Conditional Standard Errors of Measurement for Performance Ratings from Ordinary Least Squares Regression

Mark R. Raymond and Irina Grabovsky
National Board of Medical Examiners

National Council on Measurement in Education; April, 2011; New Orelans, LA.

Correspondence: Mark Raymond, NBME, 3750 Market Street, Philadelphia, PA 19104.
mraymond@nbme.org

Abstract

Although numerous scholars and publications advocate the use of conditional standard errors of measurement (*SEM*s) for evaluating measurement precision, they have yet to enjoy widespread use in psychological research or large-scale testing programs. This article describes methods for computing conditional *SEM*s and an overall index of reliability for performance ratings based on ordinary least-squares (OLS) regression. The proposed computational approach is straightforward and provides indices of measurement precision similar or identical to those based on generalizability theory. While the methods are illustrated within the context of performance assessments, the indices also apply to dichotomously scored responses from multiple-choice tests.

Conditional Standard Errors of Measurement for Performance Ratings from Ordinary Least Squares Regression

Traditional methods for computing reliability and standard errors of measurement (*SEM*) produce a single index that represents an average for all examinees. However, it is widely known that measurement error varies with, or is conditional on, examinee ability (Brennan, 2001). Lord's (1955) method for computing conditional *SEM*s based on the binomial error model was one of the early methods to gain popularity, probably due to its simplicity. It can be computed for any score by knowing only the number of items on a test. Over the past half century, several other methods for computing conditional *SEM*s have been proposed, including those based on analysis of variance, item response theory, and generalizability theory (Brennan & Feldt, 1989; Feldt, Steffen, & Gupta, 1985; Qualls-Payne, 1992). Despite their obvious utility, conditional *SEM*s have yet to be used on a wide scale by operational testing programs, perhaps because some types of conditional *SEM*s are computationally intensive.

Conditional *SEM*s seem particularly important within the context of performance assessment, where errors of measurement are often larger than desirable (Baker, O'Neill, & Linn, 1993) and where errors can vary for individual examinees depending on the particular raters or tasks they happen to encounter. The purpose of this paper is to introduce simple methods for computing conditional *SEM*s within the framework of ordinary least-squares (OLS) regression. The use of OLS regression to reduce systematic measurement error (i.e., rater leniency) in performance assessment has become increasingly common over the past two decades (Braun, 1989; Harik, Clauser, Grabovsky, Nungester, Swanson, Nandakumar, 2009; Houston, Raymond & Svec, 1991; Raymond, Harik, & Clauser, 2011), and it would be convenient to have an approach for computing conditional *SEM*s and reliability coefficients within that same framework. This paper describes such an approach.

The next section reviews methods for obtaining conditional *SEM*s based on generalizability theory. We then we describe a rater effects model based on OLS regression, present equations for computing conditional errors and reliability coefficients based on OLS and show their equivalence to those based on generalizability theory, and demonstrate the computation of conditional *SEM*s using both simulated and actual rating data. We also illustrate its application to dichotomously scored responses.

Conditional *SEM*s from Generalizability Theory

Generalizability theory provides an appealing framework for computing conditional *SEM*s because it of its flexibility in conceptualizing and quantifying sources of error for various types of measures given under complex administration designs. For completely crossed single-facet rating designs, the model can be expressed in deviation form (Brennan, 2001, p. 22):

$$X_{pr} = \mu + (\mu_p - \mu) + (\mu_r - \mu) + (X_{pr} - \mu_p - \mu_r + \mu) \tag{1}$$

where X_{pr} is the rating given to examinee p by rater r;

μ is the grand mean over all examinees and raters;

$(\mu_p - \mu)$ is the examinee effect for examinee p,

$(\mu_r - \mu)$ is the rater effect for rater r; and

$(X_{pr} - \mu_p - \mu_r + \mu)$ is residual effect (examinee-rater interaction).

