Essay Selection Methods for Adaptive Rater Monitoring

Applied Psychological Measurement
2017, Vol. 41(1) 60–79
© The Author(s) 2016
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621616672855
apm.sagepub.com

Chun Wang¹, Tian Song², Zhuoran Wang¹,
and Edward Wolfe³

Abstract

Constructed-response items are commonly used in educational and psychological testing, and the answers to those items are typically scored by human raters. In the current rater monitoring processes, validity scoring is used to ensure that the scores assigned by raters do not deviate severely from the standards of rating quality. In this article, an adaptive rater monitoring approach that may potentially improve the efficiency of current rater monitoring practice is proposed. Based on the Rasch partial credit model and known development in multidimensional computerized adaptive testing, two essay selection methods—namely, the D-optimal method and the Single Fisher information method—are proposed. These two methods intend to select the most appropriate essays based on what is already known about a rater's performance. Simulation studies, using a simulated essay bank and a cloned real essay bank, show that the proposed adaptive rater monitoring methods can recover rater parameters with much fewer essay questions. Future challenges and potential solutions are discussed in the end.

Keywords

Rasch partial credit model, essay selection, Fisher information matrix, interim scoring

Introduction

In educational testing, constructed-response items are widely used. For example, the Partnership for Assessment of Readiness for College and Career (PARCC) uses prose constructed responses in their English language arts/literacy assessment to measure the students' ability to read and analyze passages from real text, as well as their writing skills. In Program for International Student Assessment (PISA), there are also several types of constructed-response items in mathematics, reading, science, problem solving, and financial literacy assessments. Those items require students to produce an answer, and the answers are typically scored by human raters. A common concern about the scores assigned by human raters is the degree to which these scores contain measurement errors due to the subjectivity of human raters' judgments. To address this

¹University of Minnesota, Minneapolis, MN, USA

²Pearson VUE, Bloomington, MN, USA

³Educational Testing Service (ETS), Princeton, NJ, USA

Corresponding Author:

Chun Wang, Quantitative/Psychometric Methods, Department of Psychology, University of Minnesota, N658 Elliott Hall, 75 East River Road, Minneapolis, MN 55455, USA.

Email: wang4066@umn.edu

concern, raters are generally trained to apply the scoring rubric and must meet certain qualifications subsequent to trainings. Moreover, rater monitoring procedures are often employed in operational settings to ensure that raters' scores are as accurate as possible. Several procedures are typically used to maintain acceptable rating quality in operational scoring, including rater training, qualification, rater feedback, recalibration, and back reading (Wolfe, 2014). The authors of the present study do not focus on these methods but instead focus on quantitative methods that can be used to monitor and document inadequacies in rating quality that are attributed to the rater.

Several indices have been developed and are commonly used in operational settings to evaluate rater quality. One of the most common indices is the percentage of perfect and/or perfect-plus-adjacent agreement, defined as the percentage of times that the scores assigned by a rater are in exact agreement with or within one point of the true scores (i.e., scores assigned by experts or other raters). A problem with percentages of agreement is that those indices fail to take into account that agreement can occur by chance. As a result, its value can be misleadingly high. To solve this problem, Cohen introduced the kappa coefficient (Cohen, 1960) and weighted kappa coefficient (Cohen, 1968), which take into account the agreement by chance. A third set of indices currently used in practice are interrater correlations (e.g., Pearson's correlation, intraclass correlation). Those indices measure the strength of the relationship between a rater's scores and the true scores. All of these indices are more appropriate for use as general indicators of the quality of a pool of raters because they provide only limited diagnostic information about individual raters (e.g., whether raters consistently assign lower scores, or whether raters tend to assign scores in middle score categories). That is, the authors could not provide corrective feedback to raters based on these indices—They could only tell raters that their scores are generally in agreement or are not in agreement with the target scores. One remedy to this problem is offered by latent trait modeling.

Detecting Rater Effects Using Latent Trait Models

The term “rater effects” refers to patterns of scores that are associated with measurement error contributed by a rater (Wolfe, 2014). Three most common types of rater effects are Severity/Leniency, Centrality/Extremity, and Inaccuracy/Accuracy. Severity/Leniency refers to a systematic shift of the average score assigned by a particular rater. Severe raters assign scores that are lower than true scores, whereas lenient raters assign scores that are higher. Centrality/Extremity occurs when raters tend to shift scores toward the middle (centrality) or tails of the rating scale (extremity). It is statistically evidenced by a decrease/increase in the standard deviation of the scores assigned by a rater while maintaining a relatively high degree of agreement. Inaccuracy/Accuracy is the degree of randomness in the scores assigned by a particular rater. Inaccurate raters assign scores that deviate from true scores in a seemingly random manner. Hence, inaccurate raters are expected to have a low correlation between their scores and true scores. In contrast, accurate raters would be expected to have a high correlation between their scores and true scores.

In the literature, there is considerable research that applies latent trait models to the detection of rater effects, and several indices have been constructed to detect those effects (DeCarlo, Kim, & Johnson, 2011; Lunz & Stahl, 1993a; Myford & Wolfe, 2003, 2004, 2009; Wolfe, 2004, 2005; Wolfe & Song, 2015). In this article, the authors of the present study focus on one of the most commonly used models, the Rasch partial credit model (RPCM, Masters, 1982). Wolfe, Jiao, and Song (2015) showed that RPCM is able to detect all three types of rater effects. In the context of monitoring rater effects, the indexing and the interpretation of parameters in this model need to be slightly modified. Let p_{ijk} denotes the probability that rater i assigns a score in

rating category k to essay j , and let $p_{ij(k-1)}$ denotes the probability that rater i assigns a score in rating category $k - 1$ to essay j . The log ratio of these events takes the following form (Muraki, 1992):

$$\log \left[\frac{p_{ijk}}{p_{ij(k-1)}} \right] = \theta_j - \lambda_i - \tau_{ik}, \quad (1)$$

where θ_j is the examinee ability, λ_i is the rater location, and τ_{ik} is the category threshold which indicates the relative difficulty of the two adjacent rating categories.

