Methods for Monitoring Rating Quality:

Current Practices and Suggested Changes

White Paper

Edward W. Wolfe

May 2014

ALWAYS LEARNING PEARSON

About Pearson

Everything we do at Pearson grows out of a clear mission: to help people make progress in their lives through personalized and connected learning solutions that are accessible, affordable, and that achieve results, focusing on college-and-career readiness, digital learning, educator effectiveness, and research for innovation and efficacy. Through our experience and expertise, investment in pioneering technologies, and promotion of collaboration throughout the education landscape, we continue to set the standard for leadership in education. For more information about Pearson, visit http://www.pearson.com/.

About Pearson's Research Reports

Our network **mission** is to spark innovation and create, connect, and communicate research and development that drives more effective learning. Our **vision** is students and educators learning in new ways so they can progress faster in a digital world. Pearson's research papers share our experts' perspectives with educators, researchers, policy makers and other stakeholders. Pearson's research publications may be obtained at: http://researchnetwork.pearson.com/.

Abstract

This white paper discusses current rater monitoring practices and suggests several ways that rater monitoring can be improved. Within that discussion, I emphasize that the purpose of rater monitoring should be to identify rater effects. That is, scoring leaders should seek to differentiate specific patterns of ratings assigned by individual raters (e.g., severity/leniency, centrality/extremity, and accuracy/inaccuracy) so that they can provide raters with diagnostic and corrective feedback. I illustrate how several raw score and latent trait modeling indices can be used to accomplish this goal, and I further identify how many indices currently used for rater monitoring (e.g., percentage of agreement, coefficient kappa, intraclass correlation) fail to provide such information. Specifically, I demonstrate how application of a Rasch partial credit measurement model allows scoring leaders to identify rater severity/leniency (through the rater location parameter estimates), centrality/extremity (through the standard deviation of the rater threshold estimates), and accuracy/inaccuracy (through the correlation of a rater’s scores and the examinee ability estimates). I conclude by suggesting several changes in the manner in which scoring leaders monitor raters in operational scoring projects: (a) employing latent trait measurement modeling procedures that allow for diagnostically specific and automated flagging of raters, (b) selecting validity responses for individual raters adaptively in order to more precisely evaluate hypotheses that scoring leaders have about individual raters and rater effects that they may be exhibiting, (c) and expanding rater monitoring to jointly consider scores assigned to validity sets and operational scores as well as the possibility of utilizing automated scoring technologies.

Keywords: rater monitoring, rater effects

Methods for Monitoring Rating Quality: Current Practices and Suggested Changes

In educational settings, scores are assigned to student responses by human raters when assessment items cannot be machine scores through a formulaic process. These judgment-based processes may be employed when scoring essays, performance assessments, artistic or athletic performances, and demonstrations of processes to name a few of the more common assessment contexts. In addition, human raters are employed in a variety of settings outside of education (e.g., medical certification, artistic and athletic performances, employee evaluations, etc.), so many of the forthcoming discussion topics apply to those contexts as well, even though we restrict our conversation to examples in educational testing. Similarly, many applications of automated scoring technology rely on scores assigned by humans in order to train the scoring engine, so the issues we discuss have broader implications than simply contexts in which scores assigned by humans are interpreted.

A common concern about the scores assigned by human raters in these contexts is the degree to which those scores contain measurement error due to the fallibility of raters' judgments—they may have different interpretations of the content of the assessment response, they may be differentially influenced by the scoring context, or they may emphasize different features of the scoring criteria. Due to these concerns, the quality of scores assigned by humans is routinely monitored during the scoring process and is documented in technical manuals and research reports that employ those scores. This white paper provides a summary of existing quantitative methods for monitoring the quality of human ratings and suggests how these procedures might be improved. Specifically, I introduce a set of general terms related to the task of rater monitoring, identify several rater effects that should be of concern to those who monitor the quality of human ratings in operational scoring projects, describe how the quality of scores assigned by human raters is currently monitored, and identify alternatives that would improve this process. I also demonstrate how these procedures could be applied to an example data set.

Definitions

At a broad level, this white paper is concerned with the scores (ratings, marks) that human raters (judges, markers, evaluators, scorers, readers) assign to student responses (papers, essays, constructed responses, mathematics computations) to assessment prompts (items, performance assessments) that are sufficiently complex or unstructured (open-ended) to prevent formulaic scoring of those responses. Typically, the scoring criteria employed by raters is documented in a scoring guide, which includes a rubric containing written descriptions of varying levels of performance, coupled with annotated examples of student responses or performances at each of those levels. In educational settings, the number of score categories on the rating scale used to identify these levels of performance is typically relatively small in the U.S. (2 to 6 categories), although larger numbers of categories may be used in some settings, particularly outside of the States. The rubrics may require raters to make holistic judgments of student performance (i.e., a single judgment that takes into account multiple aspects of the response) or analytic judgments (i.e., multiple judgments that each results in a separate score depicting a single aspect of the response).

The assigned scores are the product of a judgmental process through which human raters determine which of the levels of performance depicted by the scoring rubric is most appropriate for describing the quality of a given student response. Rating quality is a general term that refers to the degree to which a set of scores is precise and unbiased. That is, higher levels of rating quality are defined by scores that contain little or no measurement error so that assigned scores are consistent with the scoring rubric and, therefore, the intentions of the assessment designers. Several procedural techniques are used to maintain acceptable rating quality in operational scoring projects—rater training, qualification, rater feedback, recalibration, and back reading—and I do not focus on those methods in this paper. Rather, I focus on quantitative methods that can be used to monitor and document the quality of scores during or subsequent to an operational scoring project. I also focus only with inadequacies in rating quality that can be attributed to the rater. Specifically, I focus on rater effects—patterns of scores that are associated with measurement error that is contributed by a rater (i.e., rater error).

Rating quality and rater effects can be depicted in either of two frames of reference. A rater accuracy frame of reference portrays rating quality in terms of the deviation of a set of scores from another set of scores that is assumed to be valid. These "gold standard" scores are typically ones that were assigned by expert raters or assessment designers through a consensus process, and those scores are sometimes referred to as true scores, implying that they contain no measurement error. Hence, rater accuracy refers to the degree to which a particular set of scores is consistent with a set of true scores. When rater accuracy is high, we are confident that the assigned scores are good indicators of the "true" performance suggested by the student response. In this sense, indicators of rater accuracy are indicators of the validity of the inference that assigned scores represent the trait being measured. In fact, it is common to refer to the agreement between operational scores and gold standard scores as "validity agreement."

A rater agreement frame of reference, on the other hand, portrays score quality in terms of the deviation of scores assigned by one rater from another set of scores that is not necessarily assumed to be perfectly accurate. For example, rater agreement may be evaluated by comparing the scores assigned by one rater to the scores assigned by another single rater, the scores assigned by a set of randomly chosen raters, the average of the scores assigned by an entire sample of raters, or the score assigned by an automated scoring engine. In some settings, it may be reasonable to assume that a high level of rater agreement is equivalent to a high level of rating accuracy, particularly when the pool of raters is well-trained and individual raters are paired with a relatively large number of randomly chosen raters so that one can assume that, on average, scores collapsed across raters are accurate and that scoring errors cancel each other. Additionally, because rater agreement concerns the replicability of the scoring process across multiple raters, it is considered to be an indicator of reliability, and it is common to refer to quantitative indicators of rater agreement as measures of inter-rater agreement (replicability of the exact scores) or reliability (replicability of the rank ordering of scores).

It is important to note that high levels of rater agreement do not necessarily equate to high levels of rater accuracy. As an example, in a context in which all raters assign scores that are biased toward the low score categories, we would expect levels of rater agreement to be high, but levels of rater accuracy to be low. That is, generally, the scores assigned by multiple raters would be consistent with one another, but they would not likely be consistent with true scores. On the other hand, we would typically expect high levels of rater agreement when levels of rater accuracy are high. That is, provided the scores assigned by all raters tend to agree with true scores, we would expect those scores to also agree with each other fairly well.

I close this listing of definitions by introducing one final distinction. Specifically, I want to differentiate two levels at which measures of rating quality are made. The most common level at which rating quality is measured is at the pool of raters. These measures of rating quality are employed as statistical summaries of the typical or average degree of rater accuracy or rater agreement, collapsing that information across all raters within a pool or across all raters within subgroups of the pool, such as rater gender.

Although measures of rating quality that reference the pool of raters are easy to interpret and are convenient for summaries like those that appear in technical manuals, they provide no useful diagnostic information about the performance of individual raters. Measures that are made at the level of individual raters, on the other hand, provide snapshots of the quality of the scores of a particular rater and are, therefore, useful for providing diagnostic feedback to raters during operational scoring. Hence, the difference between the two reference levels for depicting rating quality lies in their purposes—while measures made at the level of the rater pool are typically used to summarize the general quality of the assigned scores at the conclusion of a scoring project, measures that are made at the level of the rater are used to monitor and provide corrective feedback to raters during a scoring project.

Related to this notion of focusing on the performance of individual raters is the diagnostic utility of the index used to capture rating quality. Simply put, some rating quality indices provide relatively little actionable information about a particular rater while other rating quality indices provide diagnostically useful information. The percent of exact agreement between the scores assigned by a particular rater and the scores assigned by other raters is an example of an index that provides little diagnostically useful information regarding the quality of a particular rater's scores. That index simply tells us how frequently the rater assigns an identical score without telling us about trends in the remaining scores. As someone providing feedback to the rater, if we encounter a rater with a low percentage of agreement, we can only say "score better," without any specifics about how to accomplish that improvement. On the other hand, the average score assigned by a particular rater, especially when compared to the average score assigned by scoring leaders, is a rating quality index that does provide diagnostically useful information. Again, as someone providing feedback to the rater, if we are faced with a low average assigned score, we can inform the rater to be less stringent or less critical when evaluating student responses. In the remainder of this manuscript, our primary focus will be on diagnostically specific measures that are made at the level of the rater.

