CaMLA Working Papers

2015–01

A Practical Guide to Investigating Score Reliability under a Generalizability Theory Framework

Chih-Kai Lin
Center for Applied Linguistics (CAL)
United States

A Practical Guide to Investigating Score Reliability under a Generalizability Theory Framework

Author

Chih-Kai Lin
Center for Applied Linguistics (CAL)

About the Author

Chih-Kai (Cary) Lin is a Psychometrician at the Center for Applied Linguistics (CAL). He has a PhD in Educational Psychology from the University of Illinois at Urbana-Champaign (UIUC), with a research specialization in educational measurement. His research focuses on psychometric approaches to supporting operational testing programs. He is also interested in the use of multilevel modelling in teacher performance evaluation. The research reported in this paper was conducted during his graduate studies at UIUC.

Table of Contents

Table of Contents

Abstract

The current study aims to compare the precision of two analytical approaches to estimating score reliability in performance-based language assessments. The two methods operate under a generalizability theory framework and have been successfully applied in various language assessment contexts to deal specifically with sparse rated data. Given the advantages of working with fully crossed data, the two methods were designed to transform a sparse dataset into variants of fully crossed data. The rating method conceptualizes individual ratings, irrespective of the raters, as a random facet. The rater method identifies all possible blocks of fully crossed subdatasets from a sparse data matrix and estimates score reliability based on these fully crossed blocks. Results suggest that when raters are expected to have similar score variability, the rating method is recommended for operational use given that it is as precise as the rater method but much easier to implement in practice. Nevertheless, when raters are expected to have varying degrees of score variability, such as a mixture of novice and seasoned raters rating together, the rater method is recommended because it yields more precise reliability estimates. Informed by these results, the current study also demonstrates and carries out a step-by-step analysis plan to investigate the score reliability of the speaking component of the Examination for the Certificate of Proficiency in English (ECPE).

Background

Expert rated assessments of actual test performances are common in a plethora of contexts, such as academic departments at universities that rely on placement tests to assess incoming students, regional and national governments that administer achievement tests to measure student growth, and large-scale testing programs that offer academic and workplace qualifications. The popularity of performance-based tests is partly driven by validity concerns regarding the extent to which assessment tasks resemble real-world tasks and the degree to which test performances can be safely generalized to non-test contexts, which are in accord with the modern paradigm of test validation (Kane, 2006; Messick, 1989).

Given the emphasis on performance tests, rater-mediated measurement has become typical in many assessment contexts. Many testing programs continue to rely on a time-honored scoring paradigm: expert raters with rigorous training and calibration. However, scoring test performances by human raters comes with a set of stress factors. For example, even in a well-designed rating system, certain practical realities might mitigate the effectiveness of rater training, such as time pressure due to a short turnaround timeline for scoring. Furthermore, some raters may resign or be ill, forcing test administrators to use a smaller pool of trained raters or to turn to a wider pool of former raters, some of whom have not been fully or recently recalibrated. All of these factors result in score fluctuations for reasons other than the intended construct being measured and thereby affect the reliability of human scoring.

Reliability of rater-mediated measurement here is not interpreted as the internal consistency of items/tasks in a typical item analysis—the degree to which items/tasks correlate with each other and jointly measure a defined construct (Allen & Yen, 2001), nor is it conceptualized as a layperson's interpretation of trustworthiness—the extent to which the measurement is accurate (Ennis, 1999). Rather, reliability in rater-mediated measurement is about the extent to which raters are consistent in giving scores across the objects of measurement (e.g., examinees) according to a rating rubric (Stemler & Tsai, 2008). Rater-mediated measurement is a product of raters' understanding of the intended construct being measured, their interpretations of the rating rubric, and their use of the rubric in making their judgments. High inter-rater reliability is desirable so that raters can be considered interchangeable; that is, a score awarded would not be contingent upon any specific rater who is assigned to make the judgment.

In light of the fact that the utility of any rated measurement is contingent upon its score reliability, generalizability theory or G theory (Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972) provides a powerful analytical framework that allows investigators to assess the relative magnitude of construct-irrelevant variability and to factor these variations into the estimation of score reliability. G theory is a random facet measurement model which conceptualizes an observed score as a composite of various sources, or *facets* in G-theory terminology, in addition to the objects of measurement. Essentially, G theory decomposes total score variability, via analysis of variance (ANOVA)

techniques, into variance components associated with the objects of measurement and with various facets involved in the measurement. In other words, G theory conceptualizes observed score variability as a linear combination of the true variation in the objects of measurement and other variations pertaining to different measurement sources that are anticipated by or of interest to an investigator. For instance, in a speaking exam for a group of English as a second language (ESL) students, the objects of measurement are students' oral proficiency levels, and one potential source of measurement variation is score variability introduced by different raters scoring the spoken responses. Ideally, one would like to see true differences in students' oral proficiency reflect observed score variability as much as possible, not differences among rater severity/leniency.

