The Impact of Statistically Adjusting for Rater Effects on Conditional Standard Errors for Performance Ratings

Mark R. Raymond, Polina Harik and Brain E. Clauser

National Board of Medical Examiners
The Impact of Statistically Adjusting for Rater Effects on Conditional Standard Errors for Performance Ratings

Objectives

Many of the skills required for successful performance in a profession are not well measured with written examinations. Although the use of performance assessments to measure such skills has become increasingly common, achieving adequate measurement precision remains a challenge (Swanson, Norman & Linn, 1995). To reduce measurement error, some have proposed the use of statistical models to detect and correct for the error associated with rater and task effects. Studies have shown that adjustments based on ordinary least squares (OLS) regression can substantially improve the reliability of ratings in a variety of contexts (e.g., Braun, 1988; 2008; Harik, et al., 2009). These studies have used average measures of total group reliability, such as $\rho^2$ and $\Phi$, as evaluative criteria. However, the effects of OLS adjustments at various points of the score distribution have not been studied. If the use of adjusted scores does not improve measurement precision near the cut score, then little is gained.

The primary purpose of the present study was to obtain conditional SEMs for both observed and adjusted performance ratings, and evaluate changes in score precision at specific points in the score distribution. A cross-validation (i.e., resampling) was employed to evaluate the stability of rater calibrations and changes in conditional SEMs across different samples of examinees. Generalizability theory was used as the basis for computing overall indices of measurement precision and conditional SEMs.

Source of Information

Data. The study included ratings for 7,447 first-time examinees completing one part of a multistage licensing examination between July 2007 and June 2008. The exam requires that each examinee interact with 12 standardized patients (raters) for a period of 15 minutes each. Following each encounter, the rater completes three rating scales assessing communication skills (interviewing, information-sharing, and professional rapport). The three ratings are summed to obtain a single communication score, which is subsequently averaged across raters to obtain an overall score.

Examinees are crossed with cases and raters are completely nested within cases (each rater portrays a different case). That is, cases and raters are confounded, and are referred to as raters in this paper. Over time, different groups of examinees encounter different but overlapping sets of raters. Prior research has shown that indices of rater leniency drifts over time (Harik, et al., 2009; Hoskens & Wilson, 2001). Therefore, raters are calibrated every several weeks using the OLS scoring model described below.

**OLS Scoring Model.** Adjusted ratings were obtained using the following linear model:

$$X_{pr} = \alpha_p + \beta_r + e_{pr}$$  \hspace{1cm} (1)

where $X_{pr}$ is the rating given to examinee or person $p$ by rater $r$; $\alpha_p$ is the examinee’s true rating or score; $\beta_r$ is the rater effect for rater $r$, defined as the 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).

Note that $\alpha_p$ corresponds to the adjusted mean across all ratings. Because the generalizability analyses requires adjusted ratings for each examinee-rater encounter, adjusted ratings, $X'_{pr}$, were obtained by:

$$X'_{pr} = X_{pr} - \beta_r.$$  \hspace{1cm} (2)

**Cross-Validation Design.** As previously noted, estimates of $\beta_r$ require periodic reestimation due to rater drift. Therefore, examinees were split into cohorts spanning a 3-month interval. For the cohort comprising the first 3-month interval, equation (1) was used to obtain estimates of $\beta_r$ for all raters based
on a 50% random sample of examinees. Adjusted ratings for these examinees were obtained using equation (2). This sample was designated as the estimation sample (ES). The same estimates of $\beta_i$ were then used to obtain adjusted ratings for the remaining 50% of examinees. This group was designated as the cross-validation sample (CS).

This process was repeated for cohorts in the next 3-month interval. Although many raters are common across cohorts, the overlap is not complete. To place all examinee ratings onto the same scale, data were structured such that each cohort overlapped with the cohort immediately preceding it. This common examinee design allowed us to carry out mean equating (i.e., calibration or scale alignment) across all cohorts. The full paper elaborates on the design.

This estimation and cross-validation cycle was completed for 100 replications. It produced for each examinee 100 adjusted ratings (times 12 raters). For about one-half the replications the examinee was part of the estimation sample, while for the remaining replications he or she was part of the cross-validation sample. For this particular design, the number of ratings in the estimation sample available to calibrate individual raters ranged from 7 to 482, and averaged 173.