In this model, the parameter λ_i explicitly measures rater Severity/Leniency. Positive values of λ_i indicate rater severity, and negative values indicate leniency. Rater Centrality and Extremity can be detected as a high or low standard deviation of raters' category thresholds [$SD(\tau_{ik})$], respectively. Because central raters tend to have more ratings in middle categories, they are expected to have a higher standard deviation (i.e., a wider dispersion of thresholds). To detect rater inaccuracy, the correlation between the observed ratings and the estimated examinee abilities [$R(X, \theta)$] was used as an indicator. The correlation would be expected to be closer to .00 for inaccurate raters but closer to 1.00 for accurate raters. Because rater effects are quantified as functions (either mean or standard deviation) of rater parameters in Equation 1, the primary focus of this article is to propose methods to best recover rater parameters. Only if the rater parameters are precisely estimated then aberrant raters can be detected by comparing the rater effect indices with proper cutoff values, but this second-stage decision making is not the focus of this article.

Adaptive Rater Monitoring

In the current rater monitoring processes, validity scoring is used to ensure that the scores assigned by raters do not deviate severely from the standards of rating quality. Ratere are required to score fixed sets of validity essays which have been assigned a consensus score by expert raters, and such a consensus score is often referred to as the true score of the essay. These validity essays are usually administered to raters by seeding them into the raters' queue blindly. Examining the correspondence of assigned scores and true scores helps scoring leaders identify questionable raters. However, this procedure has shortcomings. First, validity scoring adds time and expense to the operational scoring project. It requires the additional cost of obtaining true scores, which have to be assigned by expert raters ahead of time. Second, raters are paid to assign scores that are for the sole purpose of monitoring rather than to score operational responses. Therefore, to reduce cost, one needs to minimize the number of validity responses used for rater monitoring. Unfortunately, a small set of validity responses may not adequately represent the full range of examinee performance covered by the scoring rubric, and hence limits the amount of information about the performance of a particular rater one can collect. In addition, using a fixed set of validity responses requires raters to score responses that may provide no useful information about that particular rater's performance. For example, if one suspects that a particular rater is scoring leniently (i.e., assigning scores that are generally too high), then asking this rater to assign validity scores to responses that demonstrate a high level of performance on the assessment task provides a scoring leader with little useful information about this rater. Most raters would be expected to assign high scores to that response. To obtain useful information about this rater, one would need to ask this rater to assign scores to responses that one would expect most raters to assign low scores to. If this rater assigns scores that are too high to these responses, then the authors are in a better position to conclude that this rater is indeed scoring leniently.

An adaptive rating monitoring system could help reduce the cost of rater monitoring and improve the efficiency of that process. In this system, validity responses are selected for each rater in a manner that takes into account what is already known about the rater. That is, if a rater has consistently assigned scores in the upper score categories, that rater would appear to be a lenient rater. Hence, it would be better to obtain scores from that rater on responses with average or low true scores. This adaptive system allows validity responses to be selected adaptively so that scoring leaders can obtain more precise and more diagnostically informative feedback about each rater at a faster pace.

In adaptive rater monitoring, the list of pregraded responses to an essay serves as the analogue to the adaptive testing “item bank,” and each rater replaces the “test taker.” As is the case with traditional computerized adaptive testing (CAT) in which an item bank is composed of items which have been precalibrated prior to the test administration (e.g., Chang, 2004), one needs to assume that there are J pregraded responses in the adaptive rater monitoring essay bank, with examinee’s ability, θ_j ($j = 1, \dots, J$), known beforehand. The goal is then to estimate the parameters related to rater effects, such as λ_i (indicating rater Severity/Leniency) and τ_{ik} (the *SD* of which indicates rater Centrality/Extremity). To best achieve this goal, one would need to select the most suitable responses that are based on the rater’s past performance. Compared with the current rater monitoring procedures in which each rater rates the same set of responses, the proposed adaptive rater monitoring system intends to reduce the number of essay responses each rater needs to score while maintaining the precision of detecting rater effect.

There are three key components of an adaptive rater monitoring system: interim latent trait estimation (aka interim scoring), adaptive essay selection algorithm, and stopping rule. In the Method section, it is described how these key components of adaptive testing can be applied to detecting rater effects. It is also emphasized how these key components differ from those in traditional RPCM-based CAT (e.g., Dodd & Koch, 1987; Koch & Dodd, 1985). Then, two simulation studies are conducted using a simulated and a cloned real essay bank to evaluate the performance of two essay selection algorithms. In our studies, rater Severity/Leniency and Centrality/Extremity are mainly focused, although the same method could be employed for other rater effects that could be detected using an index derived from a latent trait model.

Method

In the context of adaptive rater monitoring, to simplify notations hereafter, Equation 1 is rewritten as follows:

$$p_{ijk} = \frac{\exp \left[\sum_{m=0}^k (\theta_j - \delta_{im}) \right]}{\sum_{k=0}^r \exp \left[\sum_{m=0}^k (\theta_j - \delta_{im}) \right]}, \quad (2)$$

where $\delta_{im} = \lambda_i + \tau_{im}$ for all raters. Given the model identifiability constraints, it is customary to set $\delta_{i0} = 0$ for all i ’s (Muraki, 1992). As a result, for each rater, the number of parameters that need to be estimated is r , whereas the total number of score categories is $r + 1$. By definition, the values of δ_{im} ’s ($m = 0, \dots, r$) can take any order, and the possible reverse order of these values leads to the phenomenon of overlapping. For well-constructed essays and well-trained raters, the δ_{im} values are ordered (e.g., Sébille, Challa, & Mesbah, 2007; or see our real data analysis). Subscript “ i ” is dropped in Equation 2 from here onward because in adaptive rater

monitoring, the interim estimation and essay assignment are conducted for each individual rater separately.

Adaptive Essay Selection

In adaptive rater monitoring, it is preferred that both rater Severity/Leniency and Centrality/Extremity are detected with high precision. As a result, the vector of rater parameters needs to be accurately recovered, leading to a scenario analogous to multidimensional computerized adaptive testing (MCAT; for example, Mulder & van der Linden, 2009; Segall, 1996, 2001; Wang & Chang, 2011; Wang, Chang, & Boughton, 2011). In what follows, two essay selection algorithms based on the Fisher information matrix will be introduced. Note that in the past, researchers who used RPCM in adaptive testing context focused on recovering the unidimensional θ assuming known item parameters (e.g., Dodd, De Ayala, & Koch, 1995; Dodd, Koch, & De Ayala, 1993; Koch & Dodd, 1985, 1989; Ostini & Nering, 2010). Also item selection algorithms in MCAT are well established in the literature (e.g., Mulder & van der Linden, 2009; Segall, 1996, 2001; Wang & Chang, 2011; Wang et al., 2011). However, estimating multidimensional vectors from RPCM requires deriving the Fisher information matrix for the vector, which, to the authors' best knowledge, is not discussed in previous RPCM-based CAT studies.

