Article

Applied Psychological Measurement
2015, Vol. 39(4) 314–329

DOI: 10.1177/0146621614563067
apm.sagepub.com

Methods for Evaluating Composite Reliability, Classification Consistency, and Classification Accuracy for Mixed-Format Licensure Tests

Tim Moses[1] and Sooyeon Kim[2]

Abstract

The purpose of this study was to propose extensions of reliability estimation methods that could be used to determine the conditions under which single scoring for constructed-response (*CR*) items is as effective as double scoring in mixed-format licensure tests. Multivariate generalizability theory methods traditionally used to estimate overall composite score reliability were extended with simulations so that classification consistency and classification accuracy estimates could also be obtained. Composite score reliabilities, classification consistencies, and accuracies were estimated based on the double and single scoring of the *CR* items of three licensure tests. Composite score reliabilities, classification consistencies, and accuracies were also estimated in decision studies considering varied testing situations such as different numbers of *CR* items and different *CR* section weights.

Keywords

generalizability theory, rater effects, performance assessment, classification, reliability

The estimation of scoring consistency for mixed-format licensure tests can be as complex and multi-faceted as the initial construction of the tests themselves. For the composite scores of licensure tests, which are formed as weighted combinations of multiple-choice (*MC*) and constructed-response (*CR*) scores, there are multiple sources of unreliability that could be evaluated, including the measurement heterogeneity of the *MC* items, the measurement heterogeneity of the *CR* items, and rating errors in the *CR* scores. Additional complexities arise in licensure tests, as the use of classification processes can mean that classification accuracy and consistency at specific cutscores are of greater interest than reliability estimates that summarize errors across all composite scores. The estimation of composite score reliability is often addressed with multivariate generalizability theory (G-theory) analyses (Brennan, 2001; Clauser, Balog, Harik, Mee, & Kahraman, 2009; Jarjoura, Early, & Androulakakis, 2004). G-theory investigations of

[1]College Board, Newtown, PA, USA
[2]Educational Testing Service, Princeton, NJ, USA

Corresponding Author:
Tim Moses, College Board, 661 Penn Street, Suite B, Newtown, PA 18940, USA.
Email: tmoses@collegeboard.org

composite score reliability tend to be distinct and not directly informative to the analyses and investigations of classification consistency and classification accuracy (Brennan & Wan, 2004; Livingston & Lewis, 1995).

The purpose of the current study was to propose extensions of multivariate G-theory that can be used not only in the estimation of composite score reliability and the implications of different testing conditions for score reliability (e.g., *CR* weights, numbers of *CR* items, etc.) but also to consider estimates of classification consistency and classification accuracy. The proposed methods are demonstrated in the estimation of gains in composite score reliability, classification consistency, and classification accuracy that result when the *CR* tasks of a test are obtained from two ratings (double scoring) rather than one rating (single scoring).

Method

The methods used to address the licensure testing questions of double versus single *CR* scoring implications were based on multivariate G-theory models reflective of the data collected from recent administrations of three licensure tests. The multivariate G-theory models were used to produce classification consistency evaluations and decision studies where reliability implications could be assessed with respect to the manipulation of exam features of interest (i.e., the numbers of *CR* items and the percentage contributions to composite scores). The models, methods, and reliability estimates are presented after the data and structure of three licensure tests are described.

Data and Structure of the Three Licensure Tests

The empirical data for this study came from three operational tests used to make occupational licensing decisions for aspiring primary and secondary school teachers. Tests A and B are designed to measure subject knowledge in English and Mathematics for middle school teachers. Test C is intended primarily for prospective teachers of young children. It is based on a teaching approach that emphasizes the active involvement of young children in a variety of play and child-centered activities. Table 1 presents general information for each of the tests, such as numbers of *MC* and *CR* items, proportions of *MC* and *CR* components, and statistical characteristics of the scores from recent test administrations. In typical exam administrations, two independent raters from the rating groups of each test scored 1, 2, or all of the *CR* items of an exam (i.e., raters are not strictly assigned to score only 1 or all *CR* items of an exam).

For each *CR* item, the written responses are scored by two qualified raters trained to score the responses to that item according to pre-specified scoring rubric and score scales. A scoring rubric is developed by educators who are specialists in the subject area. All raters receive training before they score operational responses. The rating scales for each item are established as 0 to 3 point scales intended to reflect the evidence an item is designed to elicit from test takers. Assessment developers are responsible for the creation of scoring guides, the selection of sample responses for training purposes, and the training of scoring leadership in test content and scoring standards and procedures.

Like many large-scale tests with *CR* items scored by human raters, an unfortunate characteristic of the data collection from operational scoring is that there was little control over the raters assigned to score responses to one or more *CR* items of an exam. The result is that only one or very few of the *CR* scores summarized in Table 1 were obtained such that the *CR* items were all scored by a specific pair of raters. As described in subsequent sections below, the collection of the test data limited the G-theory design and models that could be considered.

Table 1. General Information for the Three Tests Used in This Study's Empirical Analyses.

	Test A		Test B		Test C	
		$CR_{SingleScoring}$		$CR_{SingleScoring}$		$CR_{SingleScoring}$
	MC	*CR* ratings 1,2	*MC*	*CR* ratings 1,2	*MC*	*CR* ratings 1,2
Total items	90	2	40	3	60	6
Max. Poss. Unwgt. Score	90	6	40	9	60	18
MC and *CR* section weights	1	2.5	1	1.11	1	1.67
% Composite	75	25	67	33	50	50
Total *CR* raters		25		31		68
Examinee sample size		1,211		1,803		715
M	65.8	4.00, 4.01	25.2	4.13, 4.12	45.3	13.2, 13.3
SD	10.5	1.06, 1.07	6.6	2.40, 2.39	5.8	2.27, 2.31
Skewness	−0.53	0.05, 0.02	−0.08	0.16, 0.17	−0.69	−0.29, −0.41
Kurtosis	−0.11	−0.53, −0.59	−0.46	−0.88, −0.86	0.66	−0.08, 0.033
Corr. w/*MC*	1.00	.37, .39	1.00	.70, .70	1.00	.40, .38
Corr. between *CR* ratings		.86		.98		.68

Note. *MC* = multiple choice; *CR* = constructed response; Max. Poss. Unwgt. Score = maximum possible unweighted score; % Composite = each score contribution to weighted composite score; Corr. w/*MC* = correlation with *MC* score.