Sparse Data as a Given in Operational Settings

The full potential of G theory is realized when fully crossed designs are employed. For example, a fully crossed ($p \times r$) design requires that each response or person (p) be rated by all available raters (r). The design is ideal in that it allows G-theory analysis to separately assess variability due to the main and interaction effects of the objects of measurement and the facet(s) of interest, resulting in a more straightforward interpretation of variance components corresponding to the main and interaction effects, which in turn aids the interpretation of score variability. The relationship between variance components and score reliability can be illustrated by a one-facet random effect model under the G-theory framework:

$$X_{pr} = \mu + \alpha_p + \beta_r + \varepsilon_{pr,e} \quad (1)$$

where the speaking score (X_{pr}) of person p given by rater r is the sum of an overall mean (μ) and the three components pertaining to persons (α), raters (β) and errors (ε). Observed score variability due to the three random components is represented by the estimated variance components $\hat{\sigma}_p^2$, $\hat{\sigma}_r^2$, and $\hat{\sigma}_e^2$, respectively. Generally, score reliability is interpreted in an absolute sense (Brennan, 2001) in performance-based assessments because the rating rubrics on which examinee responses are scored are usually criterion-based, describing the skills and performances associated with different levels of proficiency. Given the absolute interpretation of score reliability, the estimated phi-coefficient is computed under the G-theory framework as follows:

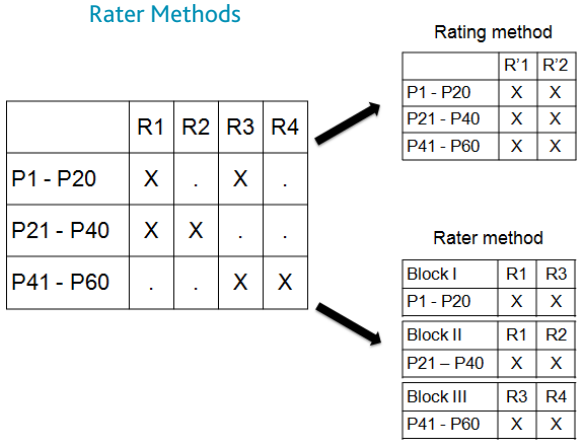
$$\text{phi} - \text{coefficient}\ (\hat{\Phi}) = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \frac{\hat{\sigma}_r^2}{n_r'} + \frac{\hat{\sigma}_e^2}{n_r'}} \quad (2)$$

where n_r' refers to the number of raters or ratings given to each response. From Equation (2), one can observe an inverse relationship between score reliability and score variability due to the measurement facets; that is, if all else is equal, the higher the estimated variance components associated with raters and/or errors are, the lower the estimated phi-coefficient becomes. This relationship is clear when the variance components can be estimated independently of one another, which is the main advantage of working with fully crossed datasets. Nevertheless, in an operational speaking-assessment setting, fully crossed designs are not practical, if not impossible, due to the tremendous scoring load for each rater if such ideal designs were to be implemented. Alternatively, many testing programs resort to a double-rating design, where each spoken response is rated by any two qualified raters with possible score adjudications from a third qualified rater if the discrepancy between the first two ratings is large.

In light of the advantages of working with fully crossed designs, two methods have been applied under the G-theory framework in performance-based language assessments. The two methods take sparse data as a given and transform the sparse datasets into some variants of fully crossed ones. First, *raters* are treated as a random facet (e.g., Xi, 2007); henceforth referred to as the rater method. Second, *ratings* are treated as a random facet (e.g., Bachman, Lynch, & Mason, 1995; Huang, 2012; Lee, 2006); henceforth referred to as the rating method. Figure 1 gives a visual representation of how the two methods break down a hypothetical sparse dataset, in which each response from sixty persons/examinees (P1–P60) is double-rated among a panel of four raters (R1–R4), into fully crossed dataset(s). As such, the rater method first identifies a total of three blocks of fully crossed subdatasets in this example. Next, variance components are to be estimated within each block (see Shavelson & Webb, 1991, p.29 for variance-component estimates). These variance-component estimates are then averaged across the three subdatasets by giving weights according to the number of examinees in each block (Chiu, 2001; Chiu & Wolfe, 2002). Finally, score

reliability is calculated based on the average estimated variance components since variance-component estimates are the building blocks of score reliability. The rating method forces a sparse dataset into a fully crossed one by treating individual ratings, irrespective of which raters, as a random facet. For example in Figure 1, the rating method transforms the 60-by-4 sparse data matrix into a 60-by-2 fully crossed dataset. The variance components and score reliability are then estimated based on the transformed fully crossed data.