Common Rater Effects

In this section, I identify and describe some of the most common rater effects. My intent is not to present an exhaustive list of rater effects. Rather, I attempt to lay the foundation for differentiating measures of rating quality that have diagnostic utility from measures that have little to no diagnostic value. If a measure of rating quality has diagnostic value, it indicates a precise pattern in the ratings of an individual rater that deviates from an accurate rating pattern (i.e., the measure indicates a rater effect). A working assumption is that corrective information can be provided to the rater based on these diagnostic measures. On the other hand, if a measure of rating quality is constructed in a manner that precludes diagnosis of individual rater characteristics or is influenced by multiple rater effects, then that measure is said to have little diagnostic value—non-diagnostic indices, for short.

I use the term rater effect to refer to readily identifiable patterns of ratings that are indicative of trends of scoring errors. Although I present a dichotomy of labels for the extreme positions for each continuum, in reality, most of these rater effects can be measured in terms of degree. That is, rather than measuring the existence or non-existence of the rater effect, we would measure the strength of the effect in question. However, in most operational settings, scoring leaders would identify a cut score beyond which a rater would be flagged as exhibiting the rater effect and then take corrective actions based on that flag. Weaker levels of that rater effect would be ignored and treated as tolerable. Further, in my discussion, I group several previously defined rater effects based on similarity and identify alternative labels for some rater effects. Finally, I should also mention that I assume a rater accuracy frame of reference in my discussion (i.e., we depict rater effects as deviations from true scores), and we focus on rater effects as static characteristics (i.e., I do not describe changes in rater effects over time—differential rater functioning over time, or DRIFT) (Myford & Wolfe, 2009).

Leniency/Severity. One of the most commonly studied pairs of rater effects result in assigned scores that are higher (leniency) or lower (severity) than true scores. I refer to this as the leniency/severity continuum of rater effects. In the context of statistical analysis, leniency/severity would be evidenced by an increase/decrease in the average score associated with a rater. Although it is possible that leniency or severity may be perfectly systematic for a particular rater (i.e., all assigned scores are the same absolute distance from true scores), it is more likely that the leniency or severity effect constitutes a tendency on the part of the rater. Leniency/severity raises concerns for decision makers when scores are interpreted relative to a cut score, such as during college admissions or placement, determining graduation qualification, or awarding professional certifications. If leniency/severity exists in the scores, then some respondents will be incorrectly classified in decision making contexts such as these. Figure 1 displays three score patterns that illustrate leniency and centrality. Specifically, the first row of Figure 1 contains the true scores associated with each of 17 simulated examinees (columns). These true scores range from a low of zero to a high of six and are sorted from low scores on the left to high scores on the right. The second and third rows of the table contain the scores assigned by a simulated lenient and severe rater, respectively. Note that, on average, the lenient rater assigns scores that tend to be higher than the true scores while the severe rater assign scores that tend to be lower than the true scores.

True Score	0	1	1	2	2	2	3	3	3	3	3	4	4	4	5	5	6
Leniency	2	2	2	3	4	4	3	5	5	4	6	6	5	5	6	6	6
Severity	0	0	0	1	2	1	1	3	2	2	3	3	2	2	4	4	5

Figure 1. Simulated Rater Leniency and Severity.

Centrality/Extremity. Another common concern in operational scoring projects is whether the distribution of assigned scores is compressed (centrality) or pushed into the tails (extremity) when compared to true scores. I refer to this as the centrality/extremism continuum of rater effects. In the context of statistical analysis, centrality/extremity would be evidenced by a decrease/increase in the standard deviation of the scores associated with a rater. As was the case with leniency/severity, we expect the shrinkage or expansion of the rater’s distribution of scores to exhibit some randomness, rather being consistent across all responses. Of these two rater effects, in educational settings, centrality is more common. Extremity may be observed when raters are asked to make polytomous distinctions when raters are only comfortable making dichotomous distinctions due to a lack of variability in the responses being scored. Some researchers have used the term central tendency to refer to centrality, and some researchers differentiate centrality from restriction of range by using the former to refer to condensed distributions that are centered on the middle of the rating scale and using the latter to refer to centrality that is coupled with severity or leniency. When centrality/extremity exists in the scores, respondents 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. Figure 2 contains score patterns that illustrate simulated rater centrality and extremity. Again, the first row contains the true scores for the 17 simulated examinees, and the second and third rows contain the simulated scores assigned by a central and extreme rater, respectively. Note that the central rater tends to assign scores that trend toward the middle rating categories, while the extreme rater tends to assign scores that trend toward the tails of the score distribution.

True Score	0	1	1	2	2	2	3	3	3	3	3	4	4	4	5	5	6
Centrality	2	2	1	2	2	3	3	3	3	3	3	3	4	4	4	4	3
Extremity	0	0	0	1	0	4	2	2	3	3	4	5	5	5	5	6	6

Figure 2. Simulated Rater Centrality and Extremity.

Accuracy/Inaccuracy. A third common concern in operational scoring projects is whether raters assign scores that deviate from true scores in a random, unpredictable manner. This rater effect pattern is referred to as inaccuracy. The opposite trend, accuracy (a high level of consistency between a rater’s scores and true scores), is a desirable outcome, but we still refer to it as a rater effect due to the fact that it is the opposite of inaccuracy and it results in a predictable pattern of scores. In a statistical analysis, we would expect there to be a very low/high percentage of perfect agreement and a near-zero/near-one correlation between the scores of an inaccurate/accurate rater and true scores. The important distinction between inaccuracy and rater leniency/severity or centrality/extremity is the level of randomness in the data. While the previous two rater effects produce a rank ordering of responses that is similar to that based on true scores, inaccuracy introduces undue randomness into that rank ordering. As a result, decision makers might attribute observed score differences to individual differences when they are, indeed, random errors. Some researchers have differentiated the terms inaccuracy and rater inconsistency by defining the latter as contextual randomness (e.g., inaccuracy over time, across prompts, across subgroups of examinees, etc.). Figure 3 contains simulated score patterns that illustrate accuracy and inaccuracy. Again, the first row represents true scores while the second and third rows represent accurate and inaccurate scores, respectively. Note that the accurate scores are fairly consistent with the true scores while the inaccurate scores are inconsistent but do not tend to be consistently higher, lower, more extreme, or more central than the associated true scores.

True Score	0	1	1	2	2	2	3	3	3	3	3	4	4	4	5	5	6
Accuracy	0	1	0	2	2	3	3	3	3	3	4	4	4	5	4	5	5
Inaccuracy	1	5	3	3	3	0	3	3	2	3	4	6	2	3	4	4	6

Figure 3. Simulated Rater Accuracy and Inaccuracy.

Illusory Halo & Logical Error. The mention of contextual inaccuracies in the previous paragraph is important because, beyond these previous three rater effect continua, there are several rater effects that may manifest themselves only in certain contexts or as a function of the context. The first of these is commonly referred to as the halo effect, which may be observed when raters assign multiple scores to a single response, as may occur when scores are assigned according to an analytic scoring rubric. I prefer the term illusory halo, so that we can differentiate unintended similarities between these multiple scores from the natural covariance that occurs between related traits (true halo). Illusory halo may occur because raters allow some prominent features of a response (positive or negative) to cloud their judgment about other, less prominent, features which is sometimes referred to as a general impression halo. Similarly, raters commit a logical error when they fail to differentiate two distinct traits, believing that these two traits are conceptually similar or identical. In either case, the multiple scores assigned by raters are more consistent with each other than are the associated true scores. Statistically, this means that the between-trait Pearson Product Moment correlation for a rater will be closer to 1.00 for a rater exhibiting an illusory halo effect or a logical error than will be the correlation for true scores which indicates true halo. When either of these rater effects exists, decision makers will fail to differentiate the relative strengths and weaknesses of a particular examinee, believing instead that the examinee is generally strong or weak in all areas addressed by the scoring rubric. Figure 4 contains score patterns that illustrate a halo effect for a two-trait scoring rubric. In these simulated data, each column represents either a true score for one of two traits (columns two and three) or scores from a rater simulated to exhibit a halo effect (columns four and five). Each row of the figure represents the scores assigned to a particular rater on a scale ranging from zero to four. In these simulated scores, the correlation between Trait 1 and Trait 2 true scores equals .51, while the analogous correlation between halo scores equals .63. That is, the scores assigned by the rater simulated to exhibit a halo effect are more highly correlated than are the true scores.

Examinee	Trait 1 True Score	Trait 2 True Score	Trait 1 Halo Score	Trait 2 Halo Score
1	0	2	0	1
2	1	1	1	0
3	1	0	2	1
4	2	1	1	2
5	2	2	2	2
6	2	3	2	3
7	2	4	3	3
8	3	2	2	4
9	3	4	3	3
10	4	3	4	3

Figure 4. Simulated Rater Halo.

Contrast & Sequential Error. Another type of rating error arises due to the immediate context associated with a particular rater. That is, it may be that a rater, instead of assigning scores that are strictly reflective of the criteria espoused by the scoring rubric, assigns scores that compare the examinee to some context-based standard. For example, a rater commits contrast errors by comparing the examinee to himself or herself, and assigns scores based on how similar or different the rater perceives the examinee to be. Similarly, sequential errors (also known as order effects) occur when a rater compares an examinee to another examinee, rather than strictly adhering to the scoring rubric. Rather than between-examinee comparisons, primacy and recency effects may occur within an examinee. When primacy/recency occurs, a rater allows the first/last portion of the response to weigh more heavily in the evaluation decision. Statistical methods for detecting these rater effects vary with the type of effect. For example, while order effects could be detected by correlating subsequently scored pairs of responses, a contrast effect would require detailed information about the rater. In general, these types of rater effects are problematic, because the assigned scores do not reflect individual differences. Rather, the scores contain seemingly random error that is unique for each rater. Hence, each of these rater effects could easily be the cause of observed rater inaccuracy.