Generalizability theory distinguishes between two types of measurement error: absolute and relative, designated as Δ and δ. The absolute error for examinee p on the encounter with rater r is defined as the difference between the examinee's observed score and true score:

$$\Delta_p = X_{pr} - \mu_p. \tag{2}$$

Absolute error is important for criterion-referenced score interpretations where an examinee's score is typically compared to a fixed standard of performance (e.g., passing score). In contrast, relative error provides information about the precision of examinee's score in relation to the group's performance and is appropriate for norm-referenced interpretations. It is defined as the difference between an examinee's observed deviation score and true deviation score:

$$\delta_p = (X_{pr} - \mu_p) - (\mu_r - \mu) \tag{3}$$

or, equivalently as:

$$\delta_p = X_{pr} - (\mu_r - \mu) - \mu_p \ . \tag{4}$$

The important feature of relative error is that it ignores, or removes, the systematic variability associated with the particular group of raters who serve as evaluators for a particular assessment occasion or test form. That is, absolute error includes the rater effect while relative error does not. Variances of these two errors of measurement are denoted as $\sigma^2(\Delta)$ and $\sigma^2(\delta)$, and their square roots are the conditional absolute and relative *SEM*s. The computation of both variances is described below; for a more thorough treatment see Brennan (1998; 2001).

Absolute Conditional SEM. From equation 2, the variance of the absolute errors for examinee *p* can be expanded using the notation of conditional variance as follows:

$$\sigma^2(\Delta_p) = var(\Delta_p|p) = var(X_{pr} - \mu_p|p) \ , \tag{5}$$

which can be calculated from observed ratings:

$$\sigma^2(\Delta_p) = \frac{\Sigma_r (X_{pr} - X_{p.})^2}{n_r(n_r - 1)} \ ,$$

where n_r is the number of ratings received by examinee *p* and $X_{p\bullet}$ is the mean of those ratings. The square root of this provides an estimator of conditional absolute *SEM*:

$$\sigma(\Delta_p) = \sqrt{\frac{\Sigma_r (X_{pr} - X_{p.})^2}{n_r(n_r - 1)}} \tag{6}$$

Relative Conditional SEM. The variance of the relative errors for examinee *p* as shown in equation 4 is conceptually defined as:

$$\sigma^2(\delta_p) = var(\delta_p|p) = var(X_{pr} - (\mu_r - \mu) - \mu_p)|p) \ . \qquad (7)$$

Because this definition involves the variance of a linear combination, it implicitly includes a covariance term, which is evident when terms are rearranged and expanded. The result is the following computational formula for the relative conditional *SEM* (Brennan, 1998, eq 36):

$$\sigma(\delta_p) = \sqrt{\frac{n_p+1}{n_p-1}\sigma^2(\Delta_p) + \frac{\sigma^2(r)}{n_r} - \frac{\left(\frac{n_p}{n_p-1}\right)\left(2\sum_1^{n_r}(X_{pr}-X_{p.})(X_{.r}-X_{..})\right)}{n_r(n_r-1)}} \ , \qquad (8)$$

where n_p is the number of examinees, $\sigma^2(r)$ is the variance component for the rater effect, $X_{\bullet r}$ is the mean of ratings for rater *r*, and the right most term represents the covariance over raters between each examinees' ratings and the mean of all ratings (i.e., the person-total covariance). The value of $\sigma^2(r)$ is obtained from a variance component or generalizability analysis, and $\sigma^2(\Delta_p)$ is obtained using equation 6. It can be seen that $\sigma^2(\delta_p)$ is equal to $\sigma^2(\Delta_p)$ plus a rather complicated adjustment. This adjustment is negative for most examinees; therefore, the relative conditional *SEM* is less than the absolute conditional *SEM* for most examinees. Brennan (1998; 2001) suggests two simplifications. First, since the two ratios involving the number of examinees approach 1 as $n_p \rightarrow \infty$, those terms can be eliminated with large sample sizes. Even so, the computation still requires some effort and the covariance term contributes to instability of estimates. The second simplification eliminates the covariance term. While this shortcut is practical, it provides only a rough approximation (Brennan, 1998; 2001). The OLS framework described next provides a simplified and conceptually appealing approach to computing $\sigma(\delta_p)$.

Modeling Rater Effects with OLS Regression

While OLS regression can be used to estimate the parameters for completely crossed rating designs, it is particularly useful for unbalanced nested (incomplete) designs where groups of examinees are evaluated by different but overlapping panels of raters. For incomplete designs, the mean of observed ratings for a given examinee will be biased to the extent that some raters are more or less lenient than other raters. Parameter estimates obtained from OLS statistically control for leniency error. Simulation studies and those based on real data indicate that OLS-adjusted ratings result in considerable increases in reliability (Braun, 1988; Houston et al., 1991; Harik et al., 2009). The following linear model can be used to estimate examinee ability with the rater effect removed:

$$X_{pr} = \alpha_p + \beta_r + e_{pr}\,, \tag{9}$$

where X_{pr} is the rating given to examinee or person p by rater r;

α_p is the examinee's true rating or score;

β_r is the rater effect for rater r, defined as the true mean of rater r across all examinees minus the grand mean of all raters and all examinees; and

e_{pr} is random error (examinee-rater interaction).