Figure 1. A Visual Representation of the Rating and Rater Methods



	R1	R2	R3	R4
P1 - P20	X	.	X	.
P21 - P40	X	X	.	.
P41 - P60	.	.	X	X

	R'1	R'2
P1 - P20	X	X
P21 - P40	X	X
P41 - P60	X	X

Block I	R1	R3
P1 - P20	X	X
Block II	R1	R2
P21 – P40	X	X
Block III	R3	R4
P41 - P60	X	X

Motivation for the Current Study

Investigators have both the rating and rater methods at their disposal in examining score reliability in performance-based assessments. Both methods have been applied in the field of language testing, and both seem to be satisfactory for the purpose of estimating score reliability. The two methods may yield similar estimates of score reliability from the same dataset in some operational contexts; however, given that the two methods differ not only in the specification of the random facet but also in the estimation procedures of variance components, the results based on the two methods may not always converge. When the estimates of score reliability differ, a natural follow-up question is which estimate to report. In the absence of true score reliability in operational contexts, choosing the higher estimate may run the risk of falsely inflating score reliability when in fact the lower estimate is more precise, whereas choosing the lower estimate may unduly underestimate score reliability when the higher estimate is actually more precise. This is an operationally driven question, but it cannot be answered empirically using operational data at hand because true score reliability is not known from operational data, and therefore an investigator has no way of knowing which estimate based on the two methods is more reflective of the true reliability.

Method

To address the issue of not being able to operationally determine the precision of different reliability estimates based on the rating and rater methods, a Monte Carlo simulation study was conducted to compare the estimation precision of the two methods. The aims of the simulation study are twofold. First, it seeks to evaluate the precision of the rating and rater methods in estimating score reliability under various simulated conditions, whose designs are informed by operational contexts. Second, results from the simulation study serve to guide the analysis plan for investigating score reliability of a large-scale language speaking assessment, the speaking component of the Examination for the Certificate of Proficiency in English (ECPE) developed by CaMLA (Cambridge Michigan Language Assessments).

Simulated Conditions

Three target sample sizes (n_p) (50, 100, and 200), three numbers of raters (n_r) (4, 8, and 16), two compositions of variance components, and two scenarios of rater score variability were chosen; hence, a total of 36 conditions were considered in the simulation study. The two variance-component (VC) compositions were: (a) 65%, 5%, and 30% of total score variance were accounted for by persons, raters and errors, respectively, and (b) 25%, 35%, and 40% of total score variance were due to persons, raters, and errors. The two rater scenarios were: (a) all raters exhibited similar variability in their scoring, corresponding to raters having similar training and/or rating experience, and (b) some raters had greater score variability than the others, reflecting realistic settings in which a mixture of novice and experienced raters were deployed in a single rating session.

The relative magnitudes of the variance components for VC composition (a) were informed by previous G-theory research on speaking assessments (Akiyama, 2001; Bachman et al., 1995; Lynch, & McNamara, 1998; Xi, 2007), in which a large proportion of score

variability was usually due to persons, a small proportion of score variability was accounted for by raters, and some variability was expected to be attributable to measurement error. It should be noted that in a simulation study, true parameters are to be selected from values that seem reasonable according to previous research (Mooney, 1997). Some G-theory simulation studies adopted values from a single empirical study (e.g., Nugent, 2009). The current study attempts to arrive at reasonable parameters for variance components by taking the average of total score variance across multiple empirical studies. The average total score variance across the aforementioned studies was 1.123. Given the VC composition (a), this translates to 0.7300 (65%) for σ_p^2, 0.0561 (5%) for σ_r^2, and 0.3369 (30%) for σ_e^2. In published research, the relative magnitude of score variability attributed to raters was usually small due to rigorous rater training. Given that, it would be informative for the current simulation study to also consider situations in which raters are not fully trained and are therefore likely to exhibit a larger relative magnitude of variance component. VC composition (b) mirrors such a context, where σ_p^2 = 0.2808 (25%), σ_r^2 = 0.3930 (35%), and σ_e^2 = .4492 (40%).

Data Generation

Data associated with rater scenario (a) were simulated according to Equation (1). Take VC composition (a) as an example. The speaking score (X_{pr}) of person p given by rater r was the sum of an overall mean (μ) and the three random components pertaining to persons, raters and errors. These three random components were generated independently from three normal distributions, where the person effect (α_p), the rater effect (β_r), and the error component ($\varepsilon_{pr,e}$) followed a normal distribution with a mean of zero and variance of σ_p^2 = 0.7300 (65%), σ_r^2 = 0.0561 (5%), and σ_e^2 = 0.3369 (35%), respectively. The true score reliability (or phi-coefficient) is then calculated by plugging the true parameters for these variance components into Equation (2). The same procedures were applied to generate data for VC composition (b), except that the three random effects followed a normal distribution with a mean of zero and variance of σ_p^2 = 0.2808 (25%), σ_r^2 = 0.3930 (35%), and σ_e^2 = 0.4492 (40%), respectively. Data were simulated to be scored on a scale of 0 to 4 by setting the overall mean (μ) at 2. Given the current setup for data generation, the true score reliability for VC composition (a) is expected to be higher than that for VC composition (b), which allows the simulation study to evaluate the precision of the rating and rater methods in estimating low and high score reliability.

