Providing Evidence for the Generalizability of a Speaking Placement Test Scores

Payman Vafaee[1] , Behrooz Yaghmaeyan[2]

Received: 15 April 2015 Accepted: 10 August 2015

Abstract

Three major potential sources of construct-irrelevant variability in the test scores can be the tasks, rating rubrics and rater judgments. The variance caused by these facets of assessment in the test scores is a threat to the dependability, and in turn, generalizability of the test scores. The generalizability of the test scores need to be empirically investigated; otherwise, no evidence can support the generalizability inferences made based on the test scores. The current study employed univariate generalizability theory to investigate the different sources of variance in the test scores and the dependability of the scores obtained from the Columbia University placement (CEP) speaking test (N=144). Moreover, this study used multivariate generalizability theory to look at the dependability of the individual scores obtained from the four scales of the analytic rubric of the test. Finally, justifiability of combining scores from the four analytic rubric scales to make a composite score was investigated. The univariate results revealed that the dependability of the scores of CEP speaking test is high enough to be taken as a consistent measure of the speaking ability of the test takers. The multivariate results showed that the high correlation between the four rating scales of the test and their almost equal effective weights in the composite score makes it justifiable to combine the four scores of the four scales to report a composite score. The present study can contribute to the understanding of L2 assessment researchers of the application of generalizability theory in their reliability and generalizability investigations.

Key words: *Generalizability theory, Performance assessment, Reliability*

1. Introduction

The current study employed generalizability (G) theory to investigate the dependability of scores obtained from a placement speaking test. The test is administered at the Community English Program (CEP) at Teachers College, Columbia University. CEP provides instruction to adult learners of English as a second language (L2) and caters to a diverse student population. While the diverse background of CEP learners have enriched the program greatly, it has also posed a challenge to placing these learners into appropriate levels.

To make accurate placement decisions, CEP administers a comprehensive placement test consisting of five sub-sections which assesses learner’s English grammatical knowledge and

[1] Columbia University and University of Maryland (Corresponding Author), Email: payman.vafaee@gmail.com
[2] Payam-e-Noor University, Mashhad, Iran.

ability in four skills of listening, reading, writing, and speaking. A composite score is calculated using scores obtained from the five sub-sections, based on which and predetermined criteria (cut-off scores), learners are placed into appropriate levels. The placement decisions are absolute ones (i.e., learners are not placed with regards to their relative standing) with relatively low stakes. In case of misplacement, the level of learners can be modified in consultation with their teachers in the first week of classes.

In spite of low stakes for CEP placement decisions, in order to save time and resources, the *dependability* of CEP placement test scores still needs to be empirically investigated. To serve this purpose, the current study conducted a set of G theory analyses on the CEP speaking test scores. The results of these analyses can be used as evidence for (or against) the plausibility of assumptions underlying the *generalization* inference, which is one of the inferences in an interpretive argument connecting evaluation to explanation inference (Chapelle, Enright & Jamieson, 2008). Following building an interpretive argument by collecting evidence for and against the plausibility of assumptions underlying its inferences, a *validity* argument for the interpretation of test scores can be built (Kane, 2012). The following sections provide a more-detailed account of the background of the study. Then, the details of the study will be presented.

2. Review of Literature

2.1 Performance Assessment

Performance assessment is becoming increasingly commonplace in the fields of L2 acquisition and assessment (Bachman, Lynch & Mason, 1995; Brown, 1995, Lynch & McNamara, 1998; McNamara, 1991; Norris et al., 1998). The primary purpose of performance assessment is to present test takers with authentic test tasks which require them understand and produce authentic language samples similar to non-test language situations. Performance assessment aims to create a correspondence between test performance and non-test language use by creating test tasks which present and elicit authentic language samples. This fundamental aim of performance assessment can be partially achieved by incorporating features of non-test use into the process of test task design and administration (Bachman et al., 1995; Bachman, 2002; Norris et al., 1998).

Creating the close link between the test situation and authentic language use in performance assessment is thought to enhance the validity of the inferences made based on the test scores (Lynch & McNamara, 1998). Authenticity of the test tasks can be considered as a piece of evidence used in the validity argument for the interpretation of test scores. However, because the main purpose of the test tasks is to assess, authenticity of the test tasks and argument for the validity of inferences on content grounds is not sufficient. Test developers and test score users still need to demonstrate that scores derived from the test tasks are reliable measures of the test takers' language ability. For this reason, the reliability of the test scores need to be empirically investigated.

Reliability investigations produce evidence for or against the generalizability of the test scores, which is an important piece in the validity argument presented for the inferences made based on test scores. In L2 performance assessment, a threat to reliability, and therefore, validity of scores' interoperations is the variability in test scores caused by interactions between characteristics of the test takers and characteristics of the assessment procedure (Bachman, 2002). These interactions can yield construct-irrelevant variance in test scores, which is a major threat to the validity of performance assessment results (Messick, 1996). Two major potential sources of construct-irrelevant variance in performance assessment test scores are tasks and rater judgments (Bachman et al. 1995; Lynch & McNamara, 1998; McNamara & Adams, 1994; Shohamy, 1983, 1984). Therefore, before generalizing inferences made on the basis of test scores, these sources of variability (i.e., test tasks and rater judgments) must be investigated empirically.

