four-factor skill model	9917.328	5349	384	0.915	0.026 (0.025 – 0.026)
Higher-order skill model	9915.883	5351	382	0.915	0.026 (0.025 – 0.026)
Bifactor skill model	No convergence				
Three-factor content model	10618.339	5352	381	0.902	0.027 (0.027 – 0.028)
Higher-order content model	10618.339	5352	381	0.902	0.027 (0.027 – 0.028)
Bifactor content model	8236.484	5247	486	0.944	0.021 (0.020 – 0.022)
Two-factor content model	10665.279	5354	379	0.901	0.028 (0.027 – 0.028)

Note: CFI = comparative fit index; RMSEA = root mean square error of approximation.

reporting three content scores. We hoped that through additional analysis, further evidence regarding the distinctiveness of the content factors could be investigated.

To summarize, two plausible ways of modeling the skill relationships emerged from the previous analysis: the higher-order model and the four-factor model. Regarding the content-specific factors, we decided to model their relationships based on either the three-factor model or the two-factor model. We therefore hypothesized four plausible baseline models, as shown in Figure 3.

Table 3 reports the results of testing the global fit of these four competing models. The only model that converged successfully was the model with higher-order skill relationships and three content factors. This model fitted very well at the global level, and none of the correlations among the content factors exceeded 0.9. The correlation between the M and U factors reduced from 0.953 in the three-factor content model to 0.751 in this model, suggesting that when modeling simultaneously with the skill factors, the M and P factors appeared to be distinct. One potential reason could be that any commonalities shared by the observed variables associated with the M and P factors were absorbed by the skill structure. This model was chosen as the baseline model for the subsequent analysis.

To examine the unique influence of the skills, we compared the baseline model to the three-factor content model previously tested in which the configuration of the content factors was the same as the one in the baseline model but no skill factors were included. The baseline model performed significantly better than the three-factor content model ($\Delta\chi^2$ (Δdf) = 2531.132 (109); $p = 0.000$; ΔCFI = 0.055), providing evidence that the influence of the skills on test performance was not negligible.

To examine the unique influence of the content factors, the baseline model was compared to the higher-order skill model previously tested in which the configuration of the skill factors was identical to the one in the baseline model but no content factors were specified. The baseline model fit significantly better than the higher-order skill model ($\Delta\chi^2$ (Δdf) = 2015.672 (108); $p = 0.000$; ΔCFI = 0.042), demonstrating the influence of the content factors on test performance.

At the global level, the baseline model satisfied all the criteria to be considered a well-fit model. However, not all loadings on the content factors were significant, whereas the loadings on the skill factors were all significant. Comparing the two standardized loadings of each item, one on the skill factor and the other on the content factor, we found that the majority of the items, 92 out of 105, had a stronger relationship with its skill factor than with its content factor, suggesting that the performance on the test items was more strongly influenced by the skill factors than the content factors.

In sum, the final MTMM model with both skill and content factors fitted significantly better than the skill-only and content-only models, indicating the impact of both skill and content on test performance. The MTMM results also suggested that skill played a more prominent role in accounting for test performance than content did.

Testing Models With Parcel Scores

The outcome of the MTMM analysis intimated the possibility of modeling skill and content at different levels. We noticed that when skills and content factors were modeled simultaneously, the vast majority of the items had a stronger tie to their respective skill factor than to the content factor. We speculated that skill and content might affect test performance at different levels. We therefore decided to experiment with models based on parcel scores to capture the test's internal structure.

In testing the parcel-level models, we hypothesized that skills were represented at the latent level as factors, whereas content areas were represented at the observed level as indicators of the latent skill factors. In other words, latent skill