Data associated with rater scenario (b) were also simulated according to Equation (1) for VC compositions (a) and (b). Nevertheless, what was different from rater scenario (a) lay in the true parameter for the rater variance component, such that the scoring variability for novice raters was simulated to be 2 times larger than that for experienced raters. Two raters were designated as novice raters across all simulated conditions under rater scenario (b); as a result, novice raters constituted 50%, 25%, and 12.5% of the raters for n_r = 4, 8, and 16, respectively.

Next, two constraints were imposed on the data generation to create sparseness in the simulated data. First, each examinee response was assigned to two raters only. Second, all raters shared an equal amount of scoring load. As a result, the levels of sparseness were directly linked to the numbers of raters (n_r) in the simulated conditions. Take n_p = 200 and n_r = 16 as an example. A fully crossed 200-by-16 dataset with complete data was first generated. The first examinee was randomly assigned to two raters out of the sixteen raters, and therefore the simulated data for this examinee associated with the other fourteen raters were removed. The next examinee was randomly assigned to two raters, and so on until the constraint of equal scoring load for each rater was met. This resulted in a sparse level of 87.5%, leading to eight 25-by-2 crossed subdatasets under the rater method and one 200-by-2 crossed dataset under the rating method. The three numbers of raters—that is, 4, 8, and 16—corresponded to sparseness levels of 50%, 75%, and 87.5%, respectively.

Evaluation of Estimation Precision

The estimated score reliability (or phi-coefficient) based on the rater and rating methods was evaluated against the true score reliability with respect to average bias based on 1,000 replications for each of the 36 simulated conditions. Bias here is defined as the degree to which an estimate deviates from its true parameter; hence, the lower the bias is, the higher the estimation precision will be. For a true phi-coefficient (Φ) associated

with a particular simulated condition, the average bias of its estimated phi-coefficient ($\widehat{\Phi}$) was obtained by

$$\text{average bias} = \frac{1}{1000}\sum_{h=1}^{1000}(\widehat{\Phi}_h - \Phi_h), \qquad (3)$$

where h refers to the hth replication. Comparisons between the two methods were possible in that their respective estimation procedures were performed on the same sparse data for each simulated condition. The data generation and score reliability estimation were performed in the R statistical software, version 2.15.2. Independent of the current study, estimated phi-coefficients were validated against the true phi-coefficients by analyzing simulated datasets with no missing data.

Simulation Results

Results Based on VC Composition (a)

Tables 1 and 2 are associated with VC composition (a), in which the relative magnitude of score variability due to raters is small, and therefore the true score reliability is expected to be high. The two tables present averages and average biases of estimated phi-coefficients across the nine simulated combinations between the numbers of persons and the numbers of raters. Within each row of n_p, the upper row shows results from the rating method while the lower row represents those from the rater method.

Table 1 shows results based on rater scenario (a), where the raters are expected to have similar training and/or experience. It can be observed that the two methods yield very similar reliability estimates that are also close to their respective true phi-coefficients. For

Table 1. Estimated Phi-Coefficient: Rating (upper) vs. Rater (lower) Methods Based on VC Composition (a) and Rater Scenario (a)

	Raters = 4 (True Phi = 0.8814)		**Raters = 8 (True Phi = 0.9369)**		**Raters = 16 (True Phi = 0.9674)**	
n_p	Average Phi	Average Bias	Average Phi	Average Bias	Average Phi	Average Bias
50	0.8730	-0.0084	0.9329	-0.0040	0.9655	-0.0019
	0.8739	-0.0075	0.9330	-0.0039	0.9656	-0.0018
100	0.8785	-0.0029	0.9339	-0.0030	0.9664	-0.0010
	0.8788	-0.0026	0.9344	-0.0025	0.9665	-0.0009
200	0.8802	-0.0012	0.9356	-0.0013	0.9669	-0.0005
	0.8808	-0.0006	0.9358	-0.0011	0.9671	-0.0003

Table 2. Estimated Phi-Coefficient: Rating (upper) vs. Rater (lower) Methods Based on VC Composition (a) and Rater Scenario (b)

	Raters = 4 (True Phi = 0.8739)		**Raters = 8 (True Phi = 0.9348)**		**Raters = 16 (True Phi = 0.9669)**	
n_p	Average Phi	Average Bias	Average Phi	Average Bias	Average Phi	Average Bias
50	0.8645	-0.0094	0.9297	-0.0051	0.9641	-0.0028
	0.8662	-0.0077	0.9303	-0.0045	0.9639	-0.0030
100	0.8701	-0.0038	0.9321	-0.0027	0.9654	-0.0015
	0.8714	-0.0025	0.9324	-0.0024	0.9656	-0.0013
200	0.8719	-0.0020	0.9347	-0.0001	0.9663	-0.0006
	0.8730	-0.0009	0.9350	0.0002	0.9663	-0.0006

instance, in the case where n_p = 200 and n_r = 8 in Table 1, the estimated phi-coefficient is 0.9356 based on the rating method and is 0.9358 based on the rater method, which both converge to the true phi-coefficient at 0.9369. As a result, the average biases of each estimated score reliability based on the two methods do not differ much from each other and are fairly small, suggesting that the two methods are equally precise in estimating score reliability when the rates are expected to have similar score variability.

