
Simulated Oral Proficiency Test

Validity Report

Payman Vafae, Ph.D.

COLUMBIA UNIVERSITY

SECOND LANGUAGE TESTING INC.

August 2019

Introduction

The simulated oral proficiency interview (SOPI) is a performance-based speaking test that is commonly used as a surrogate to the oral proficiency interview (OPI). Unlike many semi-direct tests, the SOPI contextualizes all tasks to ensure authenticity. The tasks in a prototypical SOPI assess the examinee's ability to perform different functions at the American Council on the Teaching of Foreign Languages (ACTFL) Intermediate, Advanced, and Superior levels. The prototypical SOPI format is tape-mediated and follows the same four phases as the OPI: warm-up, level checks, probes, and wind-down. The prototypical SOPI format includes picture-based tasks that allow examinees to perform tasks such as asking questions, giving directions based on a simple map, describing a place, or narrating a sequence of events based on the illustrations provided.

Given the flexibility of the SOPI format, it can be tailored to the desired level of examinee proficiency and for specific examinee age groups, backgrounds, and professions. The SOPI format has many other practical benefits. For example, any teacher, language lab technician or aide can administer the SOPI. This has proved to be an advantage in locations where a trained interviewer is not available or in languages that lack ACTFL-certified testers. In addition, the SOPI can be administered simultaneously to a group of examinees by one administrator. Thus, the SOPI may be preferable when many examinees need to be tested in a short time frame.

The SOPI may also offer psychometric advantages in terms of reliability and validity, particularly in standardized testing situations. In several studies, the SOPI proved to be a valid and reliable surrogate to the OPI (Kenyon & Tschirner, 2000; Stansfield & Kenyon, 1992, 1993). The SOPI offers the same quality of interview to all examinees, and all examinees respond to the

same questions. By recording the test for later scoring, it is possible to ensure that examinees will be rated by the most reliable raters and can be rated under controlled conditions. Raters who have scored both a live interview and a SOPI report that it is often easier to score a SOPI (Kenyon & Tschirner, 2000). This may be due in part to the SOPI's ability to produce a longer speech sample and to allow each examinee to respond to the same questions. Therefore, distinctions in proficiency may appear more obvious to the rater.

SLTI SOPI

Two events motivated Second Language Testing Inc. (SLTI) to develop a new format of SOPI that is different from its prototypical version. One event was that computers became ubiquitous. This allows the development of a computer-mediated SOPI that facilitates its administration and rating. Changing the format of the test from tape-mediated to computer-mediated increases the ease of the test administration and the quality of recorded speech from the test takers. This also facilitates the process of rating, especially because the industry standards require double-blind scoring by raters who are not necessarily present at the test administration site. The other event was the increasing popularity of Common European Framework of Reference (CEFR). Unlike ACTFL, which is used primarily by US governmental agencies and formal educational institutions, CEFR is commonly used all around the world in various educational and professional settings.

Therefore, to respond to the two above events and broaden the use and impact of SOPI, SLTI developed a computer-mediated version of SOPI that aims to estimate the English proficiency of test takers based on the [CEFR level descriptors](#).

Currently, SLTI SOPI is being used by several multinational companies to screen applicants during the recruiting process. These recruits will typically work in the customer-service call centers of these companies. In order to accommodate the needs of these clients, and based on the results of a needs analysis, the test tasks (or items) of SLTI SOPI are situated within a professional-setting of language use that is similar to the intended target language use (TLU) domain of the test. The SLTI SOPI tasks prompt test takers to respond to questions and/or situations that are closely tied to the use of language in a work setting, which enables test score users to make inferences about the ability of the test takers in the actual TLU domain.

Also, based on the results of the needs analysis, the minimum level of English proficiency for being able to effectively carry out the work-related tasks was deemed to be CEFR B1 level. This means that SLTI SOPI should be able to reliably rate the performance of the test takers to be at B1 level or lower. However, to make the test scores more informative and to broaden their use for a wider range of contexts, the SLTI SOPI includes six tasks to elicit spoken responses that are ratable at the B1 to C1 levels and their respective sub-levels, i.e, low-B1 (B1.1), high-B1 (B1.2), low B2 (B2.1), high B2 (B2.2), low-C1 (C1.1), and high-C1 (C1.2).

Accordingly, three out of the six SLTI SOPI tasks aim to elicit responses that CEFR describe as B1. However, responses to these B1 tasks are ratable within the range of lower than B1, B1.1, and B1.2 levels. One of the SLTI SOPI tasks was developed with the aim of eliciting responses at the B2-level, but the responses to this task are ratable within the range of lower than B1, B1.1, B1.2, B2.1, B2.2 and C1.1 levels. Finally, the two remaining SLTI SOPI tasks aim to elicit responses at the C1 level, but responses to these two tasks are ratable within the range of lower than B1, B1.1, B1.2, B2.1, B2.2, C1.1, and C1.2 levels.

This way, the responses of the test takers can be rated at each of the levels (except for C1.2) at least three times, which is the minimum requirement for the reliable assessment of spoken performances at each of the above levels (Kenyon & Tschirner, 2000). Also, to allow for the administration of the test in massive numbers, and to minimize the chance of cheating, SLTI SOPI has been developed to include eight equal forms. Each form includes six different, but similar, test tasks with similar difficulty levels.

The Current Validation Study

The purpose of the current validation study was to examine several psychometric qualities of SLTI SOPI. The current study was a follow-up to an earlier validation study on the test, the results of which were used to make several revisions to the test. SLTI continually monitors the quality of its tests both qualitatively and quantitatively, and the current study was intended to examine whether recent revisions to SLTI SOPI have led to improvements in the psychometric qualities of the test.