**Analyses.** Variance components and overall indices of measurement precision (SEM, $\Phi$) were estimated for both observed and adjusted ratings using formulas provided in Brennan (2001). Conditional absolute SEMs were obtained using equation 5.31 from Brennan (2001). Descriptive statistics for observed ratings and the OLS parameters corresponding to equation (1) were also obtained. The presented findings are based on 100 estimation and cross-validation samples.

**Results**

Observed total scores ranged from 9 to 24, with a mean of 19.9 and an SD of 1.6. Adjustments resulting from application of the OLS model were notable. The mean absolute change was one-third of a point for both the estimation sample and cross-validation sample, while the largest adjustments approached 1.3 points. Adjusted ratings correlated 0.97 with observed ratings for both samples. The full paper includes tables summarizing calibration results for both examinees and raters.

Table 1 summarizes the generalizability results for observed scores and for the two samples of adjusted scores over 100 replications. One feature of these data is that $\sigma^2_i = 0$ for the estimation sample. This outcome is not surprising because the linear model used to eliminate the rater effect is essentially the same model used to estimate variance components. Similarly, the decrease in $\sigma^2_\lambda$ and SEM, and corresponding increase in $\Phi$ for the estimation sample is predictable. The key finding in Table 1 is that most of the improvement in overall score precision obtained in the estimation sample was also obtained for the cross-validation sample. Given that the experimental design effectively halved the number of examinees available to calibrate each rater, the results overestimate the error due to cross-validation.

Figure 1 presents the conditional SEMs. Because of the small number of examinees with low scores, the scores at or below 16 (n=181) were placed into the same category. Conditional SEMs for observed and adjusted ratings are largest at the low end of the scale and get smaller with higher scores, except that SEMs for adjusted scores reach a minimum at about 21 and then increase slightly. The adjusted scores have smaller conditional SEMs than observed scores at every score level, although the difference is not uniform throughout the scale. Again, results for the cross validation are comparable to those for the estimation sample. Although adjustment helps, it is still somewhat disconcerting that errors are larger in the region where pass-fail decision are likely to be made.

To better gauge the changes in measurement precision associated with adjustment, we computed the difference between conditional SEMs for the observed ratings and adjusted ratings. Given the similarity in SEMs for the estimation sample and cross-validation sample, results are presented only for the latter
groups in order to simplify the display. Figure 2 presents the mean SEMs and 90% confidence intervals for each score level. It is evident from the nonhorizontal line in Figure 2 that the change in error associated with using adjusted scores varies as a function of score level. The decrease in conditional SEMs is greatest in the cut-score region, which is a positive finding. The pattern of change in conditional SEMs is intriguing. The full paper discusses known reasons for this pattern (e.g., adjusted scores partially overcome the ceiling effects that are imposed on observed scores), and speculates about other possible explanations (e.g., low performing examinees are more difficult to assess).

**Educational Importance**

Previous studies indicate that the reliability of scores on performance assessments can be improved through the use of statistical modeling. This study extends previous work by demonstrating that the improvement in measurement precision associated with using OLS-adjusted scores varies across the score distribution, and that the improvement generalizes to other samples of examinees from the same population. The results not only support the use of statistical adjustment, but also identify portions of the score distribution for targeting efforts to improve reliability.

**References**


**Table 1**

<table>
<thead>
<tr>
<th>Type of Rating</th>
<th>Variance Components</th>
<th>Measurement Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\sigma^2_p$</td>
<td>$\sigma^2_r$</td>
</tr>
<tr>
<td>Observed</td>
<td>2.123</td>
<td>2.110</td>
</tr>
<tr>
<td>Adj Estimation Sample</td>
<td>2.119</td>
<td>*-0.014</td>
</tr>
<tr>
<td>Adj Cross-Val Sample</td>
<td>2.129</td>
<td>0.053</td>
</tr>
</tbody>
</table>

* Note: set to 0 for computations in accordance with common practice.
Figure 1: Conditional SEMs for communication ratings (observed ratings, estimation sample, cross-validation sample). Total communication score is obtained by summing three ratings (1 – 9 scale) assigned by each rater and then averaging over all raters.

Figure 2: Change in conditional SEMs (observed – adjusted) for cross validation sample. Each point represents the mean change in conditional SEM, while the bars indicate the 90% confidence interval.