Improving Performance Assessment Score Validation Practices: An Instructional Module on Generalizability Theory

Oksana Naumenko

University of North Carolina at Greensboro
o_naumen@uncg.edu

Abstract

In developing and validating measures of language abilities, researchers are faced with the challenge of balancing statistical and consequential qualities of language tests. With the growing use of performance assessments, or test types that require examinees to demonstrate the mastery of a specific complex skill by performing or producing something, so grows the need for quantitative tools that can disentangle variability in language assessment scores due to language ability from those due to irrelevant factors. A well-known, but underutilized, technique in the validation of language tests for such a purpose is Generalizability Theory. Generalizability Theory (G-theory) extends Classical Test Theory (CTT) in providing a mechanism for examining dependability of behavioral measurements (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). One of the main advantages of using G-Theory in establishing evidence of measurement soundness is that in this framework, the observed test score can be partitioned into components other than the true test score and random error. Examining the relative and absolute magnitudes of variance components related to factors of study design can uncover the sources of unreliability or imprecision in the data and make possible the estimation of reliability under yet unstudied conditions. The present paper serves as an instructional module for a set of analyses under the G-Theory framework and showcases an analytic study that exemplifies the various inferences that can legitimately be made from G-study results.

Introduction

Language testing and psychometrics, the study of measurement of psychological phenomena, have been burgeoning concurrently for the latter part of the 20th century (McNamara & Roever, 2006). Recently revived interest in the use of complex performance assessments has resulted in an expansion and utilization of psychometric theories and models in the quest of language test validation (Bachman, 2002; Chappelle, 1999; Messick, 1995). Performance assessments, or task-based language performance assessments (TBLPA; Bachman, 2002), are tests of language ability that promote the view of language proficiency as encompassing both the examinee's knowledge and ability to use language (Brindley, 1994; 2001) in determining proficiency. TBLPAs differ from the more traditional objective (i.e., multiple choice) tests in that they present additional psychometric challenges in the process of validation of inferences made from test scores. The classical notion of reliability (Cronbach & Meehl, 1955), for instance, no longer applies when consistency in test scores is not based on the correct/incorrect response dichotomy. Instead, test developers are faced with providing evidence for consistency relative to raters, testing occasions, and testing prompts simultaneously. Nevertheless, the additional challenges in providing validity evidence for TBLPAs yield fruitful advantages in strengthening the fidelity between test content and language behavior outside the testing context (Fitzpatrick & Morrison, 1971; Kane, Crooks, & Cohen, 1999). As articulated by McNamara (1996), TBLPAs "still

require us to address the fundamental problem in test validation, that is, the question of justifying inferences from test performance".

Language skills like reading and oral comprehension, spoken and written production, and spoken and written interaction are elusive, multifaceted constructs that require careful consideration in the process of test development. Increasingly, the manifestation of the influence of such constructs is elicited through the use of complex performance tasks that yield constructed responses. The process of validation requires the production of evidence to support inferences made from test scores. Among the many psychometric techniques that allow for inference validation is G-Theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). The purpose of this paper is to showcase the richness of validation evidence that can be acquired through the use of generalizability studies (G-studies) in the context of performance assessments, such as TBLPAs, in which human judgment may lead to differences in test scores based on human judgment.

The Process of Gathering Validity Evidence

While dissenting opinions about validity still persist (e.g., Borsboom, Mellenbergh, & van Haerden, 2004), the educational measurement field leans toward conceptualizing validity as the degree to which arguments support the interpretations and uses of test scores (Kane, 1992; 2006; Messick, 1989). Kane's interpretive argument is one such process of score validation that explicitly calls for consideration of test interpretation and use consequences. In this paper, an interpretive argument guides the validation of inferences from hypothetical language writing test scores. It should be noted that the components of an interpretive argument vary with the context of the testing situation. Thus, the focus here is on general statements pertinent to performance assessment validation.

Next, assumptions associated with initial steps in the "language writing" interpretive argument are listed and potential validity evidence is presented. While further steps in the interpretive argument exist, in this paper, the discussion is limited to domain description, evaluation/scoring, and generalization. For a more complete account of an interpretive argument for TBLPA scores on a test of foreign language, the reader is encouraged to consult Chapelle, Enright, and Jamieson (2011).

Domain description. The first step in Kane's interpretive argument is domain description, which involves the explication of the target domain of examinee performances to which eventually we wish to generalize. Kane, Crook and Cohen (1999) assert that two assumptions are embedded in the definition of performance assessment itself. First, the definition assumes that *the interpretation of examinee scores will emphasize levels of skill in some performance domain.* To this end, research on specific elements of the writing rubric should be presented. These studies should bolster the inclusion of certain rubric elements deemed important by language writing experts. Second, *the observations used to draw inferences about skill in this domain involve performances on tasks from the domain of interest.* The second assumption is supported by the fact that the product (i.e., essay in some language) is the same that would be composed for the domain of interest. Important skills needed for quality writing in any language can be identified and translated into a hypothetical language writing rubric. The literature on the evaluation of language writing skills is external to the scope of this paper.

Evaluation/Scoring. A description of the rubric development process by content experts can be used as evidence for the evaluation inference. At this stage, studies of rubric element analysis can be presented (Chapelle et al., 2008). Kane et al. (1999) describe the appropriateness of making inferences from performances to observed scores as dependent on one relevant assumption: *the criteria used to score the performance must be appropriate and have been applied as intended.* This

assumption mandates consideration of the appropriateness of the criteria that make up the scoring rubric in order to prevent construct-irrelevant variance (Chapelle, 2010; Kane, 1992). Conversely, failure to include some pertinent criteria can lead to construct underrepresentation. When score interpretation is required, rater selection and training and/or clear communication of the nature of the rubric criteria is essential. Scoring criteria present an additional source of measurement error over and above that of objective tests. To this end, it is important to review the characteristics of written language task components.