Based on the findings of the previous validation study, although they were indicating acceptable psychometric qualities for the test, the following revisions were made on the test content, its delivery, and rating to make its psychometric qualities ideal.

1. A qualitative review of all items led to the revision and/or replacement of items with non-ideal psychometric qualities. For example, based on the findings of the previous study, the B1-level items needed revision to become more difficult to be able to elicit higher-level responses.

Therefore, these items were revised for their difficulty level.

2. All of the items were revised to remove any, even negligible, content errors (e.g., typographical and grammatical errors), and to enhance their language accessibility for the intended audience.
3. The contexts of items were revised to make them more relevant to the intended purpose and audience of the test.
4. The item images were updated to make them more culturally and ethnically diverse.
5. The overall test directions were updated to make them clearer.
6. An interactive sample item was added to each form to help test takers prepare for the actual test items better.
7. The rating rubric was revised to make its language clearer for the raters, and the raters were retrained to get familiar with the revised rubric.

To examine if the above revisions have led to the intended improvements in several psychometric qualities of the test, the following research questions guided the current study:

1. Are scores from the eight SLTI SOPI forms comparable in their score distributions? Or, are the eight forms at a comparable difficulty level?
2. To what extent do the facets of examinee, tasks and ratings contribute to total score variance of SLTI SOPI? Or, are the SLTI SOPI scores reliable and dependable?
3. To what extent do SLTI SOPI tasks and ratings fit into a many-facet Rasch model? Or, do the tasks and ratings perform as expected by the test designers and examiners?
4. To what extent do SLTI SOPI data fit single-factor confirmatory factor analysis models? Or, do all the test tasks in each test form tap the same underlying construct (i.e. speaking ability)?

Analyses and Results

Analysis of Variance

The SLTI SOPI has eight different forms, each of which was designed based on the same table of specifications. There are no linking tasks shared among the eight forms of the test, and each of the eight forms are taken by separate, randomly assigned groups of test takers. These test takers have assumed similar background characteristics, including English speaking ability. The assumption of the similarity of the different groups of test takers, especially in terms of speaking ability, who take the different forms of the SLTI SOPI stems from the information the clients have shared with us. The test takers, who are job applicants for similar positions in the customer-service call centers, take the SLTI SOPI after being screened by a face-to-face interview. Applicants who are deemed qualified and who can manage work-related conversations in English are then required to take the SLTI SOPI for the official verification of their English ability. This means that the population of the SLTI SOPI test takers has already been screened for their qualifications including their English ability, which makes this population relatively homogenous.

The purpose of the current analysis (research question 1) was to examine to what extent the eight forms of the test have led to comparable distribution of the scores, which is an indication of test-form equivalency or lack thereof. Because test takers are randomly assigned to each of the SLTI SOPI forms, group speaking ability is assumed to be equal, and any differences in the score distributions are identified as difficulty differences among the test forms. One of the methods of examining test-form equivalency within this “equivalent-groups” data collection

design is examining form difficulty differences as estimated by the mean and variance differences across the forms (e.g., Beglar & Hunt, 1999; Brandt, 1991).

To examine the difficulty differences across the eight SLTI SOPI forms taken by different randomly selected test takers, a one-way Analysis of Variance (ANOVA) was conducted. Table 1 summarizes the means and standard deviations (SD) of the scores of the test takers across the eight test forms. As seen in Table 1, sample sizes for each form are variable but similar. Also, sample sizes are large enough (5206 to 5485) for obtaining a stable estimate of the proficiency level of test takers across the eight forms.

Table 1. Means and SDs of the Eight SLTI SOPI Forms

Form	Form 1 (N=5485)	Form 2 (N=5450)	Form 3 (N=5391)	Form 4 (N=5384)	Form 5 (N=5274)	Form 6 (N=5206)	Form 7 (N=5345)	Form 8 (N=5244)
Mean	5.98	5.9	5.93	6.01	5.99	5.94	6.02	5.98
SD	1.64	1.68	1.73	1.69	1.63	1.7	1.67	1.64

The results of the one-way ANOVA (i.e. the omnibus test) was statistically significant [$F(7, 42771) = 2.35, p = 0.02$]; however, the Tukey post hoc test results showed no statistically significant differences between any of the eight test forms.

Generalizability theory

This section summarizes analyses that were conducted to respond to research question 2. Generalizability theory (G-theory) is used for specifying and estimating the relative effects of different facets of measurement on test scores. G-theory conceptualizes that in addition to a test taker's ability, other factors or facets of measurement influence the test taker's score. For example, in a test consisting of different tasks, the facet of concern would be differential levels of task difficulty. In a test with different tasks and different raters, the two facets of concern are

differential levels of task difficulty and differential levels of raters' severity. In tests with multiple facets, the interaction between different facets is also a point of concern because if some tasks are differentially difficult for different groups of test-takers, this may be a source of bias. Similarly, if raters score the performance of different groups of test-takers differently, then this could be an indication of rater bias. Interactions such as these cannot be examined by reliability estimates, such as internal consistency and inter-rater consistency, which are derived from classical test theory. Since interactions between persons and tasks, between persons and raters, and between raters and tasks are frequently the major sources of error variance, traditional approaches to estimating inter-task and inter-rater consistency or reliability are clearly inadequate.

In order to estimate the effects of the facets of concern, G-study can be conducted in which the effects of these different facets are clearly distinguishable. This provides sample statistics indicating the magnitude of these effects in the sample used in the G-study, and these sample statistics provide the basis for a decision study (D-study), in which the population values for the different sources of variation are estimated.