A closely related issue to the variability in performance assessment test scores is the issue of rater consistency in using rating rubrics. Elicited language samples from learners are rated using different kinds of rating rubrics. Two commonly used kinds of rating rubrics in L2 performance assessment are *holistic* and *analytic* (Xi, 2007). The use of each of these rating rubrics has its own advantages and disadvantages.

By considering advantages and disadvantages of using holistic versus analytic rating rubrics, the use of analytic rubrics for the placement decisions is preferred. This preference is due to the fact that placement decisions need more diagnostic information, (Xi, 2007), and analytic rubrics serve this purpose better.

However, in order for the analytic scoring to be useful, raters must be trained to be able to reliably distinguish between different dimensions or components of performance as defined in an analytic rubric. Also, the test taking population must have enough variability in different dimensions of their performance to justify the use of more costly and complex analytic scoring system. More importantly, when analytic rubrics are employed, multiple scores are reported as a language ability profile, and then these scores are averaged or added together to create a composite score. However, empirical evidence should support the use of composite scores in four ways:

A) Empirical evidence should show an interrelationship among analytic rating scales. This interrelationship shows that scores from different domains of the rubric are related. B) On the other hand, there must be empirical evidence which indicates the rating scales are distinct enough to provide information about unique aspects of the test takers' language ability. C) When a composite score with equal weighting of individual rating scales is reported, there must be empirical evidence for the relative importance of different aspects of language ability, which in turn justifies creating a composite score with equal weighting. D) Ratings from each of the scales must be dependable (Sawaki, 2007).

To summarize, in the context of L2 performance assessment, in order to provide evidence of the generalizability of scores in a validity argument, the potential sources of variability in the test scores, which can be construct-irrelevant sources of variance, should be empirically investigated. Two major sources of variability, tasks and rater judgments, have been investigated in the field of language testing by employing different statistical approaches. One of these approaches is G theory (Brennan, 1983; Shavelson & Webb, 1991). G theory is a statistical and analytical approach for investigating the relative effects of variation in test tasks and rater judgments on test scores (Bachman et al., 1995; Lynch & McNamara, 1998). Moreover, G theory has been employed to investigate score dependability, dimension separability, and variability of score profiles when analytic scoring rubrics are used (e.g., Sawaki, 2007; Xi, 2007).

2.2. Generalizability (G) Theory

G theory (Brennan, 1983; Shavelson & Webb, 1991) has been extensively used in L2 performance assessment (Bachman et al., 1995; Jianda, 2007; Lee, 2006; Lynch & McNamara, 1998; Sawaki, 2007; Xi, 2007). By employing G theory, researchers can examine the degree by which the observed measurement, made under a specific set of conditions, will generalize to all other sets of similar conditions. For example, one can ask how an examinee's score obtained from carrying out a certain task, which is graded by a specific rater, will generalize to all other scores obtained from performing all other tasks graded by any other rater for this particular individual. Given the observed variation in a set of scores, G theory enables an L2 assessment researcher to estimate what the true or universe score would have been over an infinite number of observations under various conditions of different factors or facets of measurement.

In other words, G theory provides a means to fulfill the objectives that are not adequately addressed in classical test theory framework. A G theory analysis permits researchers to learn about multiple sources of variation in test scores simultaneously (Shavelson & Webb, 1991). By using G theory, researchers are able to estimate comparatively the sources of error variation and their interactions and calculate dependability coefficients that can be used for both relative and absolute decisions (Brennan, 2000).

Tasks and rater judgments are examples of different facets of measurement in a G theory analysis. Obviously, an infinite number of raters cannot assess the data on an infinite number of tasks. Therefore, the assumption is made that the actual sampling of selected conditions is representative of this infinite set or universe of admissible observations (Shavelson & Webb, 1991). G theory allows univariate and multivariate analyses.

A univariate G theory analysis is conducted at two stages: a G-study and a D-study. At the G-study stage, variance components for different sources of score variation are estimated. These variance components are used to compute two dependability coefficients: generalizability (G) and Phi coefficients (Φ), for relative and absolute decisions, respectively.

Generalizability (G) coefficient is used for relative decisions and indicates how reliable scores are for rank ordering the examinees when compared with one another. The Phi coefficient (Φ) is used to examine the dependability of the scores for absolute decisions. Phi coefficient (Φ) shows how consistent a test taker's scores are across different tasks, raters and rater-task combinations, when rank-ordering is not the prime concern of the assessment procedure (Brennan, 2001).

In addition, G-study stage provides a basis for the alternative D-studies. The results of D-studies inform researchers on the needed improvements in the present design of the assessment to create an optimal measurement design (Brennan, 2000; Shavelson & Webb, 1991).

In G-theory, the examinee is considered as the object of measurement and in univariate G-theory, the object of measurement has one universe score. In univariate G-theory, one universe score for the test takers (object of measurement) represents the source of variability in the test scores which arises from systematic difference among test takers in terms of their language ability.

