GTEC

2020

A Validity Investigation on the Speaking and Writing Sections of GTEC

SECOND LANGUAGE TESTING INC.

PAYMAN VAFAEE, PH.D.

Contents

List of Tables

List of Figures

Introduction

To demonstrate accountability, test providers should have transparency about their test content and quality so that stakeholders can interpret the meaning of the scores accurately and use them appropriately (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Benesse Corporation (BC) develops and administers the Global Test of English Communication (GTEC). GTEC is used to assess the English communicative abilities of junior high and high school students in Japan, and its scores are used for either formative purposes (e.g., diagnosing the EFL learners' strengths and weaknesses) or summative purposes (e.g., university admissions purposes). To show accountability to the public, BC continually carries out validation studies on GTEC to provide the stakeholders with convincing evidence in support of the usefulness of GTEC scores for the intended purposes.

Among a series of other validation studies on GTEC, the current paper reports the results of an empirical investigation that aimed at providing support for the plausibility of the underlying assumptions of the warrants that license several inferences in the validity argument that is being developed for the usefulness of GTEC scores. The current study focused on the speaking and writing sections of GTEC Advanced version (see below for information about different versions of GTEC) and provides backing for the plausibility of the assumptions that underly the warrants that authorize the inferences of evaluation, generalization, and explanation[1].

[1] An explanation of the validation framework and inferences listed here will be provided in the later sections of the current paper.

In the current paper, first, more information about GTEC is in order. Then, the validation framework within which the current study was completed will be laid out. After that, the aim of the study is elaborated, and then the research questions are introduced. Following the research questions, the details of the study, including the analyses and results, are presented. Finally, the answers to the research questions are discussed. This discussion is meant to contribute to the validity argument for the usefulness of GTEC scores for the intended purposes.

Global Test of English Communication

GTEC comes in four versions, or difficulty levels: Core, Basic, Advanced and CBT. GTEC assesses the level of reading, listening, writing, and speaking abilities among EFL learners in Japan. GTEC includes test-task types and content that simulate real-life communicative tasks and their content in the intended target language use (TLU) domain, and its scores are aligned with the Common European Framework of Reference (CEFR). As seen in Table 1, the scores of each of these four GTEC versions are computed on the common scale with different maximum score.

Table 1

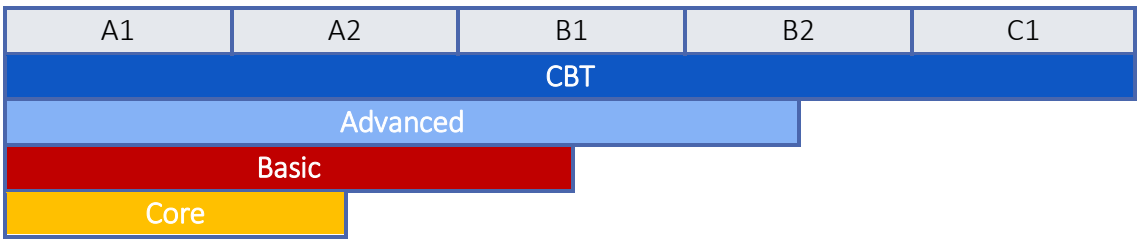
Scoring Scales for the Four Versions of GTEC

	Core	Basic	Advanced	CBT
Maximum Score Per Skill	210	270	320	350
Maximum Total Score	840	1080	1280	1400

The scores of the four versions of GTEC are mapped onto the CEFR scale, and Table 2 shows the correspondence between the levels of the four versions of GTEC and the levels of CEFR.

Table 2

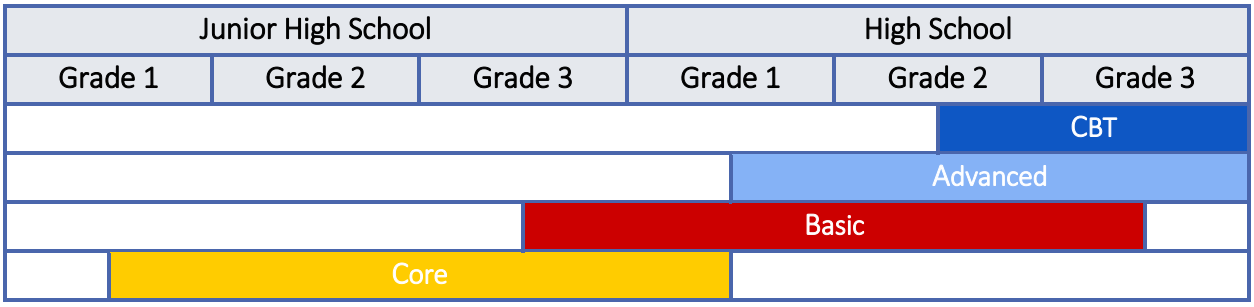
Correspondence between CEFR and GTEC



Theoretically, Japanese junior high school and high school students of any school grade can take any of the four GTEC versions. However, the most appropriate version of the test for any given learning environment should be chosen based on the students' level of English proficiency and the needs of their teachers and other stakeholders. Table 3 illustrates the recommended versions of GTEC for junior high school and high school grades.

Table 3

Recommended GTEC Versions for Different School Grades in Japan



The Speaking Section of GTEC Advanced version

The speaking section, which is delivered on a tablet, includes a total of eight questions/items that are presented in four parts. Table 4 summarizes the information about these items. Part A consists of two monologues. For each monologue, test takers look at the monologue for 30 seconds and then read it aloud within 40 seconds. These two items measure test takers' ability in Fluency and Pronunciation, and for each item, the score ranges from zero to three.

Table 4

Overview of Speaking Section of GTEC Advanced version

Part		Item	Goal Achievement	Language Use	Fluency and Pronunciation
A	Monologue Read-aloud	1			Score 1: 0, 1, 2, 3
		2			Score 2: 0, 1, 2, 3
B	Visual-based Listening and Responding	1	Score 3: 0, 1		
		2	Score 4: 0, 1		
		3	Score 5: 0, 1		
		4	Score 6: 0, 1		
C	Four-picture Storytelling	1	Score 7: 0, 1	Score 11: 0, 1, 2, 3, 4	Score 12: 0, 1, 2, 3
			Score 8: 0, 1		
			Score 9: 0, 1		
			Score 10: 0, 1		
D	Expressing Your Opinion	1	Score 13: 0, 1	Score 15: 0, 1, 2, 3, 4	Score 16: 0, 1, 2, 3
			Score 14: 0, 1, 2		

In Part B, there are two visual-based Listening and Responding tasks with two items/questions each. For these items, test takers look at some visuals (e.g., a schedule) for ten seconds, and then listen to a question about the visuals. After that, they have 15 seconds to respond to the question. The performance of test takers on these items is evaluated in terms of Goal Achievement. If they answer each question by communicating a meaningful message, they are awarded a score of one for each item. If they don't, they will get a score of zero.

Part C includes a four-picture storytelling task that prompts test takers to look at an illustrated story for 30 seconds and then tell a one-minute story based on the four sequenced pictures. For this task, in addition to Fluency and Pronunciation, as well as Goal Achievement for

each picture, the performance of test takers is scored from zero to four for Language Use, which entails the range of vocabulary and grammar, the level of control over their use, as well as how well they show the relationship between ideas.

Finally, in Part D, test takers are asked to read a prompt and then to address it. In their response, in addition to their opinion, they should provide at least one reason to support their opinion. For this task, test takers are given one minute to prepare and one minute to record their responses. For Goal Achievement, test takers get a score of zero to one for expressing their opinion, and a score of zero to two for providing a reason/reasons to support their opinion. Like Part C, for Fluency and Pronunciation and Language Use, they get a score of zero to three and a score of zero to four, respectively.

In short, as seen in Table 4, test takers receive a total of 16 different scores for their performance in four speaking parts. These 16 scores fall in different ranges (i.e., 0–1, 0–2, 0–3, 0–4) for the three components (i.e., Goal Achievement, Language Use, and Fluency and Pronunciation) of the speaking construct.

The Writing Section of GTEC Advanced version

The writing section has two parts. In Part A, test takers are given five minutes to read an email and compose a response to two questions included in the email. Their responses to the questions should be at the sentence level, rather than the single-word level. This part of the test is intended to measure written interactional skills in the context of exchanging emails. Part A of the writing section is scored for Task Achievement, Style, and Accuracy. In this part, and for each of the two questions, for Task Achievement, test takers get a score of zero or one for

providing adequate responses to the questions. For Style, which pertains to the naturalness of the email, test takers receive a score of zero to three. For Accuracy, which assesses the level of control over the accuracy of vocabulary and grammar, the score range is also between zero and three. Table 5 summarizes the information about Part A of the writing section.

Table 5

Overview of Part A of Writing Section of GTEC Advanced version

Part		Item	Task Achievement	Style	Accuracy
A	Reading and Responding to an Email	1	Score 1: 0,1	Score 3: 0, 1, 2, 3	Score 4: 0, 1, 2, 3
		2	Score 2: 0,1		

In Part B, the task prompt asks test takers to write an argumentative essay to express their opinion about a topic. They have 20 minutes to complete this task. The goal of this part is to measure test takers' ability to clearly communicate their opinion, as well as arguments and supporting reasons, on a given topic. The performance of test takers is rated based on four criteria: Goal Achievement (a score of zero or one), Vocabulary, Grammar, and Organization. For the last three criteria, the scores range from zero to four with half-point increments. Table 6 provides a summary of the information about Part B of the writing section.

Table 6

Overview of Part B of Writing Section of GTEC Advanced version

Part		Item	Goal Achievement	Vocabulary	Grammar	Organization
B	Expressing an Opinion	1	Score 5: 0, 1	Score 7: 0, .5, 1, 1.5, 2, 2.5, 3, 3.5, 4	Score 8: 0, .5, 1, 1.5, 2, 2.5, 3, 3.5, 4	Score 9: 0, .5, 1, 1.5, 2, 2.5, 3, 3.5, 4
	Supporting the Opinion	2	Score 6: 0, 1			

To sum up, as seen in Tables 5 and 6, the performance of each test taker in the writing section receives nine different scores, which fall in different ranges (i.e., 0–1, 0–3, 0–4) for the seven components (i.e., Task Achievement, Style, Accuracy, Goal Achievement, Vocabulary, Grammar, and Organization).