Rubric construction. Rubrics can be designed to be holistic or analytic (Popham, 1997). Whereas a holistic rubric requires raters to score the process or product as a whole without judging the component parts separately, an analytic rubric requires raters to score specific parts of the performance and then to sum the sub-scores to a total score (Moskal & Leydens, 2000; Nitko, 2001). With analytic rubrics, examinees receive performance feedback with respect to each of the individual scoring criteria (Nitko, 2001). In an analytic rubric, an implied continuum of proficiency is established for each of the individual scoring criterion. This is typically achieved by listing the criteria in the first column of the rubric table, then describing the continuum of proficiency at ordered intervals from unacceptable to acceptable via behavioral anchors. In this way, each criterion is scored separately using definitions of quality at each interval. Thusly, evaluative criteria are used to distinguish acceptable responses from unacceptable responses and can be weighted equally or differently. Quality definitions, which make up the behavioral anchors, describe the way that qualitative differences in examinee responses are to be judged. For example, in a writing situation "mechanics" and "style" are popular evaluative criteria. For each qualitative level, a description for each qualitative level defining the range of possible performances must be present. Quality definitions for each criterion are also called behavioral descriptions or anchors. The recommended number of such levels is four to five (Popham, 1997), to ensure enough differentiation between writing quality. Usually, these levels are associated with a quantitative label to facilitate quantitative analyses of performance and a general label associated with that performance (e.g., advanced).

Rater effects. One of the most egregious sources of error in language tests that use human raters is rater bias. Due to various factors such as inadequate training, misalignment of scoring and the worldview of the rubric or distractions while scoring, raters may provide unexpected ratings given true examinee language ability (Myford & Wolfe, 2002; 2004). The appropriateness of evaluative rubric use depends partially on the degree of possible rater effect. Saal, Downey, and Lahey (1980) outlined four major categories of rater errors: severity, central tendency, restriction of range, and halo effects. The first category of rater error, severity, or leniency, is a rater's tendency to consistently provide ratings that are lower or higher than is warranted by student performance (Myford & Wolfe, 2002; Saal, Downey, & Lahey, 1980). For example, it is possible for two raters to rank-order a set of writing samples by quality in the same exact way with one rater consistently rating all writing samples lower than the other rater. The second category of rater error, central tendency, represents raters' tendency to assign scores closer to the middle of the performance scale despite the true level of mastery (Engelhard, 1994; Myford & Wolfe, 2004; Saal, Downey, & Lahey, 1980). This bias results in ratings in the middle of the scale regardless of examinee performance. Scores concentrated in the middle of the scale will result in a reduction in variability, and thus low reliability coefficients. This phenomenon is known as restriction of range; such an effect introduces artificial dependency in rating (Saal, Downey, & Lahey, 1980). Restriction of range creates issues in conducting generalizability studies because when true score variability is restricted, identified rater or item biases can appear inflated in comparison. The fourth category of rater errors, halo, is described by Engelhard (2002) as a type of rater bias that occurs when raters do not discriminate between conceptually dissimilar and independent aspects of examinee performance. For instance,

if a rater was impressed by the accurate use of French adverbial pronouns *y* and *en*, it is possible that the rater would then inflate ratings of other writing rubric elements. Halo effects obscure an examinee's true score (Farokhi & Esfandiari, 2011), threatening the validity of inferences we can make from assessment results. Overall, rater inaccuracy results in low levels of consistency between assigned ratings and expected ratings, and thus contributes to issues with generalizability. To counterbalance this problem, it is best practice to use as many trained raters as possible when producing TBLPA scores (Myford & Wolfe, 2002).

Generalization. In language writing situations, test developers are interested in inferences about examinee language writing. Unfortunately, measurement error associated with TBLPA measures is difficult to eliminate. By understanding the sources and amount of error, one has a better idea of the precision associated with TBLPA scores. In transitioning from theory to measurement practice, Kane's (2002) terminology is used to reflect the language used in G-Theory (Brennan, 1997; Cronbach et al., 1972). Generalization involves inferring the quantity of expected scores over all aspects of the measurement situation (Shavelson & Webb, 1991). Among others, variability due to raters presents the concern of inter-rater reliability. To make the generalizability inference relative to raters, at this stage one needs to provide evidence that *raters produce consistent scores relative to each other* and that *raters consistently agree on an examinee's score relative to the behavioral anchors of the rubric*. That is, confidence in the inference that observed scores will be observed under similar rating conditions is increased to the extent that rater consistency is shown. Because samples of observations combine to produce all possible observations under the conditions of the current rating situation (i.e., the population, or universe of observations), *the universe score* is an appropriate term for the desired summary score based on all possible observations under specific rating conditions under study. Further, an individual's universe score is the expected (i.e., the average) score "over the universe of generalization" (Kane et al., 1999). The universe of generalization is one of the subdomains of what is termed the *target domain*, which contains all possible observations under *all* theoretical rating conditions. The inferences connecting the universe of observations and the target domain can be examined at a more advanced stage of Kane's validation scheme and is not discussed here. To support generalizations from the observed score to the universe score, researchers can conduct G-studies, which evaluate the consistency of scores across samples of observations (Kane, Crooks & Cohen, 1999). Next, Classical Test Theory and G-Theory are reviewed in relation to written language tests.

Classical Test Theory

In Classical Test Theory (CTT) a behavioral measure, X, is composed of the true underlying ability score, T, and error, e, which is considered to be due to random causes: $X = T + e$. In CTT, the error term is undifferentiated and considered random ($X = T + e_r$). CTT provides reliability coefficients that allow the estimation of the degree to which the T component is present in a measurement. Reliability can be defined as a correlation between true scores and all possible observed scores that could be calculated from a person taking a test infinitesimally (Lord, Novick, & Birnbaum, 1968). Reliability may also be described in terms of the proportion of variance in true scores to the variance in the observed scores (Mellenbergh, 1996). Thus, variance in observed scores, the denominator of this proportion, can consist of factors other than true score variance. From this perspective, reliability can be viewed as a complex term that is affected by many types of variance associated with a particular measurement situation; it is a sample-dependent estimate of measurement precision for a population.