The D-study provides two types of information: First, an estimate of the relative importance of the effects associated with the different facets, and the interactions among these facets, in the form of *variance components*. These variance components provide estimates of the relative influence of different facets of measurement on a test taker's score. For example, in the context of the current study, the following facets influence a test taker's score: 1) her speaking ability; 2) the relative difficulty of the tasks; 3) the relative severity of the raters; and 4) the interactions between persons (test takers) and tasks, persons and raters, and tasks and raters.

Second, a D-study provides estimates of the reliability and dependability of the test scores, taking into account the effects of the different facets and interactions. These estimates are Generalizability coefficient (G) and Phi (ϕ). The former is a reliability estimate for norm-reference assessment (i.e., rank-ordering test takers), and the latter is an estimate for the dependability of the scores for criterion-reference assessment (i.e., categorizing test takers into different groups).

In the current study, in addition to the test taker's speaking ability (person), there were two other facets of measurement: tasks and ratings. In context of the current study, because not all the test takers were rated by the same two raters, the effect of the two *ratings* was examined instead of the effect of the two *raters*. This way, the design of the current study was *fully-crossed* because in each form, all the test takers took all six tasks and received two sets of ratings.

For the current study, the computer program of GENOVA (Crick & Brennan, 1982) was used, and Table 2 summarizes the results of G-theory analysis (the results of D-study) for each of the forms. The table summarizes the percentage of variance explained by each of the facets of measurement and the interaction between them.

In an ideal situation, the biggest amount of variance should be explained by the test taker's ability or person (*p*) facet. This indicates that the variability in the test scores is mainly as the result of variability in the test takers' ability levels. Substantial amount of explained variance by other facets of the measurement and the interaction between them is a sign of bias and lack of reliability and dependability. The tables also include the estimates for Generalizability coefficient (G) and Phi (ϕ). As the SLTI SOPI scores are used in a criterion-reference assessment framework, the value of Phi is more relevant than G.

As seen in Table 1, for all of the forms, the percentage of the variance explained by the person (*p*) facet is higher than the acceptable level of 80%, and the percentage of the variability in the test scores explained by other facets and the interaction between them is negligible. In addition to the person facet (*p*) information, Table 2 summarizes information about the amount of variance in the total scores explained by the task (*t*) and ratings (*r'*) facets, as well as the interaction between these facets and random error (*e*).

Table 2. Summary Statistics for G-theory Analysis

Facet/Percentages	Form 1	Form 2	Form 3	Form 4	Form 5	Form 6	Form 7	Form 8
<i>p</i>	89.1	89.93	88.27	88.48	91.23	88.5	89	89.42
<i>t</i>	2.5	2.04	4.01	3.98	1.25	3.11	2.79	2.34
<i>r'</i>	0	0	0	0	0	0	0	0
<i>pt</i>	6.54	6.37	6.13	5.93	3.99	6.81	6.6	6.6
<i>pr'</i>	.58	.43	.31	.31	0	.28	.4	.32
<i>tr</i>	0	0	0	0	.78	0	0	0
<i>ptr,e</i>	1.28	1.23	1.28	1.3	2.75	1.3	1.21	1.32
G	.91	.91	.92	.92	.93	.91	.92	.91
Phi	.91	.9	.88	.88	.91	.88	.89	.89

Table 2 also summarizes the Generalizability coefficient (G) and Phi (ϕ) for the eight forms of SLTI SOPI. All of these values are larger than the acceptable value of .8 (Hoyt, 2010, p.152). It should be noted that Phi coefficient offers a more stringent estimate of score dependability than G coefficient, and it is typical of G-theory analysis results that Phi coefficients are smaller than the G coefficients (Sawaki & Xi, 2019).

Many-Facet Rasch Analysis

This section summarizes the results of the analyses that were conducted to respond to research question 3. Many-facet Rasch measurement represents an extension of the one-parameter Rasch model, one of several models developed within item response theory (IRT). These IRT models conceptualize a person's expected performance on a test item or task as a

function of his or her ability and characteristics of the test task, such as the difficulty of the item, or task, the task's capacity for discriminating between high and low scorers and the effect of guessing. The Rasch model is a one-parameter model because it uses only one task characteristic — task difficulty — in the estimation of person ability. Alternatively, person ability and task difficulty may be considered as facets in the measurement process, as is done in the many-facet extension of the Rasch model.

When the measurement is obtained using a rating scale, many-facet Rasch measurement can make use of the additional facet of rater severity to refine the estimation of the test taker's ability. Therefore, in many-facet Rasch analysis, separate estimates of task difficulty and rater severity can be obtained, and these in turn are used to estimate each test taker's ability. That is, if a particular rater is unusually severe (or lenient) in her ratings, this would be taken into account when estimating the ability of a test taker who had been rated by that rater.

Similarly, if a particular task is relatively difficult (or easy), this would be taken into account in estimating the ability of all test takers who responded to that task. This provides the potential for estimating the abilities of different individuals on the same scale of measurement, even though they may have been rated by different raters and may have responded to different tasks.

Table 3 summarizes the separation reliability estimates for the three facets of measurement in the current study: person, task and ratings.

Table 3. Many-Facet Rasch Analysis Separation Reliability Estimates

	Form 1	Form 2	Form 3	Form 4	Form 5	Form 6	Form 7	Form 8
Person Reliability	.93	.93	.93	.94	.92	.93	.93	.93
Number of Levels	3 or 4	3 or 4	3 or 4	3 or 4	3 or 4	3 or 4	3 or 4	3 or 4
Task Reliability	1	1	1	1	1	1	1	1
Number of Levels	3 or 4	3 or 4	3 or 4	3 or 4	3 or 4	3 or 4	3 or 4	3 or 4
Ratings Reliability	0	0	0	0	0	0	0	0
Number of Levels	1	1	1	1	1	1	1	1

These separation reliability estimates show how many different levels were identified for each of the facets of assessment. In a test administration with test takers with varying levels of ability, a high level of separation is expected for the person facets. SLTI SOPI is expected to categorize test takers into at least four levels of proficiency: lower than B1, B1, B2, and C1. This means that at least four levels of separation for the test takers should be observed. For the tasks (items), we should expect at least three levels of separation because the SLTI SOPI tasks were designed to tap B1, B2, and C1 levels of proficiency. For ratings, however, only one level of separation is expected, which indicates that the ratings were approaching the ideal of being interchangeable (Eckes, 2019).