The weighted composite scores of each test were calculated for the $p = 1$ to N_p examinees on the $j = 1$ to N_j *MC* items scores and the $i = 1$ to N_i *CR* items. For weighted composite scores based on the *CR* scores from the two ratings (i.e., double scoring as in current operational practice), weighted composite scores are calculated as,

$$\text{Composite}_{p,\,\text{DoubleScoring}} = w_{MC}MC_p + w_{CR}CR_{p,\,\text{DoubleScoring}}, \quad (1)$$

where $CR_{p,\,\text{DoubleScoring}}$ denotes examinee p's score on the *CR* section from $h = 1$ to $N_h = 2$ raters obtained as, $\sum_{h=1}^{Nh=2}\sum_{i}^{Ni} CR_{phi}$, MC_p denotes examinee p's score on the *MC* section obtained as, $\sum_{j}^{Nj} MC_{pj}$, w_{MC} is the weight of the *MC* section scores and equals 1 for the three tests of this study, and w_{CR} is the weight of the *CR* section scores (Table 1).

Weighted composite scores based on *CR* scores from one rating of each *CR* item response (i.e., single scoring) are not obtained in current operational practice. Their likely computation for rating $h = 1$ (and similarly for $h = 2$) would be,

$$\text{Composite}_{p,\,\text{SingleScoring}(h=1)} = w_{MC}MC_p + 2w_{CR}CR_{p,\,\text{SingleScoring}(h=1)}, \quad (2)$$

where $CR_{p,\,\text{SingleScoring}(h=1)}$ denotes examinee p's *CR* section scores from a single rater obtained as, $\sum_{i}^{Ni} CR_{p(h=1)i}$.

For each of the three tests, the value of the *CR* weight in Equations 1 and 2 and in Table 1 was selected such that the maximum possible weighted *CR* score from double scoring (or from single scoring multiplied by 2) would be a desired percentage of the maximum possible score on the weighted composite. The percentages were defined by assessment developers' desires for content, item formats, and time allocation rather than on psychometric interests like reliability maximization. For Tests A and B, examinees' weighted *CR* scores made up 25% and 33% of the weighted composite. For Test C, examinees' weighted *CR* scores made up 47% of the weighted composite.

For each of the three licensure tests, examinees are classified into proficiency categories on the basis of cutscores on the weighted composite scales. Because each test has multiple test users who do not use the same cutscores for the tests, each test has many cutscores. A review of these cutscores across the tests revealed that several were approximately one standard deviation below the mean of the testing population. Thus, in this study, the cutscores of interest were set at one standard deviation below the population mean for the composite score of each test form and administration group. Examinees were classified into the passing group if their composite scores met or exceeded the cutscore. If not, examinees were classified into the failing group.

Multivariate Generalizability Theory, Models, and Composite Score Reliability Estimates

The three licensure tests considered in this study reflect the data collection design defined in Brennan (2001) as the $p^{\bullet}\times i^{\circ}\times h^{\circ}$ multivariate design, presented here as a $p^{\bullet}\times(i^{\circ}\text{ or }j^{\circ})\times h^{\circ}$ design to convey that the *i* and *j* items are nested within format (i.e., the *i* nested in the *CR* section and the *j* in *MC*), where ratings (*h*) occur only for the *CR* items/section, and where individual examinees (*p*) obtain scores from all *MC* items, *CR* items, and *CR* ratings. In G-theory terminology, the examinees (*p*) represent objects of measurement, which are realized in the G-Study as random samples from a population, whereas the items (*i* and *j*) and ratings (*h*) represent facets or conditions of measurement, which appear as samples from universes of items and ratings. Because multivariate designs can be explained and understood in terms of their univariate parts, the description of this design begins by presenting the separate models for the *CR* and *MC* section scores.

The general model for examinee *p*'s unweighted *CR* section score that corresponds to the weighted composite score in Equation 1 and to the $p^{\bullet}\times i^{\circ}\times h^{\circ}$ design of interest is,

$$CR_{p,\,\text{DoubleScoring}} = N_h N_i t_{CR,p} + \left(N_h \sum_{i}^{Ni} v_i + N_i \sum_{h}^{Nh=2} v_h + \sum_{h}^{Nh=2}\sum_{i}^{Ni} v_{ih} \right) + \left(N_h \sum_{i}^{Ni} v_{pi} + N_i \sum_{h}^{Nh=2} v_{ph} + \sum_{h}^{Nh=2}\sum_{i}^{Ni} v_{pih,e} \right). \quad (3)$$

In Equation 3, $t_{CR,p}$ is the true score for examinee *p*, which has a mean equal to their expected score on all parallel forms (divided by $N_i N_h$) that could be assembled from randomly sampling N_i *CR* items from the universe of *CR* items, and from random ratings effects based on N_h ratings. The *v* terms are error terms, the first three indicating errors reflective of the leniency/stringency of the ratings, v_h, the difficulty of the *CR* items, v_i, and their interaction with each other, v_{ih}. The final three *v* terms indicate ''relative'' errors in their interactions with *p*. The ''*e*'' in the $v_{pih,e}$ effect indicates sources of error not directly estimated in the design but which contribute to the scores.

The implication of collecting the operational testing data such that raters were not systematically assigned to exams is that the raters comprise a random and hidden facet of the design and analyses and rater effects are confounded with the ratings and the '' . . . , *e*'' errors. With a more systematic assignment of exams to rater pairs, it might have been possible to observe specific rater combinations for several exams, estimate the variance components of the G-theory models for several specific pairs of raters and then evaluate reliability in the variance components after averaging the estimates obtained for each specific rater pairing (Brennan, 2001; see also Clauser et al., 2009). Conducting separate analyses for each separate pairing of raters was not possible

in this study due to relatively few exams being scored by each rater pairing. The limited data collection design also meant that it was not possible to consider more desirable G-theory models, such as designs with crossed rater effects where raters provided ratings on all the *CR* items or designs with nested rater effects where raters scored only one of the *CR* items in an exam. As a result, the effects involving raters are described as rating effects, which are assumed to be confounded with specific rater effects.