Several types of reliability exist in CTT because the error term (e_r) is undifferentiated. Test-retest reliability provides information about the consistency of examinee test ranks over time.

On the other hand, internal consistency measures the degree to which individual items in a test provide similar and consistent examinee scores. Further, parallel-forms reliability examines the rank-ordering of examinees by score across two alternative test forms. The error variance estimates vary depending on the reliability index of interest (Embretson, & Hershberger, 1999).

An alternative measure of consistency, the standard error of measurement (SEM) describes the standard deviation of errors of measurement associated with true score estimates derived from a sample of observed scores (Harvill 1991; Lord, Novick, & Birnbaum, 1968). SEM may be a more practical precision indicator than a reliability coefficient because it refers to a specific type of variance, caused by the fluctuations of observed scores around the true score. Thus, it is not dependent on true score variability within a sample and can be conceptualized more accurately as score precision around a certain scale point. It is precision of measurement for a given subject (Mellenbergh, 1996). SEM can be useful when examining precision of scores associated with a scale under question. Because reliability is dependent upon the variability of scores in a particular sample, comparing reliability coefficients may not be as meaningful as comparing SEM across samples.

Generalizability Theory

In contrast to CTT, in the G-Theory framework, the error term can be partitioned into systematic error and random error, $X = T + e_s + e_r$.The e_s element represents facet variability that can be further partitioned depending on the number of facets involved in the research design. These systematic variances are called *variance components*, which can be calculated and applied in determining the dependability of a measurement (Cronbach et al, 1972). In the writing sample writing assessment design, variance components are associated with raters and elements facets. Systematic variance is also calculated for the object of measurement, *person.*

Similar to variables having values, facets are comprised of levels that can be defined as *random* or *fixed* (Shavelson, & Webb, 1991). Random facets include levels that can be exchanged from the universe of generalization. Conceptually, a facet that is random indicates that the levels included in the analysis are an unbiased sample of levels that could be drawn from the universe of generalization (Cronbach et al., 1972). In the case of an objective test, an item facet is considered random if it is truly interchangeable with any other item measuring the same unidimensional trait. Conversely, fixed facet levels represent the full theoretical scope of the facet and cannot be exchanged with any other level. A facet is fixed when the number of levels in its universe matches the observed number of levels (Shavelson, & Webb, 1991). For example, imagine an alphabet test in which 26 items each represent a letter. In this case the measured levels exhaust the universe of generalization. The item facet, therefore, would be considered fixed. Fixed facets do not contribute systematic score variance to a fully-crossed design because they are held constant.

In the G-Theory framework, the object of measurement can be crossed with different facets. Crossing notation is such that if all foreign language constructed responses in the sample are reviewed by all raters, $p \times r$. In a fully-crossed design for a test of writing ability, each level of every facet and the object of measurement are crossed. For example, all essay entries can be crossed with all levels of the rater facet, which indicates that every rater provides rubric ratings for every constructed response in the measurement situation.

The object of measurement can also be nested in certain facets (Shavelson & Webb, 1991). The notation for essays being nested in raters is $p : r$. When the objects of measurement or facets are nested within the population of objects of measurement, it becomes more difficult to differentiate effects as they become confounded. For example, when sets of raters are nested within rater teams, the universe of admissible observations contains raters that are associated with only one

rater team. Crossed designs are favored in generalizability studies, although nested designs are often used for convenience or for increasing sample size (Shavelson, Webb, & Rowley, 1989). Increasing the sample size, in turn, typically reduces estimated error variance and increases estimated generalizability (Shavelson, Webb, & Rowley, 1989). Practically speaking, the optimal study design in some situations may well be the nested design.

Nested and crossed manipulations of the object of measurement, and random or fixed facets yield coefficients of dependability. Unlike CTT, G-Theory differentiates between two types of reliability or dependability: *relative* and *absolute* reliability (Shavelson & Webb, 1991). Relative dependability refers to the consistency with which examinees can be ranked based on language performance skill. For instance, constructed response scores for the test of written language can be ranked for each person across two or more raters; the consistency with which the raters rank the writing quality of each person is relative to each written sample. This type of dependability is represented by the G-coefficient. However, because in many cases considering absolute quality of writing in language is more meaningful rather than simply comparing constructed responses across examinees, absolute dependability of a measure can be more relevant. Absolute dependability is consistency with which scores occur around a particular scale point. This dependability is represented by a Φ (phi) coefficient. Thus, it is possible to determine consistency with which ratings from different raters occur around a specific quality point of writing.

To determine the magnitude of Φ- and G-coefficients, one must first calculate both the relative and the absolute variances associated with the study design. Specifically, relative and absolute variances are calculated by adding the variance components from the G-study after adjusting for the levels associated with each facet. Relative variance can be calculated by taking the sum of adjusted error variances that are directly related to the object of measurement, while absolute variance consists of all summed variances including those not due to or crossed with the object of measurement. Equations 1 and 2 respectively represent the relative and absolute variances of a fully-crossed design with a rater and element facets. It should be noted that the square root of $\hat{\sigma}^2_{Abs}$ (also $\hat{\sigma}^2_{(\Delta)}$) is the standard error of measurement (SEM).

