JEM JOURNAL of EDUCATIONAL MEASUREMENT NCME national council on measurement in education
DOI: 10.1111/jedm.12224 *Journal of Educational Measurement*

Pedagogical Considerations for Examining Rater Variability in Rater-Mediated Assessments: A Three-Model Framework

Brian C. Wesolowski
University of Georgia
Stefanie A. Wind
University of Alabama

Rater-mediated assessments are a common methodology for measuring persons, investigating rater behavior, and/or defining latent constructs. The purpose of this article is to provide a pedagogical framework for examining rater variability in the context of rater-mediated assessments using three distinct models. The first model is the observation model*, which includes ecological/environmental considerations for the evaluation system. The second model is the* measurement model*, which includes the transformation of observed, rater response data to linear measures using a measurement model with specific requirements of rater-invariant measurement in order to examine raters' construct-relevant variability stemming from the evaluative system. The third model is the* interaction model*, which includes an interaction parameter to allow for the investigation into raters' systematic, construct-irrelevant variability stemming from the evaluative system. Implications for measurement outcomes and validity are discussed.*

In the physical sciences, attributes such as length, mass, or temperature can be directly and objectively measured in fundamental base units using specified measurement instruments. The length of an object can be measured in meters with the use of a ruler, the mass of an object can be measured in kilograms with the use of a weighing scale, and the temperature of an object can be measured in Kelvin using a thermometer. In each of these examples, the attribute of the object is measured in equal base units using a measurement instrument specifically calibrated for its intended use. Attributes often under investigation in much of the educational and psychological research literature, however, are not measured in the same direct manner.

From a psychometric perspective, these attributes are considered to be hypothetical approximations of reality because they cannot be directly measured (Braun, Jackson, & Wiley, 2002). As such, the attributes can only be indirectly estimated through the direct measurement of overt, secondary criteria or behaviors representative of them. In the case of educational research, researchers seek to operationally define the attribute, otherwise referred to as a latent construct (Loevinger, 1957), using multiple criteria or behaviors that represent (i.e., operationally define) the construct itself. As Loevinger (1957) notes, "traits exist in people; constructs exist in the minds of psychologists" (p. 83). The overall quality (i.e., model fit) with which these secondary criteria or behaviors work together to define the latent construct and provide important validity evidence as to how well the construct is operationally defined.

Measuring Latent Constructs with Rater-Mediated Assessment Designs

The use of rater-mediated assessment designs is a common methodology in educational research for investigating and defining latent constructs. Rater-mediated assessments are often used to either identify differences in a rater's response to a person, or as a means for indicating differences in persons themselves. In cases where the object of measurement is the rater, variability in the raters' responses to the person is of direct interest to the researcher. In cases where the object of measurement is the person, raters are used to mediate the process of mapping scores and/or values onto the particular person of interest. In these cases, variability attributed to each of the individual raters is not welcomed, as it contributes unwanted noise and measurement error associated with the construct underlying the study.

The Role of Variability in Measuring Latent Constructs

From a measurement perspective, variability can broadly be defined as "a measure of the differences among . . . scores" (Zieky, 2016, p. 89). Two sources of variability exist in all rater-mediated assessments, specifically (a) construct-relevant variability, and (b) construct-irrelevant variability. Construct-relevant variability includes the differences in scores attributed to the construct being measured. In the case of rater-mediated assessments, this can include variability in the scores of the raters (i.e., rater severity), variability in the scores of the person (i.e., amount of the construct each person possesses), or variability in the scores of the items being used to elicit responses from the raters (i.e., item difficulty). In cases of research where the rater is the object of measurement, variability attributed to each of the individual raters is welcomed. As the variability increases and the range in raters' scoring increases, the spread in scoring provides more and varied information as to the relationship between the raters' responses and the construct investigated in the study. However, variability in the persons the raters are evaluating is not particularly welcomed as it contributes unwanted inconsistencies in the measurement process. This is the conceptual equivalent of raters shooting at a moving target. Oppositely, when the object of measurement is the person, variability attributed to each individual person is welcomed. As the variability increases and the range in persons scores increases, the spread in scoring provides more and varied information as to the relationships between the persons and the construct underlying the study. However, variability in the raters evaluating the persons is not particularly welcomed as it contributes unwanted inconsistencies in the measurement process. As an example, if the purpose of the measurement is to provide one single score to a person and rater responses are generating multiple scores for the same person, it is unclear what the most appropriate score is.