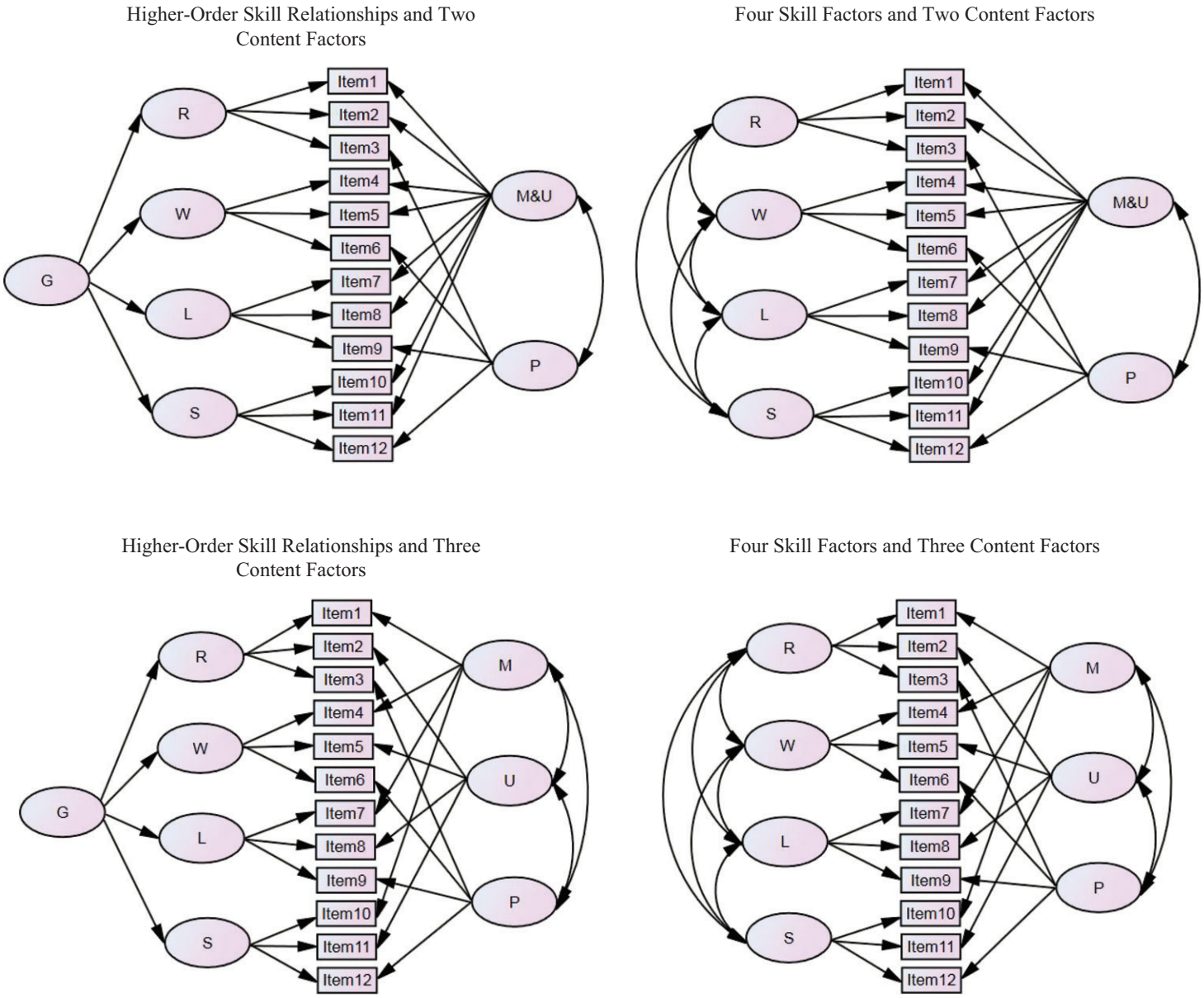

Figure 3 Schematic representations of plausible multitrait–multimethod models. G = general factor; R = reading; W = writing; L = listening; S = speaking; M = managing the classroom; U = understanding and communicating lesson content; P = providing feedback.

factors were indicated by content-based parcel scores. A parcel score is the sum of scores of the items that share commonality in terms of content-specific language use within a skill domain. Two steps were taken to create parcels by the researchers. First, tasks within each skill domain were grouped into the three broad content areas: managing the classroom, understanding and communicating lesson content, and providing feedback. Second, within each content area, further classification was attempted to categorize test tasks by the nature and mode of linguistic input and output. What follows illustrates how parcel scores were generated.