In multivariate G-theory, in contrast to univariate G-theory, object of measurement has multiple universe scores. Therefore, in multivariate G-theory analysis, variances and covariances of multiple universe scores and sources of error are decomposed. The covariance components, which are exclusive to multivariate G-theory, indicate how examinees' multiple universe scores and errors co-vary (Brennan, 2001). One of the major applications of multivariate G-theory in L2 performance assessment research is for analyzing the dependability of composite scores of tests which are made up of scores from several content domains that are correlated (Sawaki, 2007 ; Xi, 2007).

3. The Study

The first purpose of the current study was to employ univariate G theory to investigate the relative effect of variation caused in CEP speaking placemen test scores caused by test tasks and rater judgments. Moreover, because in CEP placement speaking test, an analytic rubric is used, and a composite score is generated by adding individual scores obtained from different scoring scales of an analytic rubric, the second purpose of the current study was to use multivariate G theory to investigate the score dependability of the individual rating scales. In addition, the interrelationship among analytic rating scales was investigated to find out whether there is any empirical justification for creating a composite score by adding different scores from different rating scales. High correlations among the different scales of the rubric can be taken as evidence that scores from these different scales can be combined to create a composite score[3]. Also, for justifying the use of a composite score, the weighting of individual analytic scales to the composite score was investigated.

The present study addressed the following research questions:

Q1. What are the relative contributions of persons, tasks, raters, and their interactions to CEP speaking test score variance?

Q2. How dependable are scores obtained from CEP speaking test for absolute decisions?

Q3. To what extent do changes in the CEP speaking test design contribute to its score dependability?

Q4. To what extent are the analytic scores from CEP speaking test dependable?

Q5. To what extent are the universe (or true) scores for each of the scales of the analytic rubric of the CEP speaking test correlated?

Q6. To what extent do each of the CEP speaking rating scales contribute to the composite universe score variance?

Q7. Is it justifiable to combine individual analytic scores into a single composite score?

[3] It should be noted that high correlations among scores of different scales of an analytic rubric can also suggest that the costly use of this method of scoring is not justifiable, and a holistic rubric, which has the advantage of ease of implementation, can serve the same purpose. However, as explained before, because in the context of placement assessment, more diagnostic information is required, the use of analytic rubrics is preferred. Therefore, in the present study, the high correlations among scores of different scales will be interpreted as an evidence for justifiability of creating composite scores.

4. Method

4.1 Participant

One hundred and forty four non-native speakers of English who took the CEP speaking test in a regular administration of the test were the participants of the current study. These examinees made a diverse sample in terms of their age, native language, socio-economic status, educational background, immigration status. The majority of the examinees in this study were adult immigrants from the surrounding neighborhood or were family members of international students in the Columbia University community. In terms of their first language, a large percentage of them consisted of three languages: Japanese, Korean, and Spanish.

The raters in this study were 30 CEP teachers who were MA, EdM or EdD students in the TESOL and applied linguistics programs at Teachers College. They had varying degrees of ESL teaching and testing experience. Prior to the actual rating, the raters attended a norming session in which the test tasks and the rubric were introduced and sample responses were provided for practice.

4.2 Materials

The speaking section of the CEP test was used in this study. The test consists of three tasks. These tasks were designed to measure speaking ability under various real-life language use situations. In Task 1, two test takers act as interlocutors of a conversation in which one person is looking for an apartment for rent and the other one is the landlord. Their roles are assigned to them by the examiner, and the test takers start the conversation after reading a prompt which includes the topic and some questions for their conversation. In Task 2, the same two test takers should get involved in a conversation in which they should reach an agreement about a topic which is introduced to them via a card. In Task 3, instead of two, there are three interlocutors. They should express their opinion and try to persuade the others over a topic which is given to them via a card. In all three tasks, the test takers have one minute to plan their response. In Task 1 and Task 2, they have five minutes to talk, and in Task three, this time is ten minutes.

The speech samples from the three tasks were scored based on an analytic rubric that measured four components of speaking performance: meaningfulness, grammatical competence, conversational competence, and task completion. Each of these four components was measured on a 0-5 scale.

5. Procedure

A day prior to the administration of the CEP speaking test, the 30 raters attended a norming session in which the test tasks and the rubric were introduced and sample responses were provided for practice. Time was also given for discussion of analytic scores so that the raters had opportunities to monitor their decision-making processes by comparing the rationale behind their scores with other raters’ opinions. Rating practice and discussion continued until the raters felt that they were well aware of the tasks and confident with assigning scores on different rating scales.

In the following day, the 30 raters were divided into 15 pairs, so 144 examinee's responses for all the three tasks were double-rated. 10 pairs of the raters worked in one room, each rated pairs of examinees' performance on task 1 and task 2, and 5 pairs of raters worked in another room, each pair of raters rated different numbers of the three-examinee groups' performance on task 3. Therefore, for the first two tasks (two-interlocutor conversation), each of the two examinees' performance was rated by a pair of the raters. Afterwards, the examinees were sent to another room in which another examinee joined them for task 3 (three-interlocutor conversation), and their performance was rated by another pair of raters.