$$\hat{\sigma}^2_{Rel} = \frac{\hat{\sigma}^2_{pr}}{n'_r} + \frac{\hat{\sigma}^2_{pi}}{n'_i} + \frac{\hat{\sigma}^2_{pri,e}}{n'_r n'_i} \quad [1]$$

$$\hat{\sigma}^2_{Abs} = \frac{\hat{\sigma}^2_{r}}{n'_r} + \frac{\hat{\sigma}^2_{i}}{n'_i} + \frac{\hat{\sigma}^2_{pr}}{n'_r} + \frac{\hat{\sigma}^2_{pi}}{n'_i} + \frac{\hat{\sigma}^2_{ri}}{n'_r n'_i} + \frac{\hat{\sigma}^2_{pri,e}}{n'_r n'_i} \quad [2]$$

where $\hat{\sigma}^2_r$ is the rater facet variance component, $\hat{\sigma}^2_i$ is the element facet variance component, $\hat{\sigma}^2_{pr}$ is the person by rater interaction variance component, $\hat{\sigma}^2_{pi}$ is the person by element interaction variance component, $\hat{\sigma}^2_{ri}$ is the rater by element interaction variance component, and $\hat{\sigma}^2_{pri,e}$ is the person by rater by element interaction confounded with random error variance and other unidentified sources of error.

Using variances calculated with Equations 1 and 2, relative and absolute dependability coefficients for specific measurement designs can be estimated.

$$\hat{\rho}^2 = \frac{\hat{\sigma}_p^2}{(\hat{\sigma}_p^2 + \hat{\sigma}_\delta^2)} \quad [3]$$

$$\Phi = \frac{\hat{\sigma}_p^2}{(\hat{\sigma}_p^2 + \hat{\sigma}_\Delta^2)} \quad [4]$$

Thus, if the measurement situation contained different occasions, several raters and different tasks, it is possible to obtain estimated variance components relative to each measurement element in the design. Instead of calculating different reliability coefficients as is in CTT, the variance components are used to compose an overall estimate of dependability of data, which takes into account the measurement variance accounted for in the design. Similar to conducting analyses of variance (ANOVA), it is possible to calculate the proportion of variance associated with each unwanted variance source to the overall variability in the data.

The ground for the generalization inference is the observed score. At this stage, Chapelle and colleagues (2010) recommend conducting generalizability and reliability studies. In the following sections, an analytic example that demonstrates the utility of Generalizability Theory in the development of a hypothetical performance assessment is introduced. The example provides validity evidence regarding the generalizability inferences with regard to rubric scores. The assumptions associated with this inference include that *raters provide consistent scores relative to each other and raters consistently agree on an examinee score relative to the behavioral anchors on the writing rubric*. This research also investigates the rank order of examinee abilities for the differential rater severity and differential scoring criteria difficulties. The following research questions were investigated in this analytic example:

1. How generalizable are ratings provided by readers both in terms of relative and absolute decisions?
2. What are the variance component values associated with this design?
   (a) What is the extent of variance due to rubric criteria in the writing scores?
   (b) What is the extent of variance due to rater characteristics in the writing scores?
3. What is the precision of rater's scores relative to the rubric criteria?
   (a) What is the standard error of measurement associated with the current design?
   (b) What is the typical range of standard error of measurement associated with the design in which there are three raters scoring the same set of essays?

Method

The Instrument

A hypothetical rubric was conceived to measure examinee foreign language written production, based on the various scales from the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2009). For the purposes of the example, the rubric criteria were: 1)

general linguistic range, 2) vocabulary range, 3) grammatical accuracy, 4) vocabulary control, 5) orthographic control, and 6) coherence. The analytic rubric components (or elements) were rated on a reduced CEFR scale (four scale points instead of six) supplemented with detailed behavioral anchors.

Participants

The hypothetical human readers in this study were 12 language writing content experts employed to rate written examinee responses. It is also assumed that readers were moderately trained to use the current version of the writing rubric during at least two separate training sessions. Good training practices include training sessions containing a general overview of the rubric component overview as well as several practice rating sessions. Readers may be asked to rate each of the practice written responses individually after which the responses were calibrated.

Procedures

In a common rating procedure, it is not typically possible or practical to obtain ratings from each reader for each examinee. Thus, a nested design was considered. The participants constituted four rater groups with three raters in each group. The raters within each group rated an identical set of 12 writing samples. Further, two samples were rated by all raters across all four teams. Thus, overall 42 writing samples were rated: 40 within specific groups and two common to all groups. All groups rated a common pair of samples in order to compare group leniency/harshness across groups.

Consider also a situation where examinees were able to take the written examination at two time points: once before an intervention, and once after. In such a situation it is important to ensure that raters are not assigned writing samples corresponding to the same examinees for both assessment occasions. In fact, to avoid bias that may arise when the same rater scores both the samples from the same examinee, each rater team may be assigned ten unique writing samples (five pre- and five different post- samples). While avoiding bias due to exposure effects, the test developer can still acquire evidence of examinee growth from first to last testing occasion because separate sets of raters was assigned to rate pre-intervention samples and post-intervention samples. This design is represented in Figure 1. Rater team 1 scored ten pre-intervention and post-intervention samples that corresponded to ten post-intervention and pre-intervention samples rated by rater team 2. Writing samples were arranged similarly for rater team 3 and rater team 4. Two samples were rated by all teams.

	Group 1 (n=3)	Group 2 (n=3)	Group 3 (n=3)	Group 4 (n=3)
Pre-test	1	6	11	16
	2	7	12	17
	3	8	13	18
	4	9	14	19
	5	10	15	20
	21	**21**	**21**	**21**
Post-test	6	1	16	11
	7	2	17	12
	8	3	18	13
	9	4	19	14
	10	5	20	15
	22	**22**	**22**	**22**

Figure 1. *The Hypothetical Writing Sample Study Design.*

Note. The common set of writing samples contains one pre-intervention and one post-intervention writing sample rated by each of the 12 raters.