The Validation Framework

Building on previous developments (e.g., Mislevy, Steinberg & Almond, 2002), Kane (2006) proposed an argument-based framework for validation which has provided test users with a simple-to-understand and easy-to-implement guide to work out validity arguments for the interpretations and uses of test scores. As his approach clearly defines steps to go through in the validation process, it has been labeled a "praxis step" for validation (Chapelle, 2013, p.25).

The aim of Kane's (2006) argument-based validation approach is to provide an overall evaluation of the intended interpretations and uses of test scores. This is achieved through a coherent analysis of all of the evidence for and against the proposed interpretations and uses (Cronbach, 1988; Kane, 2001). In this approach, the validation process entails building two types of arguments in two consecutive steps. In the first step, an interpretive argument is developed by specifying the proposed interpretations and uses of the test scores in detail. Then, a validity argument is formulated by evaluating the overall plausibility of the interpretations and uses outlined in the interpretive argument (Kane, 2012).

Like any other informal argument, an interpretative argument lays out the proposed interpretations and uses of test scores by specifying a network of inferences. In this network, each of the inferences refers to a link or bridge in the interpretive argument that allows test users to move from one inference to the other and finally to the intended interpretation or use of test scores (Chapelle et al., 2011). Kane (2006) defined and used the concept of inference consistent with Toulmin's (2003) description of informal or practical arguments as used in non-mathematical fields like law. In Toulmin's sense, an inference is a logical movement from one fact or proposition to another (Chapelle, 2013). Each inference or informal argument has the

following structure and components: conclusions or interpretations are drawn about, for example, a test's scores. Such conclusions follow from a chain of reasoning that starts from data or any other empirical observation. In Mislevy et al.'s (2002) terms, these conclusions are referred to as claims. Claims are made on the basis of data or observations that Toulmin (2003) referred to as grounds. The link between a claim and a ground is established or authorized by a warrant, which can be a law, generally held principle, or established procedure. The warrant has its own underlying assumptions, which should be plausible. Therefore, backing, which can be a scientific theory or well-established (empirical) evidence, should justify the underlying assumptions of the warrant. Each of the inferences in the interpretative argument has its own underlying assumptions, and each of these assumptions should be supported by backing obtained from empirical data or logical reasoning.

However, even when a warrant is supported with a solid backing, other circumstances may undermine the inferential link between the claim and the ground; therefore, rebutting is required. The rebuttal weakens the inferential link between the claim and ground, and if an inference survives a rebuttal, it creates a strong basis for movement to the next inference in the interpretive argument (Chapelle et al., 2011; Mislevy et al., 2002).

The main goal of the interpretive argument is to make the inferences, warrants, their underlying assumptions, and relevant required backings and rebuttals in the interpretation of test scores as clear as possible. It is the particular intended interpretations or uses of test scores that determine what inferences should be included in an interpretive argument and what kinds of evidence are most relevant in the validation process (Kane, 2006). For this reason, "each

interpretive argument is unique and therefore the associated validity argument will be also unique" (Kane, 2001, p. 330).

As shown in Figure 1, adopted from Chapelle et al. (2010), according to Kane's (2006) framework, and as expanded by Chapelle et al. (2008), an interpretative argument includes the following five inferences: Domain definition, Evaluation, Generalization, Explanation, and Extrapolation.

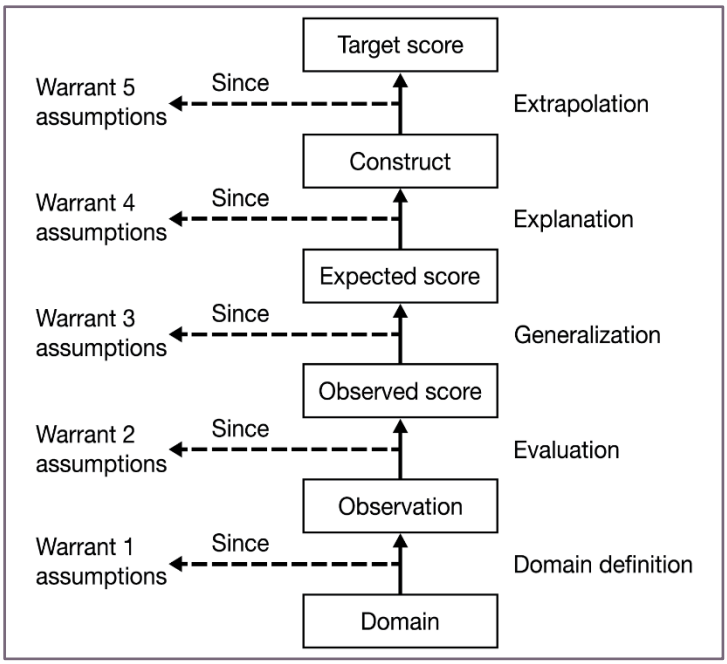


Figure 1. An outline of an interpretive argument, adopted from Chapelle et al. (2010, p. 450)

Each of these inferences is authorized with its own warrant, which has its own underlying assumptions, and each of these assumptions needs its own backing. Also, each of these inferences serves as the ground for the next one in the chain of inferences. This means that if there is not enough evidence, for example, for the plausibility of the assumptions underlying the Generalization inference, continuing validation studies to find support for the Explanation

inference is futile. In other words, the validation process terminates at any point of the interpretative argument if the inference at that point falls apart due to lack of support.

As seen in Figure 1, each interpretive argument should include the following inferences and their underlying assumptions, plus the required evidence, for them to be considered plausible:

Domain Description: This inference assumes that the test tasks elicit the same set of Knowledge, Skills, and Abilities (KSAs) that we expect to observe if learners were performing similar tasks in a real-life Target Language Use (TLU) domain. The evidence for the plausibility of the assumptions underlying the Domain Description inference comes from a rigorous domain analysis in which contexts, scenarios, themes, topics, and the tasks in the TLU domain are identified.

Evaluation: This inference connects the Observed Performance to the Observed Score. It assumes that interpretation is based on scores that are construct-relevant and free from measurement error. The assumption here is that test performance can actually generate observed scores as viable indicators of the given test construct. Evidence supporting the plausibility of the assumptions underlying the evaluation inference originates from the evaluation of conditions under which the test is administered and the care with which the scoring system has been developed and applied. Also, studies like evaluating scoring rubrics and item/task analysis can provide evidence for the plausibility of the assumptions underlying the evaluation inference.

Generalization: This inference connects the Observed Scores to the Expected Scores. This inference assumes that interpretation can reliably be based on test scores that are

interchangeable across items, tasks, raters, and other conditions. Thus, the generalization inference links the observed scores from examinees to their "expected" scores—that is, the scores anticipated to be consistently generated across different measurement conditions. This inference relies on the assumption that if examinees are given test tasks similar to those in the TLU domain, their observed scores on the test tasks should be considered evidence of their expected scores. In other words, scores should be comparable regardless of which test form or equivalent task was administered or which raters made the determinative observations. The typical studies to provide backing for the plausibility of this inference include reliability and/or generalizability studies, and scaling and equating studies.

Explanation: This inference connects the Expected Scores to the Construct-Related Scores. This inference assumes that interpretation of test scores is related to the given construct. The inference links the expected scores to the underlying test construct and is based upon the assumption that patterns in the observed scores on the test reflect the trait(s) or constructs being measured. In other words, the internal structure of a measure is a viable operationalization of a theoretical construct. Therefore, one of the types of studies that can provide backing for the assumptions of the explanation inference would be an examination of the internal structure of the test.

Extrapolation: This inference connects the Construct-Related Scores to Real-Life or Target Scores. This inference assumes that interpretation is based on the connection between the construct-related scores and the real-life performance of test takers.

Extrapolation inference assumes that performances on research instruments are favorable proxies of performance on representative criterion measures drawn from the TLU

domain. Criterion-related validity studies can determine the sturdiness of the assumptions behind this inference.

The Current Study

The current study used Multifaceted Rasch Measurement (MFRM), Confirmatory Factor Analysis (CFA), and CFA-based Multi-trait Multi-method (MTMM) to examine the plausibility of assumptions that underly the authorizing warrants of the Evaluation, Generalization, and Explanation inferences (Chapelle et al. 2011; Kane, 2006). These inferences are the building blocks of the validity argument being built for the speaking and writing sections of GTEC Advanced version. In other words, testing hypotheses related to the plausibility of the assumptions underlying these inferences provides backing for or against the validity argument being built for the use of the scores of GTEC Advanced version speaking and writing sections for the intended purposes.

Table 7 includes the research questions that guided the current study. Each of these questions targeted an aspect of the validity argument and its respective inferences and their underlying assumptions.

Participants and Data

To answer the following research questions, test data were collected from a large group (N= 8974) of Japanese first, second and third grade high school students who took GTEC Advanced version in June 2019. These students took the test in different schools and testing centers all across Japan, and by this token, can be deemed a representative sample of the typical population of students who take GTEC Advanced version. These students took the test for either gauging their progress in English proficiency or university admissions purposes.

The Research Questions

Table 7

The Research Questions of the Current Study

	Research Question	Inference	Data Analysis
1	Did the scores (criteria) have a wide range of difficulty levels?	Evaluation	MFRM
2	Were the examinees divided into distinct ability levels?	Evaluation	MFRM
3	Were the ratings different in severity?	Evaluation	MFRM
4	Were there any systematic patterns of bias in the ratings?	Evaluation	MFRM
5	Did the ratings agree with each other, and were these ratings independent from each other?	Generalization	MFRM
6	Were the criteria applied consistently throughout the test in measuring test examinees' ability level? (Did all the scores (criteria) fit the Rasch model?)	Generalization	MFRM
7	Were the ratings consistent throughout the rating sessions? (Did all the ratings fit the Rasch model?)	Generalization	MFRM
8	Did the writing and speaking sections of GTEC Advanced version have a unitary factor structure?	Explanation	MFRM
9	Did the test data structure correspond to the hypothesized structure of the scoring criteria?	Explanation	CFA
10	To what extent did the independent assessment methods in the speaking and writing sections of GTEC Advanced version diverge in their measurement of different constructs?	Explanation	MTMM
11	To what extent did different assessment methods (i.e., different items in different parts) concur in their measurement of the same construct (e.g., Fluency and Pronunciation in speaking and Accuracy in writing)?	Explanation	MTMM

Rating Rubrics, Rating Training, and Ratings

GTEC raters use analytical rating rubrics to measure different aspects of test takers' speaking and writing abilities. These rubrics are applied to sub-scores for different dimensions

(aspects) of the assessment constructs. Tables 4, 5 and 6 summarize the different dimensions of the two constructs of speaking and writing ability as measured in GTEC Advanced version.