Table 2 presents results based on rater scenario (b), which reflects situations where a mixture of novice and experienced raters participate together in scoring. Similarly, one can observe that the estimates of score reliability based on either the rating or the rater method are fairly close to their corresponding true phi-coefficients. For example, in the case where n_p = 100 and n_r = 4 in Table 2, the estimated score reliability is short by only 0.0038 on average based on the rating method, and is short by only 0.0025 on average based on the rater method. In sum, the rating and rater methods perform equally well in estimating score reliability when the relative magnitude of score variability attributed to raters is small. Given that the rating method is easier to apply in practice, the rating method is recommended for operational use in this case. Moreover, when the number of raters is fixed, the average bias is expected to decrease as the number of examinees increases.

Results Based on VC Composition (b)

Tables 3 and 4 are associated with VC composition (b), in which the relative magnitude of score variability due to raters is large, and therefore the true score reliability is expected to be low to medium. Again, within each row of n_p, the upper row presents results from the rating method while the lower row shows those from the rater method.

Table 3 shows results based on rater scenario (a), where the raters are expected to have similar training and/or experience. One can observe that the rating

Table 3. Estimated Phi-Coefficient: Rating (upper) vs. Rater (lower) Methods Based on VC Composition (b) and Rater Scenario (a)

	Raters = 4 (True Phi = 0.5714)		**Raters = 8 (True Phi = 0.7273)**		**Raters = 16 (True Phi = 0.8421)**	
n_p	Average Phi	Average Bias	Average Phi	Average Bias	Average Phi	Average Bias
50	0.5825	0.0111	0.7337	0.0064	0.8475	0.0054
	0.5783	0.0069	0.7308	0.0035	0.8453	0.0032
100	0.5763	0.0049	0.7338	0.0065	0.8428	0.0007
	0.5730	0.0016	0.7304	0.0031	0.8426	0.0005
200	0.5654	-0.0060	0.7209	-0.0064	0.8438	0.0017
	0.5703	-0.0011	0.7241	-0.0032	0.8433	0.0012

Table 4. Estimated Phi-Coefficient: Rating (upper) vs. Rater (lower) Methods Based on VC Composition (b) and Rater Scenario (b)

	Raters = 4 (True Phi = 0.5195)		**Raters = 8 (True Phi = 0.7048)**		**Raters = 16 (True Phi = 0.8344)**	
n_p	Average Phi	Average Bias	Average Phi	Average Bias	Average Phi	Average Bias
50	0.6174	0.0979	0.7964	0.0916	0.8989	0.0645
	0.5259	0.0064	0.7005	-0.0043	0.8320	-0.0024
100	0.6022	0.0827	0.7907	0.0859	0.8992	0.0648
	0.5213	0.0018	0.7005	-0.0043	0.8357	0.0013
200	0.5985	0.0790	0.7884	0.0836	0.9013	0.0669
	0.5138	-0.0057	0.7007	-0.0041	0.8367	0.0023

method consistently have a slightly higher degree of average bias in estimating score reliability than the rater method does, suggesting that the rater method is slightly more precise in this case; however, the difference may not warrant much practical concern. For instance, in the case where n_p = 50 and n_r = 4 in Table 3, the rating method overestimates the true phi-coefficient by 0.0111, whereas the rater method overestimates by 0.0069. Given the slight difference, the two methods can still be considered satisfactory in estimating score reliability when raters are expected to have similar score variability.

Nevertheless, the picture is less optimistic in Table 4, which presents results based on rater scenario (b), reflecting situations in which some raters are expected to have more score variability than the others. Clearly, the rating method consistently overestimates the true phi-coefficient across the simulated conditions, whereas the average bias based on the rater method remains small. In some cases, the undue inflation of score reliability based on the rating method may raise practical concerns. For example, in the case where n_p = 50 and n_r = 8 in Table 4, the rating method yields an estimated phi-coefficient of 0.7964, whereas the rater method suggests 0.7005. If a testing program decides to set its minimum score reliability at 0.75 for quality control purposes, the use of rating method will result in a false claim about acceptable score reliability because the rating method indicates a higher estimated phi-coefficient at 0.7964 on average than the minimum score reliability at 0.75, when in fact the true phi-coefficient is 0.7048. In sum, when the relative magnitude of score variability accounted for by the facet of raters is large, the rater method is more precise in estimating score reliability than the rating method is, particularly if the raters are expected to have varying degrees of score variability, such as a mixture of novice and seasoned raters rating together.