Results

The results of this study are organized into major parts by research question (i.e., RQs A, Ba, Bb, Bc). The experimental design was $[(p\times r):t]\times i$ with three raters nested within four teams rating twelve writing samples each, two of which are common to all four teams, and therefore to all twelve raters. Each rater evaluated six pre and six post writing samples for different students across the six elements of the rubric. The calculations were made by hand and when possible, checked through a software package designed to conduct G-studies, GENOVA (Crick & Brennan, 1986).

Analyses Addressing Psychometric Properties of the Assessment Tool

Although both the argument for fixed and for random rubric elements may be valid, $\hat{\rho}^2$ (G) and Φ (Phi) coefficients were calculated for a design that treats the item facet as fixed for brevity. The D-study modeled four levels of the random team facet, three levels of the random rater facet, and six levels of the fixed element facet.

Using generalizability notation, the full design of the dataset is $[(p\times r):t]\times i$, which indicates that each student writing sample, *p,* was crossed with raters, *r,* that both were nested in teams, *t*, and each student writing sample was rated on each of the fixed elements, *i.* Because writing samples and raters were crossed, but nested in teams, RQ A (dependability of writing sample ratings in terms of relative and absolute decisions) and RQ B (contributions of variance components associated with the partially nested design) were examined by conducting a partially nested analysis. Variance components for each facet and object of measurement (i.e., writing samples) are presented in Table 1. Because GENOVA does not calculate dependability coefficients for nested models, $\hat{\rho}^2$ and Φ coefficients were hand-calculated using formulae 3 and 4 where:

$$\hat{\sigma}^2_{\delta} = \frac{\hat{\sigma}^2_{pr,prt}}{n_r} \quad \text{and} \quad \hat{\sigma}^2_{\Delta} = \frac{\hat{\sigma}^2_{t}}{n_t} + \frac{\hat{\sigma}^2_{r,rt}}{n_r} + \frac{\hat{\sigma}^2_{pr,prt}}{n_r}$$

[5,6]

Dependability coefficients were in line with acceptable ranges in applied research contexts, $\hat{\rho}^2$ = .91 and Φ = .89. Although guidelines for acceptable values of Φ- and $\hat{\rho}^2$ -coefficients have not been established in the literature, it is justified to use the familiar Cronbach's α cutoffs. According to George and Mallery (2003), one can interpret Cronbach's α greater than .90 as "excellent", about .80 as "good", about .70 as "acceptable", about .60 as "questionable", about .50 as "poor," and anything less than .50 as "unacceptable." With these guidelines in mind, a $\hat{\rho}^2$ -coefficient value of .80 was interpreted as good relative dependability, whereas a Φ -coefficient of the same value was considered representative of very good absolute dependability. Thus, an answer to RQ A is that using the fixed elements design, the rubric ratings exhibited excellent relative and absolute dependability.

Next, focusing on conclusions RQ Bb (rater variance component magnitude), variance components associated with the partially nested design are described. The writing samples-nested-within-teams variance component (i.e., $\hat{\sigma}^2_{p,pt}$, proxy for object of measurement) was .2685, and

accounted for approximately 35.0 percent of total within-team variance in writing sample ratings. In other words, approximately a third of the average total rating variability within teams was between writing samples. In contrast, the raters-nested-within-teams variance component ($\hat{\sigma}^2_{r,rt}$ = .0199), which represented the amount of variability due to differences in rater leniency/harshness within teams, accounted for only 2.6 percent of the average total rating variability within teams. The writing sample by rater-nested-within-teams variance component ($\hat{\sigma}^2_{pr,prt}$ =.0700) indicated that 9.1 percent of the average total rating variability within teams was due to differences in the relative rank order of writing samples by raters. Thus in response to RQ Bb, systematic rater harshness and leniency was a relatively small issue within teams, on average.

The rubric element x writing sample-nested-within-teams variance component ($\hat{\sigma}^2_{pi,pit}$ = .2115) represented 27.6 percent of average total rating variability within teams. This indicated that a large proportion of within-team variability was due to differential average relative element difficulty across writing samples. In other words, within each team on average, rater groups scored some writing samples higher than average on some elements and other writing samples higher on other elements. The rater by element-nested-within-teams facet ($\hat{\sigma}^2_{ri,rit}$ =.0154), which illustrated the relative rater leniency/harshness on particular elements within teams, accounted for only 2.0 percent of the average total rating within-team variability. The final variance component related to within-team variance was the mixture of the variance attributable to the writing sample by rater by element interaction within teams and random error still remaining in the model. This confounded variance component, or "noise", ($\hat{\sigma}^2_{pri,prit}$ = .1818) accounted for 23.7 percent of average within team variability.

In addressing RQ Ba (rubric element variance component magnitude) the following variance components were relevant. The remaining variance components represented various facets of average variability between writing sample scores in general. The variance component representing variability of average team scores ($\hat{\sigma}^2_t$ = .0136) made up 1.3 percent of total variability. In contrast, variability due to element difficulty ($\hat{\sigma}^2_i$ = .2674) comprised 25.4 percent of total variability. Finally, the element by team variance component ($\hat{\sigma}^2_{ti}$ = .0003) represented .3 percent of total variability, indicating little dependence of element difficulty on team membership. Thus, in response to RQ Ba, there were substantial differences between element scores across all raters and teams.

The estimate for score precision (standard error of measurement, SEM) was calculated by taking the square root of the absolute error variance component associated with the fixed rubric elements design. The absolute SEM associated with this design was .18, indicating that raters within teams were on average about .18 points away from the universe score (RQ Ca, precision associated with the current design).

