Psychological Test and Assessment Modeling, Volume 61, 2019 (2), 127-148

# Human rater monitoring with automated scoring engines

*Hyo Jeong Shin*[1], *Edward Wolfe*[2] & *Mark Wilson*[3]

**Abstract**

This study presents a case study that applies mixed-effects ordered probit models for the purpose of utilizing scores from automated scoring engines (AE) to monitor and provide diagnostic feedback to human raters under training. Using the experimental rater training study data, we illustrate a statistical approach that can be used for analyzing three types of model-based rater effects – severity, accuracy, and centrality of each rater. Each of the rater effects is related with model parameters and compared for cases (a) when the AE is considered as the gold standard and (b) when the human expert (HE) is considered as the gold standard. Results showed that AE and HE scoring approaches agreed maximally (100%) in detecting severity. The agreement rate was somewhat lower for centrality (93.1%) and considerably lower for accuracy (66.4%). As a targeted case study, this examination concludes with practical implications and cautions for rater monitoring based on the AE.

Keywords: automated scoring, human scoring, rater effects, rater monitoring

[1] *Correspondence concerning this article should be addressed to:* Hyo Jeong Shin, PhD, University of California at Berkeley, Educational Testing Service, 660 Rosedale Road, 13-E, Princeton, NJ 08541, USA; email: hshin@ets.org

[2] Pearson, USA

[3] University of California at Berkeley, USA

## Introduction

Due to the relatively high costs associated with human scoring, interest in applying automated scoring to supplement or supplant human scoring has increased in recent years, and considerable effort has been directed toward researching and improving the automated scoring process (Attali & Burstein, 2006; Clauser, Kane, & Swanson, 2002; Landauer, Laham, & Foltz, 2003; Williamson, Xi, & Breyer, 2012). Little attention has been directed toward whether what we learn from automated scoring can be used to improve the human scoring process. In most current applications, automated scoring engines are calibrated using data from human scorers, so, clearly, obtaining ratings from humans that are the highest quality possible is paramount to the success of these automated scoring efforts. Additionally, the general public remains skeptical of the validity of automated scoring, so full implementation of automated scoring is increasing slowly. For the foreseeable future, human scoring will remain important; therefore, it is essential to continually improve the quality of human scores while making the scoring process as efficient as possible.

One common concern for those who manage and monitor human scoring projects is how to monitor rating quality in real time and over time (Myford & Wolfe, 2009). There are many potential procedures for doing this, but one that is used extensively is the administration of validity papers. These papers are student responses that have been assigned consensus scores by expert raters before an operational scoring project. After the raters complete their training and begin operational scoring, validity papers are occasionally blindly seeded into the raters' scoring queues, and the scores that raters assign to the validity papers are recorded. When a rater has assigned scores to a sufficient number of validity papers, the scoring leaders review the scores, and compare them to the consensus scores assigned by experts. If large differences are observed between the scores assigned by a particular rater and the scores assigned by experts, the scoring leaders may choose to provide feedback to that rater and/or take corrective actions.

This paper describes a study designed to evaluate the possibility of replacing scores from human experts (HE) on validity papers with scores assigned by an automated scoring engine (AE) using the experimental data consisting of ratings on the texts written by middle-school students. Calculating the agreement (either by exact agreement rates or Cohen's Kappa) or correlations between AE and HE has been a popular choice to demonstrate the correspondence of AE to the HE[4]. This correspondence between observed scores is a simple and straightforward way to communicate and might be sufficient to defend the use of AE to replace at least one of the scores from human raters. However, this method may not provide diagnostic feedback to the human raters during training for the purpose of monitoring. Therefore, rather than directly comparing the scores, we focus on the rater monitoring situation and explore multiple aspects of model-based rater effects. The potential cost savings of implementing a rater monitoring system that relies on scores assigned by an automated scoring engine is considerable. A typical

[4] See Yang, Buckendahl, Juszkiewicz, and Bhola (2002) for a recent comprehensive review of strategies for validating automated scoring.

model for monitoring raters in an operational project includes convening a group of scoring leaders who select and assign scores to papers to construct validation sets. Although this process may be undertaken as part of the process of selecting training materials, this activity results in costs beyond those required to produce the scores reported to students because convening that meeting may require the expert raters to travel and will also likely require paying the experts for their time. In addition, administering validity papers to raters during operational scoring projects adds additional cost beyond that required to assign a score to each student response. The validity papers typically need to be entered into scoring queues in a manner that prevents the assigned scores from being recorded as operational scores. In addition, the scoring distribution system must determine when and how frequently the validity responses are administered to each rater, and must then redirect the assigned scores to the rater monitoring system. Administering the validity sets to raters during operational scoring, generating reports that summarize the raters' performance on the validity sets, and reviewing and then providing feedback to raters based on the information in the reports, introduce additional costs within a rating project beyond the costs associated with producing reported scores. As a result, a very low rate of validity paper administration is employed, say, one validity paper for every 20 to 100 operational papers scored. This results in a very small amount of data and a very slow accumulation of information regarding the performance of individual raters.

If validity paper administration could be replaced or reduced by the use of automated scores to monitor and evaluate raters, many of these costs could potentially be eliminated. Although automated scores would still require pulling papers to train the engine, assigning expert consensus scores would not necessarily be required for the validity papers. In addition, the entire process of having raters score responses that are not reported back to students for the sake of evaluating rater performance could be eliminated because automated scores can be assigned to every operationally scored response. This would have the added benefit that every score assigned by a rater could be fed into the rater monitoring system. As mentioned previously, a very small number of validity papers is administered to each rater relative to the number of operational papers, due to the added cost of administering the validity papers. Thus, because the score of every response that a rater scores could be used to evaluate and monitor raters, that process could be considerably more precise and faster than it is currently possible.

The purpose of the research reported here is to determine the effectiveness and efficiency of utilizing scores from automated scoring engines to monitor and provide feedback to human raters compared to the use of validity sets that are selected and assigned consensus scores by human scoring leaders. This study is a special case study that illustrates a statistical approach that can be used for analyzing three types of model-based rater effects – severity, accuracy, and centrality of each rater to answer the question, "Are depictions of the quality of scores assigned by human raters comparable when monitoring is based on scores from an automated engine (AE) versus human experts (HE)?" To that end, we analyze a real experimental data consisting of ratings on texts written by middle-school students, taken from a rater training study, and apply the mixed-effects ordered probit models (Rabe-Hesketh & Skrondal, 2012). Specifically, the analysis starts from a simple model that estimates a single rater effect (e.g., rater severity) and develops into a

more complicated model that estimates multiple rater effects including the rater severity and inaccuracy. We then interpret how the model parameters and their transformations are related with the rater effects of interest (model-based rater effects) to illustrate the human rater monitoring using the AE for three aspects of the rater effects compared to the HE.