Subgroup/Feature Biases. If a rater is provided with or infers examinee characteristics, subgroup biases may become a rater effect in the assigned scores. Specifically, if a rater infers the examinee's gender based on handwriting appearance or based on personal information contained in the response, and that rater has beliefs or makes incorrect assumptions about gender differences on the trait being measured, then that rater may assign higher or lower scores to examinees who are assumed to belong to one group versus the other. Similarly, a rater may have incorrect beliefs about certain response features as they relate to the trait being measured, and those beliefs may also introduce error into the assigned scores. Detecting subgroup or feature biases is difficult because the relevant group membership or response content would need to be captured in some way and taken into account while creating rater effect indices. The problem with these types of biases is that the assigned scores reflect perceived group or response characteristics in addition to true individual differences, and those extraneous influences will be idiosyncratic for each rater.

Static versus Dynamic Effects

These rater effect continua have been presented as though they are static rater effects. That is, I have depicted them as being constant and stable characteristics of the rater. In order to be diagnosed, they must be stable across a useful period of time. Otherwise, no recognizable pattern would emerge in the assigned ratings. The notion of a perfectly static and independent rater effect is naïve for three reasons. First, it is likely that rater effects, when they exist, exert an influence on some, but not all, of the ratings that a rater assigns. Hence, it may be that a rater who is severe only manifests that severity on, say, half of the ratings that are assigned. Second, it is likely that a particular rater may exhibit a mixture of rater effects within a given time period. That is, a particular rater may produce ratings that contain some level of leniency but also has some amount of inaccuracy as well. Third, it is also likely that none of these rater effects exist in a completely static state. It is more likely that rater effects are dynamic over a given time span, and that the intensity of a particular effect may wax and wane with variations in the rating context (e.g., ambient noise), rater state of mind (e.g., fatigue), or rating task (e.g., trends in the true quality of consecutively rated responses).

This differentiation of static and dynamic rater effects is important for a couple of reasons. First, in order to be identifiable, the rater effect must be stable and strong enough over a usefully defined period of time in order to produce ratings that have an identifiable pattern in them. Clearly, if the pattern appears too infrequently or too weakly, it is likely not to be detectable and, therefore, will have minimal diagnostic utility for providing feedback to raters. Second, this differentiation points out the importance of identifying what constitutes a useful period of time for monitoring and providing feedback to raters. If the rater effect is not stable for a period of time within which it can be measured and then feedback can be given to the rater, then the ability to detect that effect is again of questionable utility. What can we do about an effect that we cannot detect in a timely manner or changes too rapidly for us to provide corrective feedback to the rater? Third, this differentiation also illustrates that it may be useful to break the rater monitoring process down into time segments and to monitor the stability and magnitude of identified rater effects across those time segments. It may be useful to know, for example, that raters are initially inaccurate, as a group, but that they become more accurate after a few hours of operational scoring. In this case, corrective action may include extending training or providing opportunities to raters to practice the scoring process prior to recording scores that are reported to students.

One of the examples above (i.e. trends in the quality of consecutively rated responses) is worth special mention because it reflects an entire class of rater effects that are not represented by the cases presented thus far. Specifically, there may be rater effects that manifest themselves as a function of the sequence of responses that a particular rater encounters. In operational settings, it is unlikely that individual raters will see the same sets of responses in the same sequence. However, it is possible that a rater’s judgments may be influenced by the relative quality or other characteristics of pockets of responses that the rater rates. For example, it may be that, if a rater encounters several student responses that are of a very low quality, that rater will interpret a higher quality response as being of higher quality than it truly is due to it being juxtaposed by the string of low quality responses. Similarly, there may be interactions between the content of a response and the manner in which individual raters interpret the quality of those responses. For example, a rater may be influenced by inferred gender of the student that is based on handwriting features. Detecting either of these types of rater effects, sequential effects and content effects both of which are special cases of dynamic effects, is beyond the intended scope of this paper. However, it is important to note that the rater effects presented in this section are only a sampling of commonly encountered effects and that the methods that are presented in the following sections could be extended to consider a much larger range of potential rater effect concerns.

Quality Assurance Practices

Many activities take place prior to the process of monitoring raters during an operational scoring project to minimize the opportunity for rater effects to manifest themselves in the scores that raters assign. To begin with, in most operational scoring projects, raters are carefully selected to meet predetermined minimum qualifications. Typically, these requirements include having attained a minimum level of education (e.g., a Bachelor's degree) and having studied a field related to the scoring task. Requirements may also include having prior scoring and/or teaching experiences. Once raters have been selected for the scoring project, they are trained. In high stakes assessment contexts, training activities may extend over several days. During this time, raters are introduced to the nature and purpose of the assessment and the nature of the item or prompt, and they undertake a detailed review of the scoring rubric and the process through which scoring will take place. An important part of training includes reviewing and discussing the scoring rubric; reading, scoring, and discussing pre-scored example responses; and practicing applying the scoring rubric while receiving feedback from scoring leaders. Typically, rater training concludes with a qualifying test, on which the rater must demonstrate mastery of the scoring rubric by achieving a predetermined level of agreement with consensus scores assigned by a panel of scoring leaders.

Following the initial rater training, recalibration training is typically employed in order to prevent or correct for idiosyncrasies that may develop in the interpretation or application of the scoring rubric by individual raters. In addition, several rater monitoring and feedback procedures are implemented so that scoring leaders can determine whether corrective actions should be taken regarding the scores assigned by a particular rater. For example, a process of back reading (i.e., having an expert rater review the scores assigned by a particular rater) may be employed. If inaccuracies are identified through back reading, then the rater's scores may be overwritten by those assigned by the scoring leader, and feedback may be provided to the rater concerning trends in the assigned scores. Similarly, scores assigned to validity responses (i.e., responses scored by scoring leaders ahead of time and then scored by raters, blindly, during operational scoring) may be examined for potential inaccuracies, and feedback and retraining may be applied to the scores assigned by suspect raters.

To be effective, that process (i.e., monitoring raters' operational or validity scores) requires useful statistical summaries of trends in the assigned scores. It would be nearly impossible and would, regardless, be extremely inefficient to review and discuss the score assigned by each rater to every scored response. Instead, scoring leaders rely on summaries of score distributions to determine whether a rater should be provided with feedback, retraining, score correction, or, in extreme cases, be released from the scoring project. There are several broad considerations relating to the potential usefulness and interpretability of rating quality indices.

First and foremost, to be diagnostically useful, the index must depict the performance of individual raters, rather than providing a summary of the typical performance of raters at the pool level. Second, we must consider whether the data collection design and/or statistical index depict rating quality in rater accuracy or rater agreement frameworks. Recall that a rater accuracy framework emerges from data collection designs in which the scores assigned by a rater are compared to true scores while a rater agreement framework compares the scores assigned by a rater to scores assigned by other raters. Clearly, I prefer a rater accuracy framework, but the cost and logistical challenges associated with collecting rater accuracy data may preclude doing so. Third, the index that we employ can be depicted as providing specific and actionable information that can be communicated to a rater. In a previous section, we referred to this characteristic of the rating quality index as its diagnostic utility. I adopt the stance that, for an index to be diagnostically useful, it must indicate that a trend or pattern exists in the scores assigned by a particular rater, and the index should not be sensitive to patterns of a different type so as to confound communicating corrective advice to the rater. That is, the index should, ideally, be sensitive to one and only one rater effect.

Measuring Rating Quality in Human Scores

This brings me to the primary emphasis of this white paper—current and future methods for measuring the quality of scores assigned by humans. In this section, I first provide an overview of several methods used for the purpose of quality assurance. I then identify several rater effects that have been documented in operational data. Next, I discuss several commonly employed raw score statistical indicators of rating quality. I conclude by explaining and illustrating how specific rater effects can be detected using latent trait modeling procedures.

Common Raw Score Rating Quality Indices

Scoring leaders commonly employ several statistical indicators of rating quality in order to monitor the assigned scores during an operational project. Unfortunately, as I will discuss in this section, few of those indices satisfy our definition of diagnostically useful. That is, most of the indices currently employed for rater monitoring do not allow scoring leaders to identify specific rater effects and, therefore, do not allow scoring leaders to provide diagnostic feedback to raters. Rather, many of these indices provide only general information about the quality of the assigned scores without differentiating the reason for any observed deviations from true scores.

Distribution Statistics. In some scoring projects, scoring leaders monitor raters during the project by reviewing statistical summaries of the scores assigned by each rater. By themselves, these distributional summaries are not very useful, so having a target distribution helps establish a frame of reference for interpreting those statistics. Specifically, scoring leaders might compare the distribution of scores produced by an individual rater to the distribution of target scores assigned by all raters (an agreement frame of reference). Alternatively, scoring leaders might compare a rater’s distribution of scores to validity scores assigned to those same responses by expert scorers (an accuracy frame of reference).

Scoring leaders commonly review three statistical summaries: frequency distributions, measures of centrality, and measures of dispersion. Frequency distributions indicate the percentage of scores that a rater assigns that fall into each category of the rating scale. These distributional summaries are useful, when compared to the distribution of target scores, because gross trends in the rater's scores may be apparent upon visual inspection. Especially if the distribution is summarized graphically, it will be apparent if the rater is assigning many more scores in the upper/lower score categories (leniency/severity) or the middle/tails of the rating scale (centrality/extremism). Other odd patterns may also be apparent, such as a tendency to over-utilize oddly or evenly numbered score categories.

The problem with reviewing percentages for the sake of monitoring the performance of individual raters is that determining the amount of deviation from the target distribution that is acceptable is a subjective judgment. In addition, reviewing the entire distribution for each rater may be time consuming. As a result, scoring leaders also may utilize descriptive statistics of a rater's assigned scores during rater monitoring. One common statistic is the mean or average score assigned. This measure of central tendency approximates the most typical score assigned by a rater and, thus, can be interpreted as an indicator of leniency/severity when its value is lower/higher than the analogous statistic for the target scores. In order to further remove subjectivity from interpreting a rater's mean score and as a possible means of automating the rater monitoring process, one can determine a level of tolerable deviation from the target mean by employing statistical significance testing and/or effect size indices. Other measures of central tendency exist (e.g., the median and mode), but these indices are used less frequently for the purpose of monitoring raters. The standard deviation of assigned scores is another index that scoring leaders use to monitor raters. This measure of dispersion indicates the amount of variability in a rater's scores, so values that are larger/smaller than the analogous index for target scores may indicate extremity/centrality. As is true for the mean, statistical significance testing and/or effect size indices may be useful for comparing the observed value for a particular rater to the analogous value for the target scores.

