Article

LANGUAGE TESTING

Language Testing
2017, Vol. 34(2) 271–289

DOI: 10.1177/0265532216638890
journals.sagepub.com/home/ltj
SAGE

Working with sparse data in rated language tests: Generalizability theory applications

Chih-Kai Lin
Center for Applied Linguistics, USA

Abstract

Sparse-rated data are common in operational performance-based language tests, as an inevitable result of assigning examinee responses to a fraction of available raters. The current study investigates the precision of two generalizability-theory methods (i.e., the rating method and the subdividing method) specifically designed to accommodate the technical complexity involved in estimating score reliability from sparse-rated data. Examining the estimation precision of reliability is of great importance because the utility of any performance-based language test depends on its reliability. Results suggest that when some raters are expected to have greater score variability than other raters (e.g., a mixture of novice and experienced raters being deployed in a rating session), the sub-dividing method is recommended as it yields more precise reliability estimates. When all raters are expected to exhibit similar variability in their scoring, both the rating and sub-dividing methods are equally precise in estimating score reliability, and the rating method is recommended for operational use, as it is easier to implement in practice. Informed by these methodological results, the current study also demonstrates a step-by-step analysis for investigating the score reliability from sparse-rated data taken from a large-scale English speaking proficiency test. Implications for operational performance-based language tests are discussed.

Keywords

Generalizability theory, Monte Carlo simulation, performance-based assessment, reliability, sparse-rated data

Expert-rated assessments of actual language test performances are common in many contexts, such as academic departments at universities that use language placement tests to assess incoming students, regional and national governments that administer language proficiency tests to measure student growth, and large-scale language testing programs that

Corresponding author:
Chih-Kai Lin, Center for Applied Linguistics, 4646 40th St. NW, Washington, DC 20016, USA.
Email: clin@cal.org

offer academic and workplace qualifications. The advent 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 generalized to target language use in non-test contexts (Bachman & Palmer, 1996; Chapelle, Enright, & Jamieson, 2008), which are in accord with the modern paradigm of test validation (Kane, 2006).

Given the emphasis on performance tests, rater-mediated measurement has become typical in many language assessment contexts. Many testing programs continue to rely on a time-honored scoring paradigm: expert raters with rigorous training and calibration. However, scoring performance 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 as a result of a short turnaround timeline for scoring. Furthermore, some raters may be unavailable when they are needed, 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 re-calibrated. All of these factors result in score fluctuation for reasons other than the intended construct being measured, thereby affecting the reliability of the scores.

Score reliability in rater-mediated measurement is defined as the extent to which raters are consistent in giving scores across the object of measurement (e.g., examinees or persons), 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 reliability is desirable so that score interpretations can be trusted (AERA, APA, & NCME, 2014). This paper investigates the precision of different methods in estimating score reliability from sparse-rated data. The following two sub-sections introduce the different analytical methods and give a brief description of sparse-rated data.

Generalizability theory

Language-testing researchers can estimate the relative magnitude of construct-irrelevant variability in rated-test scores and factor the variability into the estimation of score reliability. Moreover, they can identify which construct-irrelevant factors (such as different raters or different tasks – things that should not contribute to score variations) account for the overall construct-irrelevant variability. They can do this by using a measurement model called generalizability theory or G theory (Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972), which is a powerful analytical framework. G theory is a random facet measurement model that conceptualizes observed score variability as a composite of the true variation in the object of measurement and other variations pertaining to different measurement sources (called *facets* in G-theory terminology) that are anticipated by or are of interest to an investigator. Essentially, G theory decomposes total score variability, via statistical techniques such as analysis of variance, Bayesian inference (e.g., Davis, 1974), and latent variable modeling (e.g., Schoonen, 2005), into variance components associated with the object of measurement and with various facets involved in the measurement. This decomposition provides information about how much variation is explicable by each component. For instance, in a speaking test for a group of English as a second language (ESL) students, the object of measurement is students'

speaking proficiency. One potential source of score variation is variability introduced by different raters scoring the responses. Ideally, one would like to see true differences among students' speaking proficiency reflect observed score variability as much as possible, not differences due to raters.

Sparse-rated 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 examinee or person (p) be rated by all available raters (r). A fully crossed design is ideal in that it allows G-theory analysis to separately assess variability due to the main and interaction effects of the object of measurement and the facets of interest, resulting in more straightforward analysis of variance components, which in turn aids the interpretation of score reliability. The relationship between variance components and score reliability is illustrated below by a one-facet random effect model under the G-theory framework:

$$X_{pr} = \mu + \alpha_p + \beta_r + \varepsilon_{pr,e}, \tag{1}$$

where the 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 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. For the absolute interpretation of score reliability under the G-theory framework, the estimated phi-coefficient or generalizability coefficient for absolute decisions (Cronbach et al., 1972) may be presented as:

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

where n_r' refers to the number of ratings per examinee response. Equation (2) shows an inverse relationship between score reliability and score variability owing to measurement facets; that is, all else being 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 these variance components can be estimated independently of one another, which is one of the main advantages of working with fully crossed datasets.