Based on properties of OLS estimators, estimates of α_p can be shown to be equal to the mean score for examinee p over all raters after correcting for rater leniency. Estimates of β_r are deviation scores that correspond to the stringency index for each rater, with positive values indicating that the rater is more lenient than average. Expressing all parameters in deviation form results in equation 1. That equivalence is noted below because it provides the basis for computing $\sigma(\delta_p)$ from the OLS model. Specifically,

$$\mu + (\mu_p - \mu) \quad = \quad \mu_p \; = \; \alpha_p$$

$$(\mu_r - \mu) \quad = \quad \beta_r$$

$$(X_{pr} - \mu_p - \mu_r + \mu) \quad = \quad e_{pr}$$

In practical applications it is common to use the results of the OLS regression to compute an adjusted rating for each rater-examinee encounter. The adjusted rating, X^*_{pr}, can be obtained by:

$$X^*_{pr} = X_{pr} - \beta_r \, , \tag{10}$$

which is the observed rating with the effect for each rater removed. The adjusted rating can also be computed from:

$$X^*_{pr} = \alpha_p + e_{pr} \, , \tag{11}$$

where the adjusted rating is expressed as the examinee's true rating plus random error. It is evident that the mean of X^*_{pr} across raters *r* equals α_p. We denote the mean of adjusted ratings as $X^*_{p.}$.

Conditional *SEM*s using OLS Regression

Absolute Conditional *SEM*

Equation 6 is used without modification to compute $\sigma(\Delta_p)$ within the OLS framework. It can be viewed as the square root of the within-person variance of observed ratings divided by the number of raters, or as the standard error of the mean rating for examinee *p*.

Relative Conditional *SEM*

The principal difference between $\sigma(\Delta_p)$ and $\sigma(\delta_p)$ is that the former contains systematic error variance due to the rater effect plus random error, while the latter includes only random error variance. We propose two methods for estimating $\sigma(\delta_p)$: one based on the adjusted ratings, X^*_{pr}, and the other based on the residuals, e_{pr}. The two methods are algebraically equivalent; the practitioner can use either approach to greatly simplify the computation of the relative conditional *SEM*.

Equation 7 defined the conditional relative variance as $var(\delta_p|p) = var(X_{pr} - (\mu_r - \mu) - \mu_p)|p)$. Recall from the OLS regression model that $(\mu_r - \mu) = \beta_r$, and that $X_{pr} - \beta_r = X^*_{pr}$. Substituting these values into Equation 7 gives the following equivalent representation for relative error variance:

$$\sigma^2(\delta_p) = var(X^*_{pr} - \mu_p|p) \tag{12}$$

This is identical in form to the absolute conditional error defined in equation 5, suggesting that equation 6 also be used for estimating *relative* conditional error, assuming that adjusted ratings, X^*_{pr}, are used. Given that $\mu_p = X^*_{p.}$ the following can be used to compute relative conditional errors:

$$\sigma(\delta_p) = \sigma(\Delta_p)^* = \sqrt{\frac{\sum_r (X^*_{pr} - X^*_{p.})^2}{n_r(n_r - 1)}}, \tag{13}$$

The key feature of this equation is that $\sigma(\Delta_p)^*$ is computed only from adjusted ratings obtained from OLS regression; in particular, the covariance term is not explicitly required.

There is yet an alternative approach to computing the conditional relative error. From the OLS methodology we know that $\alpha_p = X^*_{p.}$ Therefore, relative error as defined in equation 12 can be rewritten as:

$$\sigma^2(\delta_p) = var(X^*_{pr} - \alpha_p|p)\ . \tag{14}$$

This is the variance of the differences between observed scores and the examinee ability estimate from the OLS model. That difference is designated as e_{pr} in equation 11. In other words, the quantity $(X^*_{pr} - X^*_{p.})$ in equation 13 can be replaced by e_{pr}, which is the residual from an OLS analysis. Thus, the relative conditional *SEM* can be conveniently computed by:

$$\sigma(\delta_p) = \sigma(\Delta_p)* = \sqrt{\frac{\sum_r(e_{pr}^2)}{n_r(n_r - 1)}} \quad , \tag{15}$$

which corresponds to the mean-squared residual taken over raters for each examinee, divided by the number of raters minus 1.