Here is the rule of thumb for interpreting the many-facet Rasch separation reliability estimates. Person, task and ratings reliability estimates of lower than .5 show that there was only a single level of separation. Separation reliability estimates of .5 show that there were one or two separate levels of separation. Separation reliability estimates between .5 to .8 show that two or three separate levels were identified, and the separation reliability estimates of .8 and above represent three or four separate levels (Boone, Staver, & Yale, 2013, p. 230).

As seen in Table 3, the test data shows SLTI SOPI performs as expected. In all eight forms, the test takers are placed in at least three levels of proficiency (i.e., B1, B2, and C1), the

tasks are at least three levels of difficulty (i.e., B1, B2, and C1), and the ratings are at a single level of severity (or leniency).

Figures 1 to 8 present Wright maps for each of the eight forms of SLTI SOPI. These maps show the relative standing of each of the facets of assessment on a single scale. In these maps, the higher-level ability learners are shown towards the top of the map. For tasks, the more difficult tasks are located towards the top, and for ratings, the more severe ones are located towards the top of the map.

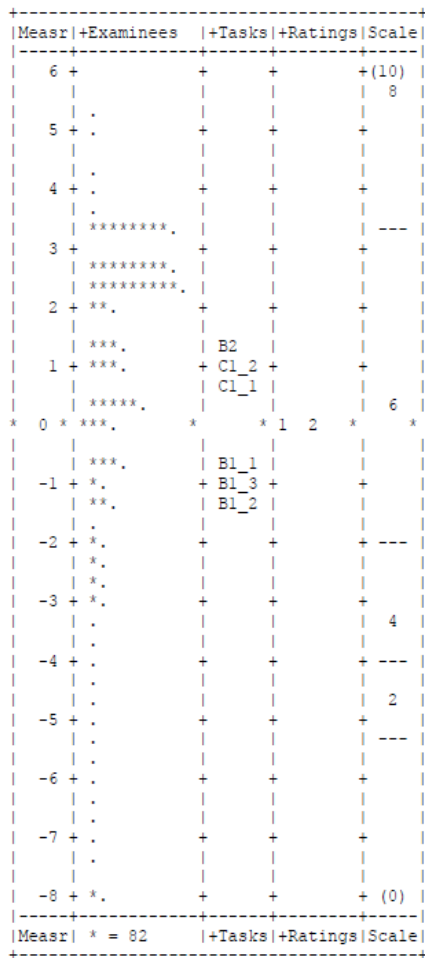


Figure 1. Wright Map for Form 1

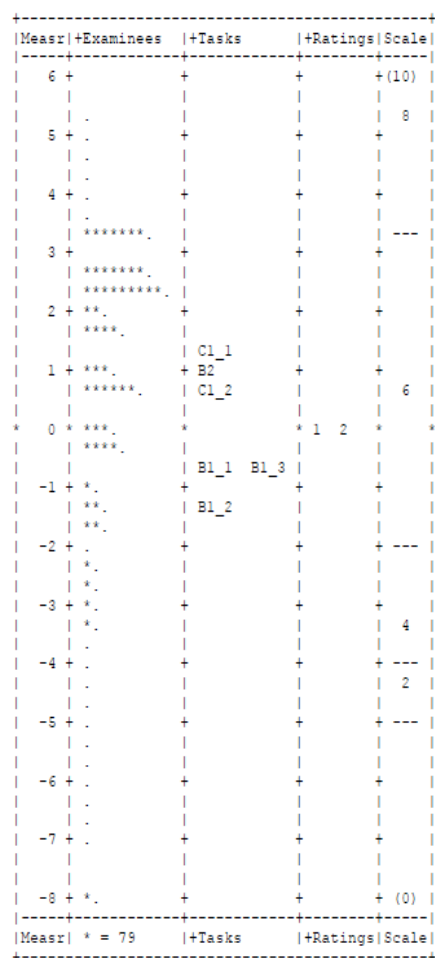


Figure 2. Wright Map for Form 2

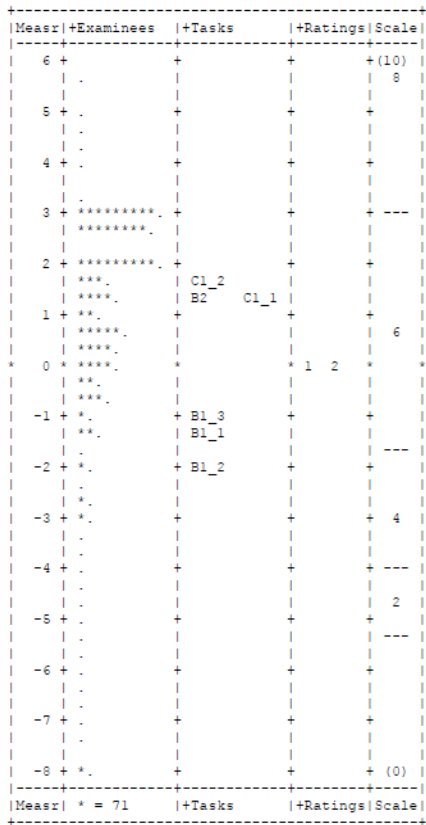


Figure 3. Wright Map for Form 3

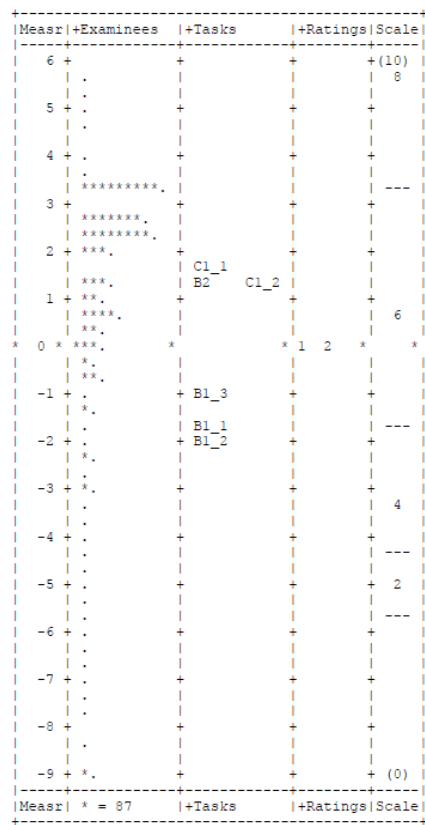


Figure 4. Wright Map for Form 4

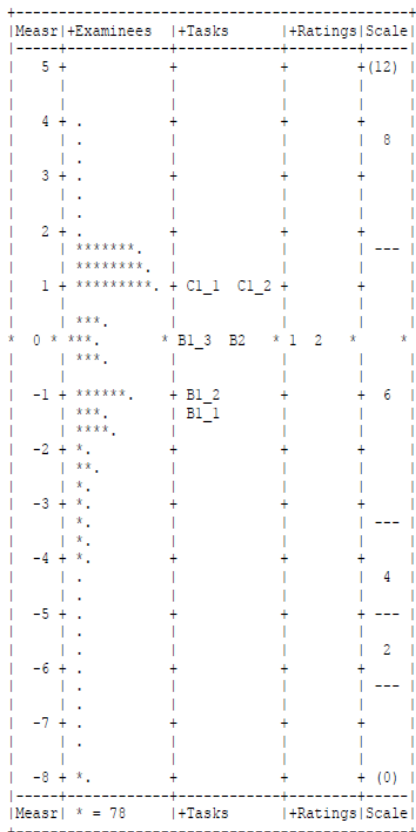


Figure 5. Wright Map for Form 5

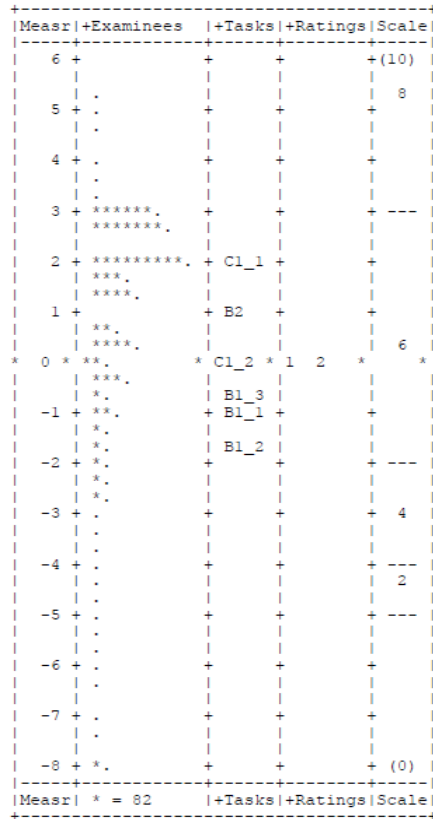


Figure 6. Wright Map for Form 6

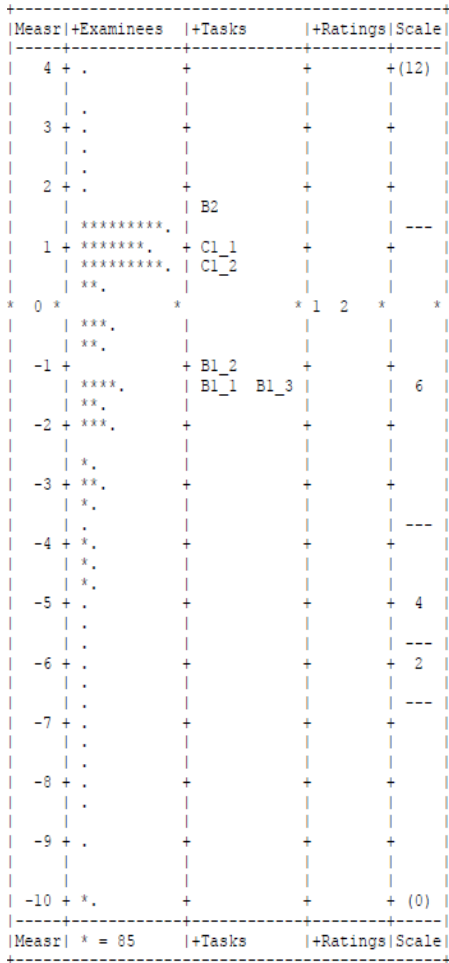


Figure 7. Wright Map for Form 7

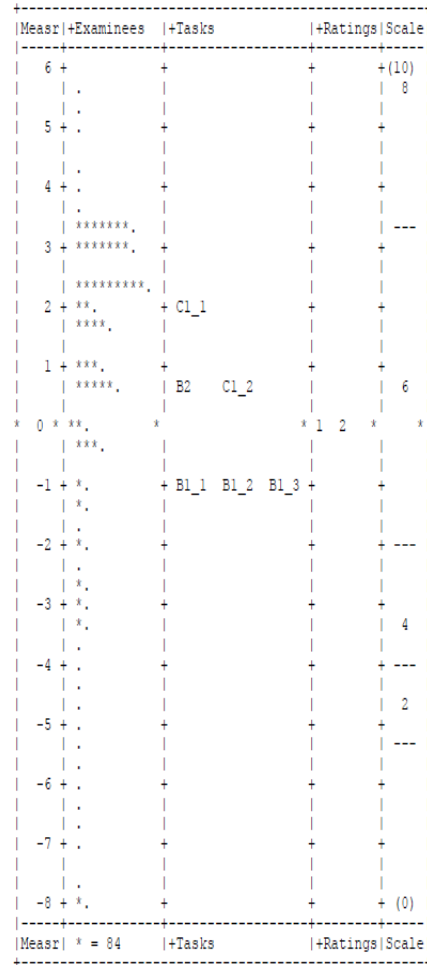


Figure 8. Wright Map for Form 8