The second type of variability is construct-irrelevant variability, or the random variation in scores not attributed to the construct being measured. Construct-irrelevant variability is often referred to as measurement error or error variance, and can threaten the validity of the assessment context (Messick, 1989). Measurement error can be categorized as either random or systematic. Some random error is expected in measurement systems. However, systematic error variances (see Interaction Model below), or patterned responses related to construct-irrelevant characteristics

of an assessment, introduce important substantive implications into the measurement system that warrant consideration.

The Lens Model and Rater Variability

One popular conceptual framework for investigating rater variability in the context of rater-mediated assessments is Brunswik's lens model (Brunswik, 1952). Brunswik's original conception was in the context of human visual perception; in particular, evaluating the relationship between an organism and its environment. Ecological validity, as Brunswik noted, posits that a relationship exists between an organism and its environment mediated specifically through a probabilistic function of environmental cues. In other words, human perception is based upon the environmental cues a human sees, rather than the human's cognitive understanding of what the human sees.

The lens model was first applied to human judgment and decision making by Hammond (1955). Ever since, the lens model has been widely used in psychological and social science research as a method for quantifying human decision-making processes (Karelaia & Hogarth, 2008). Engelhard (2013) indicates that the lens model is an important conceptual consideration for investigating rater behavior in the context of rater-mediated psychological, behavioral, and social science research:

> First of all, the concept of cues stresses the need for assessment developers to carefully select and monitor aspects of the environment that raters should pay attention to in the rating process. For example, rubrics and other guidelines used for scoring are viewed as cues to the raters. Second, the lens model is grounded in ecological psychology, and it is important to recognize that rater-mediated assessments must be viewed as contextualized within specific assessment environments and systems. And finally, lens models raise issues related to both precision and accuracy of ratings obtained from a fallible and potentially noisy rating process from human raters. (p. 193)

In rater-mediated assessments, investigations into rater variability indicate that raters can negatively affect the overall measurement of any criteria or behaviors representing a latent construct of interest, thereby negatively influencing inferences made by the investigator. Therefore, without proper psychometric considerations toward the treatment of rater-mediated scoring data, concerns for the overall validity of studies are warranted. The purpose of this article is to provide a pedagogical framework for examining rater variability in the context of rater-mediated assessments using three distinct models (see Figure 1). The use of the three-model framework, in the context of rater-mediated assessment, can aid in the development, refinement, and quality control of inferences underlying any research in which it is applied. The first model is the *observation model*, which includes three important ecological/environmental considerations for the evaluation system: (a) construct-relevant factors including the evaluative context (i.e., the type of persons being evaluated), the evaluators (i.e., the type of raters doing the evaluating), and the evaluative cues (i.e., the structure and content of the measurement instrument); (b) the collection of observed observations stemming from rater interactions with the evaluated persons and evaluative cues; and (c) consideration of the structure, or connectivity, between