## Statistical modeling of rater effects

*Rater effects* can be defined simply as "patterns of ratings that contain measurement error" (Wolfe & McVay, 2012). Raters may introduce errors into examinee scores for various reasons – unfamiliarity with or inadequate training in the use of the rating scale, fatigue or lapses in attention, deficiencies in some areas of content knowledge that are relevant to making scoring decisions, or personal beliefs that conflict with the values adopted in the scoring rubric (Myford & Wolfe, 2003, 2004; Saal, Downey, & Lahey, 1980). Wolfe and McVay (2012) identify several continua of rater effects that are commonly studied in rating applications. They define "severity" as when a rater consistently assigns a lower score than the target scores. In contrast, "leniency" is defined as when a rater consistently assigns a higher score than the target scores. Commonly, severity/leniency has been evidenced by a decrease/increase in the average score associated with a rater (Wolfe, 2014). If severity/leniency exists in the scores, then some examinees will be incorrectly classified in decision making contexts such as during college admissions or placement or determining graduation qualification. Wolfe and McVay (2012) define "accuracy" as when a rater exhibits little random variability in their scores, compared to an assumed-to-be perfect indicator (e.g., either HE or AE) (Wolfe, 2014). "Inaccuracy," on the other hand, occurs when the scores assigned by a rater exhibit a large amount of variation, relative to the assumed-to-be-perfect comparison standard. Finally, Wolfe and McVay (2012) define "centrality" as when a rater consistently assigns scores in the middle categories of the rating scale. The distribution of assigned scores can be compressed (centrality) or pushed into tails (extremity, the opposite of centrality) when compared to gold standard (e.g., HE or AE). For example, with four categories, a rater with centrality would likely use 2 or 3 more often and use 1 or 4 less often than the target scores. When centrality/extremity exists in the scores, examinees in the tails of the distribution may be misclassified and/or decision makers may believe that examinees are less or more homogeneous than is actually the case. Commonly, centrality/extremity has been evidenced by a decrease/increase in the standard deviation of the scores associated with a rater (Wolfe, 2014).

A body of literature describes how rater effects may be detected in rating data. In the family of Rasch modeling, the multi-faceted Rasch model (MFRM), which considers person, item, and rater facets, is a popular approach in item response theory (IRT) modeling. In fact, the MFRM has the same mathematical form as the linear logistic test model (LLTM) (Fischer, 1973). The MFRM and the LLTM incorporate a rater severity parameter in an additive extension of the Rasch model. For example, based on the partial credit model (PCM; Masters, 1982) for polytomous scores, the MFRM can be written as

$$\text{logit}\left[P(X_{nir} = k \mid \theta_n, X_{nir} \in \{k, k-1\})\right] = \theta_n - \beta_i - \tau_{ik} - \rho_r \quad (1)$$

where $X_{nir}$ is the polytomous score among the $k$ categories, given to examinee $n$ on item $i$ by rater $r$, $\theta_n$ is the latent proficiency of examinee $n$, $\beta_i$ is the difficulty of item $i$, $\tau_{ik}$ is the $k^{th}$ step difficulty for item $i$, and $\rho_r$ is the severity of rater $r$.

In rater monitoring contexts, it is common to focus on the scores assigned to a single writing task even though it may be more efficient from a measurement perspective to base rater evaluations on scores assigned to several writing tasks. The reason for this is logistical – raters commonly are trained and assign scores to student responses using a single scoring rubric that was written for a specific prompt. Similarly, it is also common to have students respond to only a single prompt due to the amount of testing time required for open-ended assessment items. Therefore, we focus our attention to scoring contexts in which each rater assigns scores to student responses to a single prompt or item. In these contexts, the item facet in Equation (1) can be eliminated, which is analogous to the PCM as

$$\text{logit}\left[P(X_{nir} = k \mid \theta_n, X_{nir} \in \{k, k-1\})\right] = \theta_n - \rho_r - \tau_{rk} \quad (2)$$

where $\tau_{rk}$ is the $k^{th}$ step severity for rater $r$, and $\rho_r$ is the overall severity of rater $r$. Similar to the PCM, we can allow different step severity for each rater. Under the generalized linear model framework, the PCM is a special case of a multinomial logit model, namely, an adjacent category logit model with logit link and step difficulties associated with category $k$ of item $i$ (Agresti, 2002; Skrondal & Rabe-Hesketh, 2004). Although there can be many link functions that can be acceptable, the logit link has been more commonly used in the psychometric literature.

In this study, we decide to use the probit link due to a simpler estimation of the model parameters of our interest. Although the logit and probit functions are practically identical except that logit curve has slightly flatter tails, up to our knowledge, no software allows us to estimate the model parameters (i.e., heteroscedestic measurement error for individual raters) we are interested in via a logit link. Thus, we chose to apply mixed effects ordered probit models.

Furthermore, it should be noted that the scores from human raters (HRs) are qualitatively different from the scores from HE and AE. More specifically, scores from HRs are given independently by individual HRs for each essay, while the HE score is a consensus score agreed by a group of human experts through discussion or expert panel. Thus, HE scores are different from independently observed HR scores. Moreover, the AE score is a predicted score using a mathematical algorithm based on usefully predictive features of the text (e.g., essay length), and the algorithm is validated based on separate data (Landauer, et al., 2003a, 2003b). Clauser and colleagues (2000) and Landauer and colleagues (2000) noted that AE scores were generated to be more consistent due to the mechanical nature of its scoring processes. In other words, the two types of target scores, HE and AE, are not exchangeable with the HRs (Raudenbush, 1993); thus, they can respectively serve as a reference point in two consecutive analyses. For example, we have no a priori basis to

predict how the parameter of human rater *r* will differ from that of another human rater *r'*, but we may indeed have prior information about the parameter of HE and AE.

Thus, in this study, considering the exchangeability principle, we apply two consecutive analyses anchoring the target scores (HE or AE): a) HR + HE where we fix the model parameter for the HE and estimate them for individual HRs and b) HR + AE where we fix the model parameter for the AE and estimate them for individual HRs. If HE and AE depict the rater effects in the same way, we can expect that the parameter estimates for individual HRs from two consecutive analyses would show consistent and similar patterns.