The above Wright maps indicate that the SLTI SOPI meets the expected psychometric qualities. As seen in these figures, the B1 tasks are easier than the B2 and C1 tasks, and the test takers are located all across the scale, with some falling at or below the B1 tasks and some below, at or above the B2 and C1 tasks. This means that the test has been able to successfully categorize the test takers in the lower than B1, B2, and C1 and above levels. Also, as shown by the maps, the two ratings for all of the eight forms are exactly at the same level of severity and right at zero indicating that the variability in the severity of the ratings had no impact on the variability of the test scores.

Table 4 summarizes the infit and outfit means squares (MnSq) for each of the six tasks across the eight forms. These fit statistics show the size of the randomness (i.e., the amount of distortion of the measurement system) in the test results. Their expected value is 1.0. Values less than 1.0 indicate observations are too predictable (redundancy, data overfit the model), and values greater than 1.0 indicate unpredictability (unmodeled noise, data underfit the model).

Table 4. Summary of Mean-Square Fit Statistics for the Test Tasks

Task	Form 1		Form 2	
	Infit MnSq	Outfit MnSq	Infit MnSq	Outfit MnSq
B1-1	1.17	1.13	1.23	1.13
B1-2	1.5	1.39	1.41	1.32
B1-3	1.49	1.35	1.76	1.7
B2	1.24	1.21	.88	.72
C1-1	.95	.8	1.03	.99
C1-2	1.06	1.01	1	.99
	Form 3		Form 4	
	Infit MnSq	Outfit MnSq	Infit MnSq	Outfit MnSq
B1-1	1.3	1.2	1.28	1.16
B1-2	1.22	1.14	1.33	1.24
B1-3	1.14	1.06	1.13	1.06
B2	.95	.85	.89	.74
C1-1	1.21	1.28	1.32	1.76
C1-2	1.39	1.06	1.17	1.31
	Form 5		Form 6	
	Infit MnSq	Outfit MnSq	Infit MnSq	Outfit MnSq
B1-1	1.17	1.07	1.23	1.1
B1-2	1.5	1.39	1.28	1.19
B1-3	1.13	1.09	1.87	1.76
B2	1.14	1.08	.77	.61
C1-1	.85	.69	1.17	1.25
C1-2	1.31	1.69	1.1	1.13
	Form 7		Form 8	
	Infit MnSq	Outfit MnSq	Infit MnSq	Outfit MnSq
B1-1	1.22	1.11	1.31	1.2
B1-2	1.68	1.55	1.3	1.2
B1-3	1.46	1.38	1.68	1.6
B2	1.05	.97	1.01	.9
C1-1	.94	.86	1.07	1.06
C1-2	1.23	1.32	1.03	.92