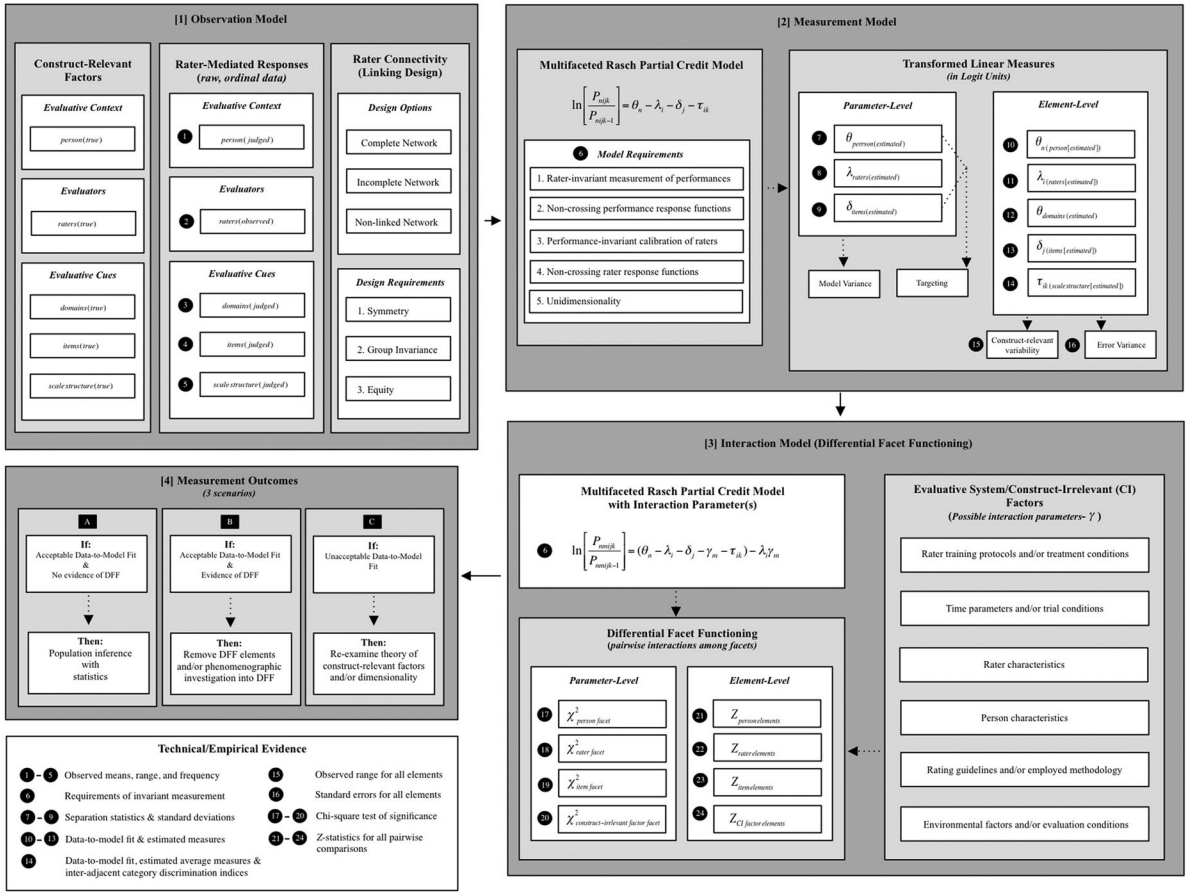

Figure 1. Three-model framework.

raters' responses to persons. The second model is the *measurement model*, which includes the transformation of observed, rater response data to linear measures using a measurement model with specific requirements of invariant measurement. Here, the logistic transformation of observed-score rater data provides an opportunity to investigate construct-relevant variability stemming from the evaluative system. The third model is the *interaction model*, which includes the addition of an interaction parameter to the previously used measurement model, allowing for the investigation into raters' systematic, construct-irrelevant variability stemming from the evaluative system. This can include the investigation into any number of interactions between facets in the model, including rater-by-person, rater-by-item/domain, or person-by-item/domain, for example.

Model I: Observation Model

The first model in this framework is the observation model. In the context of rater-mediated assessments, raters most often interact simultaneously with a person (i.e., an object of measurement) and set of evaluative cues, such as a rating scale, for example. The result of these interactions is a three-facet set of observed scores representing raters' (i.e., facet 1) use of the evaluative criteria, such as items (i.e., facet 2), when evaluating persons (i.e., facet 3). At the intersection of each rater, each item, and each person is a single, observation. In an assessment system where multiple raters evaluate multiple persons, the result is a set of observations that can be ordered in a meaningful way. Using the sums of their responses, the raters can be ordered from high to low, representing observed, ordinal levels of construct-relevant rater variability (i.e., rater severity). The items' set of scores can be summed and ordered from high to low, representing observed, ordinal levels of construct-relevant item variability (i.e., item difficulty/rater agreeableness). Finally, using sums of the raters' ratings of each persons, the persons can be summed and ordered from high to low, representing observed, ordinal levels of construct-relevant person variability (i.e., the amount of the construct each person possesses). This three-facet framework provides set of patterns that lay the foundation for measurement to begin and variability to be investigated.