In line with that comparison, we introduce two models applied in this study, a mixed effects ordered probit model for rater severity (S-HE, S-AE), and another model for rater severity and rater-specific measurement error variances (SA-HE, SA-AE). We also relate the model parameters to the rater effect indicators of our interest (i.e., rater severity, rater accuracy) and introduce how their transformations can be used for the third rater effect indicator of our interest (i.e., rater centrality). In summary, four types of analyses are conducted depending on the target score and model specifications, as summarized in Table 1. Note that rater centrality is not estimated as a model parameter but is determined from the thresholds transformed based on the resulting parameter estimates.

Elsewhere, Wolfe and colleagues (Myford & Wolfe, 2004; Wolfe & McVay, 2012; Wolfe, 2004) provided a summary of rater effect indicators. However, those indicators are grounded and derived from a different modeling strategy, which was the Rasch rating scale model (Andrich,1978) or the many-facet Rasch model (Linacre, 1994), and they used the residuals after those models are fitted. Since the models analyzed in this study use different link function and different estimation methods, it is not yet known whether those indicators are directly applicable to the model employed in this study. Instead, we identified rater effect indicators relating to the estimated model parameters and their transformations. In this process, we utilize HE or AE as "target scores" by fixing their estimates and use them as the basis for making decisions about individual raters (HR). We then compare the decisions that are made utilizing HE and AE scores as targets to determine whether HE and AE produce different depictions of the performance of individual human raters.

**Table 1:**
Four types of analyses

| | | Mixed effects ordered probit models | |
|---|---|---|---|
| | | Rater severity | Rater severity + Rater accuracy |
| Target Score | HE (HR+HE) | S-HE (severity estimated using HE anchoring) | SA-HE (severity and accuracy estimated using HE anchoring) |
| | AE (HR+AE) | S-AE (severity estimated using AE anchoring) | SA-HE (severity and accuracy estimated using AE anchoring) |

## Mixed-effects ordered probit model for rater severity

We can specify models for ordinal scores by using either a generalized linear mixed model formulation or a latent-response formulation (Agresti, 2002; Rabe-Hesketh & Skrondal, 2012). There are three ingredients for a generalized linear mixed model formulation; link function (i.e., logit, probit), linear predictor (i.e., set of independent variables), and conditional distribution of the responses (i.e., multinomial distribution for ordinal responses).

First, we consider a cumulative ordinal probit model with a random intercept for person proficiencies, $\theta_n \sim N(0,\psi)$. Mixed-effects ordered probit regression is ordered probit regression containing both fixed effects and random effects. In the absence of random effects, mixed-effects ordered probit regression reduces to ordered probit regression. The first model for the ordinal score $X_{pr}$ assigned by rater $r$ to person $p$'s essay is

$$\Pr(X_{pr} > s \mid \theta_p) = \Phi(\theta_p - \kappa_s) \tag{3}$$

where $\Phi(\cdot)$ is the standard normal cumulative density function and $\kappa_s$ is the threshold for score category $s$. This model can also be written using the latent-response formulation, with the latent-response model and the threshold model specified as

$$X^*_{pr} = \theta_p + \varepsilon_{pr}, \qquad \theta_p \sim N(0,\psi), \qquad \varepsilon_{pr} \mid \theta_p \sim N(0,\sigma) \tag{4}$$

$$X_{pr} = s \qquad \text{if } \kappa_{s-1} < X^*_{pr} < \kappa_s, \quad s = 1,...,S \tag{5}$$

respectively, with $\kappa_0 = -\infty$ and $\kappa_S = \infty$. This corresponds to a classical test theory model for $X^*_{pr}$ when $\theta_p$ represents truth and $\varepsilon_{pr}$ represents measurement error. The model assumes that all $r$ raters evaluate the same truth $\theta_p$ with the same measurement error variance $\sigma$ and assign scores using the same thresholds $\kappa_s$ ($s$=*1,...,S-1*). We can allow for rater severities to be different by including rater-specific fixed-effects $\rho_r$. However, one of the intercepts must be set to zero to identify all the thresholds. Retaining the threshold model (5), we extend the latent-response model (4) to

$$X^*_{pr} = \theta_p + \rho_1 x_{1r} + \rho_2 x_{2r} + \cdots + \rho_{(R-1)} x_{(R-1)r} + \varepsilon_{pr} \tag{6}$$

where $\mathrm{X}_r = (x_{1r}, x_{2r}, ..., x_{(R-1)r})'$ are dummy variables for raters from 1 to *R-1*. This Equation (6) corresponds to the S-HE and S-AE in Table 1 with use of different target scores $R$. The corresponding regression coefficients $(\rho_1, \rho_2, ..., \rho_{R-1})$ represent how much more severe or lenient each rater is than the last rater $R$ (chosen arbitrarily). We arrange the last rater $R$ as the HE or AE (i.e., the comparison target in a rater monitoring application) and fix the intercept of HE or AE as zero respectively in each analysis.

### Mixed-effects ordered probit model for rater severity and rater-specific measurement error variances

Although the above model accommodates rater severity, it is relatively restrictive because it still assumes that all raters *r* have the same measurement error variance $\sigma$. We can relax this homoscedasticity assumption by retaining the previous models (5) and (6) except that we now also allow each rater to have a rater-specific residual variance or measurement error variance $\sigma_r$, $\varepsilon_{pr} \mid \theta_n \sim N(0, \sigma_r)$. This can be accomplished by specifying a linear model for the log standard deviation of the measurement errors using *gllamm* function in Stata, the software we chose to use (Rabe-Hesketh & Skrondal, 2012):

$$\ln(\sqrt{\sigma_r}) = \ln(\sigma_r)/2 = \delta_1 x_{1r} + \delta_2 x_{2r} + \cdots + \delta_{(R-1)} x_{(R-1)r} \quad . \tag{7}$$

In this model for level-1 heteroscedasticity, we have again omitted the dummy variable for the last rater *R* (again, chosen arbitrarily) corresponding to HE or AE, which amounts to setting the standard deviation of the measurement error for this rater to 1 because exp(0)=1. A constraint like this is necessary to identify the model because all thresholds $\kappa_s$ (*s=1,2,3* in our data) are freely estimated. In terms of the above parameterization, the measurement error variance $\sigma_r$ for rater *r* becomes exp( $2\delta_r$ ). In this model, each rater has his/her own mean and variance,

$$X^*_{pr} \mid \theta_p \sim N(\theta_p + \rho_r, \sigma_r), \qquad \rho_R = 0\,, \tag{8}$$

but applies the same thresholds to the latent responses to generate the observed ratings. The cumulative probabilities are

$$\Pr(X_{pr} > s \mid \theta_p) = \Pr(X^*_{pr} > \kappa_s \mid \theta_p) = \Pr\left(\frac{X^*_{pr} - \theta_p - \rho_r}{\sqrt{\sigma_r}} > \frac{\kappa_s - \theta_p - \rho_r}{\sqrt{\sigma_r}}\right) = \Phi\left(\frac{\theta_p + \rho_r - \kappa_s}{\sqrt{\sigma_r}}\right). \tag{9}$$

This model can be thought of as a generalized linear model with a scaled probit link, where the scale parameter $\sqrt{\sigma_r}$ differs between raters, *r*. The covariate effect $\rho_r$ is constant across categories, a property sometimes referred to as the parallel-regression assumption because the linear predictors for different categories are parallel.