Reliability Index Based on OLS Regression

The results of OLS regression can also be used to obtain the reliability of adjusted ratings. The traditional (overall) *SEM* is given by the square root of the mean of the individual values of $\sigma^2(\delta_p)$ or $\sigma^2(\Delta_p)$. Overall *SEM*s computed in this manner will correspond to the *SEM*s based on classical test theory (i.e., KR-20 and KR-21). The formula for the traditional *SEM* from classical test theory can be rearranged to compute a reliability-like coefficient based on $\sigma^2(\Delta_p)*$ and the variance of the adjusted mean ratings, SD_p^2*, over all examinees. That coefficient can be designated as:

$$R_{OLS} = 1 - \frac{\frac{1}{n}\sum \sigma^2(\Delta_p)^*}{SD_p^{2*}} \quad . \tag{16}$$

This quantity is one minus the ratio of error variance to observed variance and is essentially identical to the KR-21 formula derived by Lord (1955) for dichotomously scored items.

Application to Rating Data

Completely Crossed Design

Table 1 presents simulated ratings for a completely crossed single-facet design where n_p = 40 and n_r = 8 (the partitioning of the table and italic font are explained later). These ratings contain levels of systematic error (rater leniency) and random error comparable to those found in operational settings where performances are being judged. Relative conditional *SEM*s were first computed for observed ratings in Table 1 using equation 8. Then, α_p, β_r, and X^*_{pr} were estimated using OLS regression, and conditional *SEM*s were computed from equation 13 or 15. As shown in Figure 1, $\sigma(\delta_p)$ and $\sigma(\Delta_p)^*$ are very similar. The mean difference is 0.003, with differences ranging from -0.006 to 0.010, which are negligible from a practical perspective.

The top two panels of Table 2 provide overall indices of measurement error based on generalizability theory for both observed and adjusted ratings. As expected, the overall *SEM*s for observed ratings are larger for absolute error than for relative error, and the overall reliability is less for absolute error. For adjusted ratings, the relative indices of measurement precision for observed ratings are identical to the absolute indices for adjusted ratings. This is because the systematic error due to the rater effect has been removed from the adjusted ratings. The bottom panel of Table 2 shows the indices of measurement precision for ratings based on OLS regression. Given that the regression model removes the rater effect, there is no distinction between relative and absolute error for the OLS model; therefore the bottom portion of the table does not include separate columns for relative and absolute error. Consistent with the data depicted in Figure 1, the indices of measurement precision for OLS are comparable but not identical to the indices based on generalizability theory (e.g., $\Phi = .8756$; $R_{OLS} = .8787$).

The data structure in Table 1 was replicated 100 times. The indices of measurement precision based on OLS were slightly more favorable and less variable than those based on generalizability theory, but the differences were minor. The mean and *SD* of the generalizability coefficient (ρ^2) for adjusted ratings over the 100 replications were .9096 and 0.0112, while the mean and *SD* for R_{OLS} were .9119 and 0.0109. The value of R_{OLS} was slightly but consistently larger than ρ^2 for each of the 100 replications, with the magnitude of the difference ranging from 0.0015 to 0.0031, and averaging 0.0023.

Incomplete (Nested) Design

While the OLS framework simplifies the computation of the relative conditional *SEM,* relative error is not relevant to many large-scale performance assessments. This is because logistic constraints (e.g., examinee volume; security) require a rating design for which examinees are nested within raters, and relative error and absolute error are indistinguishable for nested designs (Brennan, 1998; 2001). Nonetheless, the methods presented here still have application to large-scale programs that require nested designs because some testing programs use OLS or a similar model to adjust scores by removing the rater effect (Braun, 1989; Harik et al., 2009). Technical reports for such programs typically report an *overall* index of dependability (i.e., Φ) and an *overall SEM* for both observed and adjusted scores based on the variance components for X_{pr} and X_{pr}^{*}, respectively. These indices can be used to illustrate the improvements in measurement precision realized by using adjusted scores. By extension, we propose that it is also beneficial for large-scale programs to compute absolute conditional *SEM*s for both observed and adjusted scores and to compare the two to evaluate reductions in measurement error at each score level throughout the distribution.