After the scoring session, data for all the 144 test takers from the three tasks each double-rated was entered into computer and got cleaned and prepared for analysis. There was no missing data.

6. Data Analysis

The computer program GENOVA (Crick & Brenna, 1983) was used to estimate the variance components and the score Phi dependability coefficient (Φ) in the univariate analysis. The computer program mGENOVA (Brennan, 1999) was used for the multivariate analysis to estimate the variance and covariance components and the reliability coefficient for the individual scores obtained from the four domains (scales) of the analytic rubric and composite scores.

For the univariate analysis, a two-facet crossed design ($p \times t \times r'$) with tasks (t) and ratings (r') as random facets was used. As in this study, each pair of examinees were not rated by only one pair of raters across all the three tasks, and since G-theory assumes that ratings and raters were drawn from a homogenous universe, and are therefore randomly parallel (Bachman et al., 1995), a balanced fully crossed design with ratings (r') instead of raters (r) was used. This univariate analysis was conducted to investigate the relative effects of tasks, ratings and their interactions to the CEP speaking test total score variance, and the dependability of the composite scores for the absolute decision making in the placement process.

For the multivariate analysis, a balanced two-facet crossed design with tasks and ratings as random facets ($p^{\bullet} \times t^{\bullet} \times r'^{\bullet}$) was used to estimate the variance and covariance components of four domains (scales) of the analytic rubric. In this design, persons (p), tasks (t) and ratings (r') were crossed with the four domains (scales) (ν) of the analytic rubric.

7. Results and Discussions

7.1 Univariate Analysis ($p \times t \times r'$)

Table 1 displays the estimated G-study variance components, standard error of estimated variances (SE), and percentage of each variance component contributing to the total score variance in the ($p \times t \times r'$) design.

Table 1 shows the seven variance components estimated for the ($p \times t \times r'$) design in this G-study.

Table 1. D-study variance components for univariate analysis ($p \times t \times r'$)

Effect	Variance component	Standard error	Total variance (%)
Person (*p*)	9.000	1.306	82
Tasks (T)	0.000	0.005	0
Ratings (R′)	0.000	0.002	0
*P*T	1.196	0.130	11
*P*R′	0.507	0.102	5
TR′	0.001	0.002	0
*P*T R′,e	0.330	0.027	2
Total	11.034		

The results of this section answer the first research question: *What are the relative contributions of persons, tasks, raters, and their interactions to CEP speaking test scores variance?*

Among the seven variance components, the largest variance component was the examinee variance component [σ^2 (*P*)] which accounted for 82 % of the total variance in the G-study. This tells us that, in the CEP speaking test, a relatively high proportion of the test score variance can be dependably associated with the test takers' ability in speaking ESL. Ideally the variance component associated with *Persons* should be by far the largest variance component in an assessment test which is the case for CEP speaking test. The huge variance component associated with persons in the CEP speaking test indicates that the test scores obtained for this test can be interpreted as a dependable score of the test takers' speaking ability.

The second largest variance component was the one for the examinee-by-task interaction [σ^2 (Pt)] which explained 11 % of the total variance. This indicates that a relatively large number of examinees were not rank-ordered consistently across different tasks. For person-by-task interaction, as the tasks are tapping into the same underlying language ability (speaking ability in this study), generally, a small variance component is expected. However, as tasks are usually context dependent, depending on the level of the familiarity of the test takers to the context of the tasks, the test takers may be rank-ordered differentially. This deferential rank ordering does not necessarily indicate that the difficulty levels of the tasks are different, but it can show that different test takers were familiar with the context of the test differentially.

Accordingly, in second language performance assessment, a moderately large variance component estimate for person-by-task interaction is not uncommon. However, in order to increase the dependability of the test scores obtained from the CEP speaking test, the three tasks should be closely scrutinized in order for necessary revisions which can make their context equally familiar to a general population of ESL test takers, and decrease the amount of variance in the test scores explained by person-by-task interaction. For example, in a G theory analysis conducted by Bachman et al. (1995) on a Spanish test used by the University of California for a study abroad program, the reported amount of variance explained by person-by-task interaction was negligible (only 1% of the total amount of variance).

The third largest interaction component was for the examinee-by-rating interaction [σ^2 (Pr′)] interaction variance which accounted for 5 % of the total variance suggesting that small portion of the examinees were not rank-ordered consistently across the two ratings. As this variance component did not account for much of the total score variance, it cannot be taken as a point of major concern; nonetheless, improving the norming session for the raters can help reduce this source of variance to an even smaller proportion. As the raters come from various backgrounds in ESL teaching and learning, a more comprehensive norming and training session can help making their scoring more consistent.

The variance components attributable to the other sources of measurement in the current study were negligible indicating they did not contribute much to the total score variance. This indicates, except for the examinee-by-task interaction and examinee-by-rating interaction sources of variance, the rest of the facet of the measurement in the CEP speaking test did not impose a threat to the reliability and dependability of the scores.