### Rater effects in relation to model parameters and transformations

Previously, we define severity/leniency as when a rater consistently assigns a lower/higher score than the target scores (HE or AE). In our study, we depicted severity/leniency by specifying the target scores to be zero, so estimates from the two separate analyses that utilize HE and AE in this manner depict a rater as being severe if his/her rater severity/leniency ( $\rho_r$ ) estimate is significantly smaller/greater than zero. We also defined accuracy/inaccuracy as when a rater exhibits large/small random variability in

their scores, compared to the target score. In this study, we focus on the measurement errors associated with individual raters for rater accuracy. Thus, using the resulting estimates from the two sets of analyses that utilize HE and AE, a rater was considered accurate/inaccurate if his/her rater-specific measurement error variance ( $\sigma_r$ ) estimate is significantly lower/higher than the fixed value of HE or AE (1). Finally, we defined centrality/extremity as when a rater consistently assigns scores in the middle/extreme rating scale categories. Using the resulting estimates from the two analyses, we calculate estimates of the "reduced-form" thresholds (Rabe-Hesketh & Skrondal, 2012).

$$\frac{\kappa_{sr}}{\sqrt{\sigma_r}} = \left\{ \begin{array}{ll} \dfrac{\alpha_{s1} - \rho_r}{\exp(\delta_r)} & \text{for r = 1} \\ \dfrac{\alpha_{s1} + \alpha_{sr} - \rho_r}{\exp(\delta_r)} & \text{for r > 1} \end{array} \right\} \tag{10}$$

After transforming to the reduced-form thresholds, we decide which raters exhibit centrality using the thresholds of the HE and the AE as the basis. In detail, we compute the differences between the thresholds and compare those values to the corresponding differences computed from the HE and the AE. For example, if the gap between the thresholds for a certain rater is smaller than the corresponding gap from the HE, that rater is considered exhibiting centrality compared to the HE.

## Estimation

Because our modeling requires us to handle the rater-specific heteroscedastic variances, we utilized the *gllamm* command (Rabe-Hesketh, Skrondal, & Pickles, 2004) running in the widely available statistical package Stata (StataCorp., 2013). Maximum likelihood estimation was implemented in the software gllamm using adaptive Gauss-Hermite quadrature with eight integral points for the mixed-effects ordered probit models (Rabe-Hesketh, Skrondal, & Pickles, 2005). Adaptive quadrature appears to be suitable when the posterior distribution is close to normal and when it is highly non-normal, whereas ordinary quadrature fails in the first situation (Rabe-Hesketh, Skrondal, & Pickles, 2002). Adaptive quadrature is computationally more efficient than ordinary quadrature and other computer intensive methods such as Markov chain Monte Carlo. It also provides a value for the maximized log likelihood useful for likelihood-ratio tests. In contrast to ordinary quadrature, adaptive quadrature also appears to give good parameter estimates for linear models, and is useful for complex multilevel latent variable models that cannot yet be handled by other software, although computationally less efficient than other methods.

## Illustration

### Data

Each of 131 human raters (HRs) assigned holistic scores on a 4-point rating scale to 189 essays written by middle-school students in response to an explanatory prompt on a statewide writing assessment. Each essay was assigned a consensus score, which is considered as gold standard, by a panel of HE. Figure 1 presents the structure of the empirical data used in this study. The data has a multilevel structure, in which the ratings are nested not only within students but also within each rater. Note again that we viewed the ratings from the HE and the AE as two structurally different target scores from the HRs and fixed their estimates, treating them as the criterion for other HRs.

AE scores used in this study were obtained from an automated scoring engine, the Intelligent Essay Assessor (IEA; Foltz, Streeter, Lochbaum, & Landauer, 2013; Foltz, Laham, & Landauer, 1999) that was calibrated on a separate sample of essays from the same population in this study. The most unique feature of this IEA system is the application of latent semantic analysis (LSA) to measure the writing quality more directly. That is, LSA can judge the semantic relatedness and similarity among essays rather than relatively peripheral aspects, such as grammar and typos. To compute a total outcome score, IEA combines three kinds of components – content, style, and mechanics – by a form of constrained multiple regression based on human scores in a training sample. Landauer and colleagues developed IEA based on LSA, and reported validity and reliability of IEA scores using many sets of simulated and real data sets (e.g., Landauer, et al., 2003a, 2003b): Overall, IEA to human reliabilities were the same as human to human reliabilities within probable measurement error.

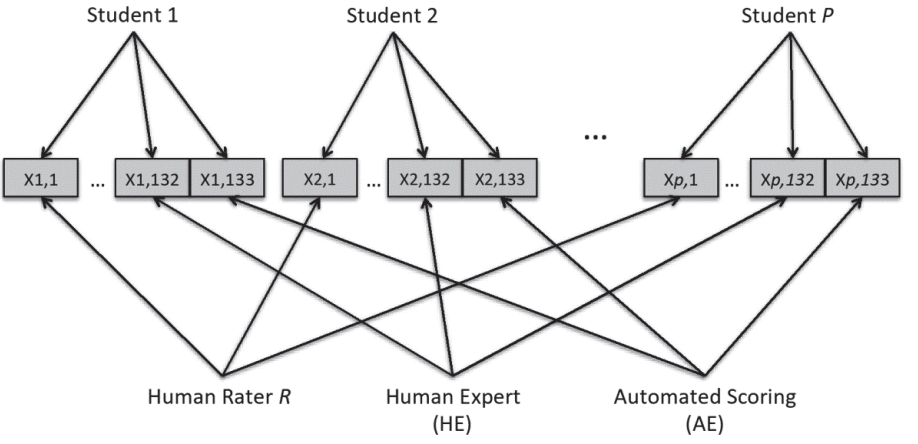


**Figure 1:**
Data Structure

## Descriptive results

Table 2 presents a direct comparison between the AE and the HE for the 189 essays. Out of the total 189 essays, the exact agreement rate between the HE and the AE was 64.02%. There was only 1 case in which the discrepancy between the AE and the HE was greater than 1 score point. When the off-diagonal elements were summed, the upper diagonal was 14.29% while the lower diagonal was 21.69%, which suggests that the AE was likely to produce lower scores compared to the HE. Interestingly, at the higher score category, the discrepancy was large: Only 0.53% when the AE gave 4 but HE gave 3, but 6.35% when the HE gave 4 but the AE gave 3.