In an operational performance-assessment setting, fully crossed designs are not practical, and may be impossible, due to the tremendous number of ratings such designs require each rater to perform (Lee, 2006). Alternatively, many testing programs resort to a

Fully crossed data

	R1	R2	R3	R4
P1–15	X	X	X	X
P16–30	X	X	X	X
P31–45	X	X	X	X
P46–60	X	X	X	X

Sparse data

	R1	R2	R3	R4
P1–15	X	.	.	X
P16–30	.	X	.	X
P31–45	.	X	X	.
P46–60	X	.	X	.

Figure 1. Illustrative examples of fully-crossed data and sparse data, where "P" refers to person/examinee, "R" refers to rater, "X" refers to a complete set of ratings awarded to a group of examinees by a rater, and "." refers to missing data by design.

double-rating scheme, by which each examinee response is rated by any two raters from a rater pool, resulting in sparse data (Chiu, 2001; DeMars, 2015). Figure 1 shows an example of fully crossed rated data in an ideal situation and an example of sparse-rated data with a double-rating scheme from 60 persons/examinees (P1–60) and four raters (R1–R4).

Language testers know the clear advantages of working with fully crossed designs, but such designs are rare or even non-existent in operational testing programs. When testers have sparse performance-based data, they can use two existing analysis-of-variance (ANOVA) methods under the G-theory framework. Both methods take sparse-rated datasets as a given and transform them into some variants of fully crossed ones. In the first method, *ratings* are treated as a random facet; henceforth referred to as the rating method. In the second method, *raters* are treated as a random facet; henceforth referred to as the sub-dividing method (Chiu, 2001). Figure 2 gives a visual representation of how the two methods break down a hypothetical sparse dataset, in which each response from 80 persons/examinees (P1–P80) is double-rated among a panel of four raters (R1–R4), into fully crossed dataset(s). As such, 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 2, the rating method transforms the 80-by-4 sparse data matrix into an 80-by-2 fully crossed dataset. The variance components and score reliability are then estimated based on the transformed fully crossed data (see Shavelson & Webb, 1991, p. 29, for variance-component estimation). The sub-dividing method first identifies a total of four blocks of fully crossed sub-datasets in this example. Next, variance components are estimated within each block, and these variance-component estimates are then averaged across the four sub-datasets by giving weights according to the number of examinees in each block. Score reliability is then calculated based on the average estimated variance components.

Both methods have been applied to analyze sparse-rated data from performance-based language tests. For instance, the rating method has been applied in a university Spanish-as-a-foreign-language placement test (Bachman, Lynch, & Mason, 1995), in an English achievement test for secondary-school ESL students (Huang, 2012), and in an English language test for immigration purposes (Lynch & McNamara, 1998). The sub-dividing method has been applied in a large-scale English language proficiency test (Xi, 2007).

In an operational setting with a double-rating scheme, a testing program can adopt a highly standardized design, by which examinees are assigned to fixed pairs of non-overlapping raters. DeMars (2015) clarified that raters can be treated as nested within examinees

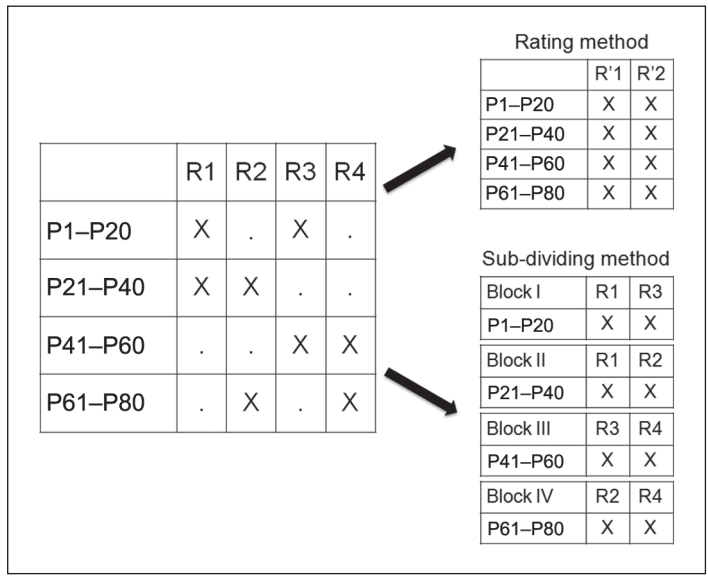

Figure 2. A visual representation of two methods to transform sparse data into fully crossed dataset variants: the rating method and the sub-dividing method.

$(r:p)$, *only if* each examinee is assigned to a unique rater pair (i.e., no sharing of raters across rater pairs). This conceptualization of raters being nested within examinees is in line with the definition of nested structure given by Shavelson and Webb (1991, p. 46). However, it would be more practical to have some overlapping raters between the rater pairs in an operational setting. Alternatively, a testing program can employ a more flexible design that allows some sharing of raters across rater pairs. In such a scenario, DeMars (2015) indicated that the correct model specification is that raters being crossed with examinees are nested within rater pairs $((r \times p):pair)$. The notion of a complete crossing of rater-by-person within each rater pair would suggest that either none or one member of a rater pair may overlap with those from other rater pair(s). DeMars further investigated the bias in estimating variance components due to the mis-specification of $(r:p)$, when in fact $((r \times p):pair)$ should have been the correct model for cases where there is some sharing of raters between the rater pairs. Results suggested that the degree of bias is small in large-scale contexts, such as state or national assessment programs. Finally, DeMars' study briefly touched on the estimation precision of the G-theory models with respect to less standardized rating designs (e.g., each examinee was randomly assigned to two raters from a rater pool), and it pointed to the lack of such research. The current study extends this line of research by focusing on sparse-rated data as a result of assigning each examinee to randomly paired raters (i.e., no fixed grouping of raters is formed). Such less standardized designs are more convenient and manageable in operational settings (Chiu & Wolfe, 2002; F. Davidson, pers. com., December 2012; J. Banerjee, pers. com., January 2015).