Empirical Analysis Plan Informed by Simulation Results

According to the simulation results, the estimation precision of score reliability of the rating and rater methods is dependent on the relative magnitude of score variability due to the facet of raters, such that when the variance component for raters (σ_r^2) is relatively small, the rating and rater methods are equally precise in estimating score reliability; however, when σ_r^2 is relatively large, the rater method is more precise in estimating score reliability. Although the true parameter of σ_r^2 is not known from operational data, it can be estimated by $\hat{\sigma}_r^2$ from the data at hand. Thus, the design of an analysis plan for examining score reliability under the G-theory framework can be informed by gauging the magnitude of estimated variance component for raters or ratings. If it is small compared to the other estimated variance components, the rating method can be readily applied because it is equally precise as the rater method but easier to implement in practice. If $\hat{\sigma}_r^2$ is relatively large, the rater method would be a better choice because it is more precise.

It should be noted that in an operational setting with a double-rating design, the structure of sparse data can be very complex, such as cross-pairing of raters and overlapping of raters for different groups of examinees. Complex data structures are an inevitable result of practical constraints with assigning each response to any two available raters from a pool of qualified raters. Such structures also increase the computational sophistication related to the rater method. Hence, when assessing the estimated variance component for raters or ratings, the rating method is recommended because its estimating procedures are very easy to implement as a practical screening tool. The following section provides a step-by-step example of an analysis plan, informed by the simulation results discussed so far, for investigating score reliability of a large-scale speaking test.

Score Reliability of the Speaking Component of the ECPE

The Examination for the Certificate of Proficiency in English (ECPE) is developed by CaMLA (Cambridge Michigan Language Assessments—CambridgeMichigan.org). It is a large-scale standardized test designed to assess the language proficiency of nonnative English language speakers. Test results are used for professional and academic purposes, such as certificates for workplace language proficiency and for school admissions. The speaking component of the ECPE consists of a multi-stage speaking task. Two to three examinees participate in a single testing session. The examinees are asked to collaborate in presenting ideas and defending their stances. Until June 2014 each examinee was rated independently by at least two trained raters on a five-point scale. The final speaking scale score was reached by a consensus process between

the trained raters. Three operational datasets from the speaking component of ECPE were analyzed in the current study. Each dataset included scores from speaking tests administered during one of the ECPE's scheduled test administrations. The datasets for tasks A, B, and C comprise 1,999, 1798, and 2,220 examinees, respectively. Each spoken response was rated by two raters. Given that not all the responses were rated by the same pairs of raters, the three datasets constitute sparse rated data.

Estimated Variance Components

Sample means, standard deviations, ranges, and coefficients of variation (CVs) of ECPE speaking tasks A, B, and C are reported in Table 5. The descriptive statistics here are based on the total valid ECPE ratings across the raters. For each speaking task, the CV is the ratio of the standard deviation of speaking scores to its corresponding mean, which serves as an index of score variation with respect to the mean. CVs function as a descriptive tool in comparing score distributions from different sources, such as the three speaking tasks in the current analysis, that are intended to measure the same construct. As can be observed from the descriptive statistics, the means and standard deviations are similar across the three speaking tasks. In addition, the three CVs for tasks A, B, and C are almost identical—0.228, 0.225, and 0.229, respectively. Assuming that the examinees were randomly assigned to the three speaking tasks and that the scoring was performed by equally qualified raters, similar descriptive statistics across the three speaking tasks suggest that differences in task difficulty are negligible.

Next, the rating method was used as a screening tool to assess the relative magnitudes of different estimated variance components (i.e., persons, ratings, and errors) for each speaking task. Table 6 presents the estimated variance components and their proportions of total score variance based on scores from the ECPE speaking tasks A, B and C, respectively. It can be observed that the compositions of estimated variance components across the three tasks are very similar in that the estimated variance component of persons (87.47%–87.84%) has the lion's share, followed by the error component (12.06%–12.49%) and then by the estimated variance component for ratings (0.04%–0.10%). Substantively, these results suggest that about 87% of observed score variability in ECPE speaking can be accounted for by true differences in examinees' oral proficiency levels. Moreover, the similarity in the patterns of estimated variance components also resonates with the descriptive statistics that the three tasks do not differ much in task difficulty.