GTEC raters work at several scoring sites in different English-speaking countries. They need to meet pre-determined requirements for English proficiency and demonstrate their ability to score before they can be hired. After these candidates participate in training sessions, in which they evaluate sample answers and actual responses from a pilot field-test, they should pass a scoring-ability test for each administration in order to be selected as official GTEC raters. If they do not demonstrate sufficiently accurate performance in the scoring-ability test, they will not be allowed to attend an operational rating session. During the operational rating session, raters are regularly monitored, and based on their performance, they are offered additional training sessions.

When doing official ratings, raters will be given online access to test responses. For the writing section, the handwritten responses are scanned and sent to the scoring system. For the speaking section, recorded audio responses are sent to the scoring system. For both the writing and speaking sections, each test response is rated blindly by two independent raters. If the two sets of scores do not match, additional review will be required. For example, in the case of the speaking test, a third, more experienced, rater will decide the final score. When finalizing the score, the third rater can refer to the scores given by the first and second raters.

Analyses

As shown for each research question in the current study, two major types of multivariate data analysis techniques, namely MFRM and CFA (including MTMM) were employed to process GTEC data. Here is a brief introduction to these two methods of data analysis.

Multifaceted Rasch Measurement (MFRM)

MFRM, which incorporates multiple facets of the measurement procedure, evolved out of the earlier dichotomous Rasch model (Bond & Fox, 2015). A facet of measurement is an aspect of the measurement procedure that affects test scores (Linacre, 2013). Common examples of these measurement facets include the severity of the raters, the difficulty level of tasks or items (or scores or criteria), and rating scale categories (e.g., Fluency and Pronunciation or Accuracy). One major advantage of MFRM is that it calibrates all measurement estimates on a single equal-interval scale (i.e., the logit scale). This creates a single frame of reference for interpreting the results of the analysis. Also, in MFRM, the facets are estimated concurrently, but the impact of each may still be examined independently.

Importantly, MFRM provides information about how well the performance of each individual test taker, rater, or task matches the expected values as predicted by the Rasch model. Therefore, MFRM can help researchers detect particular misfitting facets of the measurement. In this context, misfitting means that the behavior of a facet of measurement deviates from the predictions of the mathematical model. An example of a misfitting facet could be a rater who is unsystematically inconsistent in applying the ratings, or a task that is unexpectedly difficult, or an examinee whose responses are inconsistent (Lynch & McNamara, 1998).

In MFRM analysis, the fit statistics are also calculated from the item/person residuals and are reflected in Infit and Outfit Mean Square Values, both with an expected value of 1 (Bond & Fox, 2015). Additionally, the MFRM generates estimates of the reliability of separation index and the separation ratio. These indices quantify the amount of variance in the test scores estimated by the Rasch model for the various elements in the specified facet relative to the precision by

which these measures are estimated. The reliability of separation index for each facet ranges from 0 to 1, whereas the separation ratio ranges from 1 to infinity (Linacre, 2013). The interpretation of these two statistics, however, is different for various facets.

Low separation index for the examinee facet indicates lack of variability in the examinees' ability, which might be symptomatic of central tendency errors, meaning that the raters do not distinguish the performance of test takers at different ability levels. Conversely, low values of these two statistics for the rater facet are indicative of an unusually high degree of consistency in the measures for various elements of that facet. Once parameters of the model have been estimated, interaction effects, such as the interaction between raters and rating criteria, or between raters and examinees, can be detected by examining the standardized residuals (i.e., standardized differences between the observed and expected ratings) (Eckes, 2011).

Confirmatory Factor Analysis (CFA)

CFA is a special case of the general family of structural equation modeling (SEM), commonly used in many social sciences—including language assessment—for investigating theory-derived structural relationships or construct validity and hypotheses about a set of measured variables (Mueller & Hancock, 2008).

Conducting CFA entails several consecutive steps: (1) specifying the hypothesized model, (2) identifying and estimating the model's parameters, (3) assessing the data-model fit, (4) making possible and plausible modifications in the hypothesized model, and (5) identifying and testing alternative models that may rebut or weaken the structural inferences made in the hypothesized model.

To evaluate the fit of an SEM model, a profile of model fit tests and indices has recommended by Hu and Bentler (1999) and Mueller and Hancock (2008). This profile includes the Chi square (χ^2) of the model, with its degrees of freedom and p-value. For an acceptable model fit, the χ^2 should not be statistically significant at a .05 level. However, in large samples and complex models, the χ^2 is usually significant and not very informative. For this reason, the following descriptive fit indices were also included in this profile: the root mean square error of approximation (RMSEA< .06), the comparative fit index (CFI> .95), and the Weighted Root Mean Square Residual (WRMR< 1). To compare nested models statistically, the (scaled) chi-square difference ($\Delta\chi 2$) test should be used.

Multi-trait Multi-method (MTMM)

The MTMM framework (Campbell & Fiske, 1959) enables us to explore different validity dimensions of a measurement procedure that entails the measurement of multiple traits (i.e., constructs or criteria) by using multiple methods (e.g., items and tasks). MTMM can result in evidence in favor or against the arguments related to the convergent validity (i.e., the extent to which different assessment methods concur in their measurement of the same trait), discriminant validity (i.e., the extent to which independent assessment methods diverge in their measurement of different traits), and method effects (an extension of the discriminant validity issue and whether method effects represent bias that can derive from the use of the same method in the assessment of different traits; correlations among these traits are typically higher than those measured by different methods). In the current study, the MTMM analysis is carried out within the CFA framework.

Results

The MFRM Procedures and Findings

FACETS 3.82.3 (Linacre, 2019) was used to carry out the MFRM analyses in the current study, which employed a partial credit Rasch measurement model. Since in both the speaking and writing sections of GTEC Advanced version, each item has a set of different scores (or criteria) with different ranges focusing on different dimensions of speaking and writing abilities, the scores (or criteria), not the items or tasks, were used as one of the facets of measurement.

The other facets of the measurement included in these analyses were examinees and ratings. Because the goal of the current study was not to identify problematic raters but to evaluate the overall quality of the two independent ratings, rather than focusing on the individual raters, the ratings were treated as one of the facets of measurement. To review the structure and range of scores (or criteria) for the speaking and writing sections of GTEC Advanced version, consult Tables 4, 5, and 6[2].

The Speaking Section: First, it was examined whether the Rasch model parameters were estimated successfully. The criteria for this examination were a) that the mean of the score residuals and the mean of the standardized residuals be near zero, and b) that the standard deviation of the standardized residuals be near 1.0. In the current study, for the speaking section, the mean of the score residuals and the mean of the standardized residuals were both zero, and the standard deviation of the standardized residuals was 1.02. Therefore, it can be concluded that the parameters of the Rasch model for this section of GTEC were estimated

[2] In a separate analysis, the third score or final score was also included and the conclusions were the same as the ones presented in the current paper.

successfully in the current study, so the rest of the results can be reviewed to answer the research questions.

The results of the MFRM revealed that 68.62 % of the score variance of the speaking section of GTEC Advanced version was explained by Rasch measures, which indicates the unidimensionality of the test structure. According to Engelhard (2013), values of more than 20% attest to the unidimensionality of test data. Therefore, for the speaking section, the answer to research question 8 (i.e., Did the writing and speaking sections of GTEC Advanced version have a unitary factor structure?) is that the scores represent a unitary factor structure, which means this section of the test taps only one single underlying trait or construct, which is speaking ability.

Figure 2 illustrates the distributions of the three facets of measurement (i.e., examinees, scores or criteria, and ratings) on the Wright map or logit scale with a mean of zero, and positive values covering the examinees with higher speaking ability, and negative values with less difficult criteria, and less severe ratings.

The Chi-square test was not statistically significant ($\chi^2 = 8077.8$, $df = 8929$, $p < 0.01$), which shows that the speaking abilities of the examinees were normally distributed, and these learners came from statistically distinct ability groups.