All v effects have assumed population means of zero. All of the v and $t_{CR,p}$ terms are assumed to be independent of each other. From Equation 3, the overall variance of the double-scored *CR* section scores based on two ratings is the sum of seven variance components,

$$\sigma^2\left(CR_{\text{DoubleScoring}}\right) = N_h^2 N_i^2 \left\{ \sigma^2(t_{CR}) + \left[\frac{\sigma^2(v_i)}{N_i} + \frac{\sigma^2(v_h)}{N_h} + \frac{\sigma^2(v_{ih})}{N_i N_h} \right] + \left[\frac{\sigma^2\left(v_{pi}\right)}{N_i} + \frac{\sigma^2\left(v_{ph}\right)}{N_h} + \frac{\sigma^2\left(v_{pih,e}\right)}{N_i N_h} \right] \right\}. \tag{4}$$

The models of the *CR* item and section scores obtained from single scoring and multiplied by 2 (as shown in Equation 2) are special cases of Equation 3,

$$CR_{p,\,\text{SingleScoring},\,h} = 2\left[N_i t_{CR,p} + \left(\sum_i^{Ni} v_i + N_i v_h + \sum_i^{Ni} v_{ih} \right) + \left(\sum_i^{Ni} v_{pi} + N_i v_{ph} + \sum_i^{Ni} v_{pih,e} \right) \right], \tag{5}$$

and Equation 4,

$$\sigma^2\left(CR_{\text{SingleScoring}}\right) = 2^2 N_i^2 \left\{ \sigma^2(t_{CR}) + \left[\frac{\sigma^2(v_i)}{N_i} + \sigma^2(v_h) + \frac{\sigma^2(v_{ih})}{N_i} \right] + \left[\frac{\sigma^2\left(v_{pi}\right)}{N_i} + \sigma^2\left(v_{ph}\right) + \frac{\sigma^2\left(v_{pih,e}\right)}{N_i} \right] \right\}. \tag{6}$$

Equations 4 and 6 involve multiplications by the squared numbers of ratings and *CR* items outside of the {} brackets, multiplications which reflect the composite score summations used in operational practice (Equations 1 and 2). These multiplications are omitted in more familiar G-theory descriptions, where it is assumed that scores are averaged.

The fundamental expectation is that the reliability of *CR* scores obtained from double scoring will be higher than those obtained from single scoring, presumably resulting in more reliable composite scores. These expected reliability differences from double versus single scoring are directly reflected in the models of the variance of the *CR* section scores obtained from double scoring (Equation 4) and from single scoring (Equation 6). That is, for *CR* scores obtained from single scoring and multiplied by 2, the errors from that single rating will also be multiplied by 2 (Equation 5), and will not receive the same reduction (division by $N_h = 2$) as in double scoring. To the extent that ratings contribute error variance to the *CR* scores, the errors in the *CR* scores from single-scored *CR* ratings multiplied by 2 will be greater than those from double scoring.

The models for the *MC* scores and section variance are similar to those of the *CR* scores (Equations 3 and 4), but are simpler because for *MC* items, there is only a single rating and rating effects involving (h) are assumed to be zero. These models are expressed with j indicating the *MC* items as,

$$\begin{aligned} MC_p &= N_h N_j t_{MC,p} + \left(N_h \sum_j^{Nj} v_j + N_j \sum_h^{Nh=1} v_h + \sum_h^{Nh=1} \sum_j^{Nj} v_{jh} \right) + \left(N_h \sum_i^{Nj} v_{pj} + N_i \sum_h^{Nh=1} v_{ph} + \sum_h^{Nh=1} \sum_j^{Nj} v_{pjh,e} \right) \\ &= N_j t_{MC,p} + \sum_j^{Nj} v_j + \sum_j^{Nj} v_{pj,e}, \end{aligned} \tag{7}$$

and

$$\sigma^2(MC) = N_h^2 N_j^2 \left\{ \sigma^2(t_{MC}) + \left[\frac{\sigma^2(v_j)}{N_j} + \frac{\sigma^2(v_h)}{N_h} + \frac{\sigma^2(v_{jh})}{N_j N_h} \right] + \left[\frac{\sigma^2(v_{pj})}{N_j} + \frac{\sigma^2(v_{ph})}{N_h} + \frac{\sigma^2(v_{pjh,e})}{N_j N_h} \right] \right\}$$

$$= N_j^2 \left[\sigma^2(t_{MC}) + \frac{\sigma^2(v_j)}{N_j} + \frac{\sigma^2(v_{pj,e})}{N_j} \right]. \quad (8)$$

The complete multivariate G-theory model requires the covariances for the effects in the univariate *CR* and *MC* models (Equations 3-8). For the $p^{\bullet} \times (i^{\circ} \text{ or } j^{\circ}) \times h^{\circ}$ design, the only crossed effect is of the *p* examinees taking the *MC* and *CR* sections such that examinees' *MC* and *CR* true scores have covariance, $\sigma(t_{CR}, t_{MC})$. With this covariance, the variance of composite scores obtained as a weighted sum of the *MC* and *CR* scores from double scoring (Equation 1) or single scoring (Equation 2) is the sum of the variance components of the *MC* scores (Equation 8), plus $N_j N_h N_i w_{MC} w_{CR} 2\sigma(t_{CR}, t_{MC})$, plus w_{CR}^2 times the variance components of the double-scored *CR* scores (Equation 4) or those of the single-scored *CR* scores (Equation 6).

The overall score reliability of this weighted composite that reflects both the relative and absolute error variances that affect scores and score decisions in licensure testing can be estimated as (Brennan, 2001; Joe & Woodward, 1976),

$$\Phi(\text{Composite}_{\text{DoubleScoring}}) =$$

$$\frac{\sigma^2(t_{\text{Composite}})}{\sigma^2(t_{\text{Composite}}) + N_h^2 N_i^2 w_{CR}^2 \left\{ \left[\frac{\sigma^2(v_i)}{N_i} + \frac{\sigma^2(v_h)}{N_h} + \frac{\sigma^2(v_{ih})}{N_i N_h} \right] + \left[\frac{\sigma^2(v_{pi})}{N_i} + \frac{\sigma^2(v_{ph})}{N_h} + \frac{\sigma^2(v_{pih,e})}{N_i N_h} \right] \right\} + N_j^2 w_{MC}^2 \left[\frac{\sigma^2(v_j)}{N_j} + \frac{\sigma^2(v_{pj,e})}{N_j} \right]}, \quad (9)$$