CEP speaking test is a criterion-referenced test based on its cores absolute decisions are made. Hence, in this study, the criterion-referenced dependability index which is known as Phi coefficient (Φ) was used as an indicator of the scores dependability which was 0.82, which can be interpreted analogously to a norm-referenced reliability coefficient. This results answers the second research question of present study: *How dependable are scores obtained from CEP speaking test for absolute decisions?*

According to the above mentioned results, 82 % of the observed score variance was universe score variance which is an indicator of the test takers’ speaking ability which for low stake tests like CEP placement test, is acceptable.

In addition to obtaining estimates of variance component estimates and dependability for the actual number of facets, 3 tasks and 2 ratings, Phi coefficient (Φ) was also estimated for different combinations of the number of conditions. Thus, Phi coefficient (Φ) was estimated for bigger and smaller number of tasks and ratings. Table 2 indicates Phi coefficient (Φ) for different combinations of numbers of tasks and ratings.

Table 2. Phi coefficient (Φ) estimates for different number of tasks and ratings

No. of tasks	No. of ratings	Φ
1	1	0.58
	2	0.64
	3	0.66
2	1	0.70
	2	0.76
	3	0.78
3	1	0.76
	2	0.82
	3	0.84

4	1	0.79
	2	0.84
	3	0.87
5	1	0.81
	2	0.86
	3	0.88
6	1	0.82
	2	0.88
	3	0.90

As can be seen in Table 2, decreasing the number ratings from 2 to 1 for each number of tasks decreases the dependability of the scores for 0.05 to 0.06. On the other hand, increasing the number of ratings for each number of tasks increases the dependability of the scores for only 0.02. According to the above results, it seems decreasing the number of ratings to 1, negatively impacts the dependability of the scores; however, increasing the number of the ratings to 3, does not contribute to the dependability of the test scores much.

Accordingly, if the number of the ratings is kept as two, the effect of the increasing and decreasing the number of tasks can be investigated. As Table 2 shows, the effect of increase in the number of tasks from 1 to 3 is substantial. Increasing the number of tasks from 1 to 2, enhances the dependability of the scores for 0.18. Phi coefficient (Φ) improves for 0.06 when the number of tasks increases from 2 to 3. However, the increase in the number of tasks from 3 to 4 and up to 6 started to have a diminishing return and increases the dependability of the scores for only 2 % for the addition of each extra task.

The result of this section answers the following research question: *To what extent changes in the CEP speaking test design can contribute to its scores' dependability?*

According to the results, as increasing the number if ratings from two and tasks from three does not contribute much to the dependability of the test scores, and the fact that the score dependability of the test is already above 0.80, the current design of the test with three tasks and two ratings seem to be the optimal design. However, as discussed in questions 1 and 2, some changes in the tasks and norming sessions are required.

For the purpose of the current study, a balanced two-facet multivariate study design was employed. In this design, the four domains of the analytic rubric were treated as the fixed facet (ν). The domains of the rubric were treated as the fixed facet because they are chosen selectively from among any other writing ability components for this test, and they are not interchangeable by any other components of writing ability in this test. In this design, as the systematic differences among the test takers with regard to their performance in the CEP speaking test was the primary interest, the test takers were treated as the object of measurement (P), and the two other facets of measurement were the tasks (t) and the ratings (r') which determined the unique test methods. The tasks and ratings were considered as the random facets because they are both assumed to be part of a larger pool of equally possible tasks and ratings for this test. The object of measurement (P) and the two random facets, (t) and (r'), were crossed with the fixed facet (ν) of this study which was the domains of the analytic rubric. We considered (P) to be fully crossed with (*t*) and (*r*′) and (*v*), since all the test takers received two independent scores for each of the four domains of the rubric for the three tasks of the test. Therefore, the design of the current multivariate generalizability study can be formally written as $p^{\bullet} \times t^{\bullet} \times r'^{\bullet}$, where the superscript filled $^{\bullet}$ designates that the object of measurement and the facets were crossed with the fixed facet. In this study, the task facet (*t*) had three conditions (three tasks). The rating facet (*r*′) had two conditions (two ratings), and the fixed facet (*v*), the rubric domains, had 4 conditions (four domains).

As Table 3 demonstrates, and according to the design of the study ($p^{\bullet} \times t^{\bullet} \times r'^{\bullet}$), in the current study, different variance components that can be associated to different facets of measurement and their interactions were estimated.

Generally, a large variance component estimate is expected for the object of measurement or persons (*P*). On the other hand, as we expect differences in the raters' severity level and task difficulty levels have very little effect on the test takers' scores, near-zero or very small values for the variance component estimates of these two facets of measurement is desirable. With regard to two-way interactions of the raters (person-by-rater and task-by-rater), small variance components are expected again, as large estimates of these variance components can be interpreted as inconsistency of the raters in rank-ordering the test takers in terms of their ability , or in their understanding of the tasks in terms of difficulty.