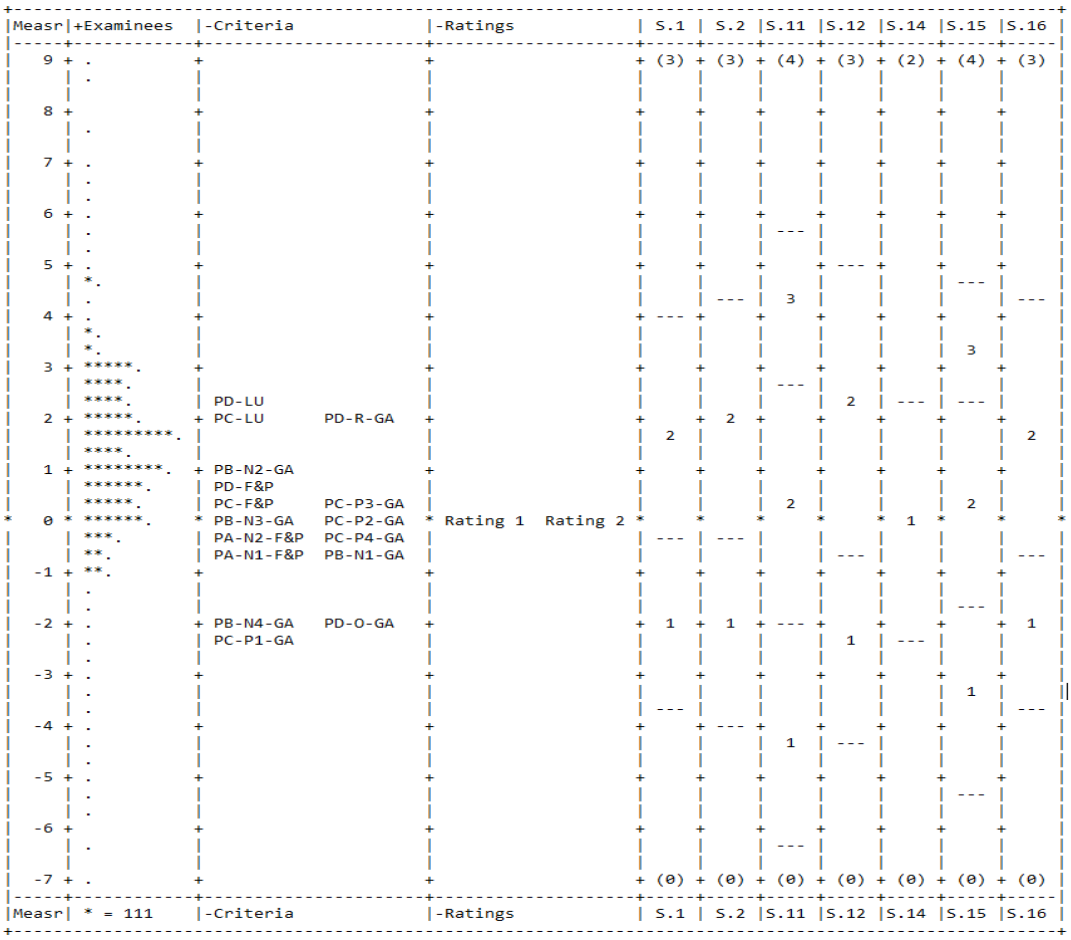


Figure 2. The MFRM Wright Map for the Speaking Section of GTEC Advanced Version

PA: Part A, PB: Part B, PC: Part C, PD: Part D, N: Item Number, P: Picture Number, O: Opinion, R: Reason, F&P: Fluency and Pronunciation, GA: Goal Achievement, LU: Language Use.

As seen in Figure 2 and Table 8, there was a wide spread of examinees' ability with a range from -7.74 to +9.91 logits (Range= 17.65). The mean of the ability level of examinees was 1.22 logits, with a standard deviation (SD) of 1.6.

Table 8

Summary Statistics for the MFRM Analysis on the Speaking Section Scores

Statistics	Examinees	Criteria (scores)	Ratings
Mean	1.22	0	0
Standard Deviation	1.6	1.44	.03
Min to Max	-7.74 to +9.91	-2.49 to 2.28	-.02 to .02
Range	17.65	4.77	.04

Also, as seen in Table 8, the mean of the score (criteria) difficulty level was zero (Min = -2.49, = Max = 2.28, SD= 1.44, Range = 4.77). This indicates that on average, the test fell closer to the easy end of the difficulty continuum in the current study. These results provide answers to research question 1 (i.e., Did the scores (criteria) have a wide range of difficulty levels?) and research question 2 (i.e., Were the examinees divided into distinct ability levels?). Although on average the test was relatively easy for the sample of test takers in the current study (or on average their ability level was higher than most of the criteria in the current study), there were still several distinct levels of difficulty associated with the scores or criteria of the speaking section of GTEC Advanced version, and this section of the test managed to categorize the examinees into several distinct levels of speaking ability.

As shown in Table 8, the results of the MFRM analysis also showed that the mean of the severity of the ratings was zero with a standard deviation of .03. (Min = -.02, = Max = .02, Range = .04). This shows that on average, the two sets of ratings fell within a narrow range of severity level. This provides the answer to research question 3 (i.e., Were the ratings different in severity?). Although the findings did not indicate identical levels of severity among the two sets of ratings, these findings did show that the ratings were not very different, and their severity levels were limited to a narrow range.

Table 9 reports values for the indices of the strata and separation with the reliability estimates associated with it.

Table 9

Summary Statistics for the MFRM Analysis on the Speaking Section Scores

Statistics	Examinees	Criteria (scores)	Ratings
Strata	4.73	86.79	3.99
Separation Index	3.3	64.84	2.74
Separation Reliability	.92	1	.88

As can be seen in the above table, the examinees were divided into a minimum of four distinct levels of speaking ability with a high level of reliability (.92). This further supports the answer to research question 2 and indicates that the test was successful in identifying different groups of test takers with distinct differences in their speaking ability. The statistics reported in Table 9 also provide further support for answers to research questions 1 and 3. As can be seen, the strata and separation indices showed that the score (criteria) had a wide range of difficulty levels with a perfect reliability estimate for their separation. As discussed before, the strata and separation indices also showed that there were about three to four distinct levels of severity in the two sets of ratings with a relatively low level of reliability for their separation (.88). However, as noted before, these differences were minimal and limited to a narrow range of Rasch logits (-.02 to .02).

Table 10 includes infit and outfit statistics for the three facets of measurement included in the speaking section of GTEC Advanced version. As seen in this table, the overall infit and outfit mean squares for all three facets of measurement were within the acceptable range of .5 to 1.5, as well as the more stringent acceptable range of .7 to 1.3 (Eckes, 2005; Linacre, 2013).

Table 10

Summary Statistics for the MFRM Analysis on the Speaking Section Scores

Statistics	Examinees	Criteria (scores)	Ratings
Infit Mean Square	.99	1	1.01
Outfit Mean Square	1.08	1.14	1.16

Additionally, the above criteria ranges for judging the infit and outfit mean squares were used to classify examinees, criteria, and ratings based on their usefulness for measurement purposes. In Table 11, the frequency of the occurrences of the mean squares within different ranges is reported. As can be seen, the vast majority of the fit indices fell within both acceptable ranges, and 83.89% of the examinees had the infit mean square of .5 to 1.5. Also, 100% of the criteria and ratings had the acceptable infit mean squares between .5 and 1.5.

Collectively, these findings about the fit of the facets of the measurement to the Rasch model provide answers to **research question 6** (i.e., Did all the scores (criteria) fit the Rasch model?) and **research question 7** (i.e., Did all the ratings fit the Rasch model?), which is that all of the criteria and ratings fit the Rasch model.

In short, the above findings showed that the criteria were applied consistently throughout the test for measuring test examinees' ability level, and the ratings were consistent throughout the rating sessions.

Table 11

Frequency Count of Infit and Outfit Mean Squares based on Fit Acceptable Ranges

	Infit Mean Square	Outfit Mean Square
Examinees		
<.7	17.89%	32.99%
<.5	6.95%	14.76%
>1.3	16.06%	20.19%
>1.5	9.16%	15.27%
Between .7 to 1.3	66.05%	46.82%
Between .5 to 1.5	83.89%	69.97%
Criteria (scores)		
<.7	0%	0%
<.5	0%	0%
>1.3	0%	31.25%
>1.5	0%	12.5%
Between .7 to 1.3	100%	68.75%
Between .5 to 1.5	100%	87.5%
Ratings		
<.7	0%	0%
<.5	0%	0%
>1.3	0%	0%
>1.5	0%	0%
Between .7 to 1.3	100%	100%
Between .5 to 1.5	100%	100%

For answering research question 5 (i.e., Did the ratings agree with each other, and were these ratings independent from each other?), the percentage of the agreements among the two sets of ratings were computed. The Rasch model expected level of exact agreement was 69.9%; however, the actual exact agreements were 82.3% of the ratings (across the two sets of rating). Using these percentages, the Rasch Kappa was computed (82.3 – 69.9)/(100 – 71) to be .43, and because this value is near zero, it can be concluded that the ratings were independent from each other. Therefore, it can be concluded that although the two sets of ratings were independent from each other, they had a high level of agreement with each other.

Finally, to respond to research question 4 (i.e., Were there any systematic patterns of bias in the ratings?) bias analysis (interaction analysis) between the three facets of measurement was

carried out using the t values of equal to or more than |2.00|. Table 12 summarizes the results of these analyses.

Table 12

Summary of Bias (Interaction) Analysis

	Examinees X Criteria	Examinees X Ratings	Criteria X Ratings
Number of empirical bias terms	142,704	17838	32
Mean for t values	-.07	0	0
SD	1.28	1.01	1.02
Min.	-7.85	-2.88	-5.38
Max.	7.33	2.68	2.36
Number of t values equal to or higher than |2.00|	8872	71	8
Total percentage of all bias terms	6.2%	.39%	25%

As seen in the table above, there were only .39% of bias patterns between ratings and examinees, which means that a few raters evaluated certain test takers more strictly or more leniently than expected. Also, there were 25% of bias patterns between ratings and the evaluation criteria, which means that some raters used certain criteria more strictly or more leniently. Finally, there were only 6.2% of bias patterns between examinees and criteria. This means that some test takers were evaluated more strictly or leniently when certain criteria were used. Although the percentage of the bias patterns among ratings and evaluation criteria were relatively high, this is not an unusual finding in the context of large-scale tests with a large number of interaction terms. For example, for this type of interaction pattern, Eckes (2005) and Koizumi, Okabe, and Kashimada (2017) reported 37% and 25.78%, respectively.

To wrap up, the answer to **research question 4** would be that the amount of bias related to the interactions between the Examinees versus Criteria and Criteria versus Ratings facets was negligible. However, the finding about the interaction between Criteria versus Ratings has

implications for ratings and rater training. For example, it has been shown that extensive rater training reduces the amount of bias associated with the interaction between the raters and criteria facets (Edler, Knoch, Barkhuizen, & von Randow, 2005).

The Writing Section: The results of the MFRM analysis showed the mean of the score residuals and the mean of the standardized residuals were zero and -.01, respectively. The standard deviation of the standardized residuals was 1.05. This means that the parameters of the Rasch model for this section of GTEC were estimated successfully in the current study, so the rest of the results can be reviewed to answer the research questions.