and

$$\Phi(\text{Composite}_{\text{SingleScoring}}) =$$

$$\frac{\sigma^2(t_{\text{Composite}})}{\sigma^2(t_{\text{Composite}}) + 2^2 N_i^2 w_{CR}^2 \left\{ \left[\frac{\sigma^2(v_i)}{N_i} + \sigma^2(v_h) + \frac{\sigma^2(v_{ih})}{N_i} \right] + \left[\frac{\sigma^2(v_{pi})}{N_i} + \sigma^2(v_{ph}) + \frac{\sigma^2(v_{pih,e})}{N_i} \right] \right\} + N_j^2 w_{MC}^2 \left[\frac{\sigma^2(v_j)}{N_j} + \frac{\sigma^2(v_{pj,e})}{N_j} \right]}, \quad (10)$$

where $\sigma^2(t_{\text{Composite}}) = N_j^2 w_{MC}^2 \sigma^2(t_{MC}) + N_i^2 2^2 w_{CR}^2 \sigma^2(t_{CR}) + N_j 2 N_i w_{MC} w_{CR} 2\sigma(t_{CR}, t_{MC})$.

Equations 9 and 10 are described as phi (Φ) coefficients, reflective of both the absolute and relative errors expected to influence absolute decisions about examinees' licensure test scores (Brennan, 2001; Feldt & Brennan, 1989; Haertel, 2006; Shavelson, Webb, & Rowley, 1989).

A Proposed Extension of Multivariate G-Theory Models That Produces Classification Consistency and Classification Accuracy Estimates

Equations 9 and 10 are indications of the overall reliability of the weighted composite scores that could be obtained from the actual testing situation and also from different testing conditions (e.g., double vs. single *CR* scoring, different numbers of *CR* items, different *CR* section weights, etc.). The extent to which the overall score reliability estimates indicate classification consistency at specific cutscores on the weighted composite of any particular test and data set is uncertain, but it is considered to be ''worthwhile checking . . . '' (Jarjoura et al., 2004, p. 34). Jarjoura et al. evaluated the stability of composite score reliability estimates obtained from their study,

data, and multivariate G-theory model by re-running their G-theory analyses using data from a subset of examinees with composite scores near a cutscore of interest and comparing the resulting estimates to the overall estimates. In this section, the authors propose a method for estimating classification reliability directly from a multivariate G-theory model that was estimated from an entire set of test data, and which can be used to consider the alternative testing conditions and D-studies of interest in G-theory. The proposed method is an integration of aspects of single administration classification consistency and classification accuracy estimation procedures (Livingston & Lewis, 1995) into a simulation from a G-theory model estimated from the administration data of a test. The steps of this simulation are as follows:

1. Generate N_p true scores on the *MC* and *CR* sections, $t_{MC,p}$ and $t_{CR,p}$, based on a G-theory model estimated from test administration data, with population means equal to the observed means divided by N_j (for the *MC*) or by $N_h N_i$ (for the *CR*), with population variances equal to $\sigma^2(t_{MC})$ and $\sigma^2(t_{CR})$, with a correlation of $\sigma(t_{CR}, t_{MC})/\sigma(t_{MC})\sigma(t_{CR})$. To reflect the classification consistency approach of Livingston and Lewis (1995), the true scores are created to have the same skewness and kurtosis as the observed score distributions. In this study, the data generation process was based on first generating independent standard normal variables, then obtaining the desired true score correlations using Cholesky decomposition, then achieving the desired skews and kurtosis using Fleishman's (1978) transformation, and finally achieving the desired means and variances by multiplications and additions of constants.
2. Generate the separate error variables for the *MC* and *CR* scores as standard normal variables and multiply these by constants, so that their variances are consistent with the corresponding variance components in the estimated G-theory model.
3. Create the observed *CR* scores from double scoring and single scoring similar to Equations 3 and 5. The equations have the form of,

$$CR_{p,\,\text{DoubleScoring}} = N_h N_i t_{CR,p} + 2\sqrt{prop(t_{CR,p})\left[1 - prop(t_{CR,p})\right]}\left[\left(N_h \sum_{i}^{Ni} v_i + N_i \sum_{h}^{Nh=2} v_h + \sum_{h}^{Nh=2}\sum_{i}^{Ni} v_{ih}\right) + \left(N_h \sum_{i}^{Ni} v_{pi} + N_i \sum_{h}^{Nh=2} v_{ph} + \sum_{h}^{Nh=2}\sum_{i}^{Ni} v_{pih,e}\right)\right], \tag{11}$$

$$CR_{p,\,\text{SingleScoring}} = 2\left\{N_i t_{CR,p} + 2\sqrt{prop(t_{CR,p})\left[1 - prop(t_{CR,p})\right]}\left[\left(\sum_{i}^{Ni} v_i + N_i v_h + \sum_{i}^{Ni} v_{ih}\right) + \left(\sum_{i}^{Ni} v_{pi} + N_i v_{ph} + \sum_{i}^{Ni} v_{pih,e}\right)\right]\right\}, \tag{12}$$

where $prop(t_{CR,p})$ is a proportional true score, $t_{CR,p} - \min(t_{CR})/\max(t_{CR}) - \min(t_{CR})$. The multiplication of the error effects by functions of $prop(t_{CR,p})$ results in a heteroskedastic error variance that has a binomial shape, such that the error variances and reliabilities for true scores near the overall mean are consistent with what is suggested by the G-theory model (Equations 9 and 10), and where smaller error variances and higher reliabilities are obtained for very low and very high true scores. This approach is similar to the classification consistency assumptions made in Livingston and Lewis (1995) and, with truncations to possible score ranges, is also consistent with the empirical findings of the general shape of test score error variances (Lord, 1958; Mollenkopf, 1949).

4. Create the observed *MC* scores similar to Equation 7 but with the multiplication of the error variables similar to Equations 11 and 12,

$$MC_p = N_j t_{MC,p} + 2\sqrt{prop(t_{MC,p})\left[1 - prop(t_{MC,p})\right]}\left(\sum_{j}^{Nj} v_j + \sum_{j}^{Nj} v_{pj,e}\right). \quad (13)$$

5. Create the observed composite scores as a weighted sum of the *MC* and *CR* scores created in Steps 3 and 4 based on Equations 1 and 2.
6. Classify examinees' weighted composite scores as passing or failing with respect to the cutscore of interest.