The ratings in Table 1 were converted to a nested design such that n_p = 40 and n_r = 3:8 (i.e., each examinee is judged by 3 of 8 raters). The italicized values within the partitions are those that were dropped, producing a matrix that was 37.5% complete. The design had sufficient overlap to permit any rating to be linked to any other rating – a requirement whether calibrating raters or items using OLS regression, item response theory, or just about any other model. OLS regression was then used to estimate α_p , β_r , and X_{pr}^{*}, and absolute conditional *SEM*s were computed for both observed and adjusted ratings.

Figure 2 plots $\sigma(\Delta_p)$ against $\sigma(\Delta_p)^*$ for the 40 examinees. Unlike the data plotted in Figure 1, we expect to see differences in the two sets of values. As suggested by Figure 2, conditional *SEM*s are generally smaller for adjusted ratings; overall they are about 30% less in magnitude. However, there are instances where observed ratings have smaller conditional *SEM*s. This is most obvious for the two examinees for whom there was perfect agreement among observed ratings, resulting in $\sigma(\Delta_p)$ = 0. There were 12 twelve examinees for whom the observed ratings disagreed only by 1 point for one rater, which resulted in $\sigma(\Delta_p)$ = 0.333. In eight of these instances the observed conditional *SEM* was less than the adjusted conditional *SEM*, while for four examinees the adjusted *SEM* was smaller. The graph illustrates that the magnitude of reduction in conditional *SEM*s for adjusted ratings is generally greatest for those with the largest error for observed ratings.

Table 3 gives variance components and various indices of measurement precision based on G-theory and on the OLS computations. Two features stand out. First, the indices of measurement precision suggest that there is considerable advantage to using adjusted scores (e.g., Φ = .4009 observed vs. Φ =.7358 adjusted). Second, the precision indices for adjusted ratings were nearly identical whether computed from generalizability theory or OLS (e.g., Φ = .7358; R_{OLS} = .7359). As some of the computations were done by hand, we suspect that this difference can be attributed to rounding error.

The nested design was also replicated 100 times. The adjusted ratings consistently exhibited overall *SEM*s that were consistently 50% smaller than SEMs for the observed ratings, The mean and *SD* of the dependability index (Φ) for *observed* ratings over the 100 replications were .5209 and .0890, while the corresponding values for the adjusted ratings were .7770 and 0.0436. This outcome is consistent with expectations. The mean Φ for *adjusted* ratings based on generalizability theory and R_{OLS} were identical to six decimal places for each of the 100 replications.

Additional Applications

We computed conditional *SEM*s from OLS regression for sets of live ratings for a large-scale performance assessment of physicians' clinical skills. This nested design involves cohorts of about 1,200 examinees who, over an 8-week period, are assigned to 12 raters from a pool of approximately 125 raters (see Harik et al., 2009 or Raymond et al., 2011 for details). Adjusted ratings exhibited substantially smaller conditional *SEM*s throughout the score distribution except for the highest ratings (i.e., where an examinee receives all 9s on a scale of 1 to 9), and values of R_{OLS} were equivalent to Φ indices based on adjusted ratings to the third and fourth decimal. Analyses were also completed using actual and simulated responses from dichotomously-scored selected response tests under circumstances for which the relative conditional *SEM* might be considered an appropriate index of measurement precision (e.g., crossed designs with norm-referenced decisions). The conditional *SEM*s based on $\sigma(\Delta_p)^*$ and $\sigma(e_{pr}^2)$ were consistently comparable to those based on $\sigma(\delta_p)$, while R_{OLS} exhibited the very small bias described earlier when compared to ρ^2.

Discussion

This paper illustrated the computation of common indices of measurement error using ratings modeled by OLS regression. Relative conditional *SEM*s for ratings from a crossed design are typically computed using equation 8, which is somewhat tedious. This paper suggested alternative methods for computing relative conditional *SEM*s based on by-products of subjecting the rating data to OLS regression. The first approach is based on computing the absolute error for each examinee for adjusted ratings, as shown in equation 13. The absolute error is much easier to compute than relative error. The second approach is based on the mean-squared residual over raters for each examinee as indicated in equation 15. Overall indices of measurement precision based on OLS were found to be comparable to those based on generalizability theory. Not only are the computations for these indices of measurement precision straightforward, but all can be computed without having to leave the OLS regression framework to run supplemental variance component or generalizability analyses.