In setting the stage for any validity arguments (Kane, 1992, 2001; Kane, Crooks, & Cohen, 1999) it is important to note that each of these construct-relevant factors can single-handedly affect the overall assessment system. Therefore, researchers' explicit assertion for the prespecified decisions made in terms of the sample of raters selected for the study, the specific cues used to elicit rater responses, and the sample of persons used in the study is necessary for sound validity arguments. Additionally, in some data collection procedures, it may not be feasible for all raters to evaluate all persons used in the study. In rater-mediated assessments, the connectivity of raters can affect the empirical results of the assessment context (Wind, Engelhard, & Wesolowski, 2016). Therefore, it is an important research design consideration that warrants researcher specification. As noted by Engelhard (1997), three categories of rater linking designs exist: (a) a complete assessment network, where all raters evaluate all persons, providing full connectivity between observations; (b) an incomplete assessment network, where some raters evaluate some persons, providing some

connectivity between observations; and (c) nonlinked assessment networks, where some raters evaluate some persons, providing no connectivity between observations. See Engelhard (1997) for full descriptions of design considerations and empirical consequences.

Model 2: Measurement Model

The second model in this framework is a *measurement model*. When researchers apply a measurement model to raters' ratings of persons, essentially they are testing the hypothesis that the ratings reflect a certain pattern of relationships between raters, persons, and items. A variety of measurement models exist that researchers can use in the context of rater-mediated assessments. One model that many researchers have found particularly useful in these contexts is a many facets version of the Rasch model (Linacre, 1989; Rasch, 1960). The Rasch model is particularly important for investigating rater variability due to its five requirements of rater-invariant measurement: (a) rater-invariant measurement of persons (i.e., the measurement of persons must be independent of the particular raters that happen to be used for the measuring), (b) noncrossing person response functions (i.e., a more able person must always have a better chance of obtaining higher ratings from raters than a less able person), (c) person-invariant calibration of raters (i.e., the calibration of the raters must be independent of the particular persons used for calibration), (d) noncrossing rater response functions (i.e., any person must have a better chance of obtaining a higher rating from lenient raters than from more severe raters), and (e) variable map (i.e., persons and raters must be simultaneously located on a single underlying latent variable) (Engelhard, 1994b, 2008, 2013). When adequate fit (see Data-to-Model Fit of Estimates below) to the measurement model is actively obtained, invariant measurement is achieved. In the case of rater-mediated assessment, adequate data-to-model fit infers that construct-relevant factors attributed to the rater do not contribute interference to the estimated item scores and person scores obtained from the measurement model.

The estimation procedure for the Rasch model includes the transformation of raters' ordinal ratings of persons to linear measures using a logistic transformation. The logistic transformation allows researchers to describe rater severity, raters' judgments of the persons, and other possible facets of interest on a linear metric. Furthermore, because the Rasch model is based on properties of invariance, one can make comparisons between raters, persons, and other facets within the same frame of reference. As a result, it is possible to explore the degree to which there is construct-relevant and construct-irrelevant variability present in raters' ratings.

Many Facets Rasch Partial Credit Model

The many facets Rasch (MFR) model is particularly useful because it allows researchers to include additional facets that may represent sources of construct-irrelevant variance in an assessment system. For example, in addition to including facets for persons, items, and raters, other characteristics of interest associated with raters (e.g., age, expertise, training, experimental condition) and persons (e.g., age, gender, experience, social economic status) can also be included in the model.