According to Linacre (1990), the mean-square values larger than .2 show the tasks distort or degrade the measurement system. Values between 1.5 and .2 mean that the tasks are unproductive for construction of measurement, but not degrading. Values between .5 and 1.5 are productive for measurement, and values smaller than .5 show that tasks are less productive for measurement, but not degrading. This means that the infit and outfit mean-square values for each of the six tasks of SLTI SOPI should fall between the .5 to 1.5 range.

As shown in Table 4, only a few of the tasks across the eight forms have the mean-square values outside of the acceptable range of .5 to 1.5. However, none of these values are larger than .2. This means although these tasks do not distort the measurement system, they are less productive for the construction of the measurement.

Table 5 summarizes the infit and outfit mean squares for each of the two ratings. As seen in the table, all of these values fall within the acceptable range of .5 to 1.5. This means that there were no noisy or muted ratings, and the two sets of ratings exhibited an acceptable rating pattern. Overall, this finding indicates a satisfactory degree of within-rating consistency (Eckes, 2019).

Table 5. Summary of Mean-Square Fit Statistics for the Ratings

Rating	Form 1		Form 2	
	Infit MnSq	Outfit MnSq	Infit MnSq	Outfit MnSq
1	1.24	1.15	1.21	1.14
2	1.24	1.14	1.2	1.13
	Form 3		Form 4	
	Infit MnSq	Outfit MnSq	Infit MnSq	Outfit MnSq
1	1.21	1.19	1.19	1.21
2	1.19	1.18	1.18	1.22
	Form 5		Form 6	
	Infit MnSq	Outfit MnSq	Infit MnSq	Outfit MnSq
1	1.19	1.18	1.23	1.17
2	1.18	1.16	1.22	1.17
	Form 7		Form 8	
	Infit MnSq	Outfit MnSq	Infit MnSq	Outfit MnSq
1	1.26	1.2	1.22	1.13
2	1.25	1.19	1.23	1.16

Confirmatory Factor Analysis

In order to examine the internal structure of the SLTI SOPI and respond to research question 4, confirmatory factor analysis (CFA) was conducted. Person ability Rasch estimates (Rasch logits) generated from the previous step were used as data for the CFA analysis. CFA is a special case of the general family of structural equation modeling, commonly used in many social sciences for investigating theory-derived structural relationships and hypotheses about a set of measured variables (Mueller & Hancock, 2008).