The current study contributes to research on the estimation precision of G-theory methods in handling sparse data (Chiu & Wolfe, 2002; DeMars, 2015) in three ways.

First of all, in addition to large-scale sample sizes, the current study also includes smaller sample sizes that may be of interest to many institutional language-placement testing programs. Second, rather than investigating the precision of variance-component estimates, the current study evaluates the precision of reliability estimates because the practical significance of bias in variance-component estimates, if any, may not easily be observable compared with bias in reliability estimates. It is true that if variance-component estimates are precise, so will be the case for reliability estimates; nevertheless, because reliability is the ratio of variance components, a slight estimation bias in a variance component or a substantial bias in a variance component of small magnitude may not bear much practical significance in the estimation of reliability. Finally, the current study attempts to evaluate the estimation precision of the rating method and of the sub-dividing method determined for sparse-rated data from less standardized rating designs. These two ANOVA-based methods have an advantage over other more advanced estimation methods (e.g., restricted maximum likelihood estimation and minimum variance quadratic unbiased estimators) for handling sparse-rated data, as the advanced methods may not be practical when the rater pool is large in operational settings (DeMars, 2015). To the author's knowledge, the rating method and the sub-dividing method are the two methods specifically designed to accommodate the technical complexity involved in estimating score reliability by transforming sparse-rated data into readily analyzable dataset(s) for the classic G-theory software, GENOVA (Crick & Brennan, 1982). In other words, the two methods are more accessible to language testers who work with sparse-rated data but who may not have the resources to utilize more advanced methods. Because of their accessibility and wide use in many operational language-testing contexts, research that sheds light on the precision of the rating method and of the sub-dividing method in estimating reliability is of great importance.

The current study

Language testers have both the rating and sub-dividing methods at their disposal in examining score reliability from sparse-rated data 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 sparse dataset in some 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 from the two methods may not always converge. When the estimates of score reliability differ, a natural follow-up question is which estimate to report. Without knowing the true reliability in any 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 by using operational data at hand because the true reliability is not known from operational data, and therefore an investigator has no way of knowing which estimate is more reflective of the true reliability. In light of this, the current study aims to:

- investigate rating condition(s) under which one method is recommended over the other via a simulation study; and
- demonstrate how this methodological research can guide the design of an analysis plan for examining the score reliability of a large-scale English speaking proficiency test.

The current study is method oriented and yet has real-world implications for language testers who work with rated performance data in that it offers methodological recommendations for analyzing sparse-rated data from language performance tests. Additionally, it demonstrates how the analysis plan for an empirical inquiry can be informed by simulation research.

Method

In order to address the issue of not being able to determine operationally the precision of different reliability estimates based on different estimation methods, the current study conducted a Monte Carlo simulation study to compare the estimation precision of the rating method and the sub-dividing method. The comparison is possible because the true reliability is predetermined and therefore known in simulation research, against which estimated score reliability based on the two methods can be evaluated. My aims in conducting this simulation study are twofold. First, I sought to evaluate the precision of the rating method and the sub-dividing method in estimating score reliability under various simulated conditions, whose designs are informed empirically. Second, I used results from the simulation study to guide the analysis plan for investigating score reliability of the speaking component of the Examination for the Certificate of Proficiency in English (ECPE) from CaMLA (Cambridge Michigan Language Assessments: www.cambridgemichigan.org).

Simulated conditions

One advantage of using simulation procedures to investigate estimation precision is that, instead of operating under a single operational setting in which an empirical study is usually carried out, investigators can purposefully choose simulated conditions that are informed by multiple realistic settings, providing useful implications for a wider audience of practitioners in diverse contexts. The current simulation study included three examinee sample sizes (n_p) (80, 360, and 1600), three rater-pool sizes (n_r) (4, 8, and 16), two compositions of variance components, and two scenarios of rater score variability, amounting to a total of 36 conditions.

Variance-component compositions. The two variance-component (VC) compositions were: (a) 65%, 5%, and 30% of total score variance is accounted for by persons, raters, and errors, respectively, and (b) 25%, 35%, and 40% of total score variance is due to persons, raters, and errors, respectively. The true reliability for VC composition (a) is expected to be higher than that for VC composition (b), because the relative magnitude of construct-irrelevant score variability (i.e., raters and errors) for VC composition (a) is smaller than

that for VC composition (b). This setup allows the simulation study to evaluate the precision of the rating method and of the sub-dividing method in estimating a range of score reliability.

Rater scenarios. The two rater scenarios were as follows: (i) all raters exhibit similar variability in their scoring, corresponding to raters having similar training and/or rating experience; and (ii) some raters have greater score variability than others, reflecting realistic settings in which a mixture of novice and experienced raters are deployed in a single rating session.