Reliability interests are typically about consistency over a large number of parallel forms. To estimate consistency, Steps 2 to 6 are replicated several times (i.e., 500 times) to create observed scores from 500 alternate and parallel forms for a single set of N_p = 10,000 examinee true scores. For estimating consistency in the pass/fail classifications at a cutscore, percent agreements between the classifications from any two observed forms are computed and averaged across all possible pairings of the 500 simulated observed forms. These consistency estimates were obtained for all the simulated double scoring and single scoring forms at the cutscores corresponding to one standard deviation below the population means from the double and single scoring distributions.

Because the proposed simulation approach included the generation of assumed true score distributions (Step 1), reliability could also be evaluated in terms of classification accuracy. For this evaluation, a weighted true score composite was generated similar to Equation 1. The weight for the *CR* true scores was selected to achieve weighted *CR* true score variances with the same proportion of the weighted true score composite variance as described in Table 1. The weighted true score composites were evaluated with respect to true cutscores at the same standardized mean differences from the true score mean as in the original cutscores on the double-scored form. Agreements in the classification of examinees' true scores with the observed scores of each double-scored form were calculated and averaged across all 500 double-scored forms. Agreements in the classification of examinees' true scores with the observed scores of each single-scored form were also calculated and averaged across all 500 single-scored forms.

Assessing Double Versus Single Rating Effects on Overall Score Reliability (Φ), Classification Consistency, and Classification Accuracy

Composite reliabilities were estimated for the three licensure tests with the goal of assessing the implications of double versus single scoring of the *CR* sections. First, the G-theory models described in Equations 3 to 8 were estimated with the empirical data for all three tests (Table 1). The univariate model estimates for each of the *MC* and *CR* scores were obtained using SAS Proc VarComp, and the covariance estimates of the *MC* and *CR* true scores were obtained using SAS Proc Mixed (Jarjoura et al., 2004). The variance component estimates obtained for each test are shown in Table 2. Second, overall composite score reliabilities (Φ) were estimated based on double and single scoring of the *CR* sections (Equations 9 and 10). Third, the six-step simulation was conducted with the estimated G-theory model(s) to obtain classification consistency and classification accuracy estimates at the cutscores for each exam.

The second and third steps of the simulation method were executed multiple times to assess double and single scoring reliability not only with respect to the structures of the actual tests described in Table 1 (i.e., numbers of *MC* and *CR* items, *CR* weights, and score distributions)

Table 2. Estimated Variance Components for the Three Licensure Tests.

Component of composite score variance	Role in composite score reliability	Test A estimates	Test B estimates	Test C estimates
$\sigma^2(t_{CR})$	Composite true score variance (*CR*)	.1158	.4188	.0636
$\sigma^2(t_{MC})$	Composite true score variance (*MC*)	.0128	.0252	.0081
$\sigma(t_{CR}, t_{MC})$	Composite true score covariance (*CR*, *MC*)	.0235	.0919	.0146
$\sigma^2(v_i)$	Absolute error from *CR* items	.0000	.0798	.0125
$\sigma^2(v_h)$	Absolute error from *CR* ratings	.0000	.0000	.0000
$\sigma^2(v_{ih})$	Absolute error from *CR* items interacting with *CR* ratings	.0000	.0000	.0000
$\sigma^2(v_{pi})$	Relative error from the examinees interacting with *CR* items	.2568	.6232	.2952
$\sigma^2(v_{ph})$	Relative error from the examinees interacting with *CR* ratings	.0000	.0006	.0009
$\sigma^2(v_{pih,e})$	Relative error from the examinees interacting with *CR* items and ratings	.0799	.0298	.1924
$\sigma^2(v_j)$	Absolute error from *MC* items	.0359	.0364	.0277
$\sigma^2(v_{pj,e})$	Relative error from the examinees interacting with *MC* items	.1480	.1725	.1499

Note. *CR* = constructed response; *MC* = multiple choice.

but also to consider reliability implications in terms of several different applications and generalizations of the measurement procedures for the tests (i.e., D-studies). The primary interest in the D-studies was comparing the reliability and classification estimates based on double versus single scoring. The double versus single scoring estimates of reliability, classification consistency, and classification accuracy were also considered with respect to tests with varied numbers of *CR* items, and tests where the *CR* sections contributed different percentages to the composite scores.

The procedures described above were conducted to consider double versus single scoring effects with respect to the G-theory estimates of each test, but with test forms composed of $N_i = 1$ to 12 *CR* items. For these shorter and longer *CR* sections, the weight values, w_{CR}, were also manipulated to produce maximum possible weighted *CR* scores that were the same proportion of the maximum possible weighted composite score as those used in the actual tests (Table 1).

The procedures were also conducted to consider double versus single scoring effects with respect to the G-theory estimates of each test, but with *CR* sections that contributed different percentages to the composite. The *CR* section weights, w_{CR}, were manipulated to obtain *CR* sections with maximum possible weighted scores that made up 0% to 100% of the maximum possible weighted composite score. It should be noted that the reliability implications of varied section weights are typically not described in terms of D-studies in G-theory, but more often in terms of a mathematical problem where section weights are selected to maximize composite reliability (Brennan, 2001; Feldt & Brennan, 1989; Haertel, 2006). The treatment of varied *CR* section weights as part of a D-study is based on the perspective that, like the licensure tests in this study, section weights may be used that do not optimize reliability.

Results

Figure 1 presents the overall score reliability (Φ), classification consistency, and classification accuracy estimates based on double versus single scoring as a function of the numbers of *CR* items on the *CR* section of Tests A, B, and C. In each subfigure, the actual numbers of *CR* items in each test appear with asterisks (*) on the horizontal axes, with solid and dashed lines representing the overall score reliability (Φ), classification consistency, and classification accuracy estimates based on double and single scoring, respectively. The results of the subfigures of Figure 1 and the additional figures to be described are presented to emphasize the distinct aspects of overall score reliability (Φ, that is, a proportion of score variance) and of classification consistency and classification accuracy (i.e., proportions of test takers). That is, the classification consistency and classification accuracy estimates correspond to the leftmost vertical axes with relatively high and narrow ranges from .60 to 1.00 that convey the high classification percentages that can occur due to chance. The overall score reliability (Φ) estimates correspond to the rightmost vertical axes, with scales ranging from .50 to 1.00, which are indicative of wider ranges that are more typical of reliability estimates than of classification estimates.