Given the probability function in Equation 2, for an individual rater, the log-likelihood function of the rater parameter, $\delta = (\delta_1, \dots, \delta_r)$ (again, the subscript i is omitted), becomes

$$\log L = \sum_{j=1}^J \sum_{k=0}^r u_{jk} \log(p_{jk}), \quad (3)$$

where $u_{jk} = 1$ if the rater assigns score k to essay j , and $u_{jk} = 0$ otherwise. J denotes the total number of essays. The Fisher information for an essay j is thus defined as follows: $I_j(\delta) = -E[(\partial^2 / \partial \delta \partial \delta') \log L_j]$, which is a r -by- r symmetric matrix.

It can be derived that,¹ for the matrix of $(\partial^2 / \partial \delta \partial \delta') \log p_{jk}$, the t th diagonal element is $-\sum_{k=0}^{t-1} p_{jk} (\sum_{k=t}^r p_{jk})$, and the (t, t') th off-diagonal element is $-\sum_{k=0}^{t-1} p_{jk} (\sum_{k=t'}^r p_{jk})$ when $t < t'$. The off-diagonal element with $t > t'$ will take the same form due to the symmetry of the matrix. Therefore, the Fisher information matrix based on J items has the t th diagonal element of

$$I^{tt}(\delta) = \sum_{j=1}^J \left\{ \sum_{k=0}^r p_{jk} \left[\sum_{k=0}^{t-1} p_{jk} \left(\sum_{k=t}^r p_{jk} \right) \right] \right\} = \sum_{j=1}^J \left\{ \sum_{k=0}^{t-1} p_{jk} \left(\sum_{k=t}^r p_{jk} \right) \right\}. \quad (4)$$

Note that the second equality holds in Equation 4 because the term $\left[\sum_{k=0}^{t-1} p_{jk} (\sum_{k=t}^r p_{jk}) \right]$ can be taken out of the summation of $\sum_{k=0}^r p_{jk}$ as it is a constant that does not depend on different values of k . Also $\sum_{k=0}^r p_{jk} = 1$.

The (t, t') th off-diagonal element is (when $t < t'$)

$$I^{tt'}(\delta) = \sum_{j=1}^J \left\{ \sum_{k=0}^{t-1} p_{jk} \left(\sum_{k=t'}^r p_{jk} \right) \right\}.$$

The determinant of the essay-level Fisher information matrix can be summarized in a succinct form as $\det(I^{tt'}(\delta)) = \prod_{k=0}^r p_{jk}(\delta)$.

Figure 1 illustrates the rater characteristic curves along with the information curves for two raters with real rater parameters from our real data section below. Rater A (i.e., the first row)

Figure 1. Rater characteristic curves and information curves.

has the most concentrated rater parameters in the sample (i.e., $\delta_1 = -.25$, $\delta_2 = 1.69$, and $\delta_3 = 2.23$), and Rater B has the most dispersed rater parameters in the sample (i.e., $\delta_1 = -5.74$, $\delta_2 = .10$, and $\delta_3 = 3.13$). In Figure 1, the left column shows the probability of a rater giving a certain score to essays with θ varying from -7 to 7 . As shown in the figure, the rater with more dispersed parameters has the characteristic curves more spread out along θ continuum, such that he or she tends to give middle scores to the majority of the essays, indicating a possible centrality effect. However, the rater with more concentrated parameters has the characteristic curves closer together, and he or she tends to give extreme scores to essays with relatively extreme θ values, implying a sign of extremity effect.

To visualize the Fisher information for each rater parameter, the second column of Figure 1 provides the Fisher information as a function of essay parameter (i.e., θ) for two selected raters. Again, each score is on a 0 to 4 scale, such that each rater has four parameters to estimate. In the figure, each right panel shows the Fisher information curve (see the dashed curves) for every rater parameter separately with the magnitude of information marked on the left bar. The determinant of the Fisher information matrix for different essays is also provided (see the solid curve) with the value marked on the right bar. Due to the unique feature of the item Fisher information function in Equation 4, the four information curves for δ_1 to δ_4 have exactly the same shape but different locations. Also, the t th ($t = 1, \dots, r$) information curve reaches its maximum at the θ value when $\sum_{k=0}^{t-1} p_{jk} = 0.5$. Moreover, comparing the two raters, it is clear that the determinant curve for Rater A is more concentrated with higher peaked value. In other words, when the four rater parameters (i.e., $\delta_1 \sim \delta_4$) are closer together, implying a potential extremity rater effect, the resulting determinant is generally high. Also the essay parameter values at which every individual Fisher information curve is maximized are close to each other, such that an appropriate essay will provide large amount of information for all rater parameters. In contrast, when the

four rater parameters are farther apart, then the resulting determinant is usually small (see the lower right panel). In this case, depending upon the essay parameter, a single essay might provide good amount of information for only one or two adjacent rater parameters.

D-optimal criterion. As a direct application of Segall's (1996) D-optimal criterion, this essay selection method intends to select the next candidate essay by maximizing the determinant of the Fisher test information matrix evaluated at a provisional estimate of $\hat{\delta}$. To elaborate, denote $S_J = \{j_1, j_2, \dots, j_J\}$ as the set of J essays that have been administered, and denote R_{-J} as the remaining essays in the pool. Then, the $(J + 1)$ th essay is selected, such that the following determinant is maximized:

$$\det\left(I_{S_J}(\hat{\delta}) + I_{J+1}(\hat{\delta})\right), J+1 \in R_{-J}.$$

Maximum Fisher information on a single parameter criterion. For an individual rater, every essay provides a different amount of information for each rater parameter, and the information accumulated for distinctive rater parameters also differs. The primary idea of this proposed essay selection method is to find the next essay, such that the Fisher information is maximized for the rater parameter on which the current standard error is highest. In other words, this method comes in two steps. In the first step, all information accrued for each rater parameter (i.e., δ_m , $m = 1, \dots, r$) needed to be added up, and a parameter that has the highest standard error, say, δ_t , $1 \leq t \leq r$, has to be identified. This decision could be made by taking the inverse of the Fisher test information computed from the previously administered essays, which is essentially a variance–covariance matrix. Then, find δ_t that corresponds to the highest diagonal element in this covariance matrix. In the second step, select the next essay that provides the highest information for δ_t , which is equivalent to maximize the equation of $\sum_{k=0}^{t-1} p_{jk}(\hat{\delta}) \left(\sum_{k=t}^r p_{jk}(\hat{\delta}) \right)$. This idea is analogous to the proposal of “maximizing information in direction with minimum information” mentioned in Reckase (2009). It will be referred to as the “Single-F” method for short hereafter.