True parameters of variance components. The relative magnitude of the variance components for VC composition (a) was informed by previous G-theory research on speaking assessments (Akiyama, 2001; Bachman et al., 1995; Lynch & McNamara, 1998; Xi, 2007). In these studies, a large proportion of score variability is due to persons, a small proportion of score variability is accounted for by raters, and some variability is expected for measurement errors; this pattern is also reported in a research synthesis by In'nami and Koizumi (2015). It must be emphasized that in a simulation study, true parameters are 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 relative magnitudes of variance components by taking the averages across the aforementioned studies in speaking assessments. The average total score variance across these studies was 1.123 after accounting for scale differences. Given the VC composition (a), this translates to 0.730 (65%) for σ_p^2, 0.056 (5%) for σ_r^2, and 0.337 (30%) for σ_e^2. In these published research studies, the relative magnitude of score variability attributed to raters is usually small due to rigorous rater training. However, 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 $\sigma_p^2 =$ 0.281 (25%), $\sigma_r^2 =$ 0.393 (35%), and $\sigma_e^2 =$ 0.449 (40%).

Sparse-data generation

Data associated with rater scenario (i) were simulated according to Equation (1). Take VC composition (a) as an example. The score (X_{pr}) of person *p* given by rater *r* is 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 $\sigma_p^2 = 0.730$ (65%), $\sigma_r^2 = 0.056$ (5%), and $\sigma_e^2 = 0.337$ (30%), respectively. Data were simulated to be scored on a scale of 0 to 4 by setting the overall mean (μ) at 2. The true reliability (or phi-coefficient) is then calculated by plugging the true parameters of these variance components and the rater-pool size 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 $\sigma_p^2 = 0.281$ (25%), $\sigma_r^2 = 0.393$ (35%), and $\sigma_e^2 = 0.449$ (40%), respectively.

Data associated with rater scenario (ii) were also simulated according to Equation (1) for VC compositions (a) and (b). Nevertheless, what was different in rater scenario (ii) lay in the true parameter for the rater variance component, such that the variability in scoring for novice raters was simulated to be two times larger than that for experienced raters. The idea that larger score variability is associated with novice raters was taken from empirical observations, which found that inexperienced raters appeared to be less consistent in their scoring than experienced raters (Weigle, 1998, 1999). Two raters were designated as novice raters across the simulated conditions under rater scenario (ii); as a result, novice raters constituted 50%, 25%, and 12.5% of the rater pools for n_r = 4, 8, and 16, respectively.[1]

Next, I imposed two realistic constraints on data generation to create sparseness in the simulated data. First, randomly paired raters from the rater pool were assigned to each examinee (i.e., random double-rating scheme). Second, all raters shared an equal amount of scoring load. Given these two constraints, the levels of sparseness were directly linked to the rater-pool sizes (n_r) in the simulated conditions. Taking $n_p = 1600$ and $n_r = 16$ as an example, a fully crossed 1600-by-16 dataset with complete data was first generated. The first examinee was assigned to two randomly paired raters out of the 16 raters, and the first examinee's simulated data associated with the other 14 raters were then removed. The same procedures were applied to the second examinee, and the first and/or the second rater in this rater pair could be the same or different from those in the rater pair for the first examinee because of random pairing of rater. For each examinee, the random rater pairing was carried out with the constraint of equal scoring load for each rater. As such, for each examinee in the examinee-by-rater matrix, data associated with 14 out of 16 raters were missing, which resulted in a sparse level at 87.5% (14/16). The three rater-pool sizes (i.e., 4, 8, and 16) corresponded to sparseness of 50% (2/4), 75% (6/8), and 87.5% (14/16), respectively. In operational settings, the examinee-by-rater data matrix is expected to be very sparse, particularly when the pool of raters is large (Chiu, 2001; DeMars, 2015).

Evaluation of estimation precision

I evaluated the estimated score reliability (or phi-coefficient), based on the rating method and the sub-dividing method, against the true reliability by examining the average bias over 5000 replications of sparse datasets for each of the 36 simulated conditions. It is the randomness involved in the data generation described in the previous section that allows the current simulation study to replicate each simulated condition for a large number of times in order to gauge the precision of these two methods in estimating score reliability. Bias here is defined as the extent 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

$$\text{average bias} = \frac{1}{5000}\sum_{h=1}^{5000}\left(\widehat{\Phi}_h - \Phi\right), \tag{3}$$

where h refers to the hth replication. Comparisons between the two methods were made possible by their respective estimation procedures being performed on the same

Table 1. Estimated phi-coefficient: Rating method (upper) vs. sub-dividing method (lower) based on VC composition (a) and rater scenario (i).

n_p	Rater pool = 4 (*True Phi = .8814*)		Rater pool = 8 (*True Phi = .9369*)		Rater pool = 16 (*True Phi = .9674*)	
	Average phi	Average bias	Average phi	Average bias	Average Phip	Average bias
80	.8760	−.0054	.9335	−.0034	.9656	−.0018
	.8764	−.0049	.9338	−.0031	.9658	−.0016
360	.8803	−.0010	.9361	−.0009	.9670	−.0005
	.8807	−.0006	.9363	−.0006	.9670	−.0004
1600	.8809	−.0004	.9366	−.0003	.9674	≈.0000
	.8814	≈.0000	.9367	−.0002	.9674	≈.0000

sparse data per simulated condition. Using the R statistical software, version 2.15.2, I carried out the data generation and score reliability estimation.