To get a sense of how the rater effect indicators compared with the target scores, three types of corresponding descriptive statistics were calculated. First, regarding severity/leniency, the percentage of scores that was higher than the target scores given by human raters was calculated. From this aggregated information, the AE appears slightly more severe than the HE, so more human raters are classified as lenient when the AE was used as the target score, ignoring the score categories. Second, regarding accuracy/inaccuracy, the exact raw score agreement rate with the target scores for each rater was calculated. When the HE was used as the target score, the exact agreement rates ranged from 38.10% to 73.55%, while the corresponding range was from 35.45% to 67.20% when the AE was used as the target score. The median rate was 56.61% for the HE and 53.97% for the AE. The correlation between the two sets of exact agreement rates was calculated as 0.72. Another way we examined the accuracy/inaccuracy was to calculate Spearman's rank-based correlation between the rating from each rater and target scores. The range of the correlation was 0.42 to 0.77 when HE was used as a target score, and the corresponding range was 0.39 to 0.77 when AE was used. Taken together, human raters appear to be classified similarly in terms of accuracy/inaccuracy when the HE score was used as the target score compared to when the AE score was used as the target score. Third, regarding centrality/extremity, the percentage of scores in rating scale categories 2 or 3 was calculated for each target rater (AE and HE), and the obtained values were 75.13% for HE and 83.60% for AE. This suggests that the AE showed more centrality than the HE. The range of the corresponding values of 131 human raters was 42.33% to 88.89%. Among the 131 human raters, 53 raters had a higher percentage of

**Table 2:**
Direct comparison between the AE and the HE

| | | AE | | | | Agreement % | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| HE | 1 | 20 | 10 | 1 | 0 | 10.58 | 5.29 | 0.53 | 0.00 |
| | 2 | 6 | 56 | 15 | 0 | 3.17 | 29.63 | 7.94 | 0.00 |
| | 3 | 0 | 23 | 41 | 1 | 0.00 | 12.17 | 21.69 | 0.53 |
| | 4 | 0 | 0 | 12 | 4 | 0.00 | 0.00 | 6.35 | 2.12 |

scores in categories 2 and 3 than the HE while only seven raters had a higher percentage than the AE, and all seven exhibited centrality when compared to AE as well. This aggregated information suggests that the AE would likely classify fewer raters as exhibiting centrality.

### Results for the mixed-effects ordered probit models

A total of four types of analyses, which include two different mixed-effects ordered probit models, one for rater severity and another for rater severity as well as rater-specific measurement error variances for two different data sets with different target scores, HR+HE and HR+AE, were conducted. First, based on the first model which specified only rater severity with the HE as a target score (S-HE), thresholds were estimated as -1.65, 0.37, and 2.26, and the variance of the person proficiency distribution ( $\psi$ ) was estimated as 2.08. Second, based on the first model which specified only rater severity with the AE as a target score (S-AE), thresholds were estimated as -1.51, 0.51, and 2.40, and the variance of the person proficiency distribution ( $\psi$ ) was estimated as 2.07. Although the threshold estimates are not directly comparable between S-HE and S-AE, the 95% confidence interval of the threshold estimates were overlapped by each other. Overall, the threshold estimates were consistently slightly higher when the AE was used as a target score. This implies that assuming the same measurement error variances across the raters, thresholds and variance of the person proficiency distribution were estimated quite similarly regardless of whether we use HE or AE as the target scores.

To our knowledge, there is no statistical software that provides $R^2$ type of statistics for the multilevel probit (and logit) models. Additionally, as a way to investigate whether the models provide sufficiently accurate fit, a link test was conducted to test the specification of the dependent variable using *linktest* function in Stata. This test was suggested by Pregibon (1979, 1980) based on an idea of Tukey (1949) that if a regression is properly specified, no additional independent variables should be significant except by chance. Both models have passed the link test, which suggests that the dependent variable was quite accurately specified in each model (S-HE: $p$=0.820, S-AE:$p$=0.822).

Next, when we allowed different measurement error variances across each rater in the second model (SA-HE and SA-AE), the results showed different patterns to some extent. When the HE was used as a target score, the thresholds were estimated as -2.93, 0.63, 3.97, while the corresponding values were -2.19, 0.71, and 3.42 when the AE was used as a target score. Unlike the results from the S-HE and S-AE, the thresholds were shrunken toward when the AE was used as a target score. However, the 95% confidence intervals associated with each threshold still overlapped each other. Moreover, the variance of the person distribution was estimated as 6.46 when the HE was used as a target score and the corresponding value was 4.26 when the AE was used as a target score. Compared with the results from the S-HE as well as S-AE, the variance estimates from both data sets became much larger, and the estimated variances were quite different. As the threshold estimates were spread wider when the HE was used as a target score, the person proficiency estimates were distributed in a wider range.

In order to compare the model fit between the two mixed-effects ordered probit models analyzed in this study, we can use a likelihood-ratio test (Rabe-Hesketh & Skrondal, 2012). Comparison between the two models tests the null hypothesis that the measurement error variances are identical for the raters, against the alternative that the measurement error variances are different for at least two raters. Under the more restricted model (S-HE and S-AE), the measurement error variances are set to 1, and the thresholds of all raters are set to be equal; that is, in model in Equation (6), the constraints $\delta_r = 0$ for *r=1,2,..., R-1* in place, but the intercepts $\rho_r$ for raters are free parameters. The more complex model is the same except that the constraints for $\delta_r$ are relaxed (SA-HE and SA-AE). For both types of target scores, the likelihood-ratio test yielded that the more complex model fitted significantly better ($\chi^2(131)=1089.62$, $p<0.001$ for HR+HE and $\chi^2(131)=1074.11$, $p<0.001$ for HR+AE). AIC (Akaike Information Criterion; Akaike, 1974) that panelizes the complexity of the models also preferred the more complex model (41190.74 vs. 40363.11 for HR+HE, and 41213.14 vs. 40401.02 for HR+AE). This suggests strong evidence that at least two participating raters do not have the same measurement error variances.