MFRM analysis also revealed that 89.74 % of the score variance was explained by Rasch measures. This finding, which provides the answer to research question 8 (i.e., Did the writing and speaking sections of GTEC Advanced version have a unitary factor structure?), shows that the writing scores represent a unitary factor structure, which means this section of the test taps only one single underlying trait or construct, which is writing ability.

Figure 3 illustrates the Wright Map for the writing section. In this map, similar to the one for the speaking section, positive values indicate examinees with higher writing ability, negative values indicate less difficult criteria, and less severe ratings. As seen in the figure, there was a wide spread of examinees' ability with a range from -7.4 to +13.9 logits (Range= 21.03). The mean ability of examinees was .7 logits, with a standard deviation of 2.06. The Chi-square test was not statistically significant ($\chi^2 = 7805.3$, df= 8981, $p < 0.01$), which indicates that the examinees in the current study belong to statistically distinct ability groups.

Figure 3 also shows that most of the examinees were located above the difficulty level of the majority of the scores (criteria). The mean ability of examinees was .7 logits, whereas the mean criteria difficulty was zero (Min = -2.41, = Max = 2.33, SD= 1.61, Range = 4.74).

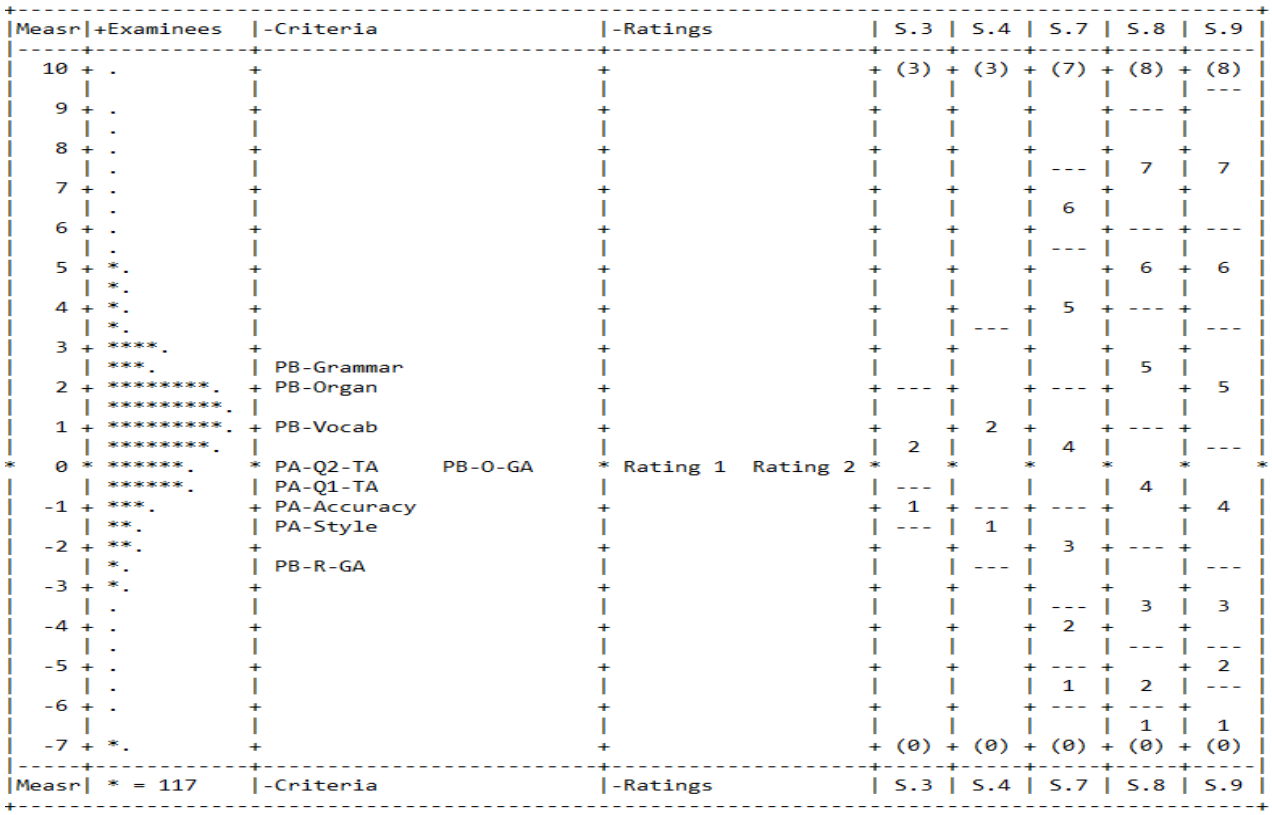


Figure 3. The MFRM Wright Map for the Writing Section of GTEC Advanced Version

PA: Part A, PB: Part B, Q: Question Number, O: Opinion, R: Reason, TA: Task Achievement, Style: Style, Accuracy: Accuracy, GA: Goal Achievement, Vocab: Vocabulary, Grammar: Grammar, Organ: Organization.

Table 13 summarizes the above statistics, which can be used to answer research question 1 (i.e., Did the scores (criteria) have a wide range of difficulty levels?) and research question 2 (i.e., Were the examinees divided into distinct ability levels?) about the writing section of Advanced GTEC version.

Table 13

Summary Statistics for the MFRM Analysis on the Writing Section Scores

Statistics	Examinees	Criteria (scores)	Ratings
Mean	.7	0	0
Standard Deviation	2.06	1.61	.04
Min to Max	-7.4 to 13.9	-2.41 to 2.33	-.04 to .04
Range	21.03	4.74	.08

Based on these findings (as shown in Table 13), it can be concluded that although on average the test was relatively easy for the sample of test takers in the current study (or on average their ability level was higher than most of the criteria in the current study), there were still several distinct levels of difficulty associated with the scores or criteria of the writing section of GTEC Advanced version, and this section of the test managed to categorize the examinees into several distinct levels of writing ability.

As seen in Table 13, the results of the MFRM analysis also showed that the mean of the severity of the ratings was zero with a standard deviation of .04. (Min = -.04, = Max = .04, Range = .08). This shows that on average, the two sets of the ratings fell within a narrow range of severity level. This provides the answer to research question 3 (i.e., Were the ratings different in severity?). Although the findings did not indicate identical levels of severity among the two sets of ratings, these findings did show that the ratings were not very different, and their severity levels were limited to a narrow range.

Table 14 reports values for the indices of the strata and separation with the reliability estimates associated with it for the writing section of GTEC Advanced version. As can be seen in this table, the examinees were divided into a minimum of five distinct levels of writing ability with a high level of reliability (.93). This further supports the answer to research question 2 and

indicates that the test was successful in identifying different groups of test takers with distinct differences in their writing ability.

Table 14

Summary Statistics for the MFRM Analysis on the Writing Section Scores

Statistics	Examinees	Criteria (scores)	Ratings
Strata	5.16	112.2	6.81
Separation Index	3.62	83.93	4.86
Separation Reliability	.93	1	.96

The statistics reported in Table 14 also provide further support for answers to research questions 1 and 3. As can be seen, the strata and separation indices showed that the scores (criteria) had a wide range of difficulty levels with a perfect reliability estimate for their separation. As discussed before, the strata and separation indices also showed that there were about six distinct levels of severity in the two sets of ratings with a high level of reliability for their separation (.96). However, as noted before, these differences were minimal and limited to a narrow range of Rasch logits (-.04 to .04).

Table 15 includes summary statistics from the MFRM analysis for the three facets of measurement in the writing section of GTEC Advanced version. As seen in this table, the overall infit and outfit mean squares for all three facets of the measurement fell within the acceptable range of .5 to 1.5, as well as the more stringent acceptable range of .7 to 1.3 (Eckes, 2005; Linacre, 2013).

Table 15

Summary Statistics for the MFRM Analysis on the Writing Section Scores

Statistics	Examinees	Criteria (scores)	Ratings
Infit Mean Square	.94	1	.95
Outfit Mean Square	1.11	1.32	1.32

Also, the criteria for judging the infit and outfit mean squares were used to classify examinees, criteria, and ratings based on their usefulness for measurement purposes. Table 16 summarizes the frequency of the occurrences of the mean squares within different acceptability ranges. As can be seen in this table, the vast majority of the fit indices fell within both acceptable ranges.

Table 16

Frequency Count of Infit and Outfit Mean Squares based on Fit Acceptable Ranges

	Infit Mean Square	Outfit Mean Square
Examinees		
<.7	37%	31.83%
<.5	17.9%	20.22%
>1.3	16.8%	22.6%
>1.5	11.81%	17.6%
Between .7 to 1.3	46.2%	45.57%
Between .5 to 1.5	70.29%	62.18%
Criteria (scores)		
<.7	0%	0%
<.5	0%	0%
>1.3	0%	44.44%
>1.5	0%	22.22%
Between .7 to 1.3	100%	55.56%
Between .5 to 1.5	100%	77.78%
Ratings		
<.7	0%	0%
<.5	0%	0%
>1.3	0%	0%
>1.5	0%	0%
Between .7 to 1.3	100%	100%
Between .5 to 1.5	100%	100%

Therefore, overall, these findings about the fit of the facets of the measurement to the Rasch model provide answers to **research question 6** (i.e., Did all the scores (criteria) fit the Rasch model?) and **research question 7** (i.e., Did all the ratings fit the Rasch model?), which is that all of the criteria and ratings fit the Rasch model.

In short, the above findings show that the criteria were applied consistently throughout the test for measuring test examinees' ability level, and the ratings were consistent throughout the rating sessions.

To answer research question 5 (i.e., Did the ratings agree with each other, and were these ratings independent from each other?), the percentage of the agreements between the two ratings were computed. The Rasch model expected level of exact agreement was 63.3%; however, the actual exact agreements were 73.2% of the ratings (across the two ratings). Using these percentages, the Rasch Kappa was computed (73.2 – 63.3)/(100-63.3) to be .27, and because this value is near zero, it can be concluded that the ratings were independent from each other. Therefore, it can be concluded that although the two ratings were independent from each other, they had a high level of agreement with each other.

Finally, to respond to research question 4 (i.e., Were there any systematic patterns of bias in the ratings?) bias analysis (interaction analysis) between the three facets of measurement was carried out using the t values of equal to or more than |2.00|. Table 17 summarizes the results of these analyses.