The calculation of the absolute conditional *SEM* is straightforward within the framework of generalizability theory; thus the methods proposed here have most utility in instances for which the more complicated relative conditional *SEM* is the suitable index of measurement precision. For example, the methods illustrated here may prove useful for small-scale crossed designs, such as low-volume certification tests, educational and psychological experiments, and various types of competitions that require subjective judgments (e.g., figure skating, artistic creations, grant reviews). Second, while not the focus of the present paper, the methods presented here also apply to any multiple-choice assessment. The importance and desirability of conditional *SEM*s has been recognized for years. While the *Standards for Educational and Psychological Testing* encourages that they be reported, that same document also acknowledges that practical constraints often preclude their use (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999. p. 29). Perhaps the method presented here will encourage wider use of conditional *SEM*s for any situation where measurement precision can be expected to vary across the objects of measurement.

For large-scale programs that require nested designs, the methods illustrated here have the most direct applicability to rating data in a two-step fashion. First, OLS regression can be used to estimate examinee ability and rater leniency parameters. Next, absolute conditional

*SEM*s and an overall index of dependability can be obtained for adjusted ratings. Conditional *SEM*s also provide a more sensitive metric for evaluating behavioral and statistical interventions intended to improve the reliability of ratings (e.g., rater training; scale refinement; rater calibration). For example, in a study of the effectiveness of modeling ratings using OLS regression, Raymond et al (2011) showed that score adjustments were largest in the region of the score distribution where measurement errors were greatest.

This paper establishes initial support for the use of conditional *SEM*s and a reliability index based on OLS regression. Given the limited scope of this study, it would be worthwhile to conduct more extensive simulations to compare the properties of OLS-based indices of measurement error to other types of indices besides those based on generalizability theory (e.g., Feldt et al., 1985; Qualls-Payne, 1992). While our results suggest that the OLS indices are at least as stable as those based on generalizability theory, additional simulations under various conditions (N_r, N_p, degree of nesting, σ^2_{rater} , σ^2_{error}) are needed to better understand the stability of different estimators of the conditional *SEM*. It also would be useful to extend the approaches suggested here to two-facet designs where, for example, both raters and tasks vary in some systematic fashion. Indices based on residuals from the OLS regression model may prove to be particularly useful for gauging the amount of conditional error contributed by raters alone, by tasks alone, and by their combination.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington DC: American Educational Research Association.

Baker, E.L., O'Neil, H.F., & Linn, R.L. (1993). Policy and validity prospects for performance-based assessments. *American Psychologist*, *48*(12), 1210-1218.

Braun, H. (1988). Understanding scoring reliability: Experiments in calibrating essay readers. *Journal of Educational Statistics*, *13,* 1-18.

Brennan, R.L. (1998). Raw-score conditional standard errors of measurement in generalizability theory. *Applied Psychological Measurement*, 22, 307-33.

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

Feldt, L.S., Steffen, M., & Gupta, N.C. A comparison of five methods for estimating the standard error of measurement at specific score levels. *Applied Psychological Measurement*, *9,* 351-361.

Harik, P., Clauser, B.E., Grabovsky, I., Nungester, R.J., Swanson, D., Nandakumar, R. (2009). An examination of rater drift within a generalizability framework. *Journal of Educational Measurement*, *46*, 43-58.

Houston, W.M., Raymond, M.R., & Svec, J. (1991). Adjustments for rater effects in performance assessment. *Applied Psychological Measurement*, *15*, 409-421.

Lord, F.M. (1955). Estimating test reliability. *Educational and Psychological Measurement*, 15, 325-336.

Lord, F.M. (1957). Do tests of the same length have the same standard error of measurement? *Educational and Psychological Measurement*, 17, 510-521.

Qualls-Payne, A.L. (1992). A comparison of score level estimates of the standard error of measurement. *Journal of Educational Measurement*, 29, 213-225.

Raymond, M.R., & Viswesvaran, C. (1993). Least-squares models to correct for rater effects in performance assessment. *Journal of Educational Measurement*, *30*, 253-268.

Raymond, M.R., Harik, P., & Clauser, B.E. (2011). The impact of statistically adjusting for rater effects on conditional standard errors of performance ratings. *Applied Psychological Measurement*, 2011, (in press).

Wilson, H.G. (1988). Parameter estimation for peer grading under incomplete design. *Educational and Psychological Measurement*, 48, 69-81.

Table 1

Ratings and Conditional SEMs for n_p = 40 and n_r = 8 completely crossed design[a].