Agreement Percentages. Several indices employed by scoring leaders focus on the percentage of times that the scores assigned by a rater match or are in proximity to either the scores of another rater (interrater agreement) or true scores (validity agreement). Most commonly, these percentages are calculated by collapsing observations across rating scale categories, so that we know the percentage of assigned scores that were in either exact agreement (commonly referred to as the percentage of perfect agreement) and/or within one score point of the target score (often referred to as "exact plus adjacent agreement" or, more simply, "adjacent agreement"). To compute a percentage of agreement, one simply adds the number of times the assigned score and target score are within the specified distance from one another and divides by the total number of assigned scores. Infrequently, due to the difficulty of obtaining sufficient sample sizes, these percentages are reported at each score point.

The problem with percentages of agreement (as well as Coefficient Kappa, which we discuss next) is that they do not provide diagnostically useful information concerning the cause of lower-than-desired values. That is, the percentage of perfect and/or adjacent agreement cannot be interpreted as evidence of specific rater effects. For example, if a rater's percent of perfect agreement is low, the rater may have assigned scores that are, on average, one score lower than the target scores (leniency). However, that rater's scores could contain a large amount of random error (inaccuracy). Similarly, the deviations for the rater could be restricted to the tails of the distribution due to a manifested centrality effect. The problem is that any of these rater effects could be the cause, so the most useful feedback that we can provide to the rater is simply to "score better."

Another problem with percentages of agreement is that the values may be misleadingly high due to chance agreement. For illustration, consider two raters, each of whom may classify each examinee into one of three rating scale categories, but each rater simply categorizes each examinee by rolling a die—assign the lowest category for outcomes of 1 or 2, the middle category for outcomes of 3 or 4, and the upper category for outcomes of 5 or 6. The probability of being assigned to each category for each examinee equals .33 (2/6) for each rater, and the probability that both raters will assign the same category equals .33 [11% ($.33^2 = .11$) of the examinees will be assigned to a particular category by both raters, and there are three categories (.11 × 3 = .33)]. Hence, through random assignment to three categories, two raters can achieve a 33% perfect agreement rate.

To compensate for this problem, Cohen (1960) introduced the kappa coefficient (κ). This index indicates the proportion of perfect agreement attained by a particular rater beyond that attainable by chance alone. κ can, potentially, take on values that range from -1.00 to 1.00. When $\kappa = 0$, the observed agreement is equal to that expected by chance, and when $\kappa = 1$, the observed agreement is perfect. Negative values of κ indicate that the observed agreement is less than that expected by chance. In the previous example, the probability of perfect agreement by chance equals .33, so the proportion of agreement that is not attainable by chance equals .67. A coefficient kappa value of .50 for this example means that the raters achieved half of that .67 in addition to the .33 that is attributed to chance. That is, when there are three rating categories, a κ value of .50 means that the raters achieved a 67% agreement rate [.33 + (.67/2)].

Several authors have suggested benchmarks for interpreting κ (Altman, 1991; Fleiss, 1981; Landis & Koch, 1977), and those guidelines generally agree that values of κ above .60 indicate good agreement; however, there are a few problems with universal application of these benchmark values for κ. First, for a given proportion of exact agreement, the maximum possible value of κ varies as a function of the prevalence of the trait in question (i.e., marginal frequencies of the trait)—the more uniform the marginal distribution, the larger the maximum value of κ. Second, for a given proportion of exact agreement, the maximum of κ also varies as a function of the bias between the raters (i.e., tendency to disagree in one direction versus the other)—the more asymmetrical the disagreement, the larger the maximum value of κ. Third, the maximum possible value is influenced by the number of categories—a greater number of categories, given a particular level of perfect agreement, the larger the value of κ. Benchmark values do not take into account these changes.

In its original form, κ does not take into account the degree of disagreement for rating scales that contain more than two categories. For rater monitoring purposes, it would be useful to differentiate raters who assign a higher proportion of adjacent ratings from raters who assign a higher proportion of non-adjacent ratings. Hence, a scoring leader might want to see higher values of κ for raters who assign a high proportion of adjacent ratings and lower values of κ for raters who assign a high proportion of non-adjacent ratings. To address this need, Cohen (1968) introduced a weighted version of κ that applies weights to the computation of κ so that disagreements that are adjacent, for example, will decrease the computed value of the weighted κ less than will disagreements that are non-adjacent. Several weighting strategies have been suggested, and the quadratically-weighted approach is more common in educational testing applications. Regardless, as indicated previously, there is little diagnostically useful information provided by κ. If a lower value of κ is observed, then it is difficult to attribute that fact to the existence of a particular rater effect, so κ is probably most useful for the purpose of summarizing the overall quality of a pool of raters, rather than monitoring the performance of individuals.

Correlation Coefficients. Another set of indices used by scoring leaders to monitor the quality of raters, correlations, measure the strength of the relationship between the rater’s scores and the target scores. I differentiate two general classes of correlation coefficients—those based on rank ordering and those based on absolute differences.
The Pearson Product Moment correlation and the Spearman Rank correlation are two correlation coefficients based on rank ordering that are used to evaluate the quality of scores assigned by individual raters. In the case of the Pearson correlation, scores are

assumed to be continuous, and the index assesses the linearity of the relationship between the rater's scores and the target scores. On the other hand, the Spearman correlation assumes scores are ranks, and the index assesses whether the two sets of scores are monotonically related. In practice, the Pearson correlation is utilized more frequently.

Values of both correlations range from -1.00 to 1.00, with greater absolute values indicating a stronger relationship and values near 0.00 indicating no relationship. In scoring applications, it is unlikely that significant negative correlations will occur, unless a rater applies the scoring rubric in the opposite direction as intended. Both correlations may be useful to scoring leaders in two ways. First, correlations that measure the consistency of rank ordering (e.g., the Pearson Product Moment correlation) may allow scoring leaders to detect inaccuracy and/or centrality. Clearly, any rater effect that alters the rank ordering of examinees or adds random error to the scores (e.g., inaccuracy) will drive the correlation coefficient closer to zero. Similarly, rater effects that decrease the variability of the ratings (e.g., centrality) will also drive the correlation closer to zero. Second, in projects that employ a single rater to assign multiple scores to a particular examinee using an analytic rubric, the correlation between the scores assigned to the multiple traits can be used to determine whether a rater exhibits evidence of an illusory halo effect. When a rater's between-trait correlation is significantly higher than the value of the between-trait correlation for target scores (true halo), one would flag that rater for evidence of a halo effect. In order to compare one correlation value to another, one can utilize a Fisher transformation of the correlations to z-score equivalents and then conduct a statistical significance test on those transformed correlations.

The intraclass correlation coefficient (ICC) is an example of an index that is sensitive to absolute differences. To be specific, the ICC is based on an Analysis of Variance (ANOVA) framework, and it is formulated as the ratio of the between-unit variance and the sum of the between-unit and within-unit variance (i.e., the total variance). Shrout and Fleiss (1979) identified three common data collection designs for using the ICC: (a) each examinee is scored by a randomly-selected rater (or sample of raters), (b) each examinee is scored by the same group of raters and those raters are a sample from a population, and (c) each rater is scored by the same group of raters and those raters are the only raters of interest. In applied settings, we would expect to see the first design in operational scoring when only one or two raters are selected to score each examinee. Alternatively, we might encounter the second or third design in operational scoring when raters respond to validity sets.

The ICC, and the related generalizability coefficient, range in value from 0.00 to 1.00, and both indices jointly measures the association and the agreement between two sets of scores. Because of this, those indices provide little useful information to scoring leaders concerning the quality of the scores assigned by individual raters. Rather, they are more appropriate for use as indicators of the overall quality of a pool of raters.

Summary. Several of the commonly used indices employed in rater monitoring are actually poor choices for that purpose because those indices do not provide useful diagnostic information concerning the occurrence of specific rater effects in the scores assigned by a particular rater—a characteristic that is essential if corrective feedback is to be provided to the rater. Specifically, the percentage of agreement, coefficient kappa, and the intraclass correlation coefficient indicate only the general quality of the assigned scores and are, therefore, better suited for documentation of score quality at the conclusion of scoring or as a general indicator of scoring quality at the level of the rater pool during scoring. On the other hand, score category frequency distributions may be an inefficient means for monitoring the quality of scores assigned by individual raters because there are no established guidelines for objectively identifying rating patterns based on those distributions, and the decision making process would take considerably more time on the part of scoring leaders during operational scoring than would be the case for the following indices.

Four statistics seem to be quite suitable for monitoring raters and providing feedback to them: (a) the score mean, which indicates leniency/severity, (b) the score standard deviation, which indicates centrality/extremity, (c) the between-rater Pearson correlation, which indicates inaccuracy/accuracy and, possibly, centrality, and (d) the between-trait Pearson correlation, which indicates illusory halo. Table 1 identifies how the first three of these indices behave in the presence of severity/leniency, centrality/extremity, and accuracy/inaccuracy. These indices are best used when the target scores are true scores, such as when raters respond to validity sets during operational scoring. A potential shortcoming of that approach is that all raters respond to a fixed set of responses which is typically of a small number. Another potential use of these indices would be in operational scoring, but, because raters are typically assigned to responses at random and because only a small number of raters typically rate each response, a considerable sample size would be required so that the samples of responses scored by each rater can be considered to be randomly equivalent. Either of these shortcomings is non-ideal, and we describe an approach that remedies this situation in the following section.