The subfigures of Figure 1 show some results that are similar across the three tests, such as that the classification accuracy estimates were usually relatively high, the overall score reliability (Φ) estimates were usually relatively low, all estimates were more influenced by the number of *CR* items than by double versus single scoring, and double scoring had larger impacts in increasing overall reliability than in increasing classification consistency and classification accuracy. The results about double and single scoring also appeared to reflect particular psychometric differences of Tests A, B, and C. For Tests A and B, which usually had relatively high overall score reliabilities (Φ) of about .85, Figure 1 shows very little differences between the overall score reliability (Φ), classification consistency, and

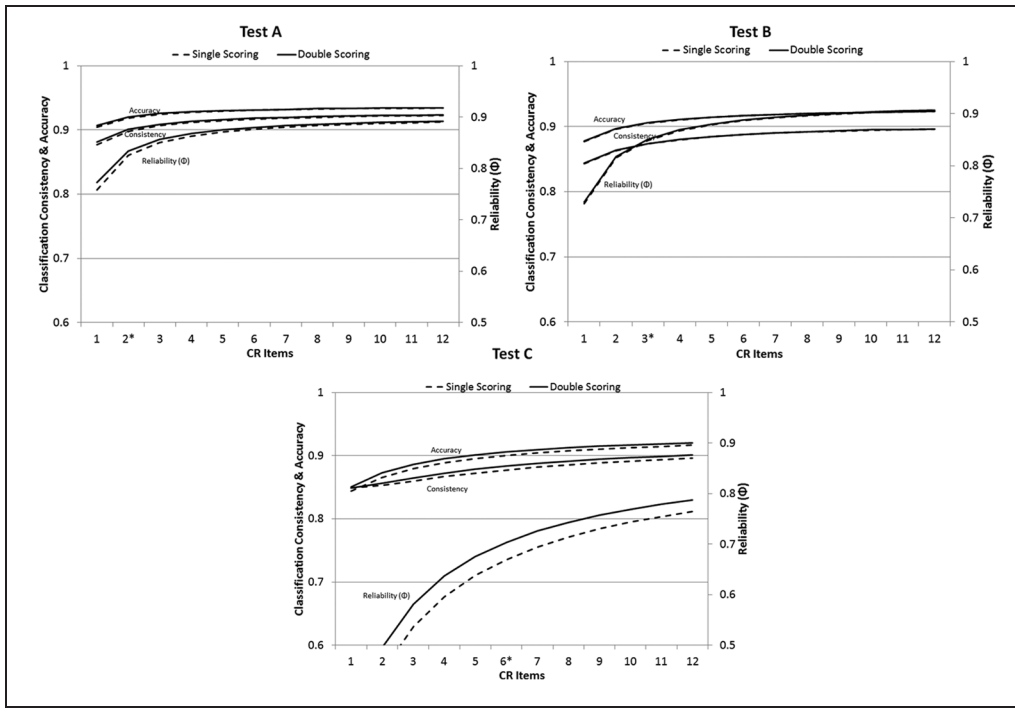

Figure 1. Implications of reliability, classification consistency, and classification accuracy for decision studies considering double and single scoring for *CR* sections with 1 to 12 *CR* items, Tests A, B, and C.
Note. The tests' actual number of *CR* items is indicated with an asterisk (*) and also in Table 1. *CR* = constructed response.

classification accuracy estimates based on double versus single scoring. Figure 1 indicates that greater improvements in overall score reliabilities (Φ), classification consistencies and accuracies can result from the addition of one or two more *CR* items to the *CR* sections than that which has resulted from double scoring.

The overall score reliability (Φ), classification consistency, and classification accuracy estimates in Figure 1 were more dramatically influenced for Test C. Test C had relatively low overall score reliability (Φ estimates often less than .80), and the influences of double versus single scoring and the numbers of *CR* items were more visible in the subfigure for this test. Test C required more single-scored *CR* items than double-scored *CR* items to achieve given levels of overall score reliability (Φ), though overall score reliability (Φ) remained below .80 even with 12 double-scored *CR* items. Another noteworthy result in Figure 1 is that for hypothetical forms of Test C with 1 *CR* item, the classification consistency estimates are nearly equal to the classification accuracy estimates (approximately 85%). This result is at odds with expectations that classification accuracy should always be noticeably greater than classification consistency. The result was due to making classifications on a composite score scale that was extremely stretched out because a very large *CR* section weight was needed to achieve the desired 50% contribution of *CR* scores to the composite with the score from a single *CR* item. The relatively low consistency of the *CR* raters' scores on Test C likely also contributes to the relatively low classification accuracy and consistency estimates for hypothetical forms based on only one *CR* item.