Interim Rater Parameter Estimation

Using the RPCM in adaptive rater monitoring, for each individual rater, a r -by-1 vector of $\hat{\delta}$ needs to be updated sequentially. Usually, the maximum likelihood estimator (MLE), obtained by maximizing the log likelihood in Equation 3, is preferred because it is a consistent, unbiased, and efficient estimator. However, MLE requires responses to fall into at least two different categories before a finite estimate can be made. The search of MLE becomes much more difficult in a high-dimensional space (Wang, 2015). In contrast, Bayesian estimator always provides a finite solution even with a few item responses, due to the imposition of a prior. Specifying informed prior distribution of δ can help reduce posterior variance in its final estimate. One widely recognized disadvantage of Bayesian estimator, however, is the tendency of estimates to regress toward the prior mean, resulting in large bias in the extreme ends of the latent trait scale. Acknowledging this limitation of Bayesian estimators does not eliminate their potential benefit, and thus in this study, the maximum a posteriori (MAP) is used with a noninformative prior to obtain interim rater parameter estimates. It is also suggested that there should be at least $K + 1$ essay ratings (similar to item responses) available, preferably with a good mixing of responses in different categories, to enable interim estimation. Lack of enough data, such as very few items or lack of enough variability in response categories, will lead to inaccurate estimates.

Stopping Rule

Two types of CAT are usually implemented in practice: the fixed-length CAT and the variable-length CAT. With a fixed-length CAT, a test stops when the number of items administered reaches a prespecified test length (Chang, 2004, 2015). For adaptive rater monitoring, this would translate to each rater scoring a fixed number of essay responses. In the case of variable-length CAT, the test stops when either a certain measurement precision is satisfied, or when there is enough evidence to make a confident classification decision. Translating this to adaptive rater monitoring, the stopping rules for a variable-length test might include ceasing administration of additional essay responses when (a) the vector of the rater's parameters, δ_i , is measured with sufficient precision (e.g., Wang, Chang, & Boughton, 2013); (b) there is sufficient information to assign a rater to one of the groups along the two rater effect dimensions; or (c) the decision of assignment stays the same even after administering a new essay response. In this study, the authors only focus on the fixed-length test.

Simulation Study I

Design

A simulation study is conducted to evaluate the performance of three essay selection algorithms, the two proposed methods, and the benchmark random selection. The essay bank size is fixed at 600, and essay parameter θ_j is generated from a normal distribution $N(0, 4)$ (Dodd & Koch, 1987; Koch & Dodd, 1989). Each essay is rated in one of the five categories from 0 to 4. The rater sample size is 1,000. For each rater, the four parameters (i.e., δ) are simulated from $N(0, 4)$ separately and ranked from smallest to largest. In particular, across 1,000 raters, the resulting four rater parameters have means of -2.04 , -0.54 , 0.61 , and 2.02 , and standard deviations of 1.93 , 1.44 , 1.51 , and 1.99 , respectively, which resemble the data generation scheme in Master (1982). The number of essays assigned to each rater varies at four levels: 20, 40, 60, and 80.

Similar to a classical CAT setting, the essay parameters (i.e., θ_j 's) are treated as known,² and the goal is to estimate rater parameters for each individual rater, δ . During the adaptive rater monitoring, δ needs to be updated sequentially, and the candidate essay is selected based on $\hat{\delta}$. An “nlm” function in the “stats” package in R (R Core Team, 2015) is used to find the interim MAP for δ . The specific prior is $N(0, 9)$ for each of the four δ s. The authors consider this a noninformative prior because the mean and standard deviation of δ s are typically around 0 and 2, respectively. In addition, one unique feature needs to be emphasized, that is, an order constraint is imposed to ensure that the elements in δ are strictly ordered from smallest to largest. This is done via pairwise comparing the values of two adjacent elements in $\hat{\delta}$, and if any order constraint is violated, the resulting likelihood value based on this point estimate of $\hat{\delta}$ will be multiplied by a very small value (i.e., 10^{-20}). This way, any $\hat{\delta}$ violating the order constraints will have much lower likelihood than the other point estimates, and thus it will be dropped. The starting values of δ estimates are selected to be the 20th, 40th, 60th, and 80th percentiles of $N(0, 4)$, which are -1.683 , -0.507 , 0.507 , and 1.683 , respectively.

The average bias and mean squared error (MSE) are used to evaluate the recovery of each rater parameter separately. Taking δ_1 as an example, the average bias and MSE are computed using the following two equations, respectively, where N denotes sample size.

$$\text{Bias}(\delta_1) = \frac{\sum_{i=1}^N (\hat{\delta}_{i1} - \delta_{i1})}{N}, \quad \text{MSE}(\delta_1) = \frac{\sum_{i=1}^N (\hat{\delta}_{i1} - \delta_{i1})^2}{N}.$$

Figure 2. Bias of rater parameters for simulated essay bank.

Note. SF = maximum Fisher information for a single parameter criterion; DO = D-optimal criterion.

In addition, considering that different δ s may show diverged recovery patterns from different item selection methods, it is beneficial to summarize the recovery of the entire δ -vector via a general distance measure, such that the comparison among different methods is easier. A general distance measure is defined as follows:

$$\text{Average distance} = \frac{1}{N} \sum_{i=1}^N \sqrt{(\hat{\delta}_i - \delta_i) \hat{\Sigma}^{-1} (\hat{\delta}_i - \delta_i)'},$$

where $\hat{\Sigma}$ is an estimated 4-by-4 covariance matrix computed from the sample estimated values.