Table 3. Estimates of variance and covariance components

		Variance-covariance Component estimates				*% variance/covariance explained*			
		MEAN (1)	*GRAM COM (2)*	*CONV COM (3)*	*TASK COM (4)*	*MEAN (1)*	*GRAM COM (2)*	*CONV COM (3)*	*TASK COM (4)*
Person	(*P*)	0.54889				77			
(1)		0.51816	0.50336			82	79		
		0.58265	0.54836	0.63848		83	84	80	
(2)		0.56936	0.54335	0.60404	0.57823	84	86	84	67
(3)									
(4)									
Task	(*t*)	0.00000				0			
(1)		-0.00029	0.00000			0	0		
		-0.00033	-0.00006	0.00035		0	0	0	
(2)		-0.00140	-0.00029	0.00024	0.00000	0	0	0	0
(3)									

(4)									
Ratings	(r')	0.00000				0			
(1)		-0.00047	0.00000			0	0		
		-0.00040	-0.00027	0.00000		0	0	0	
(2)		-0.00008	-0.00068	-0.00057	0.00000	0	0	0	0
(3)									
(4)									
Pt		0.06610				9			
(1)		0.06472	0.05000			10	8		
		0.07306	0.05234	0.07874		10	8	9	
(2)		0.08377	0.06607	0.08059	0.16073	12	10	11	19
(3)									
(4)									
Pr'		0.04145				6			
(1)		0.04069	0.03314			6	5		
		0.03339	0.03528	0.03686		5	5	5	
(2)		0.00761	0.02093	0.03037	0.05966	1	3	4	7
(3)									
(4)									
tr'		0.00000				0			
(1)		-0.00001	0.00000			0	0		
		0.00002	-0.00002	0.00000		0	0	0	
(2)		0.00044	0.00003	0.00001	0.00100	0	0	0	0
(3)									
(4)									
Prt,e		0.05515				8			
(1)		0.00898	0.04831			2	8		
		0.00962	0.01786	0.04739		2	3	6	
(2)		0.01190	0.00093	0.00683	0.06767	3	1	1	7
(3)									
(4)									

A more detailed look at the results of this section, as shown in Table 3, answers the following research question: *To what extent are the analytic scores from CEP speaking test dependable?*

Phi coefficient (Φ), which is an index for the percentage of variance explained in test scores by the object of measurement or persons (*P*), was calculated. Phi coefficient (Φ) can be interpreted analogously to a norm-referenced reliability coefficient. These values can be seen at the right top corner of Table 3. The diagonal and off-diagonal values show the percentage of variance and covariance explained in the test scores by Persons (*P*) facet of measurement, respectively. Phi coefficient (Φ) for the individual rating scales for the CEP speaking test ranged from 0.67 to 0.80 (67 % to 80%), with conversational competence with the highest and task completion with the lowest values. The other two rating scales, meaningfulness and grammatical competence, had Phi coefficient (Φ) of 0.77 and 0.79 (77% to 79%), respectively. The results show that task completion scale had a relatively low dependability compared to the other rating scales. Task completion accounted for only 67 % of the proportion of the scale's variance accounted for by the speaking ability differences in the test takers. The low Phi coefficient (Φ) for task completion demonstrates that task completion scale falls short of consistently rank ordering the test takers based on their speaking ability.

As the low task completion dependability index value is very different from the rest of the rating scales, it can be understood that different test takers with similar levels of meaningfulness, grammatical competence, and conversational competence were rank-ordered differentially in terms of their effort to complete the tasks. The lower dependability of the scores from task completion scale can also be understood by lager variance in the test scores associated with this scale in terms of person-by-task, and person-by-ratings interactions. This shows tasks rank ordered the test takers in terms of task completion less consistently, compared to the other scales. It also shows, ratings had more inconsistency in rank ordering the test takers in terms of task completion.

A part of the inconsistency in test takers rank ordering depending on which tasks they took, suggests that depending on the test takers familiarity with the context of the tasks and independent of their ability in the other components of speaking ability, test takers were rank ordered differentially. Therefore, one of the changes in the tasks which were suggested in questions 1, 2 and 3, can be some changes in the context of the tasks. If test takers are equally familiar with the context of the tasks, they will, most probably, be able to complete the tasks more consistently. With regard to inconsistency of the raters in rank ordering of the test takers based on the task completion scale, two points can be suggested. Firstly, as a part of the suggestions for questions 1, 2 and 3 for better training and norming sessions for the raters, task completion scale can be more focused on. In the norming session, task completion scale can be discussed more in depth, and more samples can be provided to help raters have a more homogenous understanding of this scale of the rubric. Secondly, the descriptors of task completion scale on the rubric can be revised. Probably more clear and straightforward descriptors for this scale can help the raters to be more consistent with it.

Covariance components in the current study provide information about how facets of measurement are contributing to variability of the test scores obtained between different domains of the rubric. The off-diagonal values on the top left section of Table 3 are the covariance components for the object of measurement or persons (*P*). On the top left section of Table 3, off-diagonal values show the percentage of covariance explained by the object of measurement or persons (*P*). Similar to variance components, covariance components for the object of measurement or persons (*P*) were the largest among other sources of variation. The high and positive covariance components for the persons which ranged from 0.51 to 0.60 (82 % to 86 %) indicate that when candidates scored high on a given domain, they also scored high on another as well.