Next, to depict the rater effects in detail, we used the resulting estimates from the second model for each type of our rater effects indicators. Note again that in two different data sets, we fixed the parameters associated with target scores, HE and AE respectively. Beyond the statistical significance of each estimate, effect sizes for each type of our rater effects indicators need to be considered. However, because conventional effect size estimation is not appropriate and the associated standard errors are incorrect in a multilevel structure setting (e.g., cluster randomized-trials, meta-analysis), we do not report the standard effect size at this moment (Donner & Klar, 2002; Rooney & Murray, 1996). Further work is needed on this topic.

*Severity/Leniency*

Figure 2 plots the rater-specific fixed effects ( $\rho_r$ ), which correspond to the rater severity/leniency when the HE or the AE was used as a target score. In general, the rater-specific fixed effects when the AE was used as a target score tended to have higher values compared to when the HE was used as a target score, particularly at the lower range. Interestingly, the range of the estimates when the HE was used as a target score was wider from -2.43 to 2.19, while the range of the estimates when the AE was used as a target score was narrower from -1.78 to 1.98. This implies that when the AE was used as a target score, the differences in severity/leniency across the raters are likely to be condensed compared to when the HE was used as a target score. The compressed pattern from AE is indicated by a skewed line away from the identity line, and this is likely due to the fact that AE scores are more centered around middle categories compared to HE scores, as shown in Table 1 (the percentage of scores 2 or 3: 75.13% for the HE and 83.60% for the AE).

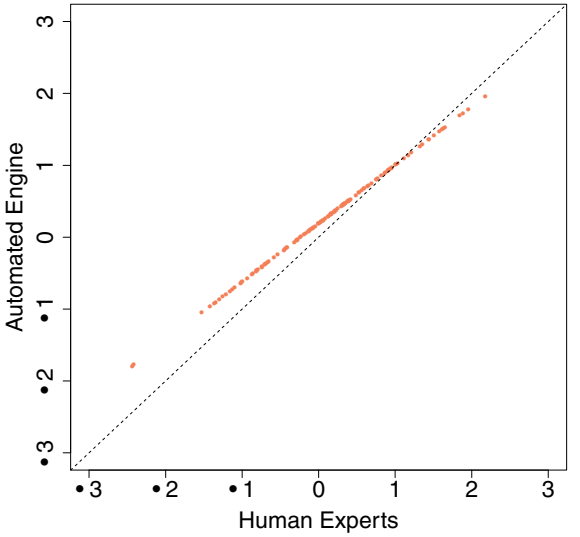


**Figure 2:**
Severity/Leniency of each HR depending on the target score

To be specific, raters were classified as severe if the estimate is statistically significantly lower than zero (i.e., fixed value of the gold standards, either AE or HE) and lenient if the estimate is statistically significantly higher than zero. Based on this criterion, 48 raters were classified as lenient raters and 35 raters were classified as severe raters using AE as the comparison target. Interestingly, classification of the raters in terms of the severity/leniency were exactly matched when the HE was used as a target score. Taken together, although the magnitude of the estimate appears slightly different, more raters were classified as lenient while fewer raters were classified as severe regardless of the type of the target score. Thus, it appears reasonable to say that the AE depicts rater severity/leniency in the same way as the HE.

*Accuracy/Inaccuracy*

Figure 3 illustrates the estimates of the rater-specific measurement error variances ($\sigma_r$) depending on the target score. Again, estimates for the HE and the AE were fixed as 1, respectively (Rabe-Hesketh & Skrondal, 2012). In general, most of the human raters among 131 raters demonstrated higher measurement error variances compared to both types of target scores. In terms of the magnitude, the measurement error variances were estimated higher across all the raters when the HE was used as a target score, compared to when the AE was used as a target score. However, given that the estimates of rater-specific measurement error variances are located in a quite higher range, using the fixed value 1 as the criterion appears very stringent. In detail, the measurement error variances were distributed wider when the HE was used as a target score ranging from 0.98 to 8.48, while the corresponding values when the AE was used as a target score ranged from 0.66 to 5.64.

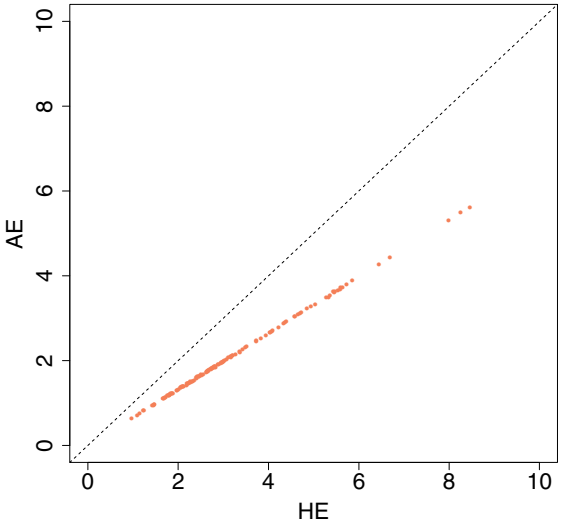


**Figure 3:**
Rater-specific measurement error variance of each HR depending on the target score

Raters were classified as inaccurate if the estimate is significantly higher than the fixed value 1, which seems like a very stringent criteria. Using the standard error associated with the rater-specific measurement error variance estimates, 95% confidence intervals for each rater were constructed and compared with the fixed value 1. Based on this criterion, as expected, most of the raters, 122 out of 131 raters, were classified as inaccurate when the HE was used as the target score. In contrast, 87 out of 131 raters were classified as inaccurate when the AE was used as the target score. These 87 raters were the subset of the 122 raters labeled as inaccurate based on the HE. Taken together, considering the limitation that we used the harsh criterion, it seems that the AE provides a different story from the HE for rater accuracy by labeling fewer raters as inaccurate.

*Centrality/Extremity*

Figure 4 displays reduced-form thresholds, red for the first threshold, orange for the second threshold, and green for the third threshold, transformed using Equation 10. The solid line represents the thresholds when the HE was used as a target score, and the dotted line represents the thresholds when the AE was used as a target score. The thresholds for each rater are indexed as a larger hollow circle when the HE was used as a target score, and smaller solid circles are used when the AE was used as a target score. As illustrated, the locations of two types of the circles were quite similar apparently due to the parallel-regression assumption. However, as discussed before, the thresholds were shrunken when the AE was used as a target score compared to when the HE was used as a target score, and this would likely affect our decision on rater centrality/extremity.