Table 1:

Raw Score Rater Effect Indices

Index	Severity/Leniency	Centrality/Extremity	Accuracy/Inaccuracy
Raw Score Mean	− / +		
Raw Score SD	$-^a$ / $-^a$	− / +	
PPMC$_{Xr,X.}$		$-^a$ / N	+ / −

Note: PPMC$_{Xr,X.}$ = Pearson Product Moment Correlation between a rater's scores and the mean score across raters, + = increases, − = decreases, blank & N = No effect, a = slightly

Latent Trait Modeling Methods

The fact that using raw score descriptive statistics for rater monitoring requires raters to either score (a) fixed sets of validity responses or (b) large numbers of randomly equivalent operational responses is inefficient for three reasons. First, a fixed set of validity responses requires the additional cost of obtaining true scores, which are assigned by expert raters. These validity scores must then be input into the scoring system, and some method of comparing the scores assigned by operational raters to the true scores must be implemented. In computer-based systems, this is not difficult. However, doing so requires an additional component to the response distribution system because those validity scores will not be reported back to students, and the system must collect and summarize those scores separately from the operational scores that raters assign. Second, administration of validity responses means that raters are being paid to be monitored rather than to score operational responses. In order to monitor raters using validity responses, those responses must be administered to raters, typically by seeding them into the rater's queue blindly. Each one of these that a rater scores is both time away from operational scoring and money being spent for an activity other than assigning operational scores. Hence, it is in the scoring leader's best interest to utilize as few validity responses per operational response as possible. Third, by requiring a relatively large number of operational responses and/or minimizing the number of validity responses that are used for rater monitoring purposes, information about the performance of a particular rater may be collected too slowly to be useful in terms of providing feedback to the rater. For example, say a rater is administered a validity response every five minutes. That means, for each hour of scoring, scoring leaders have only 12 data points upon which to base evaluative decisions. That is very little information; particularly if the responses are selected randomly and do not adequately represent the full range of student performance covered by the scoring rubric.

Ideally, responses would be selected for each rater based on what we already know about that rater's performance. That is, responses would be selected adaptively so that scoring leaders can obtain more precise and more diagnostically informative feedback about that particular rater. For example, say a rater has rated the past ten responses using one of the two lowest rating scale categories. A scoring leader might begin to suspect that that rater is rating severely. To best evaluate that suspicion, it would be best to obtain scores from that rater on student responses that would allow that rater to further express the suspected severity. That is, scoring leaders would want to see how that rater would score responses with average or, better yet, high true scores. It would provide little useful information to obtain scores from that rater on responses that have been assigned low true scores. On the other hand, we would want to see how a rater who is suspected of centrality would score responses assigned true scores in the lowest and highest score categories. Similarly, we would want to see how a rater suspected of illusory halo would score responses that have discrepant true scores assigned to the multiple traits in the case of multiple analytic scoring rubrics. In short, it would be ideal to assign responses to a rater that provide the most useful information concerning any suspicions we have about rater effects that may be exhibited by that rater.

Unfortunately, current rater monitoring procedures are not designed to do this. Instead, we assign a fixed set of responses to a validity set and then assign all of those responses to each rater without thinking about what rater effects that rater may be evidencing in the assigned scores. One reason for this is that most rater monitoring procedures do not focus on diagnostically useful indicators (i.e., we monitor for rating quality, in general, rather than specific rater effects). Another reason for this is a desire to minimize added cost of rater monitoring (i.e., the cost of obtaining true scores and the time and cost of having raters score validity responses). Both of these issues can be remedied, to some degree, by implementing an adaptive rater monitoring system, and doing that would likely require the use of latent trait measurement models, also known as item response models. Latent trait models are routinely applied in operational testing programs, and significant work has been conducted to determine how rater effects manifest themselves in applications of these models. The purpose of the remainder of this chapter is to summarize the range of models that might be applicable in the context of rater monitoring, identify the relevant rater effect indices, and illustrate their usefulness using simulated data.

Dichotomous Models. The simplest latent trait model is the Rasch (1960) dichotomous model, named for the Danish mathematician who formulated it. That model posits a linear function to explain the log-odds of observing a particular outcome versus not observing that outcome. The most common outcome modeled is an examinee answering a dichotomously scored test item correctly (versus incorrectly), but, we will focus on scores that a rater assigns. So, for example, the Rasch dichotomous model can be applied to model the log-odds of a rater assigning a score to a student response that either matches or does not match an assigned true score—a rater accuracy model (DRAM), first suggested by Engelhard (1996). That model, shown in Equation 1, expresses the log-odds of a rater matching a true score as the difference of the difficulty of correctly scoring the student response (θ_n) and the accuracy of the rater (ρ_r), commonly referred to as location parameters because different values change the location of the inflection point of the resulting logistic curve on the underlying continuum.

$$LN\left(\frac{\pi_{nr}}{1-\pi_{nr}}\right) = \theta_n - \rho_r \tag{1}$$

The DRAM can also be formulated in terms of the probability of a rater matching the true score, as shown in Equation 2. This formulation is convenient because it shows that when $\theta_n = \rho_r$, the probability of the rater's score matching the true score equals 0.50 and that the probability decreases as θ_n becomes progressively greater than ρ_r.

$$\pi_{nr} = \frac{\exp\left(\theta_n - \rho_r\right)}{1-\exp\left(\theta_n - \rho_r\right)} \tag{2}$$

The DRAM can be extended by including a parameter that changes the slope of the logistic curve (α_r) , commonly referred to as a discrimination parameter, and we refer to that model as a dichotomous two-parameter rater accuracy model (D2RAM). The three-parameter model (D3RAM) adds to this a lower asymptote (χ_r), commonly referred to as a guessing parameter when that model is applied to multiple-choice test data. De Ayala (2009) provides a summary of these models. The four-parameter model (D4RAM) (Barton & Lord, 1981), shown in Equation 3, adds to that an upper asymptote (δ_r).

$$\pi_{nr} = \chi_r + \left(\delta_r - \chi_r\right)\left[\frac{\exp\left(\alpha_r\left(\theta_n - \rho_r\right)\right)}{1 - \exp\left(\alpha_r\left(\theta_n - \rho_r\right)\right)}\right] \tag{3}$$

Recall that all of these RAMs estimate parameters based on whether a rater's assigned scores match or do not match true scores. We refer to these conventional models as the dichotomous Rasch (DR), dichtomomous two-parameter (D2), dichotomous three-parameter (D3), and dichotomous four-parameter (D4) models. Parameters for each of these models can also be estimated based on the actual scores that a rater assigns, although we do not further discuss these models in this chapter.

Polytomous Models. Additional models can be formulated that allow for polytomous data. In our examples, we will model the raw score assigned by a rater, although polytomous versions of the RAMs could also be designated (e.g., exact agreement versus adjacent agreement versus non-adjacent agreement). We adopt this focus because of limited utility of the RAMs, which we discuss in the following section. In the Rasch rating scale model (Andrich, 1978), a separate parameter is added to the log-odds formulation of the dichotomous model to account for the relative difficulties of the *k* thresholds that separate adjacent rating categories (Equation 4).

$$LN\left(\frac{\pi_{nrk}}{\pi_{nrk-1}}\right) = \theta_n - \rho_r - \tau_k \tag{4}$$

That model imposes the same threshold structure on each rater, which is a strong assumption that suggests that all raters conceptualize or differentiate score categories in the same manner. Masters (1982) introduced the Rasch partial credit model, shown in Equation 5, which relaxes that assumption, allowing each item (in most applications), or, in our case, rater, to exhibit a unique threshold structure.

$$LN\left(\frac{\pi_{nrk}}{\pi_{nrk-1}}\right) = \theta_n - \rho_r - \tau_{rk} \tag{5}$$

Muraki (1992) generalized both of these polytomous Rasch models to allow for slope parameters. Although the slope is commonly applied to items, in our examples, the slope is applied to raters, as shown in the generalized partial credit model (Equation 6).

$$LN\left(\frac{\pi_{nrk}}{\pi_{nrk-1}}\right) = \alpha_r\left(\theta_n - \rho_r - \tau_{rk}\right) \tag{6}$$

It is important to note that the generalized partial credit model (GPCM) is the most general version of these polytomous latent trait models. By constraining the thresholds to be equal across raters, the model simplifies to become the generalized rating scale model (GRSM). By constraining the slopes of the GPCM to be equal, the model simplifies to become the Rasch partial credit model (RPCM). By further constraining the thresholds to be equal across raters, the model simplifies to become the Rasch rating scale model (RRSM).

Extensions. Additional latent trait models have been formulated to address other data collection designs. The multifaceted Rasch model implements any of the previous Rasch models, but it adds additional location parameters to account for multiple sources of systematic measurement error. For example, in scoring projects that employ analytic scoring rubrics, raters may assign scores using multiple (*c*) criteria, and the relative difficulties of those criteria may vary. In that case, a multifaceted RRSM would include location parameters for examinees, raters, and criteria, as shown in Equation 7. Multifaceted components can be added to any of the polytomous or dichotomous models previously identified.

$$LN\left(\frac{\pi_{nrck}}{\pi_{nrck-1}}\right) = \theta_n - \rho_r - \chi_c - \tau_k \tag{7}$$

Another extension of these models would allow for multiple abilities, or dimensions, to be indicated by examinee responses. In multidimensional models—the multidimensional random coefficients multinomial logit model (Adams, Wilson, & Wang, 1997), for example—multiple examinee ability parameters are estimated based on a mapping of the items onto the latent abilities. Such a model would require a multifaceted extension to include a rater location parameter as well for rater monitoring applications. A final extension of these models, to include hierarchical formulations, explicitly accounts for the correlated errors that are likely to arise from the fact that the multiple ratings of a single examinee response are likely not independent of one another. Previous research has demonstrated that ignoring these dependencies may adversely impact parameter estimates and their standard errors, so several models have been proposed to compensate for this shortcoming of the more common latent trait models (DeCarlo, Kim, & Johnson, 2011; Patz, Junker, Johnson, & Mariano, 2002; Shin & Choi, 2011; Wilson & Hoskens, 2001).

Latent Trait Rater Effect Indices

In this section, I identify several indices associated with five latent trait models (RAM, RRSM, RPCM, GRSM, and GPCM) that are useful for detecting rater effect patterns in assigned scores. Specifically, we focus on rater severity/leniency, centrality/extremity, and accuracy/inaccuracy.