Reading Parcels

Three types of reading tasks were identified based on which three reading parcel scores were generated. Two are designed for the purpose of understanding and communicating lesson content. One type requires teachers to comprehend the

Table 3 Results of Testing Multitrait–Multimethod (MTMM) Models

Model	χ^2	d*f*	Parameters	CFI	RMSEA
Higher-order skill relationships and two content factors	No convergence				
Four skill factors and two content factors	No convergence				
Higher-order skill relationships and three content factors	7559.626	5243	490	0.957	0.018 (0.017–0.019)
Four skill factors and three content factors	No convergence				

Note: CFI = comparative fit index; RMSEA = root mean square error of approximation.

kind of language used in students' textbooks, such as lesson goals, instructions, explanations, exemplifications, exercises, and answer keys. The other type requires teachers to comprehend different types of reading texts used in the students' textbooks. Lastly, for the purpose of providing feedback, teachers are required to comprehend and evaluate the accuracy of students' written output based on textbook or workbook activities.

Writing Parcels

Writing items were classified into four groups, and therefore four writing parcel scores were obtained. For the purpose of managing the classroom, one task type requires teachers to write announcements, assign homework, give test and quiz instructions, and so on. We identified two kinds of teacher tasks for the purpose of understanding and communicating lesson content: One asks teachers to write instructions or explanations, and the other requires teachers to transcribe student output based on lesson content. The fourth category includes tasks that ask teachers to provide written feedback on students' output or correct errors in students' work.

Listening Parcels

Four types of teacher tasks were indentified in the listening domain. As a result, four listening parcel scores were formed. One type is for teachers to comprehend students' output (e.g., questions about classroom activities) for the purpose of managing the classroom. Similar to reading, we identified two kinds of tasks teachers typically perform for the purpose of understanding and communicating lesson content. One requires teachers to comprehend the kind of spoken language used in textbook audio materials that helps navigate through the textbook content, including instructions, explanations, exemplifications, exercise, and answer keys. The other requires teachers to comprehend the spoken texts used in students' audio materials. For the purpose of providing feedback, teachers are asked to comprehend and evaluate the accuracy of a student's oral output.

Speaking Parcels

Six types of teacher tasks were identified in the speaking domain, giving rise to six speaking parcel scores. Two were identified for the purpose of managing the classroom: One requires teachers to read aloud within-the-curriculum materials (e.g., homework assignments) intelligibly, and the other asks teachers to produce formulaic language (e.g., language used for taking attendance). For the purpose of understanding and communicating lesson content, we identified three types of teacher tasks. They ask teachers to read aloud, produce formulaic language, and repeat aloud content based on within-the-curriculum materials. The last task type identified asks teachers to orally provide feedback on student output.

Test performance pertaining to specific types of teacher tasks was summarized as parcel scores, giving rise to 17 parcel scores. Table 4 summarizes the number of teacher tasks associated with each parcel score.

In the following model testing, we used parcel scores as the level of measure. The relationships among the parcel scores were modeled by using latent skill factors. We hypothesized four competing parcel-based models: a unidimensional model, a higher-order model, a correlated four-factor model, and a bifactor model (Figure 4). We tested these models to select the one that best represented the relationships among the parcel variables.

Table 4 Summary of Parcel Scores

Skill domain	Parcels	No. of tasks
Reading	Parcel 1	12
	Parcel 2	9
	Parcel 3	12
Writing	Parcel 1	2
	Parcel 2	8
	Parcel 3	2
	Parcel 4	9
Listening	Parcel 1	4
	Parcel 2	3
	Parcel 3	17
	Parcel 4	6
Speaking	Parcel 1	3
	Parcel 2	6
	Parcel 3	3
	Parcel 4	4
	Parcel 5	2
	Parcel 6	3

The parcel scores were treated as continuous variables because, except for two listening parcels, all parcel variables have more than five ordered categories. Descriptive statistics (Table 5) showed that a couple of variables have a skewness value larger than two, indicating that distributions of these variables deviate from univariate normality. A corrected normal theory estimation method, the Satorra-Bentler estimation (Satorra & Bentler, 1994), was employed by using the MLM estimator in Mplus to correct global fit indices and standard errors for non-normality. Standardized root mean square residual (SRMR) is commonly used as a model-data fit criterion when a normal theory estimation method is employed, and it was therefore included in the analysis in addition to the criteria for testing the item-level models earlier. An SRMR value of 0.08 or below is commonly considered as a sign of acceptable fit (Hu & Bentler, 1999).