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 reliability is expected to be high. The two tables present averages and average biases of estimated phi-coefficients across the nine combinations between the examinee sample sizes and the rater-pool sizes (3×3). Each rater-pool size is followed by its true phi-coefficient in each simulated condition. Within each row of n_p, the upper row shows results from the rating method. The lower row represents those from the sub-dividing method.

Table 1 shows results based on rater scenario (i), where raters are expected to have similar training and/or experience. The results show that the two methods yield very similar score reliability estimates that are also close to their respective true phi-coefficients. For instance, in the case where $n_p = 360$ and $n_r = 8$ in Table 1, the estimated phi-coefficient is 0.9361 based on the rating method and is 0.9363 based on the sub-dividing method, which both converge to the true phi-coefficient at 0.9369. As a result, the average bias of each estimated reliability from the two methods does not differ much from each other and is fairly small, suggesting that the two methods are equally precise in estimating score reliability when raters are expected to have similar score variability.

Table 2 presents results based on rater scenario (ii), which reflects situations where a mixture of novice and experienced raters participate together in scoring. Similarly, the results show that the estimates of score reliability based on either the rating or the sub-dividing method are fairly close to their corresponding true phi-coefficients. For example, in the case where $n_p = 80$ and $n_r = 4$ in Table 2, the estimated score reliability is short by only 0.0033 on average based on the rating method, and is short by only 0.0022 on average based on the sub-dividing method. Moreover, as expected, when holding the rater-pool size constant, the magnitude of average bias decreases as the number of

Table 2. Estimated phi-coefficient: Rating method (upper) vs. sub-dividing method (lower) based on VC composition (a) and rater scenario (ii).

n_p	Rater pool = 4 (*True phi = .8739*)		Rater pool = 8 (*True phi = .9348*)		Rater pool = 16 (*True phi = .9669*)	
	Average phi	Average bias	Average phi	Average bias	Average phi	Average bias
80	.8706	−.0033	.9321	−.0027	.9653	−.0016
	.8717	−.0022	.9327	−.0022	.9654	−.0015
360	.8724	−.0016	.9335	−.0012	.9664	−.0004
	.8735	−.0005	.9339	−.0009	.9666	−.0003
1600	.8734	−.0005	.9342	−.0006	.9666	−.0003
	.8743	.0004	.9345	−.0003	.9667	−.0002

Table 3. Estimated phi-coefficient: Rating method (upper) vs. sub-dividing method (lower) based on VC composition (b) and rater scenario (i).

n_p	Rater pool = 4 (*True phi = .5714*)		Rater pool = 8 (*True phi = .7273*)		Rater pool = 16 (*True phi = .8421*)	
	Average phi	Average bias	Average phi	Average bias	Average phi	Average bias
80	.5760	.0045	.7217	−.0056	.8444	.0023
	.5742	.0028	.7257	−.0015	.8407	−.0013
360	.5692	−.0022	.7290	.0017	.8396	−.0024
	.5711	−.0003	.7269	−.0004	.8413	−.0008
1600	.5727	.0012	.7261	−.0012	.8409	−.0012
	.5718	.0004	.7262	−.0010	.8418	−.0004

examinees increases. For example. When n_r =4, the average bias based on the rating and sub-dividing methods changes from −0.0033 and −0.0022 to −0.0005 and 0.0004, respectively, as n_p increases from 80 to 1600. In sum, the rating method and the sub-dividing method perform equally well in estimating score reliability when the relative magnitude of score variability attributed to raters is small.

Results based on VC composition (b)

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

Table 3 shows results based on rater scenario (i), where raters are expected to have similar training and/or experience. The results show that the sub-dividing method consistently has a slightly lower degree of average bias in estimating score reliability than the rating method, suggesting that the sub-dividing method is slightly more precise in this case;

Table 4. Estimated phi-coefficient: Rating method (upper) vs. sub-dividing method (lower) based on VC composition (b) and rater scenario (ii).

n_p	Rater pool = 4 (*True phi = .5195*)		Rater pool = 8 (*True phi = .7048*)		Rater pool = 16 (*True phi = .8344*)	
	Average phi	Average bias	Average phi	Average bias	Average phi	Average bias
80	.6092	.0897	.8016	.0967	.8991	.0647
	.5242	.0047	.7095	.0046	.8320	−.0024
360	.6031	.0837	.7968	.0919	.8983	.0639
	.5212	.0017	.7031	−.0017	.8333	−.0010
1600	.6026	.0831	.7963	.0914	.8987	.0642
	.5171	−.0024	.7037	−.0011	.8350	.0006

however, the differences may not warrant much practical concern. For instance, in the case where $n_p = 1600$ and $n_r = 4$ in Table 3, the rating method overestimates the true phi-coefficient by 0.0012, whereas the sub-dividing method overestimates by 0.0004. Given the slight differences, the two methods can still be considered satisfactory in estimating score reliability when raters are expected to have similar variability in their scoring.