**Figure 4:**
Thresholds compared to the HE (solid) and the AE (dotted)

Specifically, we computed the differences between the thresholds and compared those values to the corresponding differences from each threshold when HE and AE were used as target scores. Note that thresholds for each target score were not fixed, but computed based on the resulting parameter estimates and fixed values: the differences between the thresholds were larger for the HE (3.56 between the first and the second threshold, and 3.33 between the second and the third threshold), compared to the value for the AE (2.90 between the first and the second threshold, and 2.72 between the second and the third threshold). As summarized in the Table 3, when the HE was used as a target score, the gap between the first and the second threshold of human raters ranged from 1.22 to 3.59 with the median of 2.13 and the mean of 2.14. The gap between the second and the third threshold ranged from 1.15 to 3.36 with the median of 2.00 and the mean of 2.01. The corresponding ranges when the AE was used as a target score showed quite similar ranges: the gap between the first and the second threshold ranged from 1.22 to 3.58 with the median of 2.13 and the mean of 2.15, and the gap between the second and the third threshold ranged from 1.14 to 3.35 with the median of 2.00 and the mean of 2.01. Taken together, the ranges from the human raters were much smaller than the ranges computed from the HE and the AE, implying that this criterion looks very stringent.

Given that the thresholds of each rater based on either HE or AE yielded similar values, more raters are likely to be classified as demonstrating centrality when the HE was used as a target score because of the criterion (i.e., threshold gaps of the HE and the AE; 3.56 and 3.33 for the HE and 2.90 and 2.72 for the AE). Based on this criterion, 130 out of 131 raters were classified as exhibiting centrality when the HE was used as a target score, while 122 out of 131 raters were classified as exhibiting centrality when the AE was used as a target score. These 122 raters were the subset of the 130 raters labeled as central based on the HE. It appears that the AE depicts rater centrality slightly different from the HE, by labeling fewer raters as centrality.

**Table 3:**
Summary of distance between thresholds

| | Mean of distances between thresholds | | Median of distances between thresholds | | Range of distances between thresholds | |
|---|---|---|---|---|---|---|
| | 1 & 2 | 2 & 3 | 1 & 2 | 2 & 3 | 1 & 2 | 2 & 3 |
| HE anchoring | 2.14 | 2.01 | 2.13 | 2.00 | 1.22 ~ 3.59 | 1.15 ~ 3.63 |
| AE anchoring | 2.15 | 2.01 | 2.13 | 2.00 | 1.22 ~ 3.58 | 1.14 ~ 3.35 |

## Conclusion and discussion

Due to the high cost associated with monitoring raters, particularly in the assignment of human consensus scores, we sought to determine whether scores from an automated scoring engine could supplant human consensus scores. Specifically, we analyzed empirical rating data using two different mixed-effects ordered probit models – one that afforded comparison of raters to a human consensus score (HE) and another that afforded comparison of raters to an automated score (AE). The important question that we sought to answer was "do we make similar decisions about raters when comparing raters to these two target scores". The answer for this question is summarized in Table 4. The table presents the number of flagged raters for each rater effect indicator when different target scores were used, and the proportion of identical decision between two different target scores.

**Table 4:**
Comparison between the AE and the HE

| | # of flagged raters (HE anchoring) | # of flagged raters (AE anchoring) | % of identical decision |
|---|---|---|---|
| Severity | 35 | 35 | 100.0% |
| Accuracy | 122 | 87 | 66.4% |
| Centrality | 130 | 122 | 93.1% |

For the data in this study, the results showed that the AE depicts the HE exactly the same in terms of the rater severity, while slightly different in terms of the rater centrality and considerably different in terms of the rater accuracy. In particular, the AE labeled raters identically as lenient and severe as the HE did. However, AE classified fewer raters as demonstrating inaccuracy and centrality. The difference was only slight for centrality (93% of the raters were identically labelled) and was probably unacceptable for accuracy labels (only 66% were identical). Unlike the consensus scores assigned by the group of human experts, previous studies reported that AE scores were generated to be more consistent due to the mechanical nature of its scoring processes (Clauser, et al., 2000; Landauer, et al., 2000). This could be the reason why AE did not depict the HE particularly in terms of the accuracy and centrality that are more related with the variability of the observed scores. Furthermore, the employed criteria for rater inaccuracy and centrality appeared to be very stringent. We fixed the model parameters associated with the HE and the AE and used them as the criteria, but the range of the parameter estimates was quite large compared to those fixed values. It is possible that this is related to the relatively smaller sample size compared to the huge number of model parameters, which led to larger standard errors. Thus, for future study, it may be interesting to explore the use of these models to larger data sets to examine the findings in a similar context, or to experiment with more reasonable and realistic criteria for rater inaccuracy and rater centrality.

Furthermore, it would be interesting to estimate effect sizes for each type of our rater effect indicators considering the multilevel ordinal data in this study. Recently, Hedges (2007) defined and proposed effect size estimation in cluster randomized trials by correcting the intraclass correlation. Larsen and Merlo (2005) also suggested calculation of the median odds ratio as a measure of heterogeneity, which can also be understood as a type of effect size. However, they were not designed for ordinal data, and it is unknown whether their methods can be directly applicable for the multilevel probit regression we used in this study. Thus, it will be worthwhile to examine and possibly modify their methods for our setting in order to develop more practical criteria and utilize them in addition to the statistical significance.

In addition, we illustrated the use of two different mixed-effects ordered probit models with an empirical example. A simpler model incorporated only the rater severity and a more complicated second model also allowed level-1 heteroscedasticity for rater-specific measurement error variances. Comparison between these two models revealed that it is necessary to relax the same measurement error variances across the raters. In particular, the modeling strategy used in this study is convenient and straightforward since it estimates the model parameters without having to set any arbitrarily defined cut scores. Those model parameters can be directly related with types of rater effects indicators of our interest, such as rater severity/leniency and rater accuracy/inaccuracy. Fixed-effects rater location estimates were used for rater severity/leniency, and rater-specific measurement error variances were used for rater accuracy/inaccuracy. Thus, the fixed values used for hypothesis testing are meaningful because they provide the basis for statistical comparison of individual HRs against each target score with respect to severity and accuracy. We also computed the thresholds using the resulting estimates to depict rater centrality/extremity. Both models assumed the proportional regression assumption that the linear predictors for different categories are parallel. Given the sample size and the number of model parameters, we were not successful in relaxing this assumption. However, relaxation of this assumption would lead to estimation of model parameters that are more directly related with the rater centrality/extremity, by allowing rater-specific thresholds.