Following covariance components for persons, is the covariance components for person-by-task interaction which was the second largest. Covariance component associated with the person-by-task interaction ranged between 0.05 to 0.08 (8 % to 12 %). This relatively large covariance component for the person-by-task interaction, which was predicted due to relatively large variance components, indicates that there were some similarities in the patterns of rank-ordering differences across the four domains.

Covariance components for person-by-rankings interaction were also sizable, ranged between 0.00 to 0.04 (1 % to 6%). This relatively large covariance component for the person-by-rankings interaction indicates that there were some similarities in the patterns of rank-ordering differences by the ratings across the four domains.

The residual covariance components were small, ranged between 0.00 to 0.01 (1 % to 3 %). This small covariance component shows small variations in the test scores as the result of systematic and unsystematic sources which are not accounted for by the facets of measurement in this study. As the variance components for the other facets of measurement and their interactions (tasks, ratings, and task-by-ratings interaction) were negligible, the covariance components for them were negligible as well.

The results of this section will be used to answer the following research question: *To what extent are the universe (or true) scores for each of the scales of the analytic rubric of the CEP speaking test correlated?*

The high and positive covariance components for the persons which ranged from 0.51 to 0.60 (82 % to 86 %) indicate that when candidates scored high on a given domain, they also scored high on the others as well. There is no significant difference in the covariance components between different domains suggesting that all of them are highly correlated, but as the correlations among them are not extremely strong, they are still distinct indicators of speaking ability. The high correlations among the four scales can be taken as a sign of the fact that they are all measures of the same underlying ability. Although task completion was a less dependable scale in rank ordering the test takers in terms of their speaking ability, it still correlated highly with the other three scales. The high correlation of task completion with the other scales suggest that task completion is a measure of speaking ability; however, with lower dependability.

The Phi coefficient (Φ) for the composite or total, which is the ratio of the composite universe-score variance to the sum of itself and the composite absolute-error variance (Φ = 0.56/

(0.56 + 0.13) was calculated which was equal to 0. 82. This result is equal to the Phi coefficient (Φ) calculated for the total score in the univariate analysis.

Table 4 shows the effective weights for the four rating scales for the universe score variance and the absolute error variance. The weights assigned to each of the rating scales by the test designers are called nominal weights. However, the nominal weights can be different from the effective weights which show the degree to which individual rating scales empirically contribute information to the composite score (Sawaki, 2007).

Table 4. Composite score analysis results

Contributions to :	Variance and covariance components			
	Meaningfulness	Grammatical competence	Conversational competence	Task completion
A priori weights Effective weights contributing to :	0.25	0.25	0.25	0.25
Universe score variance (%)	25	23	26	26
Absolute error variance (%)	24	22	25	29

In the CEP speaking test, an equal weight of 25 % is assigned to each of the four rating scales. Nonetheless, the effective weights for each of the rating scales should be obtained to investigate the extent to which each of the rating scales empirically contribute information to the composite score.

The effective weights were obtained separately for the composite universe-score and absolute-error variances as part of the D study in the multivariate G theory analysis. The effective weights of a give rating scale for the composite true score (or absolute- error) variance is calculated by: (1) the nominal weights for each of the rating scales which was 0.25 in the current study, (2) universe-score (or absolute- error) variance for the rating scale, and (3) the universe score (or absolute-error) covariance of the rating scale with the others (Brennan, 2001).

The results show that meaningfulness, grammatical competence, conversational competence and task completion accounted for 25 %, 23%, 26% and 26 % of the composite universe-score variance respectively.

Results of this section address the fowling research question: *To what extent each of the CEP speaking rating scales contribute to the composite score variance?*

The effective weights for the four rating scales for the universe score variance and the absolute error variance was calculated. The results showed that the four scales accounted for almost equal amount of the universe-score variance in the composite score. Such results suggest that each of the four rating scales contributed relatively similar amount of information to the composite universe-score variance. This similarity in the relative contribution of the four scales to the composite universe-score variance indicates that the four scales work equally well to provide information about the composite universe-score. Moreover, as the nominal weights and effective weights were almost equal in this study, it can be concluded that the same nominal weights can be retained, unless the test users want to give more weight to any of the rating scales.

The final research question was: *Is it justifiable to combine individual analytic scores into a single composite score?*

According to the results of the present study discussed for questions five and six, the universe scores obtained from the four rating scales are highly correlated, and all of the scales contributed almost equal amount of information to the composite universe-score. The high correlation between the universe scores and similar amount of their contribution to the composite universe score suggests that the four scales are measuring the same underlying ability which is the speaking ability in present study. Therefore, it can be concluded that it is justifiable that to combine the scores from the four rating scales to report a composite score for the test takers' speaking ability in the CEP test.