Nevertheless, the picture is less optimistic in Table 4, which presents results based on rater scenario (ii), reflecting situations in which some raters are expected to have more score variability than others. Clearly, the rating method considerably overestimates the true phi-coefficient across the simulated conditions, whereas the average bias based on the sub-dividing method remains small. In some cases, the undue inflation of estimated reliability based on the rating method may raise practical concerns. For example, in the case where $n_p = 360$ and $n_r = 8$ in Table 4, the rating method yields an average estimated phi-coefficient of 0.7968, whereas the sub-dividing method suggests 0.7031. If test-program directors decide to set a test's minimum score reliability at 0.75 for quality control purposes, the use of the rating method will result in a false claim about acceptable score reliability, as the rating method indicates a higher estimated phi-coefficient at 0.7968 on average than the minimum score reliability at 0.75, when in fact the true phi-coefficient is 0.7048.

A further analysis revealed that the standard errors of variance-component estimates based on the rating method are generally larger than those based on the sub-dividing method, especially when the relative magnitude of rater variance is large. In sum, when the relative magnitude of score variability accounted for by the facet of raters is large, the sub-dividing method is more precise in estimating score reliability and more stable in estimating variance components than is the rating method, particularly when 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 from the rating and sub-dividing methods is affected by the relative magnitude of score variability due to the facet of raters/ratings, such that when σ_r^2 is relatively small, the rating and

sub-dividing methods are equally precise in estimating score reliability; however, when σ_r^2 is relatively large, the sub-dividing 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/ratings ($\hat{\sigma}_r^2$). If it is small compared to the other estimated variance components, the rating method can be readily applied since it is as precise as the sub-dividing method but easier to carry out in practice. If $\hat{\sigma}_r^2$ is relatively large, the sub-dividing method would be a better choice in terms of estimation precision.

It must be emphasized here that in an operational setting with a double-rating scheme, the structure of sparse data can be complex. This is an inevitable result of assigning each examinee response to any two available raters. Each of the two methods discussed in this paper has practical constraints, but those associated with the sub-dividing method may be larger. Hence, as a preliminary screening tool for assessing the magnitude of $\hat{\sigma}_r^2$, the rating method is recommended because its estimation procedures are relatively easy to implement. The following section provides a step-by-step example of an analysis, informed by the simulation results discussed so far, for investigating the score reliability of a large-scale English speaking proficiency test.

Speaking component of ECPE

The Examination for the Certificate of Proficiency in English (ECPE) was developed and is managed by CaMLA (Cambridge Michigan Language Assessments: www.cambridgemichigan.org). It is a large-scale standardized test designed to assess the language proficiency of non-native English language speakers. Test results are used for professional and academic purposes. The speaking component of 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. Each examinee is rated independently by two trained raters on a holistic five-point scale. The two trained raters award a final speaking score by making a decision together. The holistic approach to scoring described in this paper was practiced until 2012. Since the time of this research, the ECPE Speaking Scale and scoring procedures have been revised. Raters apply an analytic approach to scoring and individually assign examinees a score of 0-5 on three criteria.

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 in 2012. Tasks A, B, and C were given to 1,999, 1,798, and 2,220 examinees, respectively. Each examinee response was rated by two raters. Given that each response was assigned to a fraction of raters from a pool of 345 raters, the three datasets constituted sparse data, to which both the rating and sub-dividing methods are applicable.

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. 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

Table 5. Descriptive statistics of ECPE speaking scores by task.

	Sample size	Total ratings	Mean	Standard deviation	Min./Max.	Coefficient of variation
Task A	1999	3998	3.061	.697	1/5	.228
Task B	1798	3596	3.092	.695	1/5	.225
Task C	2220	4440	3.019	.692	1/5	.229

Table 6. ECPE speaking: Estimated variance components and proportions of total score variance by task.

	Task A		**Task B**		**Task C**	
	Estimated VC	% of total variance	Estimated VC	% of total variance	Estimated VC	% of total variance
p	.4275	87.84	.4229	87.47	.4203	87.61
r	.0005	.10	.0002	.04	.0003	.07
e	.0587	12.06	.0604	12.49	.0591	12.32
Total	.4867	100	.4835	100	.4797	100

descriptive tool for comparing score distributions from different sources, such as the three speaking tasks in the current analysis, that are intended to measure the same construct. The descriptive statistics show that 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 at 0.228, 0.225, and 0.229, respectively. Given 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 preliminary 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. The results show that the compositions of estimated variance components across the three tasks are very similar in that the estimated variance component for 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%). About 87% of observed score variability in ECPE speaking can be accounted for by true differences in examinees' oral proficiency. Moreover, the similarity in the patterns of estimated variance components resonates with the previous analysis of descriptive statistics where the three tasks do not differ much in task difficulties.

Score reliability and standard errors of measurement

Methodologically, the simulation results in the previous sections suggest that when the relative magnitude of variance component for ratings is small, both the rating and