Severity/Leniency. Severity/leniency is the most easily detected rater effect because, with the exception of the RAMs, the latent trait measurement models we have described contain a parameter that explicitly measures these rater effects. Specifically, if parameters are estimated for data that represent the score assigned (rather than the difference between the assigned score and a true score, as is the case with the RAM), then the ρ_r estimates (i.e., the rater location estimates) capture severity/leniency trends and are non-linear transformations of the mean scores assigned by raters. Typically, scaling polarizes ρ_r estimates so that negative values are indicative of rater leniency, and positive values are indicative of rater severity. Detection of severity/leniency is not directly possible with the RAMs. Rather, the ρ_r estimates from these models are general indicators of rating quality, and they do not differentiate the causes of inconsistencies between a rater’s assigned scores and true scores. In the RAMs, severity/leniency may be one of several possible reasons for observed inconsistencies. That is, severity/leniency is confounded with other rater effects in these models.

Centrality/Extremity. The centrality/extremism effects have not been studied as closely as the severity/leniency effects, although the former is of considerable concern to scoring leaders in operational projects in education. Often, the score distributions produced by raters are leptokurtic (i.e., heavily concentrated in the central scoring categories), so scoring leaders are justifiably concerned about whether that outcome is caused by rater centrality. Others have posited that raters may adopt a "play it safe" mentality when they are being monitored, opting to assign ratings in the middle rating categories when in doubt, thus avoiding scores that will increase the risk of being flagged by rater monitoring procedures (Myford & Wolfe, 2004). As was the case with severity/leniency, the ρ_r estimates from the RAMs confound centrality/extremism with other rater effects, so they are of little use for these diagnosing rater effects.

On the other hand, there are several options for measuring centrality/extremism effects in the RRSM, RPCM, GRSM, and GPCM. First, the Pearson product moment correlation between the scores assigned by a particular rater and the examinee ability estimates in these models ($\text{PPMC}_{r,\theta}$) serves as a weak potential indicator of centrality. This index is analogous to the raw score correlation between a rater's scores and the average of the scores assigned by all other raters, so it is clear that the restriction of range imposed by the rater centrality effect ought to decrease the magnitude of the $\text{PPMC}_{r,\theta}$. However, this effect tends to be only slight for all four of these models due to the fact that rater centrality does not substantially change the rank ordering of the examinees based on the rater's assigned scores.

For the GRSM and GPCM, the rater slope estimate (α_r) is another potential, again weak, indicator of centrality. Because rater centrality causes fewer distinctions to be made between examinees in the middle of the ability continuum, rater slope, also referred to as rater discrimination, tends to be lower for raters who exhibit this effect. Conversely, rater extremity results in increased values of rater slope estimates. However, the impact of rater centrality/extremity on the rater slope also depends on whether the model explicitly accounts for these rater effects through estimated parameters as well as the characteristics of the underlying score distribution. That is, if the model accounts for centrality in some other way, such as by modeling separate threshold structures for each rater as suggested in the following paragraph, then the rater slope parameter estimates will no longer be influenced by the existence of rater centrality/extremity. Hence, detecting rater centrality/extremity with a rater slope estimate is not straightforward. For example, when parameters are estimated for the GPCM and some raters exhibit centrality, the slope estimates for those raters may actually increase. Therefore, although rater slope may be influenced by rater centrality, we refer to it as a weak indicator of that rater effect.

The best option for detecting rater centrality is to include parameters in the model that directly capture the effect, which is the case for the RPCM and GPCM. These two models allow the rating category thresholds (τ_{rk}) to vary between raters. In effect, this means that the proportion of scores assigned to each category is modeled separately for each rater, which, clearly, captures differences between raters on the centrality/extremity continuum. Hence, in the RPCM and GPCM, rater centrality/extremity can be detected as decreases/increases in the standard deviation of raters' estimated category thresholds ($SD\tau_{rk}$).

Both the RRSM and GRSM fix the rating scale category thresholds to be equal between raters, so the model does not explicitly account for variations between raters with respect to score category use. Noting this, Wolfe and McVay (2012) explain that this model-data misfit produces patterns in the residuals associated with a particular rater that are produced under the RRSM and GRSM. Specifically, raters who engage in centrality assign scores that are higher than the model-based expectations for low-ability examinees and assign scores that are lower than expected for high-ability examinees. Hence, a scatterplot of model-data residuals as a function of model-based expectations should display a negative slope for raters engaging in centrality. The opposite should occur for raters engaging in extremity. Hence, under the RRSM and GRSM, the residual-expected Pearson product moment correlation ($\text{PPMC}_{\text{res,exp}}$) will be negative when raters exhibit centrality, positive when raters exhibit extremity, and near zero when raters exhibit neither.

I should also note that previous research concerning the use of common model-data fit indices (i.e., those based on model-data residuals, such as the mean-square unweighted, MSU, and mean-squared weighted, MSW indices) has suggested that these indices may be useful for identifying raters who exhibit centrality (Engelhard, 1994). However, other researchers have demonstrated these indices to be sensitive to a variety of rater effects, making it difficult to use them in isolation (Myford & Wolfe, 2003, 2004; Wolfe, 2004, 2005). In short, mean-squared residual fit indices are no more useful than the RAMs for diagnosing rater effects. Hence, I do not recommend using either in operational settings.

Accuracy/Inaccuracy. As was the case for centrality/extremity, the accuracy/inaccuracy rater continuum has not been studied much in previous research. Fortunately, detection of rater inaccuracy is relatively straightforward. As was the case with severity/leniency and centrality/extremity, the RAMs are not particularly useful for diagnosing rater effects— the ρ_r estimates are influenced by all rater effects, so differentiation is impossible. Regardless of model, the values of $\text{PPMC}_{r,\theta}$ increase when raters are accurate and decrease when raters are inaccurate. In addition, the values of α_r also increase/decrease with rater accuracy/inaccuracy for the GRSM and GPCM. Other indices are either impacted only slightly or are impacted by multiple rater effects. For example, the values of $\text{PPMC}_{\text{res,exp}}$ are slightly elevated/decreased by rater(Linacre, 2012) accuracy/centrality. The values of $\text{SD}\tau_{rk}$ vary in response to accuracy/inaccuracy, depending on the model, and the values of MSU tend to increase with accuracy and decrease with accuracy, although, as noted previously, that index is sensitive to multiple rater effects.

Summary. Table 2 identifies five models and several indices that can be influenced by each of the rater effects previously discussed. Note that I present summaries for MSU and $\text{PPMC}_{\text{res,exp}}$ for the two Rasch models only, simply because these indices are easier to produce using commercially available software that estimates parameters for these models, such as Winsteps (Linacre, 2012). In short, I find the DRAM to be of little use in diagnosing rater effects because the value of ρ_r is influenced by all six of the rater effects we considered. Similarly, I recommend using the RPCM or GPCM, rather than the RRSM or GRSM, to diagnose rater effects because differentiation of centrality/extremity versus accuracy/inaccuracy is simpler using the former models. Specifically, under the RPCM and GPCM, centrality/extremity manifests itself as an increase/decrease in $SD\tau_{rk}$ while accuracy/inaccuracy manifests itself as increases/decreases in $PPMC_{r,\theta}$ or, in the case of the GPCM, as increases/decreases in the value of α_r. Finally, I question whether inclusion of a rater slope parameter justifies the increased complexity of the GPCM in lieu of the simpler RPCM, particularly given that (a) the values of $PPMC_{r,\theta}$ are also sensitive to accuracy/inaccuracy, and, as I shall show in the examples that follow, (b) α_r is not a particularly strong index, diagnostically, of accuracy/inaccuracy.

Table 2:

Latent Trait Model Rater Effect Indices

Model	Index	Severity/Leniency	Centrality/Extremity	Accuracy/Inaccuracy
DRAM	ρ_r	+ / +	$+^{a}$ / $+^{a}$	- / $+^{a}$
RRSM	ρ_r	+ / −		
	$PPMC_{r,\theta}$		$-^{a}$ / N	+ / −
	$PPMC_{res,exp}$		− / +	$+^{a}$ / $-^{a}$
	MSU		+ / +	- / +
RPCM	ρ_r	+ / −		
	$PPMC_{r,\theta}$		$-^{a}$ / N	+ / −
	$PPMC_{res,exp}$		$+^{a}$ / N	+ / −
	$SD\tau_{rk}$		+ / -	+ / N
	MSU		$-^{a}$ / $+^{a}$	- / +
GRSM	ρ_r	+ / −		
	$PPMC_{r,\theta}$		$-^{a}$ / $+^{a}$	+ / −
	α_r		$-^{a}$ / $+^{a}$	+ / −
GPCM	ρ_r	+ / −		
	$PPMC_{r,\theta}$		$-^{a}$/ N	+ / −
	$SD\tau_{rk}$		+ / −	$-^{a}$ / $+^{a}$
	α_r		$+^{a}$ / $-^{a}$	+ / −

Note: ρ_r = rater location, $PPMC_{r,\theta.}$ = Pearson Product Moment Correlation between a rater's scores and examinee ability, $PPMC_{res,exp}$ = Pearson Product Moment Correlation between a rater's model-data residuals and model-based expected scores, α_r = rater slope, $SD\tau_{rk}$ = standard deviation of rater thresholds, + = increases, − = decreases, blank & N = no effect, a = slightly

Depicting Rating Quality

In this section, I illustrate the use of these indices may be used to monitor the rating quality. First, I explain how several raw score and rater agreement indices function as indicators of rater effects. Then, I illustrate how rater effects can be detected using the Rasch Partial Credit Model (RPCM).

Example Data

I generated simulated data for illustrative purposes, and descriptive statistics for this data set are presented in Table 3. In this table, each row should be compared to the first column (true score) to determine how the scores of the raters in that group deviated from the true scores. Shaded cells highlight which rater effect (column) manifests itself in each index (row). To generate data, first, I created 2000 true scores on a unimodal five-point rating scale ranging from 0 to 4. The true scores were symmetrically distributed about the center rating category with a mean of 1.97 and standard deviation of 1.14. I then generated ratings for five Normal raters by introducing a small amount of random error into those true scores so that the mean and standard deviation of the ratings for the Normal raters was approximately equal to the values for the true scores and so that the Normal correlated with true scores at about .84.