8. Conclusion

The current study investigated the different sources of variance in the test scores and the dependability of the scores obtained from the CEP speaking test. Moreover, this study looked at the dependability of the individual scores obtained from the four scales of the analytic rubric of the CEP speaking test. Finally, justifiability of combining scores from the four analytic rubric scales to make a composite score was investigated.

According to the results of this study, the dependability of the scores of CEP speaking test is high enough to be taken as a consistent measure of test takers speaking ability. Certain changes in the tasks and ratings of this test can render even higher dependability in its scores. The results of the current investigation provide the generalizability piece of validity argument for the CEP speaking test.

In addition, the results of the current investigation revealed that task completion is a less consistent measure of the speaking ability in the CEP placement test. Therefore, some improvement in the tasks, rubric descriptors and rater trainings are called for.

High correlation between the four rating scales and their almost equal effective weights in the composite score makes it justifiable to combine the four scores of the four scales to report a composite score. More importantly, high correlation between the four rating scales provides another piece of evidence for the validity argument presented for the CEP speaking test. This high correlation shows the four rating scales are measuring the same underlying ability which is speaking ability in this test.

Acknowledgements
We would like to thank Professor Kirby Grabowski for teaching us the principles of language testing and the application of Generalizability theory in the context of language testing. Her extensive and constrictive feedback also improved our manuscript. All errors and omissions are, of course, our own.

References

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing*, *19*, 453-476.

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing, 12,* 238–57.

Bachman, L. F., & Savignon, S. J. (1986). The evaluation of communicative language proficiency. *Modern Language Journal*, *70,* 380-390.

Brennan, R. L. (1983). Elements of generalizability theory. Iowa City, Iowa: ACT Publications.

Brennan, R. L. (1999). Manual for mGENOVA. Iowa Testing Programs Occasional Papers Number 47.

Brennan, R. L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement*, *24*(4), 339– 353.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Brown, A. (1995). The effect of rater variables in the development of an occupation specific language performance test. *Language Testing*, *12*(1), 1-15.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). Test score interpretation and use. *Building a validity argument for the Test of English as a Foreign Language*, 1-25.

Crick, J. E., & Brennan, R. L. (1983). Manual for GENOVA: A generalized analysis of variance system (ACT Technical Bulletin No. 43). Iowa City: American College Testing.

Douglas, D., & Smith, L. (1997). *Theoretical underpinnings of the Test of Spoken English revision project* (TOEFL Monograph Series MS-9). Princeton, NJ: Educational Testing Service.

Grabowski, K. (2009). *Investigating the construct validity of a test designed to measure grammatical and pragmatic knowledge in the context of speaking*. (Unpublished doctoral dissertation). Teachers College, Columbia University, New York.

Henning, G. (1996). Accounting for nonsystematic error in performance ratings. *Language Testing*, 13, 53-61.

Jianda, L. (2007). Developing a pragmatics test for Chinese EFL learners. *Language Testing, 24,* 391-415.

Kane, M. (2012). Validating score interpretations and uses. *Language Testing*, *29*(1), 3-17.

Kondo-Brown, K. (2002). An analysis of rater bias with FACETS in measuring Japanese L2 writing performance. *Language Testing*, *19*, 1-29.

Lee, Y. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language Testing*, *23*, 131- 166.

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*, 158–80.

McNamara, T. F. (1991). Test dimensionality: IRT analysis of an ESP listening test 1. *Language Testing, 8*, 139-159.

McNamara, T. F., Adams, R. J. (1994). Exploring rater characteristics with Rasch techniques. In

Selected papers of the 13th Language Testing Research Colloquium (LTRC). Princeton, NJ: Educational Testing Service.

Mehrens, W. A. (1992). Using performance assessment for accountability purposes. *Educational Measurement: Issues and Practice, 11*, 3-9.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), 13-23.

Messick, S. (1996). Validity and wash back in language testing. *Language Testing*, *13*, 241-56.

Miller, M. D., & Legg, S. M. (1993). Alternative assessment in a high stakes environment. *Educational Measurement: Issues and Practice, 12*, 9-15.

Norris, J. M., Brown, J. D., Hudson, T., & Yoshioka, J. (1998). *Designing second language performance assessments*. (Vol. SLTCC Technical Report #18). Honolulu: Second Language Teaching and Curriculum Center, University of Hawaii at Manoa.

Sawaki, Y. (2007).Construct validation of analytic rating scales in a speaking assessment: Reporting a score profile and a composite. *Language Testing*, *24*, 355-390.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A prime*. Newbury Park, CA: Sage.

Shohamy, E. (1983). The stability of oral proficiency assessment on the oral interview testing procedure. *Language Learning*, *33*, 527-40.

Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing, 1*, 147-80.

Shohamy, E. (1995). Performance assessment in language testing. *Annual Review of Applied Linguistics*, *15*, 188-211.

Weir, C. (1990). *Understanding and Developing Language Tests*. London: Prentice Hall. City, Iowa: ACT Publications

Xi, X. (2007).Evaluating analytic scoring for the TOEFL® Academic Speaking Test (TAST) for operational use. *Language Testing, 24,* 251-286.