Table 17

Summary of Bias (Interaction) Analysis

	Examinees X Criteria	Examinees X Ratings	Criteria X Ratings
Number of empirical bias terms	79695	17710	18
Mean for t values	-.09	0	-.01
SD	.85	.83	1.85
Min.	-6.5	-4.2	-4.22
Max.	4.14	5.51	4.14
Number of t values equal to or higher than |2.00|	4392	335	4
Total percentage of all bias terms	5.5%	1.89%	22.22%

As seen in the table above, there were 1.89% of bias patterns between ratings and examinees, which means that a few raters evaluated certain test takers more strictly or more leniently than expected. Also, there were 22.22% of bias patterns between ratings and the evaluation criteria, which means that some raters used certain criteria more strictly or more leniently. Finally, there were only 5.5% of bias patterns between examinees and criteria. This means that some test takers were evaluated more strictly or leniently when certain criteria were used. Although the percentage of the bias patterns among ratings and evaluation criteria were relatively high, this is not an unusual finding in the context of large-scale tests with a large number of interaction terms. For example, for this type of interaction pattern, Eckes (2005) and Koizumi, Okabe, and Kashimada (2017) reported 37% and 25.78%, respectively.

To wrap up, the answer to research question 4 would be that the amount of bias related to the interactions between the Examinees versus Criteria and Criteria versus Ratings facets was negligible. However, the finding about the interaction between Criteria versus Ratings has implications for ratings and rater training. For example, it has been shown that extensive rater training reduces the amount of bias associated with the interaction between the raters and criteria facets (Edler, Knoch, Barkhuizen, & von Randow, 2005).

The CFA Procedures and Findings

Mplus 7.02 (Muthén & Muthén, 2014) was employed to conduct separate CFAs on test data from the speaking and writing sections of GTEC Advanced version in order to examine whether the test data structure corresponds with the test design structure. These analyses

provide answers to research question 9 (i.e., Did the test data structure correspond to the hypothesized structure of the scoring criteria?).

In the CFA models for each section, scores from different parts of the test, which were the observed variables of the models, were loaded on their respective test parts (e.g., A, B), which were the latent variables. Then, all the test parts (the latent variables) were freely correlated with each other because in the design of the speaking or writing sections of GTEC Advanced version, all the parts are considered independent but interrelated.

To evaluate the fit of the CFA models, the indices for the models' χ^2, RMSEA, CFI and WRMR were computed and examined. WRMR is suitable for models where sample statistics have widely varying variances and when sample statistics are on different scales. WRMR is also suitable with non-normal continuous outcomes (Muthén, 2004; p. 24).

The Speaking Section: Figure 4 depicts the CFA model tested for the speaking section of GTEC Advanced version. For parameter estimation, the robust weighted least squares model (WLSM) using a diagonal weight matrix with standard errors and mean-adjusted chi-square test statistic that uses a full weight matrix was used.

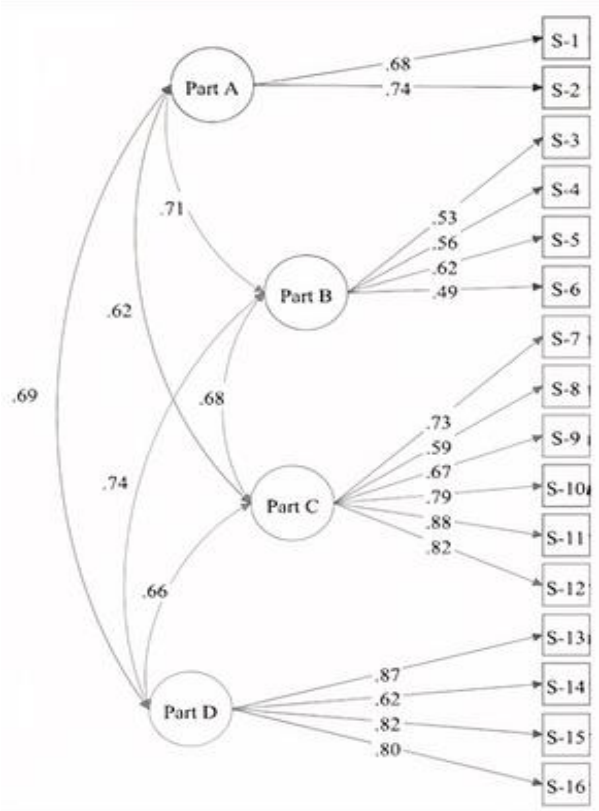


Figure 4. CFA Model for the Speaking Section of GTEC Advanced Version
S: Score (See Table 4)

The default estimator for this type of analysis is maximum likelihood with robust standard errors using a numerical integration algorithm. However, for models that include latent variables that have both categorical (dichotomous) and continuous indicators or observed variables, WLSM is the appropriate method of model parameter estimation.

As seen in Figure 4, scores 1 to 2, 3 to 6, 7 to 12, and 13 to 16 were loaded on Part A to Part D (latent variables), respectively. Also, for Parts C and D, the error variance of the scores were correlated to account for the method effects created as the result of the existence of multiple scores for different dimensions of speaking ability within a single part of the test. All the

estimated model parameters were statistically significant, and Table 18 summarizes the fit statistics for the model.

Table 18

Fit Indices for the CFA Model for the Speaking Section of GTEC Advanced Version

χ2	Df	p-value	WRMR	RMSEA	CFI
99301.59	120	0	1.72	0.03	0.99

Although the chi-square was statistically significant[3], and the WRMR was larger than 1, the CFI and RMSEA were within the acceptable ranges, which indicates sufficient model fit. As this CFA model was built based on the test specifications and how scores (criteria) were related to each other within and across different parts of the speaking section, the fit of the model to the data lets us conclude that in response to **research question 9,** it can be said that the structure of the data resulted from the administration of the speaking section of GTEC Advanced version corresponds to the hypothesized structure of the test.

The Writing Section: WLSM was also used to estimate the parameters of the CFA model tested for the data structure of the writing section of GTEC Advanced version. Similarly to the model for the speaking section, error covariances for the scores embedded within a single part of the test were also added. Figure 5 depicts the CFA model tested for the writing section of GTEC Advanced version.

Table 19 summarizes the fit statistics for the CFA model tested for the writing section of GTEC Advanced version. In this model, all the estimated parameters were statistically significant.

[3] As noted before, for large sample sizes and complex models, the χ^2 is usually statistically significant, which is not necessarily a sign of model mis-fit. That is why other fit indices such as CFI and RMSEA should be prioritized.

Table 19

Fit Indices for the CFA Model for the Writing Section of GTEC Advanced Version

χ2	Df	p-value	WRMR	RMSEA	CFI
88953.86	36	0	1.2	0.04	1

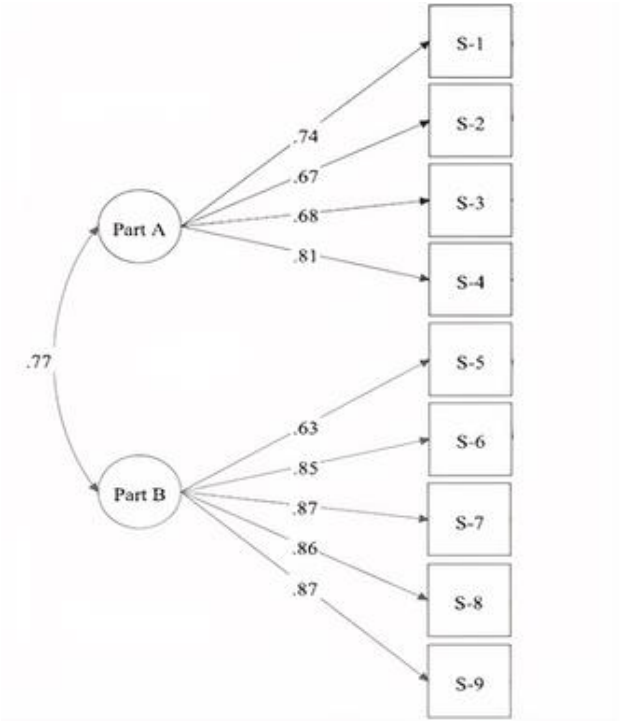


Figure 5. CFA Model for the writing Section of GTEC Advanced Version
S: Score (See Tables 5 and 6)

To sum up and in response to **research question 9** about the writing section, it can be concluded that the structure of the test data corresponded to the hypothesized structure of the test.

The MTMM Procedures and Findings

Using Mplus 7.02 (Muthén & Muthén, 2014), CFA MTMM analyses were carried out on the speaking and writing sections of GTEC Advanced version. This kind of analysis is needed to verify the extent to which these two sections of GTEC Advanced version assess English speaking and writing abilities versus assigning scores to the aspects of the students' performance that are specific to the context and nature of the test itself.

Following Campbell and Fiske's (1959) conceptualization, the MTMM approach for assessing the construct validity of a set of measures requires the measurement of two or more traits (constructs) using two or more assessment methods. The data used in the current study met this requirement. As summarized in Tables 4, 5, and 6, the speaking and writing sections of GTEC Advanced version measure different sub-constructs or sub-traits of the speaking and writing abilities (i.e., Goal Achievement, Language Use, Fluency and Pronunciation for the speaking test, and Task Achievement, Style, Accuracy, Goal Achievement, Grammar, Vocabulary and Organization for the writing test). Also, these sub-constructs are tested via different methods (i.e., different items in different parts of the two sections).

MTMM was carried out to investigate the convergence, discrimination, and method effects aspects of construct validity. In the current study, convergence refers to the extent to which the scores (criteria) in different parts of each of the two speaking and writing sections of GTEC Advanced version tap the same sub-constructs. For example, did the scores for Fluency and Pronunciation obtained from Parts A, C, and D in the speaking section converge? In other words, did the three separate scores from the three separate parts of the test indicate variable levels of the same underlying sub-construct (i.e., Fluency and Pronunciation)?