The present study rests on one real data set based on the fully crossed design, which was experimental and ideal enough to estimate multiple rater effects for individuals. Because most of the automated scoring systems, including the one used in this study, are data-driven (from specific data sets) statistical procedures that maximize the predictive accuracy of the outcome variables, AE scores as a reference might not be stable enough to estimate multiple rater effects. Furthermore, as Clauser, et al (2000) revealed, the specific algorithm represents the policy of a sample of qualified experts. Thus, even with the same automated scoring engines, the results might be different depending on the trained data that were used for building the automated scoring engine system. Ultimately, these results are not readily generalizable to other tasks or to other scoring machines, but the presented methods can be useful to investigate the potential use of AE for rater monitoring purposes.

The potential cost savings associated with a transition from using automated scores instead of the current emphasis on human expert scored validity papers is incredible. For

example, ignoring the cost of collecting the human expert scores and the cost of training the automated scoring engine, let us assume that the raters in our study would have been expected to score validity papers at a rate of 5% (i.e., every 20$^{th}$ score would have been a validity paper). Each of our 131 raters scored 189 essays – an extremely small project. That means that our raters would have scored about an additional 10 validity papers each (189/19 = 9.95) for the sake of rater monitoring. Jointly, they would have assigned 1,310 scores, which would have increased scoring time and cost by 5.3% [131 raters × (189 essays + 10 validity papers) = 26,069, 131 raters × 189 essays = 24,759, 100 × 26,069 / 24,759 = 105.29%). A 5% savings is not insignificant. In conclusion, it seems that automated scores could, potentially, supplant human expert scores as the target for rater monitoring in terms of rater severity. In our study, we determined that, at least for our data, raters were generally labelled as exhibiting severity/leniency regardless of which target scores were used. However, monitoring rater accuracy/inaccuracy or centrality did not result in sufficient correspondence that would be useful in applied settings, which should be considered and interpreted with more caution.

**Acknowledgement**

The majority of the work was done when the first author, Hyo Jeong Shin, was at the University of California at Berkeley, and the second author, Edward Wolfe, was at Pearson. Hyo Jeong Shin is now Research Scientist at Educational Testing Service, and Edward Wolfe is now Principal Research Scientist at Educational Testing Service. The authors would like to thank Emily Lubaway and Larry Hanover for their editing help. The ideas and opinions expressed herein are those of the authors alone, and endorsement by the authors' institutions is not intended and should not be inferred.

## References

Agresti, A. (2002). *Categorical data analysis*. John Wiley & Sons.

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, *19*(6), 716–723.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561–573.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. *The Journal of Technology, Learning and Assessment*, *4*(3).

Clauser,B. E., Harik, P., & Clyman, S. G. (2000). The generalizability of scores for a performance assessment scored with a computer-automated scoring system. *Journal of Educational Measurement, 37*, 245–261.

Clauser, B. E., Kane, M. T., & Swanson, D. B. (2002). Validity issues for performance-based tests scored with computer-automated scoring systems. *Applied Measurement in Education*, *15*(4), 413–432.

Donner, A., & Klar, N. (2002). Issues in the meta-analysis of cluster randomized trials. *Statistics in Medicine*, *21*(19), 2971–2980.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*(6), 359–374.

Foltz, P. W., Streeter, L. A., Lochbaum, K. E., & Landauer, T. K (2013). Implementation and applications of the Intelligent Essay Assessor. *Handbook of Automated Essay Evaluation*, M. Shermis & J. Burstein, (Eds.). pp. 68-88. Routledge, NY.

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). The Intelligent Essay Assessor: Applications to Educational Technology. *Interactive Multimedia Education Journal of Computer Enhanced Learning*. *1,*(2).

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, *32*(4), 341–370.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003a). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M.D. Shermis & J. Berstein (Eds.), *Automated Essay Scoring: A Cross-Disciplinary Perspective*, pp. 87–112. Mahwah, Nj: Lawrence Erlbaum.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003b). Automatic essay assessment. *Assessment in Education: Principles, Policy, & Practice, 10*, 295-308.

Larsen, K., & Merlo, J. (2005). Appropriate assessment of neighborhood effects on individual health: integrating random and fixed effects in multilevel logistic regression. *American Journal of Epidemiology*, *161*(1), 81–88.

Linacre, J. M. (1994). *Many-facet Rasch measurement*. Chicago, IL: Mesa Press.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4*, 386–422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, *5*, 189–227.

Myford, C. M., & Wolfe, E. W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement*, *46*, 371–389.

Pregibon, D. (1979). Data analytic methods for generalized linear models. PhD dissertation, University of Toronto.

Pregibon, D. (1980). Goodness of link tests for generalized linear models. *Applied Statistics, 29*, 15–24.

Rabe-Hesketh, S., & Skrondal, A. (2012). *Multilevel and longitudinal modeling using Stata* (3rd ed.). Stata Press.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal*, *2*(1), 1–21.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). *GLLAMM manual* (No. 160).

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics*, *128*(2), 301–323.

Raudenbush, S. W. (1993). A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *Journal of Educational and Behavioral Statistics*, *18*(4), 321–349.

Rooney, B. L., & Murray, D. M. (1996). A meta-analysis of smoking prevention programs after adjustment for errors in the unit of analysis. *Health Education & Behavior*, *23*(1), 48–64.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, *88*, 413–428.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Crc Press.

StataCorp. (2013). *Stata Statistical Software: Release 13.0*. Colleage Station, TX: Stata Corporation.

Tukey, J. W. 1949. One degree of freedom for non-additivity. *Biometrics, 5*, 232–242.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, *31*(1), 2–13.

Wolfe, E. (2014). *Methods for Monitoring Rating Quality: Current Practices and Suggested Changes* (White Paper). Pearson.

Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science*, *46*, 35–51.

Wolfe, E. W., & McVay, A. (2012). Applications of latent trait models to identifying substantively interesting raters. *Educational Measurement: Issues and Practice*, *31*, 31–37.

Yang, Y., Buckendahl, C. W., Juszkiewicz, P. J., & Bhola, D. S. (2002). A review of strategies for validating computer-automated scoring, *Applied Measurement in Education, 15*(4), 391-412, DOI: 10.1207/S15324818AME1504_04