This article focuses on a partial credit model (PCM; Masters, 1982) formulation of the MFR model. The rating scale model (RSM; Andrich, 1978) formulation is a more ubiquitous variation of the model, where all items share the same rating scale structure. However, in considering pedagogical utility, the PCM was chosen as an exemplar because it is more detail-specific. In particular, the PCM is used when each item has a unique rating scale structure or if the researcher is interested in investigating difficulty thresholds for each unique item. Linacre (2000) provides detailed considerations for appropriately choosing between the RSM and PCM. A general form of the MFR-PC model for rater-mediated assessments can be stated as follows:

$$\ln\left[\frac{P_{nmi(x=k)}}{P_{nmi(x=k-1)}}\right] = \theta_n - \lambda_m - \delta_i - \tau_{mk}. \quad (1)$$

The left side of the formula represents the log of the odds that person n is rated in rating scale category k, rather than in category $(k-1)$ by rater m on domain/item i. The other terms are defined as follows:

$\theta_n =$ the judged score of person n,
$\lambda_m =$ the severity of rater m,
$\delta_i =$ the judged difficulty of domain/item i, and
$\tau_{mk} =$ the location on the logit scale at which there is an equal chance for a rating in category k and category $(k-1)$, specific to rater m.

When researchers apply the MFR-PC model shown in Equation 1 to their data, they calculate logit-scale location estimates for every person (θ), rater (λ), domain/item (δ), and threshold (τ), respectively. These estimates provide evidence of construct-relevant variance resulting from the measurement procedure. Furthermore, one can calculate indicators of measurement precision and data-to-model fit as additional evidence to support the interpretation of the estimates. Table 1 includes a summary of several outcomes from the MFR-PC model that are relevant in the context of a rater-mediated assessment. In the next section, a brief illustration of an MFR-PC analysis of a rater-mediated assessment is presented in which each of the sources of evidence included in Table 1 is examined.

Illustration

A data set of example ratings in which 20 raters rated 100 persons on three domains using a five-category rating scale (0, 1, 2, 3, 4) was simulated to illustrate the interpretation of each component of Equation 1, where lower numbers indicate lower judged scores of the person and higher numbers indicate higher judged scores of the person. Complete assessment networks, where all raters rate all persons, are theoretically ideal and desirable; however, most large-scale operational rater-mediated assessment systems involve various forms of incomplete assessment networks due to time, money, and other administrative constraints (Wind, Engelhard, & Wesolowski, 2016). As Engelhard (1997) notes, incomplete assessment network designs, when constructed using sound data collection designs, "obtain reliable and valid links both within and between facets that are less costly in terms of examinee time and rater salaries" (p. 27). Therefore, in the example data, two randomly selected raters rated each person, and each rater rated at least one person in common with at least one

Table 1
Measurement Model Evidence and Interpretive Questions

Evidence Category	Source of Evidence	Interpretive Question
Location	Logit-Scale Location Estimates	What are the locations of individual elements of a given facet (person, raters, domains/items) on the logit scale?
Precision	Targeting	To what extent are the locations of individual elements of each facet aligned with elements of every other facet?
	Standard Errors	How much error is associated with the logit-scale estimate for each element of each facet?
	Separation Statistics	To what extent are the differences in location estimates meaningful for elements within each facet?
Fit	Numeric Fit Statistics	To what extent do the observed ratings for each element of each facet match what is expected by the measurement model?
	Graphical Fit Displays	What is the pattern of residuals (deviations of observations from model expectation) associated with a given element of a given facet?

other rater (an example of an *incomplete assessment network*; Engelhard, 1994a). The example data set was examined with the MFR-PC model using the *FACETS* software program (Linacre, 2015).

Location Estimates

The first source of evidence for interpreting the results from the measurement model is the set of location estimates (i.e., logit scores) for the individual elements of each facet (e.g., each rater score, each person score. etc.) in the model. In the example data, location estimates for individual person, rater, and domain/item were examined. As shown in Table 1, location estimates are used to address the following interpretive question: What are the locations of individual elements of a given facet (persons, raters, domains/items) on the logit scale?