The results of model testing are reported in Table 6. At the global level, both the higher-order model and the four-factor model fitted the data well. However, several factor correlations in the four-factor model exceeded the 0.9 cut-off value. We consequently decided to adopt the higher-order model as the one that best represented the relationships among the parcel scores.

The final parcel model consists of four first-order skill factors and assumes the presence of a common underlying dimension across the four modalities.

Discussion

In this study, we investigated the internal structure of the TEFT assessment, a newly developed test of EFL teachers' command of English for teaching English in classroom settings. The results can be used to evaluate the extent to which the intended score interpretation and use can be warranted. Establishing consistency between a test's internal structure and the score reporting practice provides validity evidence in support of the intended score interpretation and use.

Generally speaking, the results supported the current score reporting practice, that is, to report a total scaled score along with score information on skills and on language use in specific content areas.

We found that the influences of both skill and content factors on test performance were present in the final parcel-based model, with the former shown at the latent level and the latter manifested at the observed level. At the latent level, this model demonstrates that the ability construct measured by the test is predominantly skill oriented and hierarchical in nature, which lends support to the reporting of the total score along with separate scores for each skill. Because test items were grouped by content to generate parcel scores at the observed level, the arrival at this model also substantiates the argument that content plays a role in capturing the test's internal structure and defends the practice of reporting content scores.

The finding of this higher-order skill relationship is consistent with the consensus reached on the multicomponent nature of language proficiency by applied linguists and language testers. Language ability has been portrayed as being