Score Reliability and Standard Errors of Measurement

The simulation results from the previous sections suggest that when the relative magnitude of variance component for raters or ratings is small, the rating and rater methods are equally precise in estimating score reliability. Since the proportion of total score variance due to the estimated variance component of ratings is very small in the empirical analysis, it is therefore methodologically sound to proceed with the rating method in estimating score reliability of the speaking component of the ECPE. The estimated phi-coefficients

Table 5. Descriptive Statistics of ECPE Speaking Scores by Tasks

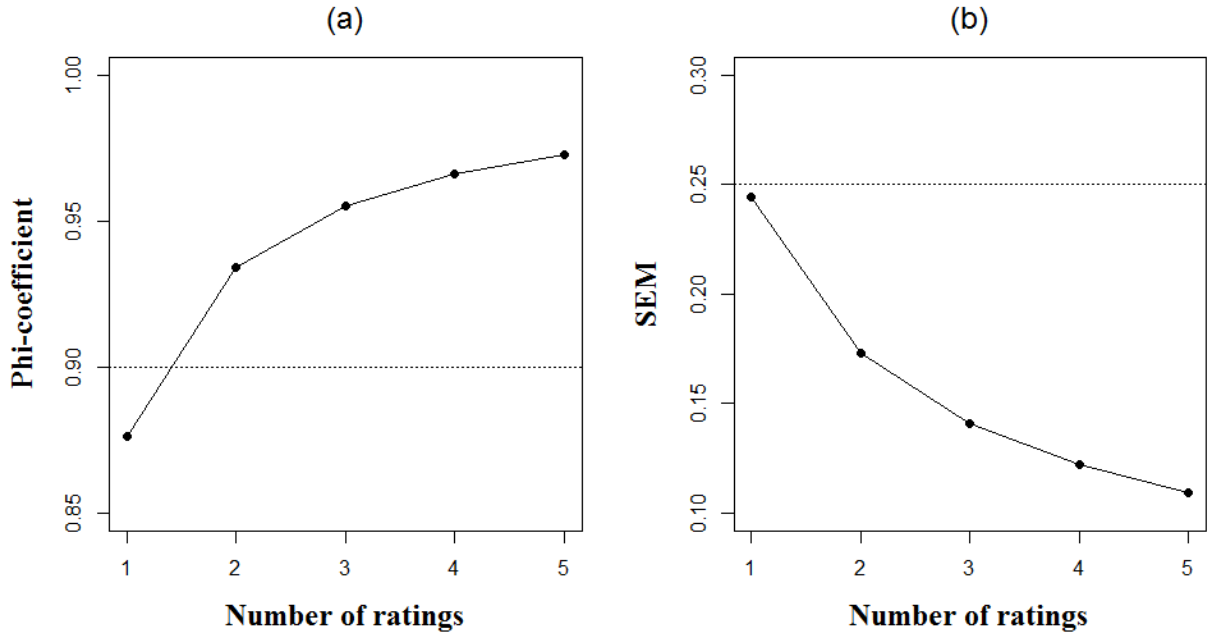
	Sample Size	Total Ratings	Mean	Standard Deviation	Min./Max.	Coefficient of Variation
Task A	1,999	3,998	3.061	0.697	1/5	0.228
Task B	1,798	3,596	3.092	0.695	1/5	0.225
Task C	2,220	4,440	3.019	0.692	1/5	0.229

Table 6. ECPE Speaking: Estimated Variance Components and Proportions of Total Score Variance by Tasks

	Task A		Task B		Task C	
	Estimated VC	*% of total variance*	*Estimated VC*	*% of total variance*	*Estimated VC*	*% of total variance*
p	0.4275	87.84%	0.4229	87.47%	0.4203	87.61%
r	0.0005	0.10%	0.0002	0.04%	0.0003	0.07%
e	0.0587	12.06%	0.0604	12.49%	0.0591	12.32%
Total	0.4867	100%	0.4835	100%	0.4797	100%

in Equation (2) were computed based on the average estimated variance components across the three speaking tasks, and the estimated phi-coefficients were also evaluated with respect to the numbers of ratings by varying the numbers of ratings from 1 to 5. In addition to phi-coefficients, standard errors of measurement (SEMs) in ECPE speaking were also evaluated with respect to the numbers of ratings. Phi-coefficients

Figure 2. Phi-Coefficients and SEMs of ECPE Speaking



provide information about the extent to which awarded scores are reliable, while SEMs indicate the degree to which imprecision exists in awarded scores. Both pieces of information are useful in making decisions about the utility of performance-based assessments (Brennan, Gao, & Colton, 1995). SEMs are computed as follows:

$$\text{SEM} = \sqrt{\frac{\hat{\sigma}_r^2}{n_r'} + \frac{\hat{\sigma}_e^2}{n_r'}}, \quad (4)$$

where n_r' refers to the number of raters or ratings per spoken response.