The essay bank usage is also explored, and the chi-square statistic (e.g., Chang & Ying, 1999) is computed to evaluate if the essay exposure rate is balanced. Larger chi-square value implies a more skewed essay exposure distribution.

Results

Estimation precision. Figures 2 and 3 present the average bias and MSE of each rater parameter, from the three essay selection methods, as a function of the number of essays. Focusing on the bias plot, all bias values are in the range of -0.05 to 0.05 , and the number of essays does not seem to have any visible impact on bias outcomes. There is no appreciable difference among three item selection methods either. For all methods, bias for the two boundary parameters, δ_1 and δ_4 , are relatively larger in absolute values than bias for the two middle parameters. This is considered reasonable because the information provided by each essay question for the boundary parameters is relatively low. This conjecture is reconfirmed by the MSE plot.

Figure 3. MSE of rater parameters for simulated essay bank.

Note. MSE = mean squared error; SF = maximum Fisher information for a single parameter criterion; DO = D-optimal criterion.

Different from bias, MSEs for all parameters decrease when test length increases. Both D-optimal and Single-F methods outperform random selection by generating smaller MSEs for all parameters. More interestingly, D-optimal method yields the smallest MSE for two middle parameters, whereas Single-F method leads to the smallest MSE for two extreme parameters. Taken together, D-optimal and Single-F methods produce comparable overall parameter recovery, as reflected by their close values of average distance in Table 1.

Essay bank usage. Also reported in Table 1 are the chi-square index values from three essay selection methods. Random selection leads to uniform essay selection, and thus the resulting chi-square values are close to 0 under different test length conditions. However, the two proposed essay selection methods both generate somewhat skewed essay exposure rate distributions, and D-optimal method yields slightly more balanced essay exposure among the two. Figures 4 and 5 show the specific essay exposure rates for different essays in the bank with essay difficulty plotted on the x -axis. It is interesting to note that for the Single-F method, essays with difficulty levels around -2 and 2 are most likely to be selected, whereas for the D-optimal method, essays with a difficulty level around 0 have the highest exposure rates. This observation is indeed consistent with the MSE plot shown in Figure 3, in that Single-F method generates the best δ_1 and δ_4 recovery, whereas D-optimal method generates the best δ_2 and δ_3 recovery.

Simulation Study II Using Real Data Parameters

Data Description

The data are obtained from a statewide writing assessment in the United States. Students were asked to write an essay in response to an explanatory prompt. Four hundred essays were scored by each of 131 professional raters on a 4-point rating scale. The ratings were then calibrated

Table 1. Average Distance and Chi-Square Index From Different Item Selection Methods (Simulated Essay Bank).

Essay selection method	Average distance				Chi-square index			
Test length	20	40	60	80	20	40	60	80
Single-F	1.14	0.91	0.80	0.73	9.92	14.90	17.32	17.47
D-optimal	1.20	0.97	0.81	0.71	9.76	13.42	13.14	11.09
Random	1.30	1.03	0.89	0.81	0.54	0.51	0.47	0.47

Figure 4. Essay exposure rate distribution for Single-F method.

using the RPCM in WINSTEPS (Lincare, 2012), and the essay and rater parameters were estimated. The results showed a good model-data fit at both essay and rater levels. The mean squared statistics (infit and outfit) in WINSTEPS were used to evaluate the model-data fit. In general, the value of 1.0 indicates a perfect fit, and values between 0.5 and 1.5 are considered to be productive (Lincare, 2002). In our data, at essay level, the infit statistics ranges from 0.52 to 2.13, and the outfit statistics ranges from 0.3 and 2.19. About 95% of essays have both fit

Figure 5. Essay exposure rate distribution for the D-optimal method.

statistics between 0.5 and 1.5. Similarly, at rater level, the infit statistics ranges from 0.52 to 1.97, and the outfit statistics ranges from 0.49 to 2.1. About 92% of raters have both fit statistics between 0.5 and 1.5.

The original essay bank contained 400 essays. However, there are only 240 essays with unique essay parameters.³ The mean and standard deviation of the essay parameters are -0.65 and 2.63 , respectively. Figure 6 (see the last panel) provides the histogram of θ with a superimposed normal curve. To get an essay bank with enough essays, a random number from $U(-0.1, 0.1)$ (serving as a small disturbance) is added to each unique essay parameter to form a “cloned” copy of essay bank. The original essay bank and the cloned copy are combined to obtain the ultimate essay bank with 480 essays. In the original real sample, there are 131 raters, each rater has three rater parameters. The means of δ_1 , δ_2 , and δ_3 are -3.04 , 0.13 , and 2.91 , respectively, and their standard deviations are 1.04 , 0.87 , and 0.83 , respectively. Figure 6 presents the histogram of these three rater parameters with normal curves overlay, and it appears that they all follow normal distributions closely. The original 131 raters were expanded to 1,048 in a similar fashion to that for expanding the essay bank. Basically, a random noise

Figure 6. Histogram of delta and theta distribution with superimposed normal curves.

(simulated from $U(-0.1, 0.1)$) is added to the true rater parameters to form jittered empirical values. Three essay selection methods are compared.

Results

Figures 7 and 8 present the average bias and MSE for the four rater parameters at varying test length levels from three essay selection methods. Compared with simulation results, the bias is slightly larger, and the values are in the range of -0.2 to 0.2 when test length is as short as 20. However, as more essays are given to the rater, the bias decays and the values approach 0 quickly. Consistent with the previous observation, the bias for extreme parameters (i.e., δ_1 and δ_3) is larger than that for middle parameters. As to MSE, the most notable finding from Figure 8 as compared with Figure 3 is that the same patterns persist for all test length conditions. In brief, both D-optimal and Single-F generate smaller MSEs for δ_1 and δ_3 as compared with random selection, whereas the difference between these two methods is almost negligible. As to the middle parameter δ_2 , the three methods are indistinguishable, implying that the advantage of the two new methods vanishes when δ_2 , one of the easiest parameters among the three, is considered. But overall, the two methods lead to a similar average distance, and they both outperform random selection, as shown in Table 2. In terms of essay exposure rate, both D-optimal and Single-F methods result in unbalanced essay usage as expected. The detailed exposure rate plot is omitted in the section to avoid redundancy because the same pattern exhibit as in Figures 4 and 5.