	Rater							
Person	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
1	4	4	3	*4*	*5*	*5*	*4*	*3*
2	5	5	4	*1*	*5*	*3*	*5*	*2*
3	4	5	5	*3*	*5*	*5*	*6*	*4*
4	6	6	6	*5*	*6*	*5*	*6*	*5*
5	5	5	4	*3*	*6*	*4*	*4*	*4*
6	*3*	5	5	3	*5*	*5*	*4*	*4*
7	*5*	6	4	2	*6*	*5*	*6*	*4*
8	*6*	4	5	4	*5*	*5*	*6*	*4*
9	*3*	4	3	5	*3*	*3*	*5*	*3*
10	*3*	3	3	1	*4*	*2*	*4*	*2*
11	*3*	*5*	5	4	5	*6*	*6*	*3*
12	*5*	*5*	7	4	6	*7*	*5*	*6*
13	*5*	*6*	5	4	4	*6*	*5*	*5*
14	*4*	*4*	4	3	6	*3*	*5*	*3*
15	*4*	*6*	6	5	7	*6*	*7*	*5*
16	*5*	*6*	*5*	4	5	5	*6*	*5*
17	*4*	*6*	*6*	4	6	4	*6*	*5*
18	*7*	*7*	*7*	5	7	6	*7*	*7*
19	*5*	*6*	*5*	3	6	5	*6*	*3*
20	*4*	*5*	*5*	4	6	7	*7*	*5*
21	*4*	*4*	*5*	*4*	4	3	6	*3*
22	*5*	*5*	*6*	*5*	5	5	6	*2*
23	*6*	*5*	*4*	*4*	4	3	5	*3*
24	*5*	*5*	*5*	*2*	5	4	4	*4*
25	*4*	*4*	*3*	*2*	6	4	5	*3*
26	*5*	*6*	*6*	*4*	*7*	6	7	6
27	*5*	*4*	*6*	*3*	*6*	7	6	3
28	*4*	*5*	*5*	*4*	*6*	4	7	4
29	*4*	*5*	*5*	*4*	*4*	4	4	4
30	*7*	*7*	*7*	*4*	*7*	7	7	6
31	4	*6*	*5*	*3*	*6*	6	7	6
32	6	*6*	*6*	*5*	*5*	6	7	4
33	4	*5*	*4*	*3*	*5*	5	7	4
34	4	*6*	*6*	*3*	*5*	5	7	4
35	5	*6*	*6*	*4*	*4*	5	5	4
36	4	5	*5*	*3*	*6*	*6*	*6*	3
37	6	7	*7*	*4*	*7*	*7*	*6*	6
38	5	6	*6*	*4*	*6*	*7*	*6*	4
39	4	6	*5*	*5*	*6*	*4*	*7*	4
40	4	6	*5*	*4*	*5*	*4*	*7*	4

[a] The italicized values within the partitions denote ratings that were later dropped to produce a nested rating design for additional simulation.

Table 2

Comparison of Indices of Measurement Precision based on Generalizability Theory with Indices Based on OLS for the Completely Crossed Design Depicted in Table 1.

Type of Rating	**Index of Precision**[a]	**Type of Error**	
		Relative	**Absolute**
Observed	G-Theory overall *SEM*	.2803	.3769
	G-Theory reliability (ρ^2, Φ)	.8756	.7957
Adjusted	G-Theory overall *SEM*	.2803	.2803
	G-Theory reliability (ρ^2, Φ)	.8756	.8756
Adjusted	OLS overall *SEM*	.2768	
	OLS reliability (R_{OLS})	.8787	

[a] Notes: Overall *SEM*s based on generalizability theory are given by $\sqrt{\sum \sigma^2 \, (\delta_p)/n_p}$ and by $\sqrt{\sum \sigma^2 \, (\Delta_p)/n_p}$, while ρ^2 and Φ are based on standard formulas (see Brennan, 2001). Equation 13 or 15 produce the overall *SEM* based on OLS, while R_{OLS} is given by equation 16.

Table 3.

Comparison of Indices of Measurement Precision based on Generalizability Theory with Indices Based on OLS for the Nested Design Depicted in Table 1.

Type of Rating	**Index of Precision**	**Absolute Error**
Observed	G-Theory overall *SEM*	.6583
	G-Theory reliability (Φ)	.4009
Adjusted	G-Theory overall *SEM*	.4534
	G-Theory reliability (Φ)	.7358
Adjusted	OLS overall *SEM*	.4533
	OLS reliability (R_{OLS})	.7359

[a] Notes: Overall *SEM*s based on generalizability theory are given by $\sqrt{\sum \sigma^2\,(\Delta_p)/n_p}$, while Φ is based on the standard formula (see Brennan, 2001). Equation 13 or 15 produce the overall *SEM* based on OLS, while R_{OLS} is given by equation 16.