Second, I generated ratings for a Lenient and a Severe rater by randomly reducing the values of ratings, resulting in a score distribution that has a mean rating that is about one point higher and lower, respectively, when compared to the Normal raters. This caused the standard deviations of ratings for the Lenient and Severe raters to be slightly lower than that of the Normal raters, but the obtained correlation with true scores was about the same for all three types of raters. Third, I generated ratings for a Central and an Extreme rater by reducing and increasing the dispersion of the simulated ratings, resulting in a distribution or ratings with a standard deviation that was about one-half (Central) and one and one-half (Extreme) the value for the Normal raters. This reduced the correlation with true scores slightly for these two effect raters due to range restriction/expansion. Finally, I simulated an Inaccurate and an Accurate rater by adding/reducing the amount of random error added to the true scores. This caused the obtained correlation with true scores to equal .69 and .95, respectively.

Table 3.

Descriptive Statistics for the Example Data

Statistic	True	Normal Average	Lenient	Severe	Central	Extreme	Inaccurate	Accurate
P(0)	.10	.10	.01	.37	.00	.33	.09	.10
P(1)	.25	.26	.09	.33	.16	.13	.26	.25
P(2)	.32	.30	.21	.21	.68	.11	.32	.33
P(3)	.23	.24	.34	.08	.16	.12	.23	.22
P(4)	.10	.09	.35	.01	.00	.31	.09	.10
Mean	1.97	1.98	2.93	1.05	2.00	1.95	1.97	1.99
SD	1.13	1.13	1.00	1.01	0.57	1.68	1.11	1.13
R(true)	1.00	.84	.81	.81	.75	0.80	.69	.95

Note: Blue shading = reference values. Orange shading = outliers. P(*n*) = proportion of ratings in category *n*. SD = standard deviation. R(true) = correlation with true scores.

Raw Score Indices

It is instructive to examine the raw score statistics for each displayed in Table 3 because these simple summary statistics give us a good bit of diagnostically useful information. That is, it is easy to determine which of these indices are influenced by each type of rater effect. It is also easy to determine which of these indices is easier to interpret quickly as would be required if a scoring leader sought to evaluate raters during an operational project. Overall, the frequency distributions for each rater provide a good bit of very specific information about each rater. However, these proportions are difficult to interpret quickly because several rules would be required to match patterns of highs and lows to specific rater effects. For example, if I see a higher proportion of ratings in categories 3 and 4, I would diagnose a rater as exhibiting Leniency. If I were to see a high proportion of ratings in the 2 score category, I would diagnose Centrality. These proportions do not, however, tell me anything about Accuracy and Inaccuracy.

If we focus on only the final three rows of the table, diagnosis of rater effects is much quicker and straightforward. If I see an increase or decrease in the mean of the raw scores, I would diagnose Leniency or Severity, respectively. If I see a decrease or increase in the standard deviation of the raw scores, I would diagnose Centrality or Extremity, respectively. If I were to see a low or high correlation with true scores (e.g., consensus scores assigned by scoring leaders in an operational setting), then I would diagnose Inaccuracy or Accuracy, respectively.

One thing that is not clear from these indices is what threshold I should use for make these diagnoses. This is a somewhat complicated topic, so I only discuss it in a cursory manner in this paper. In short, there are three approaches to identifying whether to apply a flag and take action regarding a particular rater. First, I could simply make relative comparisons. That is, I could identify the most extreme cases or some proportion of the most extreme cases for each index, flagging those raters for corrective action. Second, I could employ inferential statistics to apply a hypothesis test, flagging raters who demonstrate a statistically significant difference from a particular reference value (e.g., the average of the scores assigned by all raters or the true scores). Third, I could employ effect size transformations to these raw score statistics, declaring a particular magnitude of difference "actionable." The point of this discussion is that, through a judgment-based process, scoring leaders must identify a set of rules for flagging individual raters for corrective action, and there are several options for how to identify those thresholds for each statistical indicator. There is no single right or wrong value, so scoring leaders must engage in policy making decisions, often in consultation with stakeholders, prior to the scoring project.

Agreement & Accuracy Indices

Table 4 summarizes several rater agreement and rater accuracy statistics that are often employed in operational scoring projects for the sake of monitoring raters. In this case, the accuracy indices were computed by comparing the scores of each rater to true scores, and the agreement indices were create by comparing the scores of each rater to the score of the first Normal rater. Again, highlighted cells draw attention to which rater effects (columns) influence the values of each index (rows). Lighter shading indicates smaller deviations, while darker shading indicates larger deviations. In this case, the magnitude of deviations selected for flagging an individual rater was chosen somewhat arbitrarily [i.e., values beyond .1 and .2 for P(P), P(A), and the two kappa coefficients and values beyond .05 and .10 for P(N) and the correlation coefficients]. In general, rows that have relatively fewer highlights are more diagnostically useful, while rows with relatively more highlights are less useful because they are influenced by a larger number of rater effects and are, therefore, unable to differentiate rater effects.

Several things are clear from these indices. First, percentage of agreement indices are influenced by several rater effects, regardless of frame of reference, meaning that they are of little use for diagnostic purposes. As an example, the percent of perfect agreement is quite low for the Lenient, Sever, and Extreme raters in both frames of reference. Second, coefficient kappa and the intraclass correlation coefficient are even worse—their values are lower than the values produced by Normal raters for nearly all of the rater effects. It is also interesting to note that the values of these statistics are lowest for the Lenient and the Severe raters—the two rater effects that are the simplest to correct. Hence, these indices not only flag several rater effects, but the raters who appear to be "worst" according to these indices are the ones whose scores would be easiest to remedy. Regardless, it seems that the only message we can provide to raters who would be flagged based on percentages of agreement, coefficient kappa, and the intraclass correlation is to "score better" without specific directions about how to accomplish that. Third, the Pearson and Spearman correlation coefficients seem to provide the most diagnostically useful information concerning rater effects. Specifically, both of these indices produce a slightly lower value than that observed for Normal raters for the Central rater, and both produce a large decrease/increase for the Inaccurate/Accurate rater. Hence, of the percentage and correlation indices, the only ones that seem to be useful for the purposes of monitoring raters and providing them with diagnostic feedback are the Pearson and Spearman coefficients, which will flag Central raters when the value is slightly decreased (coupled with a decreased raw score standard deviation) and will flag Inaccurate/Accurate raters when the value is more substantially decreased/increased (in the absence of a decreased raw score standard deviation).

Table 4.

Validity and Agreement Percentages and Correlations for the Example Data

Reference	Statistic	Normal Average	Lenient	Severe	Central	Extreme	Inaccurate	Accurate
Accuracy	P(P)	0.62	0.23	0.25	0.45	0.35	0.47	0.87
	P(A)	0.37	0.57	0.58	0.52	0.52	0.45	0.13
	P(NA)	0.01	0.19	0.17	0.03	0.13	0.07	0.00
	R(true)	0.84	0.81	0.81	0.75	0.80	0.69	0.95
	RS(true)	0.84	0.82	0.82	0.73	0.82	0.68	0.95
	K(true)	0.50	0.04	0.05	0.23	0.22	0.31	0.83
	KWT(true)	0.69	0.33	0.34	0.41	0.53	0.51	0.90
	ICC(true)	0.84	0.50	0.53	0.60	0.74	0.69	0.95
Agreement	P(P)	0.52	0.27	0.29	0.42	0.32	0.43	0.61
	P(A)	0.44	0.50	0.49	0.53	0.49	0.47	0.38
	P(NA)	0.05	0.24	0.22	0.05	0.18	0.11	0.01
	R(N1)	0.76	0.73	0.74	0.67	0.73	0.63	0.83
	RS(N1)	0.75	0.74	0.74	0.66	0.74	0.62	0.83
	K(N1)	0.36	0.08	0.10	0.19	0.19	0.25	0.49
	KWT(N1)	0.58	0.31	0.33	0.36	0.47	0.45	0.68
	ICC(N1)	0.76	0.43	0.46	0.54	0.68	0.63	0.83

Note: Blue shading = reference values. Dark orange shading = outliers. Light orange shading = slightly inflated/deflated values. Accuracy statistics target true scores. Agreement statistics target the first Normal rater. Normal Average = average value of relevant Normal raters. P(P), P(A), P(NA) = proportion of perfect, adjacent, and non-adjacent agreement, respectively. R(target) = Pearson correlation with target. RS(target) = Spearman correlation with target. K(target) = Cohen's kappa coefficient with target. KWT(target) = Quadratically weighted coefficient kappa with target. ICC(target) = intraclass correlation coefficient with target.

Latent Trait Scaling

I estimated parameters for the RPCM based on these data with the Winsteps software (Linacre, 2012). Although several commercially available computer programs exist that are capable of estimating parameters for this model (*IRTPRO: User Guide*, 2011;

Muraki & Bock, 2003; Wu, Adams, Wilson, & Haldane, 2007) and although you can estimate parameters for this model using several more general statistical packages (e.g., PROC NLMIXED in SAS) as well so open-source software packages (e.g., Winbugs and R), Winsteps is convenient because all of the rater effect detection indices of interest can be readily output from the software. Appendix A contains the Winsteps command file used to produce the output summarized here. In this file, the only output required is the summary of the rater estimates (i.e., the TFILE= statement) and the rater thresholds (i.e., the SFILE= statement). If we were interested in computing the expected-residual correlation, we would also need to add an XFILE= statement, which will create a data file containing the expected scores and residuals for every rater-by-examinee combinations. It is also important to note that, for the sake of illustration, I scaled the data with the true scores included in the data file. Doing this allows us to use the values of either the true scores or the Normal raters as the basis of comparison for effect raters. If I had excluded true scores, then only Normal raters could be used for the sake of comparison. It is important to note, however, that in operational settings, scoring leaders would either use true scores as the basis of comparison or the average or typical value of all raters in the pool as the reference value for interpreting indices from each individual rater.