For SLTI SOPI, it is hypothesized that all six tasks in each form tap the same underlying construct, which is speaking ability. This means that the test data should support this hypothesized structure of the test, and single-factor CFA models should fit the data from the eight forms. This means that eight single-factor CFA models were tested for each of the forms. The Robust Maximum Likelihood (RML) was used as the method of model parameter estimation for these models.

For evaluating the fit of the CFA models in the current study, a profile of model fit tests and indices recommended by Hu and Bentler (1999) and Mueller and Hancock (2008) was used. Chi-square, with its degrees of freedom and *p*-value, was checked. For a good model fit, the chi-square should not be statistically significant at a .05 level. However, in large samples and complex models, a chi-square is usually significant and not very informative. For this reason, the following descriptive fit indices were also used: the standardized root mean square residual (SRMR < .08), the root mean square error of approximation (RMSEA < .06), and the comparative fit index (CFI > .95).

Figure 9 summarizes the CFA parameter estimates of all the eight forms of SLTI SOPI. All the estimates across the eight models were statistically significant.

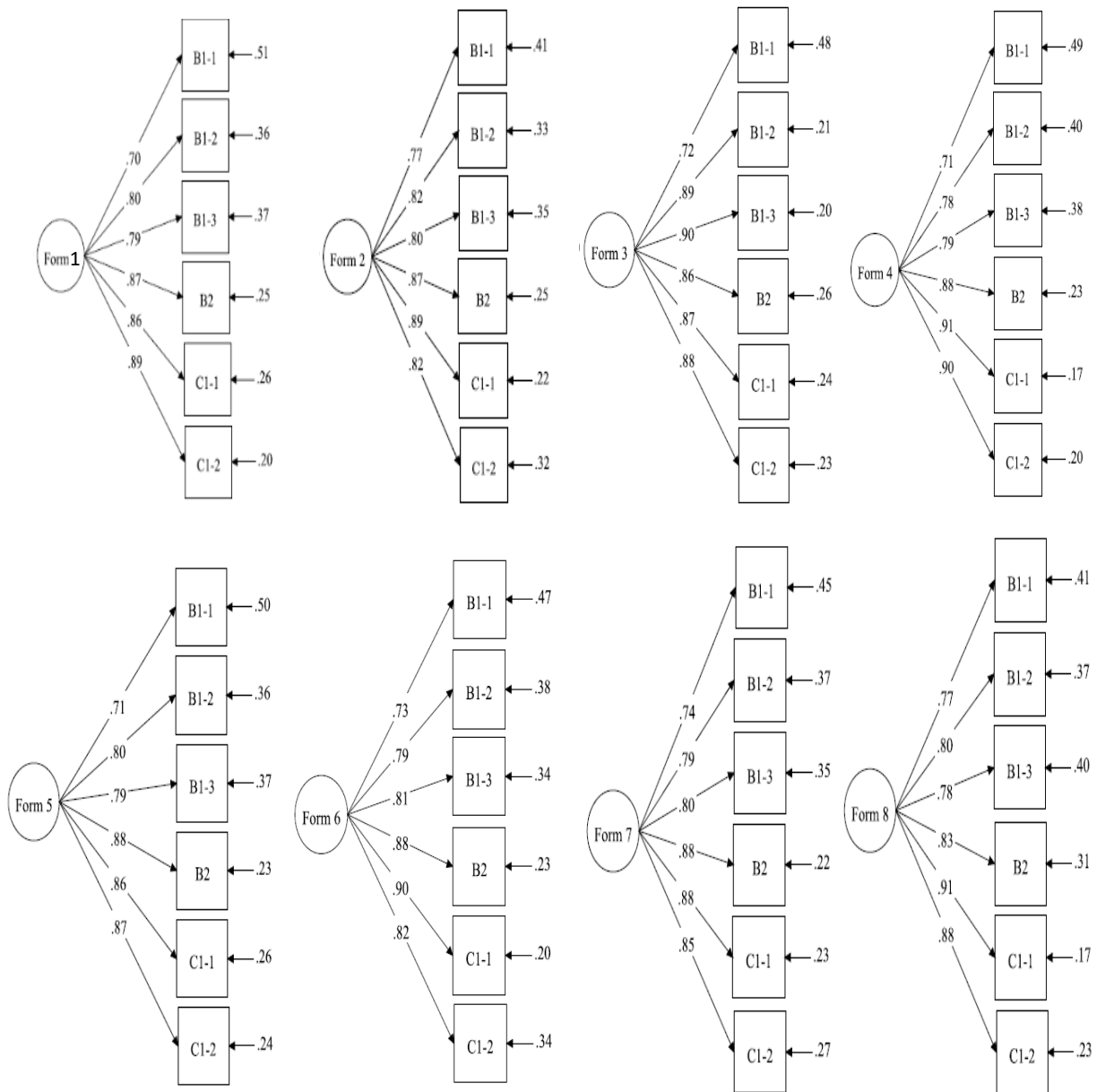


Figure 9. Single-Factor CFA Models with Parameter Estimates

Table 6 summarizes the fit statistics for the eight CFA models for the eight forms of the SLTI SOPI. As seen in the table, all the fit indices are within the acceptable range indicating that the single-factor model fit the data well. The chi-square for all the eight models was statistically

significant; however, as explained above, these significant results are probably due to the large sample sizes in the current study.