The percent variability in Table 1 was calculated differently for nested sources of variation (p:t, r:t, pr:t, pi:t, ri:t, and pri:t,e) and free sources of variation (t, i, ti) because variability within teams has a different meaning than variability between teams. Whereas the *variability within teams* for any one of the nested sources of variation is the average variability across four teams, *variability between teams* represents the actual variance between teams and items. For example, the magnitude of the "team" (t) variability denotes differences between teams' average writing sample scores. That is, variance is examined holistically. On the other hand, the "raters within teams" (r:t) source of variation represents *not* how all raters' average writing sample scores vary across the entire design, but the average of how raters vary *within each team*. Thus, percent total variance for each of the nested sources of variation (p:t, r:t, pr:t, pi:t, ri:t, and pri:t,e) was calculated separately using the denominator that is the sum of the nested variance components. The percent total variance for the free sources of variation (t, i, ti) was calculated out of the total

observed score variability. Total "nested" variability was .7671, whereas total variability was 1.0512.

Table 1. *Writing Sample Ratings Using the Partially Nested Design: Contribution of Each Facet to Score Variance*

Source of variation	Notation	Variance Component	SE	% Total Variance*
Writing samples within teams (p:t)	$\hat{\sigma}^2_{p,pt}$	0.2685	0.061	35.0
Raters within teams (r:t)	$\hat{\sigma}^2_{r,rt}$	0.0199	0.012	2.6
Team (t)	$\hat{\sigma}^2_{t}$	0.0136	0.029	1.3
Rubric Elements (i)	$\hat{\sigma}^2_{i}$	0.2674	-	25.4
ti	$\hat{\sigma}^2_{ti}$	0.0031	0.011	0.3
pr:t	$\hat{\sigma}^2_{pr,prt}$	0.0700	0.010	9.1
pi:t	$\hat{\sigma}^2_{pi,pit}$	0.2115	0.026	27.6
ri:t	$\hat{\sigma}^2_{ri,rit}$	0.0154	0.007	2.0
pri:t,e	$\hat{\sigma}^2_{pri,prit}$	0.1818	0.012	23.7

Note. *Variance components' % Total Variance for facets that were nested within the teams facet were calculated using total within team variance only. % Total Variance for non-nested components was calculated from the summed total of all variance components. The standard error (se) for the items facet (when fixed) was not calculated because the effect is not generalized to other samples, which precludes consideration of sampling distributions.

Disaggregated analyses of rater dependability by team. Although dependability coefficients were at least acceptable in the full D-studies, the variance components contributing to total variability within teams may differ depending on the team. Further, it is informative to learn about the typical variance component contributions when three raters score 12 artifacts. Four fully-crossed D-studies modeling three levels of the random rater facet (*r*) and six levels of the fixed element facet (*i*) were conducted to examine the within-team variability more closely. Dependability coefficients and absolute standard errors of measurement (SEMs) were calculated to describe dependability of writing sample ratings. Team SEMs summarize each team's rating precision. Team SEM can be interpreted as the overall team ratings' average distance from the team's universe scores. In other words, the extent to which the raters are close to the team average score is an indication of how close ratings are to each other on average within each team. Recall that the absolute SEM represents the precision of a team's writing sample set ratings relative to the rubric behavioral anchors. The absolute SEMs were calculated for each team by taking the square root of the absolute variance component from a D- study design using three raters and six fixed rubric elements. Table 2 contains the summary of variance component contributions in each of the four rater teams.

The following is the type of information available for scrutiny about each of the rater teams. On an ordinal scale of 1 to 4, Team 1 assigned an average writing sample score of 2.62 (i.e., between "below average" and "above average") to the combined set of six initial (pre) and six final (post) writing sample drafts. Team 1 members were consistent in rank-ordering writing samples by quality with each other ($\hat{\rho}^2$ = .87), and had good consistency relative to the rubric scale (ϕ = .84). On average, raters were 0.29 points away from the writing sample rating universe score. About 84 percent of rating variability was due to differences in writing sample quality and only 3.5 percent of rating variability was due to systematic rater harshness/leniency. This indicates that most of the differences in writing sample quality scores were due to actual differences in writing sample quality and not construct-irrelevant effects such as rater leniency or confusion due to rubric element ambiguity.

Table 2 describes variance component contributions for all four teams. Team members were fairly consistent with each other ($\hat{\rho}^2$ range .77 - .89), and had mostly adequate consistency relative to the scale (ϕ range .52 - .84). Three raters within each team varied in distance from the writing sample quality universe score (RQ Cb, precision associated with a three rater design). Standard error of measurement associated with the overall writing sample score ranged from .2492 to .3425. Notably, Team 2 had low total variability in writing sample quality ratings, which may have contributed to the low absolute dependability coefficient. For Team 2, only approximately 52 percent of score variability in the rubric scores was due to writing sample quality, whereas 33.9 percent were due to systematic rater leniency/harshness (RQIBb, rater variance component magnitude). That is, almost half of the differences in writing sample quality was due to rater leniency/harshness effects and misinterpretation of rubric elements or other unidentified sources of error. In such cases detecting true writing sample quality scores is difficult, and scores are considered less dependable than ratings with a higher percent of variability attributed to writing sample quality differences. It should be noted that on average, this group's average writing sample score was still similar to that of Team 1. Team 3 had the highest overall mean across all writing samples.

Table 2. *Variance Component Contributions within Each Rater Team in Four Fully-Crossed, Fixed Element Design Studies*

	$\hat{\sigma}^2_p$	$\hat{\sigma}^2_p$%	$\hat{\sigma}^2_r$	$\hat{\sigma}^2_r$%	$\hat{\sigma}^2_{pr,e}$	$\hat{\sigma}^2_{pr,e}$%	$\hat{\rho}^2$	ϕ	Total Variance
Team 1	0.4450	84.3	0.0182	3.5	0.0648	12.3	0.87	0.84	0.5694
Team 2	0.1271	52.0	0.0828	33.9	0.0345	14.1	0.79	0.52	0.2444
Team 3	0.2087	68.7	0.0336	11.1	0.0617	20.3	0.77	0.69	0.3040
Team 4	0.2005	76.4	0.0368	14.0	0.0253	9.6	0.89	0.76	0.2626

Note. Standard Error of Measurement (SEM) calculated using σ^2_Δ associated with each team's rating variability was 0.2880 for Team 1, 0.3425 for Team 2, 0.3087 for Team 3, and 0.2492 for Team 4.