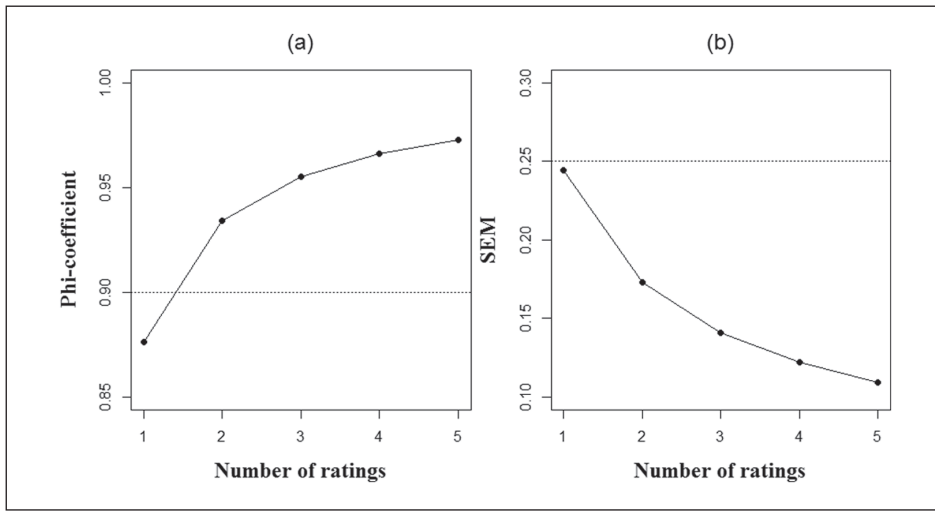

Figure 3. Phi-coefficients and SEMs of ECPE speaking.

sub-dividing methods are equally precise in estimating score reliability. Because the proportion of total score variance attributable to the estimated variance component for 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 ECPE. The estimated phi-coefficient in Equation (2) was computed based on the average estimated variance components across the three ECPE speaking tasks, and the estimated phi-coefficient was evaluated with respect to the number of ratings by varying the number of ratings from one to five. In addition to phi-coefficients, standard errors of measurement (SEMs) in the ECPE speaking component were also evaluated in relation to the number of ratings.

Phi-coefficients provide information about the extent to which awarded scores are reliable, while SEMs indicate the degree to which imprecision resides 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:

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

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

Figure 3 shows the estimated phi-coefficients and SEMs with respect to the number of ratings for the speaking component of ECPE. As expected, reliability increases as the number of ratings increases, while imprecision in awarded scores decreases as the number of ratings increases. Figure 3 (a) indicates that the increase in reliability is larger when the number of ratings increases from one to two, but the improvement shrinks when two or more ratings 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 3 (b). Additionally, Figure 3 (a) suggests that at least two ratings are required to achieve a reliability of 0.90 or higher for the ECPE speaking. This high reliability is necessary given the high-stakes use of ECPE in academic and workplace settings. Figure 3 (b) shows that when a single rating is employed, the SEM is expected to be 0.24 points, which translates to 0.96 points with a 95% confidence limit (equivalent to four SEMs). This suggests that the measurement errors in awarded scores, even with only one rating, are acceptable because the imprecision is not likely to be larger than one point on 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. As both reliability and precision are equally important in a high-stakes assessment such as the ECPE speaking, taking both the phi-coefficient and SEM into consideration would suggest that at least two ratings are needed for operational use of the speaking component of ECPE.

Conclusion

High reliability in rater-mediated measurement is desirable so that raters can be considered interchangeable; that is, a score awarded will not be contingent upon any specific rater making the judgment. Nevertheless, one cannot assume that all individual raters in the rater pool are interchangeable even in a well-designed rating system. Estimation methods for score reliability are thereby needed to check this assumption for quality-control purposes. In the current study, I evaluated the precision of the rating method and the sub-dividing method in estimating score reliability under the G-theory framework. I illustrated how simulation research can be useful in guiding the analysis plan for an operational inquiry. As such, I designed the simulation study with an eye toward reflecting realistic settings in performance-based language assessments, so that the results can inform operational analysis. Depending on the composition of variance components and the score variability across different raters, estimated score reliability can be different between the rating method and the sub-dividing method. When the relative magnitude of variance component for raters/ratings is small, the two methods are equally precise in estimating score reliability. Given that the rating method is much easier to implement in practice, the rating method is sufficient for operational use. However, when there is a sizeable variance component for raters/ratings, the rating method tends to inflate estimated reliability, particularly when raters are expected to have varying degrees of score variability; hence, the sub-dividing method is recommended for operational use in this case.

The theoretical foundation for both the rating and sub-dividing methods is built on de Finetti's (1931) theorem of exchangeability, by which sequences of independent and identically distributed random variables are exchangeable given some underlying distribution. In G theory, when the elements in a facet have not been sampled randomly from the *universe of admissible observations* but the intended *universe of generalization* is infinitely large, the facet can be treated as random and consequently the facet elements are assumed to be exchangeable even though they are fixed (Shavelson & Webb, 1981). In a rater-mediated assessment, when the relative magnitude of variance component for raters is small, treating the rater facet as random when they have not been sampled randomly does not matter. However, when the relative magnitude of rater variance

component is large, the impact of violating the exchangeability assumption seems to be more serious in the rating method than in the sub-dividing method, especially when raters are expected to have varying degrees of score variability and are thereby not exchangeable.

The simulation results informed the design of an empirical analysis of the speaking component of ECPE, following a step-by-step analysis plan. First, the rating method was used as a preliminary screening tool to evaluate the relative magnitude of score variability due to ratings. Upon discovering that the estimated variance component for ratings is small, the empirical analysis followed the recommendation based on the simulation study and resorted to the rating method throughout the analysis. Empirical results suggest that at least two ratings are necessary for operational use, in order to achieve satisfactory score reliability and control for reasonable measurement errors for the speaking component of ECPE.