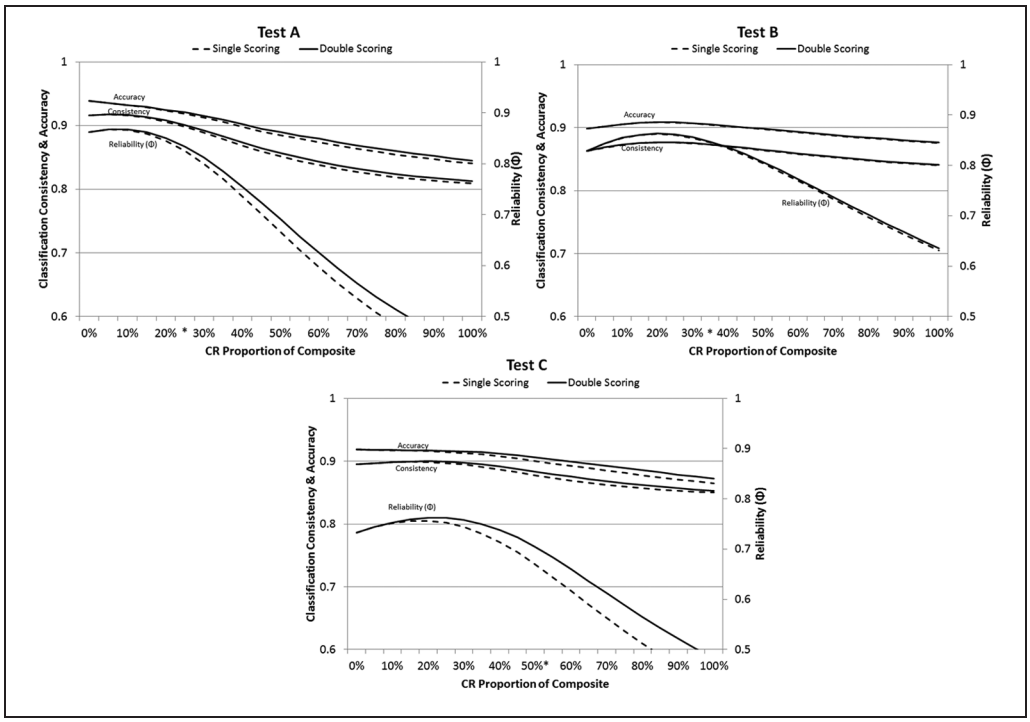

Figure 2. Implications of reliability, classification consistency, and classification accuracy for decision studies considering double and single scoring for *CR* sections that contribute 0% to 100% to the composite scores of Tests A, B, and C.

Note. The tests' actual *CR* contribution percentage is indicated with an asterisk (*) and also in Table 1. *CR* = constructed response.

The subfigures of Figure 2 present the implications of double versus single scoring on the overall score reliability (Φ), classification consistency, and classification accuracy as a function of the proportion of the *CR* section to the composite score of Tests A, B, and C. The actual proportion of the *CR* section scores in the composite of each test appears with an asterisk (*) on the horizontal axis of each figure and also in Table 1. The subfigures of Figure 2 show some results that are similar across the three tests, such as that the classification accuracy and consistency estimates were higher than the overall score reliability (Φ) estimates. Increases in *CR* proportions were usually inversely related to increases in overall score reliability (Φ), classification consistency, and classification accuracy, such that the estimates approached their highest for very low *CR* proportions where composite scores primarily reflected *MC* scores with higher reliability and larger numbers of items. For all three tests, double scoring produced its greatest improvements in overall score reliability (Φ) and in testing situations where the *CR* sections made up a larger proportion of the composite.

The overall score reliability (Φ), classification consistency, and classification accuracy estimates of Test B differed from those of the other tests, in that double scoring resulted in the least visible improvement and increased *CR* proportions result in the smallest reliability reductions (Figure 2, second subfigure). The relatively small impacts of double versus single scoring and of different *CR* proportions appear to be reflective of the high correlations between *MC* and *CR* scores and between *CR* ratings (Table 1). That is, although increasing the *CR* proportion could

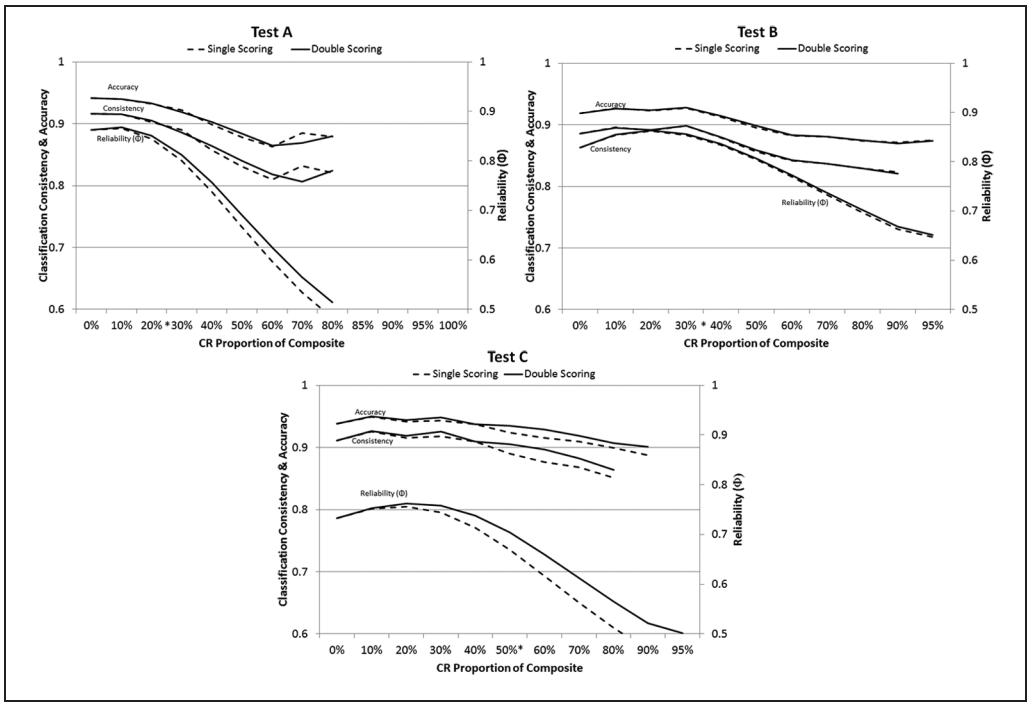

Figure 3. Implications of reliability, classification consistency, and classification accuracy for decision studies considering double and single scoring for *CR* sections that contribute 0% to 100% to the composite score for Tests A, B, and C based on an alternative approach.
Note. CR = constructed response.

result in undesirable decreases in reliability, classification consistency and classification accuracy in Test B, changes from double to single scoring may not be detrimental.

Figure 2 shows that compared with Test B, the overall score reliability (Φ), classification consistency, and classification accuracy estimates for Tests A and C exhibited more visible improvements from double scoring and also larger reliability reductions when the *CR* sections contributed proportions to the composite scores of 20% or more. Tests A and C had lower *MC* and *CR* correlations than Test B (Table 1). For Test A, which had relatively high correlations between the *CR* ratings (Table 1), Figure 2 shows that the overall score reliability (Φ), classification consistency, and classification accuracy estimates are primarily influenced by proportions of the *CR* sections, and that the potential benefits of double scoring were most visible for unrealistically large *CR* section proportions that led to reliability levels below .50. For Test C, which had lower correlations between the *CR* ratings than the other tests (Table 1), the improvements of double scoring were especially visible for *CR* section proportions greater than 20%. When the *CR* section proportion was 100% of Test C's composite, the overall score reliability (Φ) was lower than .50 regardless of scoring model, and the reliability decreases were especially salient with single scoring.