Element impact. To gain a better understanding of the role elements play in writing sample rubric rating dependability, 24 team-by-team D-studies were conducted investigating one element at a time. These studies also provided the information about the precision around rubric element scores. In investigating rubric element difficulty, rubric elements were ranked according to the element mean magnitude (see Table 3 for element difficulty rank ordering). Whereas Team 1, Team 2 and Team 4 raters agreed on the average ranking of elements by difficulty, Team 3, characterized by a large $\hat{\sigma}^2_{pr,e}$ component (i.e., writing sample by rater interaction confounded with random

error), ranked element difficulties differently. The Objective element (rank order = 5) and Related Experience (rank order = 2) elements were ranked consistently by difficulty across all four teams. It should be noted that notwithstanding similar overall average rubric scores (2.62, 2.64, and 2.46 for Team 1, 2, and 4, respectively), the dependability of rubric scores from Team 2 was relatively lower in terms of consistency relative to the rubric anchors.

One may note that the large overall mean and the lack of accord in element difficulty rank ordering could be due to sampling error. Sampling error could affect: 1) better quality writing samples being assigned to Team 3 by chance (hence, the larger team mean), a different profile of strengths and weaknesses within the writing sample set assigned to Team 3 (hence, the different element rank order of element means), or 3) both.

Table 3. *Rank Order of Rubric Elements by Mean Score*

	Team 1		Team 2		Team 3		Team 4	
Element	*M*	Rank	*M*	Rank	*M*	Rank	*M*	Rank
General Linguistic Range	2.34	5	2.36	5	2.78	5	2.33	5
Vocabulary Range	2.82	2	3.11	2	3.19	2	2.53	2
Grammatical Accuracy	2.31	6	2.08	6	2.96	4	2.33	6
Vocabulary Control	2.82	3	2.58	3	3.24	1	2.53	3
Orthographic Control	2.38	4	2.54	4	2.61	6	2.39	4
Coherence	3.03	1	3.15	1	3.00	3	2.65	1
Total	2.62		2.64		2.96		2.46	

Note. The rank order was determined by giving the highest rank to the element with the highest mean score (descending order).

Rating precision related to elements. Standard errors of measurement (SEM) due to rubric elements summarize raters' precision around each element across all writing samples. The element SEM can be interpreted as the precision of the score based on a single element. Each element's absolute SEMs were analyzed for each rater team (see Table 4). Absolute SEM represents the precision of element ratings relative to the rubric behavioral anchors. In each of the 24 D-studies the rater facet was the only modeled source of interpretable systematic error; each study described the rater variance components associated with each rubric element. In addressing RQ Cb, (precision associated with a three rater design) absolute SEMs were calculated for each element by taking the square root of the absolute variance component from a D- study design using three raters.

Teams 2 and 3 appeared to have similar standard errors across the six arbitrarily labeled elements; Vocabulary Control had the smallest standard error (SE_{Team2} = .33, SE_{Team3} = .28), whereas General Linguistic Range (SE_{Team2} = .67, SE_{Team3} = .63) and Orthographic Control (SE_{Team2} = .73, SE_{Team3} = .57) had the largest standard error for these two teams. Teams 1 and 4 had the smallest standard errors associated with Coherence (SE_{Team1} = .25, SE_{Team4} = .24) and Vocabulary Control (SE_{Team1} = .26, SE_{Team4} = .25). For these two teams, the largest standard errors were associated with Grammatical Accuracy (SE_{Team1} = .60, SE_{Team4} = .55).

Table 4 provides the absolute standard errors for all four teams on all six elements. Overall, the general trends were that Vocabulary Control (SE range .25 - .33), Coherence (SE range .24 - .36), and Vocabulary Range (SE range .32 - .43) had the smallest standard errors across all four teams. In contrast, the largest standard errors were associated with Grammatical Accuracy (SE range .55 - .60), General Linguistic Range (SE range .47 - .67) and Orthographic Control (SE range

.52 - .73). This type of information is useful when considering potential ambiguities associated with rubric element behavioral anchors. When standard errors are high, the most plausible reason tends to be unclear language that prevents raters from agreement.

Table 4. *Rank-Order of Rubric Elements by Element Absolute Standard Errors of the Mean*

	Team 1		Team 2		Team 3		Team 4	
Element	*SE*	Rank	*SE*	Rank	*SE*	Rank	*SE*	Rank
General Linguistic Range	0.591	5	0.667	5	0.631	6	0.467	4
Vocabulary Range	0.357	3	0.434	3	0.360	2	0.326	3
Grammatical Accuracy	0.599	6	0.569	4	0.547	4	0.553	6
Vocabulary Control	0.263	2	0.333	1	0.282	1	0.255	2
Orthographic Control	0.520	4	0.731	6	0.569	5	0.516	5
Coherence	0.255	1	0.340	2	0.362	3	0.238	1

Note. The rank order is determined by ascending order of SEMs.

Discussion and Conclusion

In order to make appropriate inferences regarding assessment results, language test developers must take into account rating unreliability associated with performance assessments. Although many quantitative methods are available, generalizability studies are among the most appropriate options for providing evidence for inferences made about TBLPA scores. In this paper, various indicators acquired through generalizability analyses were used to investigate the relative and absolute dependability associated with an analytic example of performance assessment rubric scores. The rich information yielded through analyses provides evidence for several assumptions and inferences that can be made about TBLPA scores.

Validity of assessment inferences was introduced through the lens of Kane's (1992) validity argument approach. The first three stages of Kane's interpretive argument-based approach to test score validation were addressed – domain description, evaluation, and generalization. The extent to which assumptions were met dictates whether or not the next stages of the language writing rubric validation should be initiated. Reliability must be established before more advanced stages of validity can be examined.