Figure 2 is a *variable map*, which is a visual depiction of the location estimates that resulted from the simulated data. The first column shows the interval-level logit scale on which all of the facets have been estimated. Higher numbers indicate higher person scores, more-severe raters, more-difficult domains/items, and more difficult rating scale categories. The following is the interpretation of each of the remaining columns.

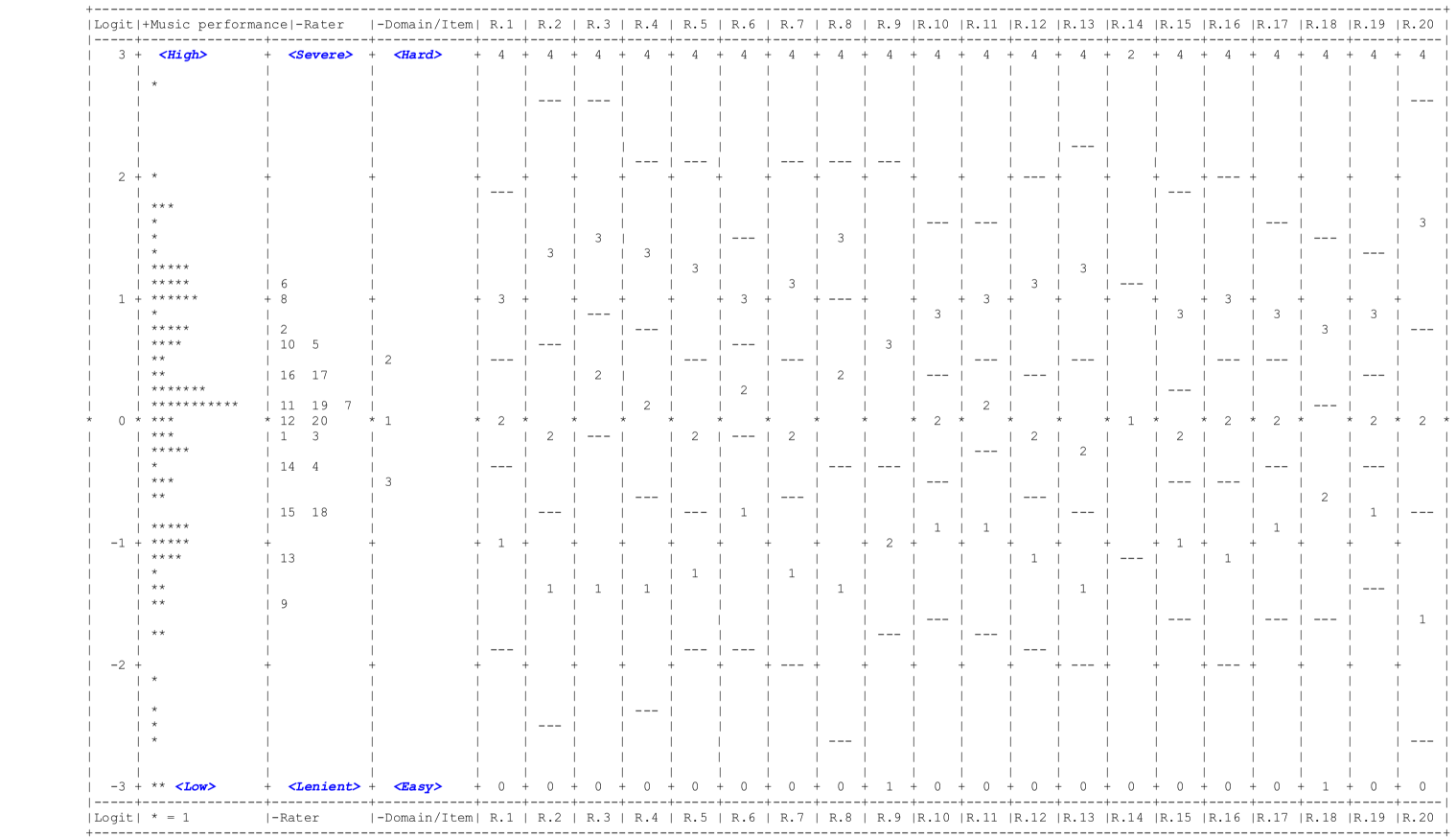

Figure 2. Variable map for the PC-MFR model based on example data. (Color figure can be viewed at wileyonlinelibrary.com)