Comparisons With Another Plausible Approach

One way to evaluate the approach proposed in this study is to compare its results with those obtained by another plausible approach. An approach suggested by a reviewer is to obtain

composite score reliabilities based on the double and single scoring of *CR* items and on specific *CR* section weights, and then use these reliability estimates with the Livingston and Lewis (1995) procedure to estimate classification consistency and classification accuracy. The results of this second approach for the tests and *CR* section weights considered in this study are shown in the subfigures of Figure 3. The similarity of the results in Figure 3 to those in Figure 2 is support for the reasonableness of the estimates of both approaches. Some limitations can be observed in the second approach, however, such as that the model-fitting used in the Livingston and Lewis approach tended to produce somewhat erratic classification consistencies and accuracies. In addition, for some extreme *CR* weights and composite score distributions, the second approach produced undefined classification and accuracy estimates due to model convergence problems. Finally, it is not clear how classification consistency and classification accuracy estimates from different numbers of *CR* items could be obtained with the second approach (e.g., Figure 1).

Discussion

Traditional reliability estimation methods can seem to provide limited answers to issues in mixed-format licensure testing, such as questions about the implications of double versus single scoring of *CR* items. The G-theory methods traditionally used to assess overall score reliability and the implications of testing conditions, such as double versus single scoring, varied numbers of *CR* items, and/or *CR* section weights, do not directly address interests in classification consistency (Brennan, 2001; Shavelson et al., 1989). Methods for assessing classification consistency and classification accuracy (Brennan & Wan, 2004; Livingston & Lewis, 1995) have had limited applications to the estimation of the reliability gains from double scoring relative to single scoring (i.e., the decision studies that are possible in G-theory). Although not extensively considered in prior studies, it would seem that combining traditional reliability estimation methods could result in methodological extensions useful for quantifying changes in reliability, classification consistency, and classification accuracy due to changes in testing conditions.

In this study, an integration of G-theory methods and single administration classification consistency methods was proposed for extending these estimation approaches and for providing a methodology to quantify the implications of double versus single scoring for the licensure tests. The application of the proposed methodology to quantify the gains from double versus single scoring in the three licensure tests reflected and extended the findings of prior literature. Double scoring appears to result in greater overall and classification reliability gains for tests where rating reliability is lower (Test C) rather than higher (Tests A and B). The classification consistencies and accuracies estimated from the model reflect prior studies' results showing that consistencies and accuracies are usually higher than overall score reliabilities, and accuracies are usually higher than consistencies (Livingston & Lewis, 1995). The finding that double scoring can result in greater reliability gains for less reliable *CR* tests containing fewer items has been described (Gao & Brennan, 2001). This study extends prior descriptions by indicating that the gains from double scoring can be greater for mixed-format tests made up of higher proportions *CR* section scores, for tests with lower *MC* and *CR* correlations, and for *CR* tests where the intercorrelation of the *CR* ratings is low. These results reflected the particular psychometric characteristics of each test, implying that potential changes to the psychometric features of a testing program should be considered not only with respect to general research findings but also by directly applying the methods proposed in this study to testing program data.

The integration of the G-theory methods with the classification consistency and classification accuracy estimation proposed in this study means that the methods will have the same flexibility as G-theory methods. Extensions of the current study's investigations might be applied

to obtain classification reliability estimates for other multivariate or univariate designs reflected by tests not considered in this study. For the tests considered in this study, extensions might be useful for quantifying the changes in classification and reliability estimates due to other scoring features of interest. In particular, the influence of equating analyses for obtaining the cutscores used in classification processes could be assessed. Equating analyses would presumably be used to adjust scores and correct for difficulty effects in *MC* and/or *CR* items (i.e., items' absolute errors) and the leniency/stringency effects of ratings (i.e., ratings' absolute errors) if trend scoring is implemented (Kim, Walker, & McHale, 2010). The influence of equating analyses on estimates of overall score reliability, classification consistency, and classification accuracy has been considered in different ways in prior studies. For example, Jarjoura et al. (2004) argued that equating results implied that certain absolute errors could be ignored in their estimates of measurement error and reliability. Other studies have introduced levels of random errors expected based on standard errors of equating in their estimation of reliability, classification accuracy and consistency (Kim & Moses, 2013). The method proposed in this study is sufficiently flexible to deal with multiple approaches to incorporating errors from equating and from other sources.

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) received no financial support for the research, authorship, and/or publication of this article.

References

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

Brennan, R. L., & Wan, L. (2004). *A bootstrap procedure for estimating decision consistency for single-administration complex assessments* (Research Report No. 7). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa.

Clauser, B. E., Balog, K., Harik, P., Mee, J., & Kahraman, N. (2009). A multivariate generalizability analysis of history-taking and physical examination scores from the USMLE Step 2 Clinical Skills Examination. *Academic Medicine*, *84*, 586-589.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105-146). New York, NY: Macmillan.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, *43*, 521-532.

Gao, X., & Brennan, R. L. (2001). Variability of estimated variance components and related statistics in a performance assessment. *Applied Measurement in Education*, *14*, 191-203.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65-110). Westport, CT: American Council on Education.

Jarjoura, D., Early, L., & Androulakakis, V. (2004). A multivariate generalizability model for clinical skills assessments. *Educational and Psychological Measurement*, *64*, 22-39.

Joe, G. W., & Woodward, J. A. (1976). Some developments in multivariate generalizability. *Psychometrika*, *34*, 183-201.

Kim, S., & Moses, T. (2013). Determining when single scoring for constructed-response items is as effective as double scoring in mixed-format licensure tests. *International Journal of Testing*, *13*, 314-328.

Kim, S., Walker, M. E., & McHale, F. (2010). Comparisons among designs for equating mixed-format tests in large scale assessments. *Journal of Educational Measurement*, *47*, 36-53.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, *32*, 179-197.

Lord, F. M. (1958). *An empirical study of the normality and independence of errors of measurement in test scores* (Research Bulletin 58-14). Princeton, NJ: Educational Testing Service.

Mollenkopf, W. G. (1949). Variation of the standard error of measurement. *Psychometrika*, *14*, 189-229.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, *44*, 922-932.