Table 6. CFA Fit Statistics

Form/Fit Statistics	CFI	RMSEA	SRMR	Chi-square
Form 1	.97	.04	.03	785.15 (9)
Form 2	.97	.04	.02	557.41 (9)
Form 3	.91	.05	.03	865.95 (9)
Form 4	.98	.02	.02	468.19 (9)
Form 5	.96	.03	.03	751.99 (9)
Form 6	.99	.02	.02	280.75 (9)
Form 7	.97	.05	.03	827.56 (9)
Form 8	.97	.02	.02	621.41 (9)

Discussion and Conclusion

The current study aimed at evaluating different psychometric qualities of SLTI SOPI. The result of this evaluation feeds into the validity argument for the usefulness of the SLTI SOPI score for its intended purpose. Currently, SLTI SOPI is primarily used by large-scale telecommunication companies to verify the minimum level of English speaking ability (CEFR B1) required for tasks in their customer-service call centers.

SLTI SOPI has six test tasks aiming to elicit spoken responses ratable in the range of B1 to C1 CEFR levels. The test has eight different forms to make administration of the test to a large number of test takers possible with minimal chance of re-taking the same test form multiple times. The current study was guided by four research questions.

A one-way ANOVA was conducted to respond to research question 1 (Are scores from the eight SLTI SOPI forms comparable in their score distributions? Or, are the eight forms at a comparable difficulty level?). The results of this analysis showed that the means and variances of scores across the eight forms were not statistically significantly different. This provides support

for the comparability of the difficulty level of the eight forms of SLTI SOPI. However, equating studies are still required for creating tables with corresponding scores across the eight forms, which allows the interchangeable use of scores from the different eight forms.

To respond to research question 2 (To what extent do the facets of examinee, tasks and ratings contribute to total score variance of SLTI SOPI? Or, are the SLTI SOPI scores reliable and dependable?), G-theory analysis was conducted. The results of this analysis suggest that the largest portion of variability in the SLTI SOPI scores is explained by differences in the speaking ability of the test takers. In other words, the SLTI SOPI scores reliably rank order and dependably categorize test takers based on their speaking ability, and in this rank-ordering and categorizing, the influence of other facets of measurement is negligible. In summary, the results of G-theory analysis indicate that SLTI SOPI scores are both reliable and dependably at a level acceptable for large-scale high-stakes tests.

Many-facet Rasch analysis was carried out to answer research question 3 (To what extent do SLTI SOPI tasks and ratings fit into a many-facet Rasch model? Or, do the tasks and ratings perform as expected by the test designers and examiners?). The results of this analysis showed that the test takers are at the expected levels of difficulty (at least three different levels of difficulty) and reliably categorized test takers into at least three different levels of speaking ability. Also, the results show that the two sets of ratings were reliably classified at only one level of severity, which means the two sets of ratings are interchangeable. Finally, all six tasks and two sets of ratings were productive for measurement of the speaking ability of test takers, except for a few tasks in the whole battery of the test that were not productive. However, none of the tasks were deemed to be disruptive for the measurement purposes.

Finally, CFA analysis was used to answer research question 4 (To what extent do SLTI SOPI data fit single-factor confirmatory factor analysis models? Or, do all the test tasks in each test form tap the same underlying construct (i.e. speaking ability)?). The results of the analysis showed that single-factor CFA models fit the data from all the eight forms. This means all six tasks within each of the eight forms of the SLTI SOPI loaded on a single latent factor, which is an indication of the unidimensionality of the test, and that all the SLTI SOPI tasks tap the same underlying construct, which is speaking ability.

All in all, the current study provided important supporting information for the validity argument that justifies the use of SLTI SOPI scores for their intended purpose. Future studies can provide further evidence to strengthen this validity argument. For example, linguistic analysis on the speaking performance of test takers on the SLTI SOPI and their speaking performance in the actual TLU domain of the test, which is a professional setting, reveals the amount of correspondence between the language performance of test takers on the test and in real world.

References

- Beglar, D., & Hunt, A. (1999). Revising and validating the 2000 word level and university word level vocabulary tests. *Language testing*, 16(2), 131-162.
- Boone, W. J., Staver, J. R., & Yale, M. S. (2013). *Rasch analysis in the human sciences*. Springer Science & Business Media.
- Brandt, J. (1991). The Hopkins Verbal Learning Test: Development of a new memory test with six equivalent forms. *The Clinical Neuropsychologist*, 5(2), 125-142.
- Crick, J. E., & Brennan, R. L. (1982). GENOVA: A generalized analysis of variance system (FORTRAN IV computer program and manual). *Dorchester, MA: Computer Facilities, University of Massachusetts at Boston*.
- Eckes, T. (2019). Many-facet Rasch measurement: Implications for rater-mediated language assessment. *Quantitative Data Analysis for Language Assessment Volume I: Fundamental Techniques*, 153.
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1–55.
- Kenyon, D. M., & Tschirner, E. (2000). The rating of direct and semi-direct oral proficiency interviews: Comparing performance at lower proficiency levels. *The Modern Language Journal*, 84(1), 85-101.
- Linacre, J. M. (1990). Many-faceted Rasch measurement.
- Mueller, R. O., & Hancock, G. R. (2008). Best practices in structural equation modeling. *Best practices in quantitative methods*, 488-508.
- Sawaki, Y., & Xi, X. (2019). Univariate 2 generalizability theory in language assessment. *Quantitative Data Analysis for Language Assessment Volume I:*

Fundamental Techniques, 30.

Stansfield, C. W., & Kenyon, D. M. (1992). Research on the comparability of the oral proficiency interview and the simulated oral proficiency interview. *System*, 20(3), 347-364.

Stansfield, C. W., & Kenyon, D. M. (1993). Development and validation of the Hausa Speaking Test with the ACTFL Proficiency Guidelines. *Issues in Applied Linguistics*, 4(1), 5-31.