Latent Trait Indices

Table 5 presents the values of the rater location (ρ_r), rater score-ability correlation ($PPMC_{r,\theta}$), standard deviation of rater thresholds ($SD\tau_{rk}$), and the unweighted mean-square fit statistic (MSU) for the true scores, the average of the Normal raters, and each of the six effect raters. The simplicity of the patterns of index values as they relate to rater effects makes it clear that using latent trait measurement models makes identification of rater effects straightforward. Deflated/inflated values of the rater location estimate (ρ_r) indicates leniency/severity. Inflated/deflated values of the standard deviation of the rater threshold estimate ($SD\tau_{rk}$) indicates centrality/extremity. Inflation of the rater score-ability correlations ($PPMC_{r,\theta}$) in the absence of inflated/deflated standard deviations of rater threshold estimates indicates rater accuracy/inaccuracy. Note that the value of the unweighted mean-square fit index (MSU) is relatively unaffected by leniency, severity, and centrality, is slightly elevated for extremity, and is considerably inflated/deflated for inaccuracy/accuracy. That is, that index does not do a good job of differentiating the various rater

Table 5.

Validity and Agreement Percentages and Correlations for the Example Data

Index	True Score	Normal	Lenient	Severe	Central	Extreme	Inaccurate	Accurate
ρ_r	-0.03	0.00	-3.28	3.49	0.09	-0.05	-0.03	-0.07
$PPMC_{r,\theta}$	0.95	0.87	0.82	0.85	0.79	0.82	0.76	0.94
$SD\tau_{rk}$	4.16	4.02	4.09	4.27	8.65	0.95	4.30	4.09
MSU	0.37	0.91	1.08	0.89	0.99	1.59	1.93	0.40

Note: Blue shading = reference values. Dark orange shading = outliers. Light orange shading = slightly inflated/deflated values. Normal = average index value for the five Normal raters.

effects.

Conclusions

In this white paper, I have attempted to stress several points concerning the current status of rater monitoring practices and identify potential improvements to that system. First, it is important to note that monitoring the quality of ratings is an important activity. Due to the numerous potential variables relating to rater characteristics, response content, and rating context that can affect a rater's decision-making process, scoring leaders need to implement rater monitoring procedures so that they can take corrective action during a scoring project, should evidence arise that raters have either failed to learn to apply the scoring rubric adequately or have, over time, drifted from the original intent of the rubric. Unfortunately, due to the relatively isolated manner in which researchers who focus on substantive issues relating to rating quality and researchers who focus on quantitative indicators of rating quality have approached those subjects, we know very little about why rater effects occur and what we can do to prevent them (Wolfe & McVay, 2012).

Second, in order to be optimal, rater monitoring procedures should utilize diagnostically specific indices. That is, the indices used to monitor raters should be sensitive to rating patterns that deviate from true scores (i.e., rater effects), and those indices should differentiate one rater effect from another. Unfortunately, many of the indices that are currently used in operational settings do not do this. Specifically, I have demonstrated that nearly all of the percentage agreement indices, coefficient kappa, and the intraclass correlation coefficient are influenced by several rater effects, so they cannot be used as the basis for providing diagnostic feedback to raters. Several other raw score statistics and related latent trait measurement models, on the other hand, are sensitive to only one particular rater effect and can, therefore, be utilized to make diagnostic decisions about whether the scores of a particular rater exhibit evidence of a specific rater effect.

Finally, due to the significant costs associated with the scoring enterprise, it is important to undertake rater monitoring in the most efficient manner possible. Scoring, in and of itself, is an expensive activity, so anything that adds cost to the system without resulting in a score that can be reported back to students increases that cost. Activities such as paying scoring leaders to identify and assign true scores to responses that will be included in validity sets, having raters score validity responses, and having scoring leaders spend time reviewing rater effect statistics for each rater will do just that—increase scoring costs without resulting in a score that is being reported to students. These are very important activities because they allow scoring leaders to monitor the quality of ratings in real time, so there is indeed a need to balance the cost of that activity with the very useful information that it provides. In fact, by implementing the procedures describe in this manuscript, that goal would be realized—scoring leaders would obtain the most focused and accurate information possible while minimizing the cost of collecting that information.

There are several features of current rater monitoring procedures that can be improved in terms of efficiency, beyond utilizing diagnostically specific rater indices. To begin with, scoring leaders can utilize diagnostic rater effect indices that allow for the development of simple rules for flagging raters. For example, although score category frequency distributions are diagnostically specific, they make it difficult for a scoring leader to quickly look at the pattern of scores for an individual rater and make an immediate diagnosis of whether that rater's scores exhibit evidence of a particular rater effect. A better choice would be something like the raw score mean and standard deviation, so that raters can be flagged by software that automates the rater monitoring process by determining whether the values of statistics associated with a particular rater exceed pre-established thresholds for those indices.

Another way that current rater monitoring procedures can be improved is by identifying ways to decrease the number of scores that a rater assigns only for the purpose of rater monitoring as well as utilizing operational scores and automated scores in the rater monitoring process. A simple way of decreasing the number of scores assigned solely for the purpose of monitoring raters is to move away from the use of fixed sets of validity papers and to begin selecting validity responses that will provide the most information about our current hypotheses about rater effects that a particular rater may be exhibiting. For example, if we believe that a rater's scores exhibit leniency, then assign validity responses that have low true scores. If we believe that a rater's scores exhibit centrality, then assign validity responses to that rater that have high and low scores. If we believe that a rater's scores exhibit inaccuracy, then assign validity responses to that rater that are easiest for other raters to score accurately. By engaging adaptive selection of validity responses, scoring leaders can maximize the amount of information they receive from that activity.

Scoring leaders can also decrease the number of responses that are scored solely for monitoring purposes by expanding the number of scored responses that are considered. For example, validity responses alone sometimes serve as the basis of rater monitoring decisions because raters are paired with other raters of unknown accuracy during operational scoring. Hence, in order to avoid making misdiagnoses during rater monitoring, scoring leaders restrict their attention to a rater's performance on validity responses.

However, by utilizing latent trait measurement models, scores assigned to validity sets and operational responses can be considered jointly, particularly in scoring projects that involve more than one rater assigning scores to a particular examinee. In such an application, scoring leaders can compare the value of a diagnostically specific index for an individual rater to the average value of that index across all other raters. To confirm that the average value for all raters is an appropriate basis for comparison, those averages can be compared to the values of the indices obtained for true scores. By engaging in this type of joint scaling, scoring leaders will be able to increase the number of observations upon which they are making rater monitoring decisions for each rater, increasing the precision of rater monitoring decisions while decreasing the cost of that increase due to the use of operational scores in the monitoring process. In a related way, automated scoring technology could be employed to further increase the number of decision points used during rater monitoring as well as allowing these proposed procedures to be extended to operational scoring projects in which only a single rater scores each response.

In closing, I want to acknowledge that current rater monitoring practices are based on many years of experience of scoring experts who have dedicated their professional lives to making the scoring enterprise the best that it can be. That knowledge has evolved over time and been transformed into a coherent set of practices that do a reasonably good job of accomplishing what they are intended to do (i.e., provide scoring leaders with accurate information). However, as new technologies have been introduced, not all of them have been incorporated into the existing rater monitoring process to the degree that they could. Implementing a system such as the one I've described will likely be expensive, so the benefits of doing so should be weighed relative to other emerging technologies that will impact the way that we assign scores to performance assessments (e.g., automated scoring) as well as the magnitude of the improvement that such a system would offer over our current rater monitoring procedures. By rethinking what we currently do, why we do it, and how it can be improved, we will be able to hold down scoring costs and increase the quality of the scores that are assigned to examinees, thus improving the quality and accuracy of the decisions that are made based on assessment results.

References

Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*, 1–23.

Altman, D. G. (1991). *Practical statistics for medical research*. London: Chapman and Hall.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561–573. Retrieved from 10.1007/BF02293814

Barton, M. A., & Lord, F. M. (1981). *An upper asymptote for the three-parameter logistic tem-response model*. Princeton, NJ: Educational Testing Service.

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 creit. *Psychological Bulletin*, *70*, 213–220.

De Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. New York: Guilford Press.

DeCarlo, L. T., Kim, Y. K., & Johnson, M. S. (2011). A hierarchical rater model for constructed responses, with a signal detection rater model. *Journal of Educational Measurement*, *48*, 333–356.

Engelhard, G. J. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*, 93–112.

Engelhard, G. J. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement*, *33*, 56–70.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed., pp. 38–46). John Wiley.

IRTPRO: User Guide. (2011). Lincolnwood, IL: Scientific Software International, Inc.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174.

Linacre, J. M. (2012). WINSTEPS. Beaverton, OR: Winsteps.com.

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

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–176.

Muraki, E., & Bock, D. (2003). PARSCALE 4: IRT item analysis and test scoring for rating scale data. Chicago, IL: Scientific Software International.

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.

Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics*, *27*, 341–384.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedogogiske Institut. (reprinted 1980 with Foreword, Afterword, and References, Chicago: The University of Chicago Press).

Shin, H. J., & Choi, J. (2011). Detecting rater effects: Comparison of many-facet Rasch model, rater bundle model, and hierarchical rater model. *National Council on Measurement in Education*.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428.

Wilson, M., & Hoskens, M. (2001). The rater bundle model. *Journal of Educational and Behavioral Statistics*, *26*, 283–306.

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

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). ACER ConQuest: Generalized item response modeling software. Melbourne, VIC, Australia: Australian Council for Educational Research.

Appendix A

Winsteps Command Files

```
&inst
title=WINSTEPS RATER EFFECT ANALYSIS WITH TRUE SCORES
ni=12
item1=1
groups=0
codes=01234
tfile=*
14
*
sfile=thresholds_true.txt
&end
True
Normal 1
Normal 2
Normal 3
Normal 4
Normal 5
Lenient
Severe
Central
Extreme
Iinaccurate
Accurate
ENDNAMES
0 0 0 0 0 1 1 0 0 0 0 0
.
.
.
444444423444
```