Discussion

In the reform of educational assessment, constructed-response items are becoming popular. Instead of selecting an answer from an array of possible options as in multiple-choice questions, this type of item requires students to produce an answer, and the answers are typically scored by human raters. To ensure that the scores are objective, fair, and accurate, one has to closely

Figure 7. Bias of rater parameters from real essay bank.

Note. SF = maximum Fisher information for a single parameter criterion; DO = D-optimal criterion.

monitor the rater performance and timely identify, and possibly eliminate or retrain, those questionable raters who might exhibit different types of undesirable rater effects. To achieve this goal, defining useful statistical indices to flag certain rater effects is pivotal. Several indices have been used in operational settings, such as percentage agreement or coefficient kappa. Even though they are sensitive to several rater effects, they do not differentiate one rater effect from another (Wolfe, 2014), so they cannot be used to diagnose whether a rater exhibits a specific rater effect. Latent trait models, however, serve as valuable alternatives.

Built upon the RPCM, the authors propose an adaptive mechanism for rater monitoring to potentially improve the efficiency of the current procedures. In this system, the most appropriate essay responses will be selected for each individual rater, so that the rater parameters could be estimated precisely with fewer essays. As a result, specific rater effects that a rater may exhibit could be diagnosed more quickly compared with having each rater score the same set of validity responses. It intends to reduce the number of validity responses each rater needs to score in rater monitoring process but not sacrifice precision of estimating rater parameters.

To our best knowledge, this article is the first attempt in the literature that applies the idea of computerized adaptive testing to rater monitoring. Note that in the context of rater monitoring, the purpose is to estimate rater parameters rather than locate examinee's true ability, so one needs to (a) estimate a large number of unknown parameters (i.e., rater location and thresholds) simultaneously for each individual rater during the interim scoring procedure and (b) select the validity essays that will provide the most information about, or minimize the measurement errors of, rater parameters. Technically, this boils down to a multidimensional CAT

Figure 8. MSE of rater parameters from real essay bank.

Note. MSE = mean squared error; SF = maximum Fisher information for a single parameter criterion; DO = D-optimal criterion.

problem, and the authors have built our adaptive rater monitoring procedure on the current knowledge of multidimensional CAT with necessary modifications.

Specifically, for adaptive essay selection, the D-optimal and Single-F methods are proposed, both of which outperform random selection, as evidenced by our simulation results. Regarding interim scoring, the authors propose to use MAP estimator with order constraints that is also proven to work successfully. An interesting phenomenon spotted from Simulation Study I is that the D-optimal method best recovers the middle parameters (δ_2 and δ_3), whereas the Single-F method best recovers the boundary parameters (δ_1 and δ_4). As a result, essay responses with medium θ s are more likely to be overly exposed in the D-optimal method, whereas essay responses with extreme θ s tend to be more often selected by the Single-F method. Hence, an exciting future direction would be to combine the two methods in such a way that the precision of all rater threshold parameters is recovered equally well and the essay exposure is further balanced.

Our study can also be generalized in different directions. First, the objective of detecting rater effects is approached by way of multidimensional CAT. In other words, once the rater parameter vector δ_i is recovered precisely, the rater effects can also be gauged from δ_i accurately. However, the objective can also be approached from a multidimensional classification CAT perspective with the intent of minimizing the classification error. This approach is preferable if the objective is to classify raters into multiple groups with varying levels of rater effects. For instance, along the severity/leniency continuum, raters can be assigned into one of the three groups (lenient, normal, and severe). Similarly, a rater can also be classified into one of the three groups along the centrality/extremity continuum (central, normal, and extreme). Combining

Table 2. Average Distance and Chi-Square Index From Different Item Selection Methods (Real Essay Bank).

Essay selection method	Average distance				Chi-square index			
	20	40	60	80	20	40	60	80
Test length	20	40	60	80	20	40	60	80
Single-F	1.12	0.93	0.82	0.74	14.79	24.62	30.04	37.11
D-optimal	1.11	0.93	0.81	0.75	12.33	19.46	24.41	27.87
Random	1.19	0.98	0.88	0.82	0.93	0.88	0.80	0.73

these two dimensions, a rater will be classified into one of the nine groups. To implement the idea of classification CAT in adaptive rater monitoring, one needs to solve the two challenges—the two-dimensional decision and more than two categories along each dimension.

The challenge with respect to classifying raters in two-dimensional space arises because practitioners must first choose a cutoff point function that separates a “mastery” region from a “nonmastery” region. Two commonly used cutoff point functions are as follows: (a) a linear combination of latent traits across dimensions must exceed a threshold, and (b) each latent trait must exceed a threshold irrespective of any other latent trait. It appears that in adaptive rater monitoring, it is reasonable to use the second function (i.e., noncompensatory function), such that different cutoff points are set along both centrality/extremity and severity/leniency continuums. Only very little psychometric research has generalized the unidimensional classification CAT methods to multidimensional scenario (e.g., Glas & Vos, 2010; Nydick, 2013; Spray, Parshall, & Huang, 1997), and all existing research only focus on classifying examinees into two categories along each dimension. With more than two categories as in the adaptive rater monitoring, complications arise because there is little simulation study showing how the essays should be selected when there are more than one cutoff point.

A second future direction is to consider variable-length version. In adaptive rater monitoring, one can stop assigning an additional essay response to a rater whenever the standard error of measurement (SEM) of δ_i is satisfactorily small. Given the multidimensional nature of δ_i , there are different ways to design the stopping criteria (Wang et al., 2013): (a) the SEM along each dimension of δ_i is below a cutoff, or (b) the generalized variance of δ_i (or asymptotically equivalently, the reciprocal of the determinant of the Fisher test information evaluated at $\hat{\delta}_i$) is below a cutoff. According to the specific feature of the Fisher information (see Figure 1), it is anticipated that raters with more concentrated threshold parameters should need fewer essays, whereas raters with more dispersed threshold parameters will need more essays to reach equal precision.

From the application perspective, the adaptive rater monitoring idea can also be extended. For instance, the authors only consider holistic ratings, whereas future research could consider analytic rubrics or multiple prompts. This type of rating is often seen in more complex scoring context, such as longer essays that measure multiple traits. In those cases, a different latent trait model should be used, and the accompanying essay selection algorithms need to be modified accordingly.