In this study the dependability associated with the partially nested design appeared to be adequate for rating purposes (i.e., overall $\hat{\rho}^2$ = .91 and $\hat{\phi}$ = .89). When separate team D- studies were conducted to examine sets of 12 examinee samples rated by three raters, some instability in the dependability coefficients was revealed, indicating that within teams specific essay scores may be more biased than the overall dependability estimate initially suggested. This additional investigation is recommended when raters are grouped into teams, as our conclusions about the trustworthiness of scores changed.

Three possible solutions could address the issue with using a nested design. First, instead of having four separate rater teams, one could conduct a fully-crossed study with twelve raters scoring twelve student writing samples. In generalizability studies it is favorable to use a fully-crossed design (Brennan, 1992), which allows for a more precise estimate of the typical rater and rubric element effects on rubric scores. The drawback to fully-crossed designs is that the

object of interest must be crossed with every facet. This would present a resource issue in that many raters would be rating only a limited number of examinee samples. This in turn increases the chance of sampling error playing a part in the conclusions. That is, if the smaller sample of essays happened by chance to include mostly high-quality artifacts, then rater and rubric element variance components would not be representative of the full possible range of writing quality. Second, one could have four rater teams rating twelve different sets of writing. This approach, however, prevents making inferences regarding relative team harshness or writing quality across teams. It is problematic that raters within any one team with the highest examinee scores could happen to be the most lenient team, or could have by chance received examinee samples that were better in quality. A third option, then, would be to include anchor essays that would be rated by all teams. With a large enough sample of anchors, this mechanism would allow comparisons of writing quality across all rater teams.

Other than the overall dependability of performance assessment scores, a major consideration of this study was how the rubric element and rater facets contributed to overall score variability. This was useful because examining raters in isolation revealed the most direct information regarding rater training. Likewise, examining elements separately contributed the most direct information about improving the language writing rubric.

An examination of standard errors of measurement associated with the rubric element-specific D-studies using three raters uncovered several areas recommended for improvement. The absolute standard errors associated with elements-specific analyses allow test developers to speak to the precision with which raters score each rubric element on average (in each team). In our hypothetical situation, across the four teams, raters agreed best on two writing elements, but were less precise in rating others. Using precision information supplied by absolute SEMs it is possible to develop strategies to improve reliability.

Well-designed scoring rubrics respond to the concern of intra-rater reliability by establishing a description of the scoring criteria in advance (Moskal & Leydens, 2000). Establishing clear scoring criteria will allow raters to refer to constant scoring rules, and emphasizing the importance of frequently revisiting the rubric criteria can maintain consistency within each rater (i.e., decrease the rater by rubric element interaction variance components). Given the rubric element precision information acquired through a G-study, it is possible to revisit the rubric design, identify potential problematic wording at the source of rater disagreement, and pilot systematic manipulations of wording in rubric studies. G-studies thus provide detailed enough information necessary to pinpoint and improve specific issues in the design of TBLPA rubrics and scoring of TBLPAs.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing (US). (2014). *Standards for Educational and Psychological Testing*.

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing*, *19*(4), 453-476.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061.

Brennan, R. L. (1997). A Perspective on the History of Generabability Theory. *Educational Measurement: Issues and Practice, 16*(4), 14-20.

Brindley, G. 1994. Competency-based assessment in second language programs: some issues and questions, *Prospect 9*: 41-85.

Brindley, G. (2001). Language assessment and professional development. *Experimenting with uncertainty: Essays in honour of Alan Davies, 11*, 137-143.

Chapelle, C. A. (1999). Validity in language assessment. *Annual review of applied linguistics, 19*, 254-272.

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

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

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

Council of Europe (2009). *Manual for relating language examinations to the Common European Framework of Reference for Languages (CEFR)*. Strasbourg, France: Council of Europe. Retrieved from http://www.coe.int/t/dg4/linguistic/Manuel1_EN.asp

Cronbach, L. J., Gleser, G. C., & Nanda, H. Rajaratnam. N.(1972). The dependability of behaviour measurement: Theory of Generalizability for scores and profiles. *American Educational Research Journal. 11*(1), 54-56.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281.

Embretson, S. E., & Hershberger, S. L. (Eds.). (1999). *The new rules of measurement: What every psychologist and educator should know*. New York, NY: Psychology Press.

Farrokhi, F., & Esfandiari, R. (2011). A many-facet Rasch model to detect halo effect in three types of raters. *Theory and Practice in Language Studies, 1*(11), 1531-1540.

Fitzpatrick, R., & Morrison, E. J. (1971). Performance and product evaluation. *Educational Measurement, 2*, 237-270.

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference. 11.0 update*. Boston, MA: Allyn & Bacon.

Harvill, L. M. (1991). Standard error of measurement. *Educational Measurement: Issues and practice, 10*(2), 33-41.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*(2), 5-17.

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). Statistical theories of mental test scores. New York, NY: Addison-Wesley.

McNamara, T. (1996). *Second language performance measuring*. New York, NY: Longman.

McNamara, T. & Roever, C. (2006). Language testing: the social dimension. *International Journal of Applied Linguistics, 16*(2), 242-258.

Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods, 1*(3), 293.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5-11.

Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741.

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research & Evaluation, 7*(10), 71-81.

Myford, C. M., & Wolfe, E. W. (2002). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386-422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet

Rasch measurement: Part II. *Journal of Applied Measurement, 5*(2), 189-227.
Nitko, A. J. (2001). *Educational assessment of students*. Des Moines, IA: Prentice-Hall.
Popham, W. J. (1997). What's wrong-and what's right-with rubrics. *Educational leadership, 55*, 72-75.
Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*(2), 413.
Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist, 44*(6), 922.
Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer* (Vol. 1). Sage Publications.