In the current study, discrimination refers to the extent to which scores (criteria) in different parts of the test measure distinct sub-constructs. For example, in the writing section, were the sub-constructs of Style and Organization measured in a distinct way? Also, by investigating the method effects, the current study aimed at identifying bias that can arise from the use of the same method in the measurement of different constructs.

The present study adopted the CFA approach to MTMM analysis as proposed by Widaman (1985). In this approach, a single model is selected at the outset as the most acceptable or best fitting structural model of the data (a.k.a. hypothesized model). Then, using (scaled) chi-square difference tests, the fit of this model is compared to the fit of alternative nested models in order to examine the degree of convergence, discrimination, and method variance.

The hypothesized model is used as a baseline against which all other alternatively nested models are compared in the process of assessing evidence of construct validity. This model represents a much more complex structure than the alternatively nested models. This complexity arises from the loading of each observed variable onto both a trait and a method factor. In addition, the model postulates that although the traits are correlated among themselves, as are the methods, any correlations between traits and methods are assumed to be zero. Testing for evidence of convergent and discriminant validity involves comparisons between the hypothesized model (Model 1) and four alternative MTMM models (Models 2 to 5).

Model 1: Correlated Traits and Correlated Methods (CTCM)

As noted earlier, the specification of this model includes both trait and method factors and allows for correlations among traits and among methods. Therefore, this model is typically the least restrictive.

Model 2: No Traits/Correlated Methods (NTCM)

Of major importance with this model is the total absence of trait factors. Evidence of convergent validity (i.e., the extent to which independent measures of the same trait are correlated) can be tested by comparing a model in which traits are specified (Model 1) with one in which they are not (Model 2). The larger the discrepancy between the χ^2 and CFI values, the stronger the support for evidence of convergent validity.

Model 3: Perfectly Correlated Traits/Freely Correlated Methods (PCTCM)

In this model, the trait correlations are perfect (i.e., they are set to be equal to 1.0) and consistent with both Models 1 and 2, and the method factors are freely estimated.

Model 4: Freely Correlated Traits/Uncorrelated Methods (CTUM)

The CTUM model differs from Model 1 only in the absence of specified correlations among the method factors. Discriminant validity is typically assessed in terms of both traits and methods. In testing for evidence of trait discriminant validity, interest focuses on the extent to which independent measures of different traits are correlated; these values should be negligible. When the independent measures represent different methods, correlations bear on the discriminant validity of traits; when they represent the same method, correlations bear on the presence of method effects, another aspect of discriminant validity.

In testing for evidence of discriminant validity among traits, we compare a model in which traits correlate freely (Model 1) with one in which they are perfectly correlated (Model 3); the larger the discrepancy between the χ^2 and CFI values, the stronger the support for evidence of discriminant validity.

Based on the same logic, albeit in reverse, evidence of discriminant validity related to method effects can be tested by comparing a model in which method factors are freely correlated (Model 1) with one in which they are uncorrelated (Model 4). Again, the larger the discrepancy between the χ^2 and CFI values, the stronger the support for evidence of discriminant validity.

Model 5: Freely Correlated Traits /No Methods (CTNM)

This model assumes that all the methods of assessment were the same and the results of the assessment are biased as the result of the specific method effect used. The larger the discrepancy between the χ^2 and CFI values of Model 5 and Model 1, the stronger the support for evidence for lack of method effects bias.

The Speaking Section: Figure 6 represents the structure of the MTMM Model 1 (CTCM) for the speaking section of GTEC Advanced version. This model was used as the baseline for comparisons with more restrictive models. The 16 scores obtained from this section of the test were mapped onto different parts of the speaking test and sub-constructs they measure. Then, several more restrictive models were tested against this model. For the model comparisons, because chi-squares from WLSM method of parameter estimation cannot be used similarly to chi-squares from Maximum Likelihood method, scaled chi-square difference test was employed.

Table 20 summarizes the information about these alternative models, their fit, and the results of comparison with the baseline model.

Table 20

CFA MTMM Model Comparisons for the Speaking Section of GTEC Advanced version

	Model	χ^2	Df	WRMR	RMSEA	CFI	Scaled $\Delta\chi^2$
1	CTCM	311.44	79	1.01	0.02	0.99	Baseline Comparison
2	NTCM	1311.52	98	2.16	0.04	0.99	$\Delta\chi^2$= 768.37, df= 19, p < .05
3	PCTCM	542.72	82	1.35	0.03	0.99	$\Delta\chi^2$= 154.21, df= 3, p < .05
4	CCTUM	1783.99	85	2.46	0.05	0.98	$\Delta\chi^2$= 970.36, df= 6, p < .05
5	CTNM	4739.29	101	4.21	0.07	0.95	$\Delta\chi^2$= 3054.22, df= 22, p < .05

As seen in Table 20, the scaled chi-square difference ($\Delta\chi^2$) tests revealed the superior fit of the baseline model, providing evidence for the divergence and discrimination validity and lack of bias created by the method effects for the speaking section of GTEC Advanced version.

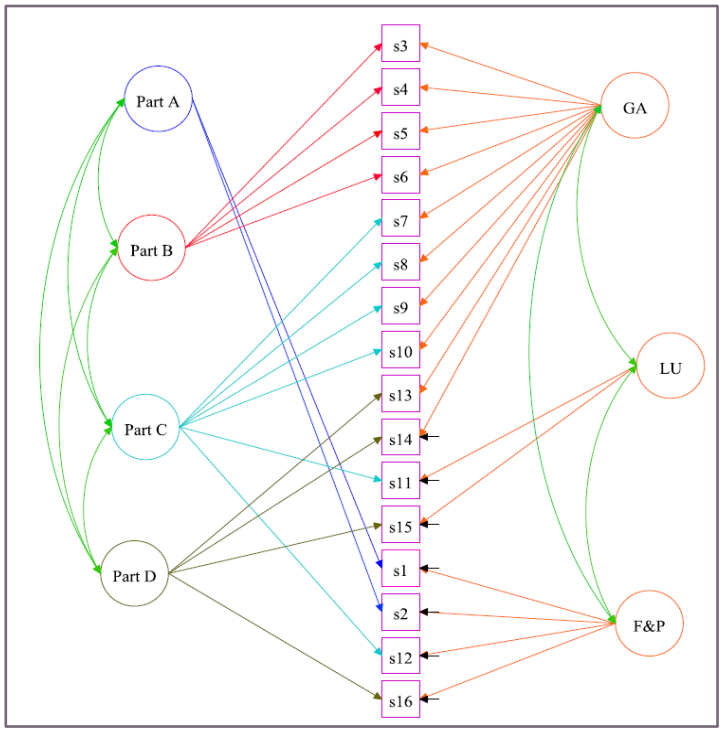


Figure 6. The Baseline CFA MTMM Model for the Speaking Section of GTEC Advanced version **S:** Score (See Table 4), **GA:** Goal Achievement, **LU:** Language Use, **F&P:** Fluency and Pronunciation

To recap and answer **research question 10** (i.e. To what extent did the independent assessment methods in the speaking and writing sections of GTEC Advanced version diverge in their measurement of different constructs?) and **research question 11** (i.e., To what extent did different assessment methods (i.e., different items in different parts) concur in their measurement of the same construct (e.g., Fluency and Pronunciation)?), it can be said that the MTMM provided evidence for both divergence and discrimination validity for the speaking section of GTEC Advanced version.

The Writing Section: The same procedure was followed for the writing section of the test. Figure 7 represents the baseline model built for the writing section. One note of clarification: because the sub-construct of Accuracy included control over both Vocabulary and Grammar, the scores for these two sub-constructs from Part B of the writing section were also loaded on the sub-construct of Accuracy. Also, because the two sub-constructs of Task Achievement and Goal Achievement entailed the same kind of underlying ability, scores 1, 2, 5 and 6 (see Tables 5 and 6) were loaded on the same latent variable, which has been labeled as GA in the model (see Figure 7).

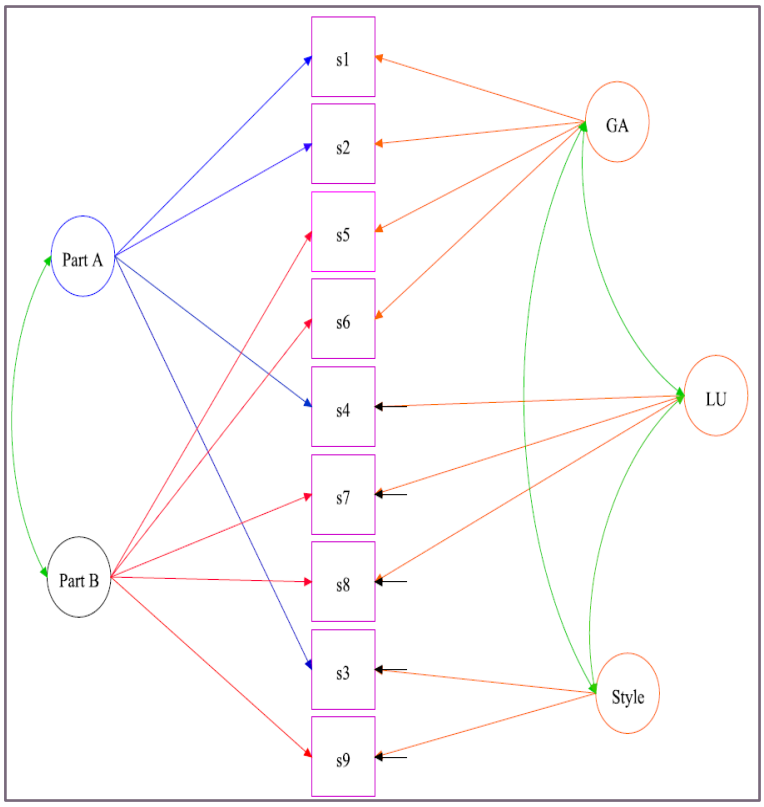


Figure 7. The Baseline CFA MTMM Model for the Writing Section of GTEC Advanced version
S: Score (See Tables 5 and 6), **GA:** Goal Achievement and Task Achievement, **LU:** Vocabulary, Grammar, Accuracy, **Style:** Organization and Style

Table 21 summarizes information about this baseline model and the alternative nested models, their fit, and the results of comparisons with the baseline model.