One important caution must be exerted in applying either the rating method or the sub-dividing method. Technically, both methods do not require missing data to be completely at random for the purposes of estimating variance components and computing phi-coefficients. However, if some random mechanism of examinee–rater distribution is not included in the assessment design, biased estimates may appear. For example, in an extreme case where examinee proficiency levels are dependent on rater characteristics (e.g., novice raters are always assigned to low proficient examinees), bias is very likely to be introduced into the estimated variance components, which will consequently result in biased reliability estimates. Although such extreme cases are not likely to occur in large-scale language scoring centers as random examinee-to-rater assignment is usually built into the scoring design, it should not be taken as a given.

Finally, the current study is limited to the comparison between two ANOVA-based methods in estimating score reliability from sparse-rated data under the G-theory framework. Future research can compare these two methods with other more advanced approaches, such as structural equation modeling (SEM) and restricted maximum likelihood (REML), in dealing with sparse-rated data (see Schoonen, 2005, for an SEM analysis; see Bouwer, Béguin, Sanders, & van den Bergh, 2015, for an REML application). Additionally, while this study is based on a one-facet model in which raters/ratings were treated as a random facet, the scope can be further expanded into a two-facet model which includes speaking tasks as another facet as the most common facets involved in performance-based language assessments are those associated with tasks and raters (Lee, 2006).

Acknowledgements

The author conducted the research reported in this paper during his time as a graduate student at the University of Illinois at Urbana-Champaign. The author wishes to thank Jinming Zhang, Jayanti Banerjee, Fred Davidson, Natalie Nordby Chen, and two anonymous *Language Testing* reviewers for their valuable feedback on earlier versions of this paper.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This paper reports on research funded through CaMLA's Spaan Research Grant Program, 2013. The opinions expressed herein are solely those of the author and do not necessarily represent those of CaMLA.

Note

1. For rater scenario (ii), in which novice raters are expected to have greater score variability than experienced raters, the true parameter $\sigma^2_{r(exp)}$ for experienced raters followed the simulated value of 0.056 and 0.393 for VC composition (a) and VC composition (b), respectively, while the true parameter $\sigma^2_{r(nov)}$ for novice raters was set to be twice of $\sigma^2_{r(exp)}$ at 0.112 and 0.786. Because of the varying degree of score variability across the raters, the true parameters σ^2_r for the overall rater variance component varies as the ratio of novice raters to experienced raters differs in each rater-pool size. The true overall rater variance component cannot be derived analytically; however, it can be approximated by simulations over a large number of replications. Take VC composition (a) and $n_r = 8$ for example. To approximate the true parameter σ^2_r by simulations, a dataset with no missing data was first generated according to Equation (1), with individual rater effects of the six experienced raters following $N(0,\sigma^2_{r(exp)} = 0.056)$ and individual rater effects of the two novice raters following $N(0,\sigma^2_{r(nov)} = 0.112)$. The overall variance component for the rater effect was then estimated from the full dataset. The above process was independently repeated 10,000 times in order to arrive at a stable approximation of the true parameter σ^2_r by taking the average over the 10,000 replications. The approximated true parameter σ^2_r for the overall rater variance component in VC composition (a) was 0.084, 0.070, and 0.063 for $n_r = 4$, 8, and 16, respectively, and the approximated true parameter for the overall rater variance component in VC composition (b) was 0.590, 0.491, and 0.442 for $n_r = 4$, 8, and 16, respectively.

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.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

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*, 238–257.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford, UK: Oxford University Press.

Bouwer, R., Béguin, A., Sanders, T., & van den Bergh, H. (2015). Effect of genre on the generalizability of writing scores. *Language Testing*, *32*, 83–100.

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

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

Chapelle, C. A., Enright, M. K., & Jamieson, J. (Eds.) (2008). *Building a validity argument for the Test of English as a Foreign Language*. London: Routledge.

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*, 321–338.

Crick, J. E., & Brennan, R. L. (1982). *GENOVA: A generalized analysis of variance system (FORTRAN IV computer program and manual)*. Dorchester, MA: Computer Facilities, University of Massachusetts at Boston.

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

Davis, C. (1974). Bayesian inference in two way models: An approach to generalizability (Unpublished doctoral dissertation). University of Iowa, Iowa City, IA.

de Finetti, B. (1931). Funzione caratteristica di un fenomeno aleatorio. *Atti della R. Academia Nazionale dei Lincei, Serie 6. Memorie, Classe di Scienze Fisiche, Mathematice e Naturale*, *4*, 251–299.

DeMars, C. (2015). Estimating variance components from sparse data matrices in large-scale educational assessments. *Applied Measurement in Education*, *28*, 1–13.

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

In'nami, Y., & Koizumi, R. (2015). Task and rater effects in L2 speaking and writing: A synthesis of generalizability studies. *Language Testing*. Advance online publication. doi: 10.1177/0265532215587390

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*, 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*, 158–180.

Mooney, C. Z. (1997). *Monte carlo simulation* (No. 116). Thousand Oaks, CA: SAGE Publications.

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*, 62–78.

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing*, *22*, 1–30.

Shavelson, R. J., & Webb, N. M. (1981). Generalizability theory: 1973–1980. *British Journal of Mathematical and Statistical Psychology*, *34*, 133–166.

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

Stemler, S. E., & Tsai, J. (2008). Best practices in interrater reliability: Three common approaches. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 29–49). Thousand Oaks, CA: SAGE Publications.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*, 263–287.

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing*, *6*, 145–178.

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