To sum up, in this article, an adaptive rater monitoring idea is proposed to improve the current rater monitoring procedure. The primary goal is to move away from the use of fixed sets of validity papers and to select validity responses that provide the most information about each individual rater. For example, if one believes that a rater’s scores exhibit centrality, then assign validity responses that have high and low scores. By engaging adaptive selection of validity responses, one can maximize the amount of information received from rater monitoring

activity. The desirable outcome is improved precision of rater parameter recovery and reduced number of validity essays needed.

Last but not least, with the booming of automated scoring, the authors want to emphasize the relationship between rater monitoring techniques and automated scoring. First, many applications of automated scoring technology rely on scores assigned by humans to train the scoring engine, so the issues of ensuring rater quality have broader implications than simply contexts in which human scores are interpreted. Second, scores from a well-trained, well-validated automated scoring engine can be used as additional decision points during rater monitoring. This is particularly useful in an operational scoring project when each rater scores each response (Wolfe, 2014). In this case, when a latent trait measurement model is used, the scores assigned to validity sets and operational responses can be considered jointly. Then, the value of a diagnostically specific index for an individual rater can be compared with the values of the indices obtained for automated scores. By engaging in this joint scaling, scoring leaders can increase the number of observations upon which they are making monitoring decisions for each rater, increasing the precision of the decision while decreasing the cost of using too many validity responses.

Appendix

Derivation of the r -by- r Matrix of $[(\partial^2 / \partial \delta \partial \delta') \log p_{jk}]$

Given the model in Equation 2, it can be first written as $\log p_{jk} = \sum_{m=0}^k (\theta_j - \delta_m) - \log \left(\sum_{k=0}^r \exp \left[\sum_{m=0}^k (\theta_j - \delta_m) \right] \right)$. The subscript “ i ” is omitted from p_{ijk} to avoid notational clutter.

1. Compute the first derivative as follows:

When $k = 0$,

$$\frac{d \log p_{jk}}{d \delta_t} = 0 - \frac{1}{\text{de}} \times (-1) \times \sum_{k=t}^r \exp \left[\sum_{m=0}^k (\theta_j - \delta_m) \right], \quad (\text{A1})$$

where “de” denotes the denominator in Equation 2, which is $\sum_{k=0}^r \exp \left[\sum_{m=0}^k (\theta_j - \delta_m) \right]$. Equation A1 is further simplified as

$$\frac{d \log p_{jk}}{d \delta_t} = \frac{1}{\text{de}} \times \sum_{k=t}^r \exp \left[\sum_{m=0}^k (\theta_j - \delta_m) \right] = 1 - \sum_{k=0}^{t-1} p_{jk}. \quad (\text{A2})$$

The second equality in Equation A2 holds simply because

$$\frac{1}{\text{de}} \times \sum_{k=0}^r \exp \left[\sum_{m=0}^k (\theta_j - \delta_m) \right] = 1 \quad \text{and} \quad \frac{1}{\text{de}} \times \exp \left[\sum_{m=0}^k (\theta_j - \delta_m) \right] = p_{jk}.$$

Similarly, when $k = 1, \dots, r$,

$$\frac{d \log p_{jk}}{d \delta_t} = -1 - \frac{1}{\text{de}} \times (-1) \times \sum_{k=t}^r \exp \left[\sum_{m=0}^k (\theta_j - \delta_m) \right] = - \sum_{k=0}^{t-1} p_{jk}.$$

Taken together, it can be written as

$$\frac{d \log p_{jk}}{d\delta_t} = \begin{cases} 1 - \sum_{k=0}^{t-1} p_{jk}, & k=0 \\ - \sum_{k=0}^{t-1} p_{jk}, & k=1, \dots, r \end{cases} \quad (\text{A3})$$

2. Compute the second derivative as below:

First consider the diagonal terms,

$$\frac{d^2 \log p_{jk}}{d\delta_t^2} = - \sum_{k=0}^{t-1} \frac{dp_{jk}}{d\delta_t}, \quad (\text{A4})$$

where

$$\frac{dp_{jk}}{d\delta_t} = \frac{0 - \exp \left[\sum_{m=0}^k (\theta_j - \delta_m) \right] \frac{d(\text{de})}{d\delta_t}}{(\text{de})^2} = -p_{jk} \times \frac{(-1) \times \sum_{k=t}^r \exp \left[\sum_{m=0}^k (\theta_j - \delta_m) \right]}{\text{de}}. \quad (\text{A5})$$

Note that Equation A5 holds because one only considers p_{jk} when $k=0, \dots, t-1$, which essentially means $k < t$. Equation A5 can be further simplified as $p_{jk} \left(\sum_{k=t}^r p_{jk} \right)$. Combining Equations A4 and A5, it can be written as

$$\frac{d^2 \log p_{jk}}{d\delta_t^2} = - \sum_{k=0}^{t-1} \frac{dp_{jk}}{d\delta_t} = - \sum_{k=0}^{t-1} p_{jk} \left(\sum_{k=t}^r p_{jk} \right). \quad (\text{A6})$$

Then, consider the off-diagonal terms. When $t < t'$, then, similar to the derivation of A6, it can be easily derived that

$$\frac{d^2 \log p_{jk}}{d\delta_t d\delta_{t'}} = - \sum_{k=0}^{t-1} \frac{dp_{jk}}{d\delta_{t'}} = - \sum_{k=0}^{t-1} p_{jk} \left(\sum_{k=t'}^r p_{jk} \right). \quad (\text{A7})$$

This is because the numerator term in p_{jk} , $\exp \left[\sum_{m=0}^k (\theta_j - \delta_m) \right]$, becomes 0 when taking its first derivative with respect to $\delta_{t'}$, simply due to $k < t < t'$. When $t > t'$, the derivation becomes slightly complex because, if when $k < t'$, the numerator in p_{jk} again becomes 0 when taking its first derivative with respect to $\delta_{t'}$ but it becomes 1 when $t' < k < t$. However, due to the symmetry of Hessian matrix, one can simply switch t' and t in the second derivative operation, such that

$$\frac{d^2 \log p_{jk}}{d\delta_t d\delta_{t'}} = - \sum_{k=0}^{t-1} \frac{dp_{jk}}{d\delta_{t'}} = - \sum_{k=0}^{t'-1} p_{jk} \left(\sum_{k=t}^r p_{jk} \right).$$

Taken together, it can be written as

$$\frac{d^2 \log p_{jk}}{d\delta_t d\delta_{t'}} = - \sum_{k=0}^{t-1} \frac{dp_{jk}}{d\delta_{t'}} = \begin{cases} - \sum_{k=0}^{t-1} p_{jk} \left(\sum_{k=t'}^r p_{jk} \right), & t < t' \\ - \sum_{k=0}^{t'-1} p_{jk} \left(\sum_{k=t}^r p_{jk} \right), & t > t' \end{cases}.$$

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research was partly funded by Pearson 2015 R&D Research Grant.