Table 21

CFA MTMM Model Comparisons for the Writing Section of GTEC Advanced version

	Model	χ^2	Df	WRMR	RMSEA	CFI	Scaled $\Delta\chi^2$
1	CTCM	80.53	14	0.58	0.02	0.99	Baseline Comparison
2	NTCM	1616.27	26	3.328	0.08	0.98	$\Delta\chi^2$= 1086.98, df= 12, p < .05
3	PCTCM	511.72	17	1.54	0.06	0.99	$\Delta\chi^2$= 314.96, df= 3, p < .05
4	CCTUM	487.11	15	1.46	0.06	0.99	$\Delta\chi^2$= 250.71, df= 1, p < .05
5	CTNM	6643.17	24	7.1	0.18	0.93	$\Delta\chi^2$= 4044, df= 10, p < .05

As seen in Table 21, the scaled chi-square difference ($\Delta\chi^2$) tests revealed the superior fit of the baseline model, providing evidence for the divergence and discrimination validity and lack of bias created by the method effects.

To recap and answer **research question 10** (i.e. To what extent did the independent assessment methods in the speaking and writing sections of GTEC Advanced version diverge in their measurement of different constructs?) and **research question 11** (i.e., To what extent did different assessment methods (i.e., different items in different parts) concur in their measurement of the same construct (e.g., Accuracy)?), it can be said that the MTMM provided evidence for both divergence and discrimination validity for the writing section of GTEC Advanced version.

Discussions and Conclusions

The purpose of the current paper was to test the plausibility of some of the assumptions underlying the authorizing warrants for three inferences of *evaluation*, *generalization* and *explanation* in the validity argument that has been developed for the usefulness of GTEC Advanced version scores for the intended purposes. This study was structured within the argument-based validation framework (e.g., Kane, 2006), and each of its research questions targeted one or two of the assumptions underlying the warrants mentioned above.

The current paper did not intend to test hypotheses about the inferences of domain description, extrapolation. Hence, the current study contributes only to certain aspects of the validity argument that pertain to the evaluation, generalization, and explanation inferences.

As explained before, the **evaluation inference** connects the observed performance to the observed score. It assumes that interpretation is based on scores that are construct-relevant and free from measurement error. The assumption here is that test performance can actually generate observed scores as viable indicators of the given test construct. Studies like evaluating scoring rubrics and item/task analysis can provide evidence for the plausibility of the assumptions underlying the licensing warrants of the evaluation inference.

In the current study, the first four research questions triggered the testing of several assumptions with regards to the quality of evaluation in the context of the speaking and writing sections of GTEC Advanced version. The answers to each of these questions are reviewed and discussed below.

Research question 1: Did the scores (criteria) have a wide range of difficulty levels? This question is related to the evaluation inference because test developers assume that the test items, tasks, and/or criteria they have included in a test prompt variable levels of performance. Without such a coverage scope, the actual range of ability levels cannot be identified. For this reason, tests should include a wide range of items, tasks, and/or criteria to gauge the ability level of test takers with no major gap at any point of the continuum upon which the ability levels are mapped. Given the findings in the results section of the current paper, we can conclude that the criteria in both the speaking and writing sections of GTEC Advanced version cover a wide range of ability levels.

Research question 2: Were the examinees divided into distinct ability levels? Similar to what was said about the assumption discussed under research question 1, test developers and users typically assume that their test enables them to identify the level of the speaking and

writing abilities of the test takers and divide these examinees into distinct groups of ability level. Without this assumption, the test scores would be totally useless. As you saw in the results section, both the speaking and writing sections of GTEC Advanced version divided the test takers into at least four distinct levels of ability. This finding is in alignment with the expectations of the test developers because the scores of the test are aligned with at least four CEFR levels ranging from A1 to B2. Therefore, the current study resulted in evidence based on which we can conclude that it is safe to assume that the test assigns distinct scores to distinct levels of speaking and writing abilities.

Research question 3: Were the ratings different in severity? One important factor that heavily influences the quality of evaluation in the context of performance assessment is the severity (a.k.a difficulty) level of the raters. Unlike the criteria and examinees facets, for which we expected several or even many variable levels, for the rater or rating facet we assume that all the ratings have an equal level of severity. Otherwise, the variability observed in the scores can be the result of variable levels of rater severity rather than variable levels of ability on the part of the test takers. Based on the findings reported in the results section of the current study, not all of the ratings had the same level of severity. However, it is unrealistic to expect to have a single level of severity among the raters, especially in the context of a large-scale test with hundreds of raters. Previous research has shown that even extensive rater training would not eliminate the variable severity levels altogether. That being said, because the number of the different severity levels in the current study was small, we can conclude that more training is required to minimize the current number of severity levels among the raters.

Research question 4: Were there any systematic patterns of bias in the ratings? If evaluation is done right, there should be no interaction between different facets of assessment. If there is an interaction between, for example, ratings and examinees, it means that ratings were more severe or less severe depending on particular individual test takers. This means raters were more severe with some test takers and less severe with the others. The test developers and users assume that there is no interaction between different facets of measurement in their assessment context, but this is something that should be verified empirically. For example, if students come from different parts of a country with different accents, the test users should make sure that no particular group of test takers were advantaged or disadvantaged because of their accent. The results of the current study revealed that this assumption can be held about the scores of the speaking and writing sections of GTEC Advanced version because there were no interactions between the examinee, criteria, and ratings facets of the measurement in the current study.

Now, with enough confidence in the plausibility of the major assumptions related to the evaluation inference, we can move on to the generalization inference and its underlying assumptions. This inference connects the Observed Scores to the Expected Scores. This inference assumes that interpretation can reliably be based on test scores that are interchangeable across items, tasks, raters, and other conditions. In the current study, research questions 5, 6, and 7 targeted this assumption. These questions prompted the examination of the fit between performance of different facets of measurement and the idealized Rasch model. If there is an acceptable level of fit between test data and the Rasch model, it can be concluded that interpretation of the test scores can be reliably based on test scores that are

interchangeable across, for example, criteria and ratings. These questions also targeted the independence and consistency of the ratings, as well as the agreement level between different ratings. As seen in the results section of the current paper, the ratings of the speaking and writing sections of GTEC Advanced version were independent from each other, they were consistent across ratings sessions, and there was an acceptable level of agreement between them. This gives us confidence about the plausibility of several important assumptions related to the generalization inference, and now we can move on to the explanation inference.

The **explanation inference** connects the Expected Scores to the Construct-Related Scores. This inference assumes that interpretation of test scores is related to the given construct. The inference links the expected scores to the underlying test construct and is based upon the assumption that patterns in the observed scores on the test reflect the trait(s) or constructs being measured. In other words, the internal structure of a measure is a viable operationalization of a theoretical construct. Therefore, one of the types of studies that can provide backing for the assumptions of the explanation inference would be an examination of the internal structure of the test. In the current study, **research questions 8, 9, 10,** and **11** targeted the assumptions related to the internal structure of the test data and its correspondence to the test structure. As seen in the results section, the speaking and writing sections of GTEC Advanced version had generated unidimensional data structure, which is the expectation of the test designers because each of these two sections of the test should test only one single construct at a time. Also, the results showed that when CFA models built based on test structure were tested against the test data in the current test, there was a good level of correspondence between test data and test structure. Additionally, the results section provided

evidence from the MTMM analysis for the convergence and divergence validity, indicating that different parts (e.g., Part A, Part B) under each section (i.e., speaking and writing) of the test tapped different sub-constructs without much of the bias resulting from different measurement methods.

All in all, the current study contributed to the validity argument that has been developed for the usefulness of GTEC scores for the intended purposes. Although this paper covered several major assumptions with regard to the interpretations and uses of GTEC scores, past and future studies should complement the current one.

References

American Educational Research Association. American Psychological Association & National Council of Measurement in Education (2014). Standards for educational and psychological testing.

Bond, T., & Fox, C. M. (2015). Applying the Rasch model: Fundamental measurement in the human sciences. Routledge.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological bulletin, 56(2), 81.

Chapelle, C. A. (2013). Conceptions of validity. In The Routledge handbook of language testing (pp. 35-47). Routledge.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2011). Building a validity argument for the Test of English as a Foreign Language. Routledge.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument‐based approach to validity make a difference? Educational measurement: Issues and practice, 29(1), 3-13.

Cronbach, L. J. (1988). Five perspectives on validity argument. Test validity, 3-17.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. Language Assessment Quarterly: An International Journal, 2(3), 197-221.

Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2005). Individual feedback to enhance rater training: Does it work?. Language Assessment Quarterly: An International Journal, 2(3), 175-196.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural equation modeling: a multidisciplinary journal, 6(1), 1-55.

Kane, M. (2001). Current concerns in validity theory. Journal of educational Measurement, 38(4), 319-342.

Kane, M. (2006). Content-related validity evidence in test development. Handbook of test development, 1, 131-153.

Kane, M. (2012). All validity is construct validity. Or is it?. Measurement: Interdisciplinary Research & Perspective, 10(1-2), 66-70.

KOIZUMI, R., OKABE, Y., & KASHIMADA, Y. (2017). A Multifaceted Rasch Analysis of Rater Reliability of the Speaking Section of the GTEC CBT. ARELE: Annual Review of English Language Education in Japan, 28, 241-256.

Linacre, J. M. (2013). Facets computer program for many-facet Rasch measurement, version 3.71. 4. Beaverton, Oregon: Winsteps. com.

Linacre, J. M. (2019). Winsteps®(Version 4.4. 1)[Computer Software]. Beaverton, Oregon: Winsteps.com. Retrieved January 1, 2019.

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. Language Testing, 15(2), 158-180.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Design and analysis in task-based language assessment. Language testing, 19(4), 477-496.

Mueller, R. O., & Hancock, G. R. (2008). Best practices in structural equation modeling. Best practices in quantitative methods, 488508.

Muthén, B. (2004). Latent variable analysis. The Sage handbook of quantitative methodology for the social sciences, 345(368), 106-109.

Toulmin, S. E. (2003). The uses of argument. Cambridge university press.