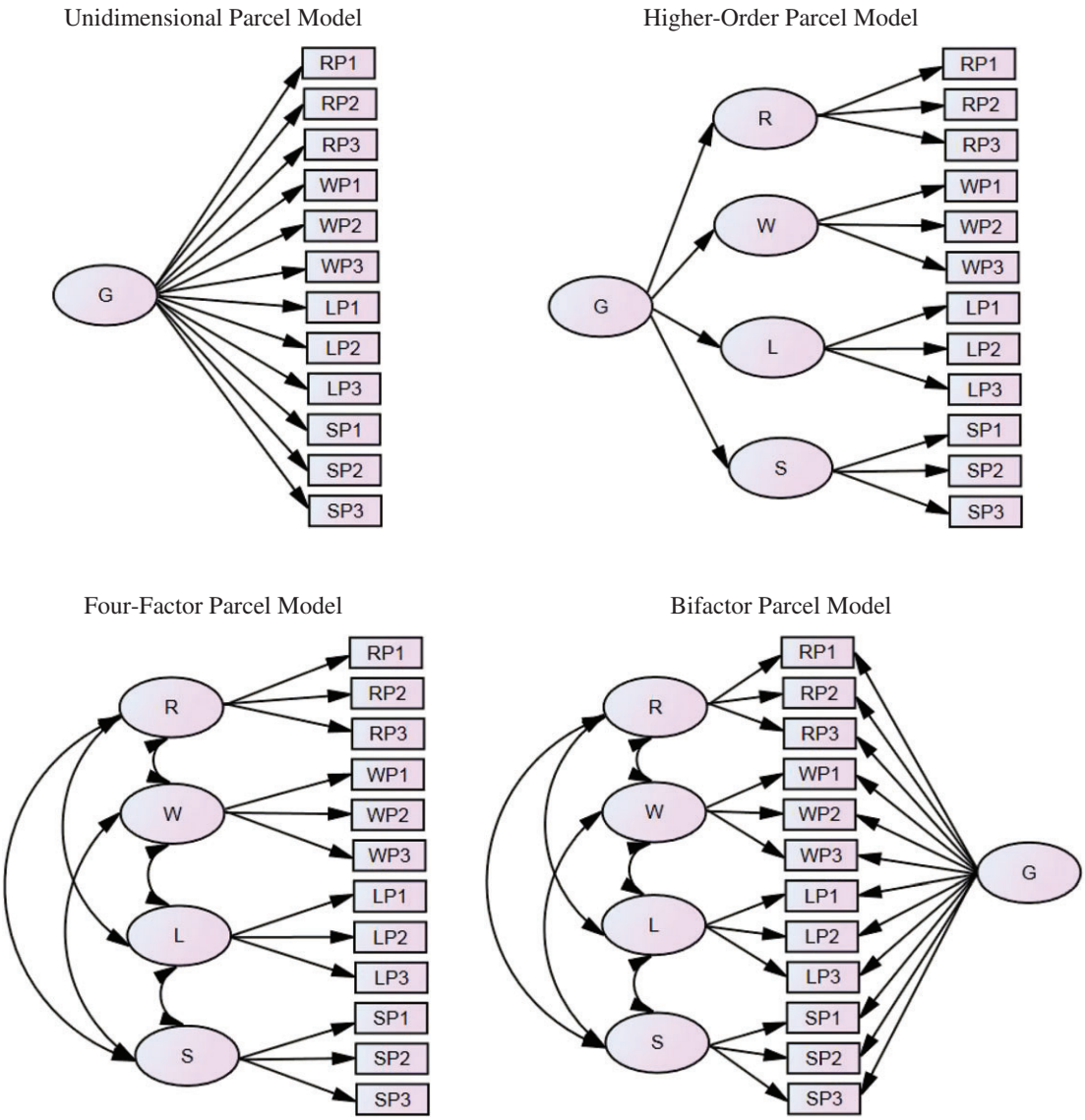

Figure 4 Schematic representations of parcel models. G = general factor; R = reading; W = writing; L = listening; S = speaking; RP = reading parcel; WP = writing parcel; LP = listening parcel; SP = speaking parcel.

hierarchical in nature by previous researchers. Fouly, Bachman, and Cziko (1990) tested the divisibility of language ability and concluded that both distinct skill factors and a general language factor existed. Bachman, Davidson, Ryan, and Choi (1995) found that the higher-order model with first-order factor best represented the construct measured by two large-scale test batteries. The higher-order model was also adopted by Stricker and Rock (2008) and Sawaki et al. (2009) in the context of *TOEFL iBT*® testing.

The finding that the first-order factors correspond to the four skills in our final model also corroborates Carroll's (1965) four-skills approach to conceptualize language skills. Carroll proposed to distinguish an integrated approach from a discrete structure-point approach to language testing. When an integrated approach is taken, the total communicative effect of an utterance is emphasized instead of specific points of structure or lexicon. Naturally, the four skills of listening, reading, speaking, and writing, which are regarded as integrated performance based on a learner's mastery of the whole array of language components, receive focus in this integrated approach to testing. He asserted that an ideal language proficiency test should make it possible to differentiate levels of performance in those integrated skill dimensions of performance.

A few limitations need to be pointed out. We acknowledge that the data came from a pilot test instead of an operational test, and hence the test takers may not have been as motivated as regular test takers and may not have been completely

Table 5 Descriptive Statistics for Parcel Variables

Variable	Range	Mean	*SD*	Kurtosis	Skewness
Reading Parcel 1 (RP1)	0 – 12	8.819	2.549	0.030	−0.801
Reading Parcel 2 (RP2)	0 – 9	8.012	1.198	5.545	−1.972
Reading Parcel 3 (RP3)	0 – 12	10.936	1.854	11.243	−3.034
Writing Parcel 1 (WP1)	0 – 6	6.080	2.874	−1.008	−0.173
Writing Parcel 2 (WP2)	0 – 24	25.652	7.569	−0.532	−0.532
Writing Parcel 3 (WP3)	0 – 6	5.911	2.845	−1.216	−0.027
Writing Parcel 4 (WP4)	0 – 27	35.357	10.302	0.579	−1.167
Listening Parcel 1 (LP1)	0 – 4	3.634	0.646	4.113	−1.929
Listening Parcel 2 (LP2)	0 – 3	2.741	0.607	7.106	−2.647
Listening Parcel 3 (LP3)	0 – 17	16.328	1.291	22.593	−3.759
Listening Parcel 4 (LP4)	0 – 6	4.949	1.149	1.395	−1.248
Speaking Parcel 1 (SP1)	0 – 9	12.544	2.616	3.283	−1.587
Speaking Parcel 2 (SP2)	0 – 18	21.286	5.957	−0.039	−0.682
Speaking Parcel 3 (SP3)	0 – 9	12.122	2.692	1.125	−1.123
Speaking Parcel 4 (SP4)	0 – 12	16.101	3.626	1.440	−1.230
Speaking Parcel 5 (SP5)	0 – 6	8.016	2.263	0.736	−1.195
Speaking Parcel 6 (SP6)	0 – 9	12.200	3.248	0.839	−1.266