Notes

1. The detailed derivation is given in the appendix for interested readers.
2. The essay parameters can be obtained from prior calibration, using the ratings from expert raters. Usually, it is expensive to use a fully-crossed design, that is, each expert rater rates every essay question. Instead, each expert rater can rate a portion of essays, then the rater-by-essay responses are entered into WINSTEPS for calibration. WINSTEPS can handle systematic missing (in this case, due to the noncrossed design) well.
3. This is because in Rasch model, number-correct score is a sufficient statistics for the ability estimate. So persons with the same number-correct score would receive the same ability estimate. This leads to the fewer number of unique ability estimates.

References

- Chang, H.-H. (2004). Understanding computerized adaptive testing: From Robbins-Monro to Lord and beyond. In D. Kaplan (Ed.), *The SAGE handbook of quantitative methodology for the social sciences* (pp. 117-133). Thousand Oaks, CA: Sage.
- Chang, H.-H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika*, 80, 1-20.
- Chang, H.-H., & Ying, Z. (1999). A-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, 23, 211-222.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.
- Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70, 213-220.
- DeCarlo, L. T., Kim, Y., & Johnson, M. S. (2011). A hierarchical rater model for constructed responses, with a signal detection rater model. *Journal of Educational Measurement*, 48, 333-356.
- Dodd, B. G., De Ayala, R. J., & Koch, W. R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement*, 19, 5-22.
- Dodd, B. G., & Koch, W. R. (1987). Effects of variations in item step values on item and test information in the partial credit model. *Applied Psychological Measurement*, 11, 371-384.
- Dodd, B. G., Koch, W. R., & De Ayala, R. J. (1993). Computerized adaptive testing using the partial credit model: Effects of item pool characteristics and different stopping rules. *Educational and Psychological Measurement*, 53, 61-77.
- Glas, C. A. W., & Vos, H. J. (2010). Adaptive mastery testing using a multidimensional IRT model. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 409-431). New York, NY: Springer.
- Koch, W. R., & Dodd, B. G. (1985, April). *Computerized adaptive attitude measurement*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.
- Koch, W. R., & Dodd, B. G. (1989). An investigation of procedures for computerized adaptive testing using partial credit scoring. *Applied Measurement in Education*, 2, 335-357.
- Lincare, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, 16, 878.
- Lincare, J. M. (2012). *WINSTEPS*. Beaverton, OR: Winsteps.com.

- Lunz, M. E., & Stahl, J. A. (1993a). The effect of rater severity on person ability measure: A Rasch model analysis. *American Journal of Occupational Therapy*, 47, 311-317.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174.
- Mulder, J., & van der Linden, W. J. (2009). Multidimensional adaptive testing with optimal design criteria for item selection. *Psychometrika*, 74, 273-296.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *ETS Research Report Series*, 1992(1), i-30.
- 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.
- Nydic, S. W. (2013). *Multidimensional mastery testing with CAT* (Doctoral dissertation). University of Minnesota, Minneapolis.
- Ostini, R., & Nering, M. L. (2010). New perspectives and applications. In *Handbook of polytomous item response theory models* (pp. 3-20). New York, NY: Routledge, Taylor & Francis Group.
- R Core Team. (2015). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Available from <http://www.R-project.org/>
- Reckase, M. (2009). *Multidimensional item response theory*. New York, NY: Springer.
- Sébillé, V., Challa, T., & Mesbah, M. (2007). Sequential analysis of quality of life measurements with the mixed partial credit model. In J.-L. Auget, N. Balakrishnan, M. Mesbah, & G. Molenberghs (Eds.), *Advances in statistical methods for the health sciences* (pp. 109-125). Boston, MA: Birkhäuser.
- Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, 61, 331-354.
- Segall, D. O. (2001). General ability measurement: An application of multidimensional item response theory. *Psychometrika*, 66, 79-97.
- Spray, J. A., Parshall, C. G., & Huang, C. H. (1997, June). *Calibration of CAT items administered online for classification: Assumption of local independence*. Paper presented at annual meeting of the Psychometric Society, Gatlinburg, TN.
- Wang, C. (2015). On latent trait estimation in multidimensional compensatory item response models. *Psychometrika*, 80, 428-449.
- Wang, C., & Chang, H.-H. (2011). Item selection in multidimensional computerized adaptive testing—Gaining information from different angles. *Psychometrika*, 76, 363-384.
- Wang, C., Chang, H.-H., & Boughton, K. A. (2011). Kullback–Leibler information and its applications in multi-dimensional adaptive testing. *Psychometrika*, 76, 13-39.
- Wang, C., Chang, H.-H., & Boughton, K. A. (2013). Deriving stopping rules for multidimensional computerized adaptive testing. *Applied Psychological Measurement*, 37, 99-122.
- Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science*, 46, 35-51.
- Wolfe, E. W. (2005). Identifying rater effects in performance ratings. In S. Reddy (Ed.), *Performance appraisals: A critical view* (pp. 91-103). Hyderabad, India: ICFAI University Press.
- Wolfe, E. W. (2014). *Methods for monitoring rating quality: Current practices and suggested changes* (White paper). Iowa City, IA: Pearson.
- Wolfe, E. W., Jiao, H., & Song, T. (2015). A family of rater accuracy models. *Journal of Applied Measurement*, 16, 153-160.
- Wolfe, E. W., & Song, T. (2015). Comparison of models and indices for detecting rater centrality. *Journal of Applied Measurement*, 16, 228-241.