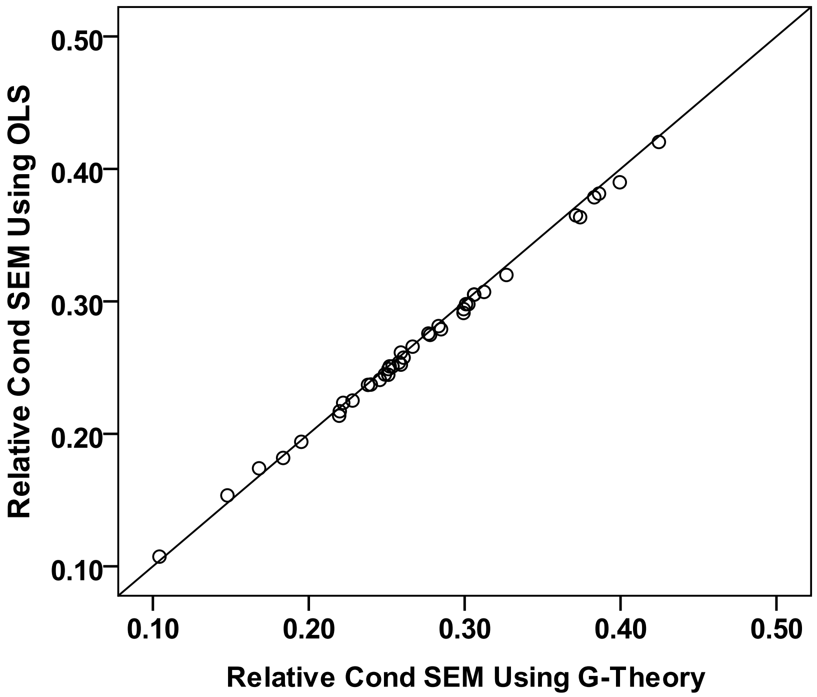

Figure 1. Plot of relative conditional *SEM*s based on generalizability theory and ordinary least squares regression for all ratings in Table 1.

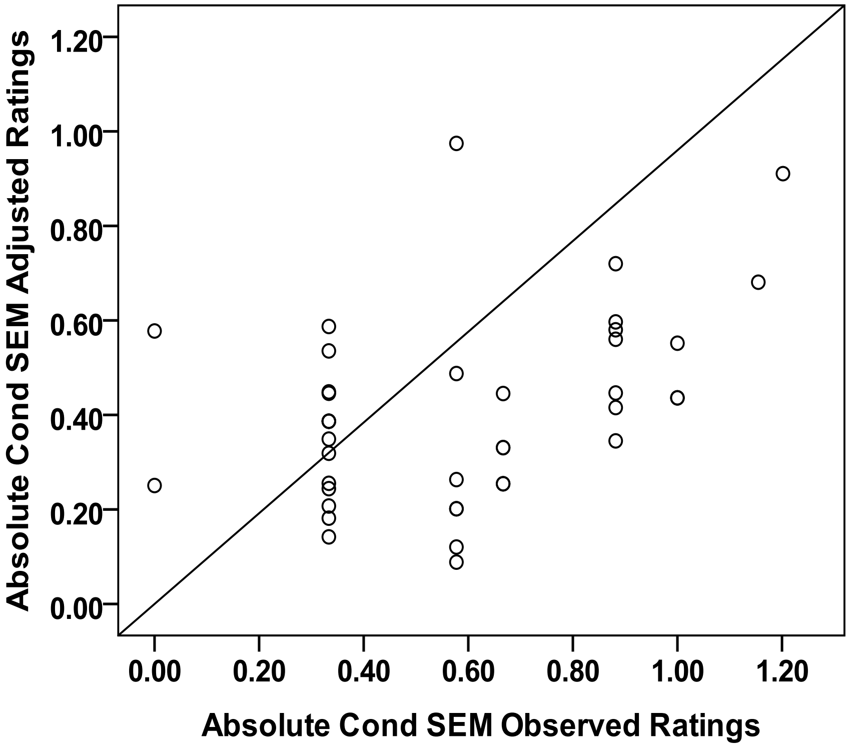

Figure 2. Absolute conditional *SEM*s based on observed ratings and on adjusted ratings for the nested design depicted in Table 1.