Table 6 Results of Testing Parcel Models

Model	χ^2	d*f*	Parameters	CFI	RMSEA	SRMR
Unidimensional parcel model	1000.254	119	51	0.896	0.075 (0.071 – 0.080)	0.044
Higher-order parcel model	718.935	115	55	0.929	0.063 (0.059 – 0.068)	0.039
Four-factor parcel model	697.948	113	57	0.931	0.063 (0.058 – 0.067)	0.038
Bifactor parcel model	No convergence					

Note: CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual.

representative in their backgrounds. Furthermore, for the item-level analyses conducted in the study, the sample size was relatively small, and therefore, the results needed to be interpreted with caution.

In conclusion, our findings are consistent with the practice of reporting a total score along with skill and content scores, which contributes validity evidence in support of the intended score interpretation and use.

References

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, *2*, 1 – 34.

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.

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

Butler, Y. G. (2004). What level of English proficiency do elementary school teachers need to attain to teach EFL? Case studies from Korea, Taiwan, and Japan. *TESOL Quarterly*, *38*, 245 – 278.

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.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). Test score interpretation and use. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 1 – 25). New York, NY: Routledge.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes of testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*, 233 – 255.

Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge, UK: Cambridge University Press.

Dudley-Evans, T., & St John, M. J. (1998). *Developments in English for specific purposes*. Cambridge, UK: Cambridge University Press.

Finney, S. J., & DiStefano, C. (2013). Nonnormal and categorical data in structural equation models. In G. R. Hancock & R. O. Mueller (Eds.), S*tructural equation modeling: A second course* (2nd ed., pp. 439 – 492). Charlotte, NC: Information Age.

Fouly, K. A., Bachman, L. F., & Cziko, G. A. (1990). The divisibility of language competence: A confirmatory approach. *Language Learning*, *40*, 1 – 21.

Freeman, D., Katz, A., Le Dréan, L., Burns, A., & Hauck, M. (2013). *ELTeach global pilot report 2012*. Retrieved from http://elteach.com/ELTeach/media/Documents/ELTeach_GPR_9-20-13.pdf

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

Hutchinson, T., & Waters, A. (1987). *English for specific purposes: A learning-centered approach*. Cambridge, UK: Cambridge University Press.

Muthén, L. K., & Muthén, B. O. (2010). Mplus (6th ed.) [Computer software]. Los Angeles, CA: Muthén & Muthén.

Nunan, D. (2003). The impact of English as a global language on educational policies and practices in the Asia-Pacific region. *TESOL Quarterly*, *37*, 589 – 613.

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

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

Stricker, L. J., & Rock, D. A. (2008). *Factor structure of the TOEFL Internet-Based Test across subgroups* (TOEFL iBT Research Report 07). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2008.tb02152.x

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement*, *9*, 1 – 26.

Young, J. W., Freeman, D., Hauck, M. C., Garcia Gomez, P., & Papageorgiou, S. (2014). *A design framework for the ELTeach program assessments* (Research Report No. RR-14-36). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12036

Suggested citation:

Gu, L., Turkan, S., & Garcia Gomez, P. (2015). *Examining the internal structure of the Test of English-for-Teaching (TEFT™)* (ETS Research Report No. RR-15-16). Princeton, NJ: Educational Testing Service. doi: 10.1002/ets2.12060

Action Editor: Heather Buzick

Reviewers: Elizabeth Stone and Spiros Papageorgiou