Figure 2 shows estimated phi-coefficients and SEMs with respect to the numbers of ratings for the ECPE speaking component. As expected, score reliability increases as the number of ratings increases, whereas imprecision in awarded scores decreases as the number of ratings increases. Figure 2 (a) indicates that the increase in score reliability is larger when the number of ratings increases from one to two, but the improvement lessens when two ratings or more are used. In a similar vein, the decrease in imprecision of awarded scores is larger when the number of ratings increases from one to two in Figure 2 (b). In addition, Figure 2 (a) suggests that at least two ratings are required to achieve a score reliability of 0.90 or higher for the ECPE speaking—the high score reliability is necessary given the high-stakes use of ECPE in academic and workplace settings. Regarding the precision of awarded scores, when a single rating is employed, the SEM is expected to be 0.24 in Figure 2 (b). This translates to 0.96 points with a 95% confidence limit (equivalent to four SEMs) and suggests that the imprecision in awarded scores (with only one rating) is not likely to be larger than one scale level, which is acceptable for the five-point scale of ECPE speaking. In sum, although one rating is recommended from a precision perspective, two ratings are required on reliability grounds. Given that both score reliability and precision are equally important in a high-stakes assessment such as the ECPE speaking component, taking both phi-coefficients and SEMs into consideration would suggest that at least two ratings are needed for operational use of the ECPE speaking component.

Conclusion

The current study evaluates the precision of the rating and rater methods in estimating score reliability under the G-theory framework. It illustrates how simulation research can be useful in guiding the analysis plan in an operational setting. As such, the simulation study was designed with an eye to reflecting realistic settings in performance-based language assessments, so that the simulated results can be readily applied to inform operational analysis. Depending on the compositions of variance components and on the score variability across different raters, estimated score reliability can be different based on the rating and rater methods. When the relative magnitude of variance component for raters/ratings is small, the two methods are equally precise in estimating score reliability, suggesting that the rating method may be a better choice for operational use given that the rating method is much easier to implement in practice. However, when the variance component of raters/ratings is relatively large, the rating method tends to unduly overestimate score reliability, and therefore the rater method is recommended for operational use.

The simulation results were then fed into an empirical analysis of the speaking component of ECPE by a step-by-step fashion. First, the rating method was used as a screening tool to assess the relative magnitude of score variability due to ratings. Upon discovering that the estimated variance component of ratings was small, the empirical analysis followed the recommendation based on the simulation study and resorted to the rating method throughout the analysis. Empirical results suggested that at least two ratings are necessary for operational use to achieve satisfactory score reliability and to control for reasonable measurement errors for the speaking component of ECPE.

It should be emphasized that the simulation results and the step-by-step analysis plan presented in this paper are applicable to different speaking assessment contexts, so long as at least two scores are given by independent raters for each spoken response. This requirement is critical in G-theory applications because gauging the relative magnitude of score variability due to any measurement facet is only possible when at least two elements are present in the facet. Nevertheless, the recommended number of ratings for the speaking component of ECPE is only applicable to the test format and scoring approach reported in this study. Updating the rating rubric and/or revising the scoring method would make it necessary to revisit the number of recommended ratings for future operational use.

References

Akiyama, T. (2001). The application of G-theory and IRT in the analysis of data from speaking tests administered in a classroom context. *Melbourne Papers in Language Testing*, *10*, 1–21.

Allen, M. J., & Yen, W. M. (2001). *Introduction to measurement theory*. Long Grove, IL: Waveland Press.

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*(2), 238–257.

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

Brennan, R. L., Gao, X., & Colton, D. A. (1995). Generalizability analyses of Work Keys listening and writing tests. *Educational and Psychological Measurement*, *55*(2), 157–176.

Chiu, C. W. T. (2001). *Scoring performance assessments based on judgments: Generalizability theory*. Boston, MA: Kluwer Academic.

Chiu, C. W. T., & Wolfe, E. W. (2002). A method for analyzing sparse data matrices in the generalizability theory framework. *Applied Psychological Measurement, 26*(3), 321–338.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. New York, NY: Wiley.

Ennis, R. H. (1999). Test reliability: A practical exemplification of ordinary language philosophy. *Philosophy of Education Archive*, 242–248.

Huang, J. (2012). Using generalizability theory to examine the accuracy and validity of large-scale ESL writing assessment. *Assessing Writing*, *17*(3), 123–139.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: Greenwood.

Lee, Y.-W. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language Testing*, *23*(2), 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*(2), 158–180.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.

Mooney, C. Z. (1997). *Monte carlo simulation*. Thousand Oaks, CA: Sage.

Nugent, W. R. (2009). Construct validity invariance and discrepancies in meta-analytic effect sizes based on different measures: A simulation study. *Educational and Psychological Measurement*, *69*(1), 62–78.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Thousand Oaks, CA: Sage Publications.

Stemler, S. E., & Tsai, J. (2008). Best practices in inter-rater reliability: Three common approaches. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 29–49). Los Angeles, CA: Sage.

Xi, X. (2007). Evaluating analytic scoring for the TOEFL® Academic Speaking Test (TAST) for operational use. *Language Testing*, *24*(2), 251–286.