Person Estimates

The second column in Figure 2 shows the logit-scale locations for each of the 100 persons included in the example data set. Specifically, an asterisk (*) represents the estimated logit-scale location for one person. These locations are estimated using the observed ratings associated with each person. Higher locations indicate that, in general, the raters judged a person as having higher levels of the construct over all of the domains/items. Likewise, lower locations indicate that, in general, the raters judged a person as having lower levels of the construct over all of the domains/items.

Because of the characteristics of the Rasch model, each of the person estimates can be compared directly to each of the other person estimates, even though all of the raters did not rate all of the persons. Based on the available information in the model, Person 64 had the highest judged score. This person had an average rating of 3.74 on the observed score scale, which corresponds to a logit-scale location of 2.80 logits. In contrast, Person 3 had the lowest judged score. This person had an average rating of .50 on the observed score scale, which corresponds to a logit-scale location of −3.13 logits.

Rater Severity Estimates

The third column in Figure 2 shows the logit-scale locations for each of the 20 raters included in the example data set. Here, rater identification numbers are used to indicate the locations of each rater. The rater severity locations are estimated using the ratings that each rater assigned to all of the persons that they rated over all of the domains/items. Higher locations indicate that, in general, a rater gave low ratings more often—thus indicating that they were a generally severe rater. Likewise, lower locations indicate that, in general, a rater gave high ratings more often—thus indicating that they were a generally lenient rater.

Similar to the student estimates, the Rasch model properties allow for direct comparisons between individual raters even though all of the raters did not rate all of the persons. In the example data set, Rater 6 was the most severe rater. This rater's average rating was .93 on the observed score scale, which corresponds to a logit-scale location of 1.09 logits. In contrast, Rater 9 was the most lenient rater. This rater's average rating was 3.13 on the observed score scale, which corresponds to a logit-scale location of −1.50 logits. All of the rater location estimates are provided in Table 2.

Domain/Item Difficulty Estimates

The fourth column in Figure 2 shows the logit-scale locations for each of the three domains/items included in the example data set. Identification numbers are used to indicate the locations of each domain/item. The domain/item difficulties were estimated using the ratings that each of the raters assigned to each person for the domain/item. Higher locations indicate that, in general, raters gave low ratings on the domain/item—thus indicating that the raters judged the domain/item as relatively difficult, or less likely to endorse. Likewise, lower locations indicate that, in general, raters gave high ratings on the domain/item—thus indicating that the raters judged the domain/item as relatively easy, or more likely to endorse.

Table 2
Results From Estimation of the Rater Facet

Rater	Average Rating	Logit-Scale Location	Standard Error	Infit *MSE*	Outfit *MSE*
6	.93	1.09	.23	.83	.97
8	1.00	.94	.27	.78	.80
2	1.90	.76	.21	1.68	1.74
5	1.80	.64	.22	.87	.82
10	1.87	.58	.19	1.62	1.75
17	1.43	.44	.19	.61	.62
16	1.70	.33	.20	1.07	1.16
11	2.30	.18	.17	1.02	1.15
19	1.97	.16	.19	.89	.75
7	1.70	.13	.22	.77	.79
20	1.83	.02	.23	1.57	1.54
12	2.57	−.05	.21	.93	.96
1	2.30	−.13	.19	.93	.93
3	2.00	−.15	.22	.91	.88
14	.83	−.36	.29	.80	1.09
4	2.27	−.43	.25	.91	.85
18	3.07	−.74	.24	1.07	1.09
15	2.70	−.77	.21	.68	.78
13	2.53	−1.14	.24	.68	.67
9	3.13	−1.50	.29	.75	.68
Mean	1.99	.00	.22	.97	1.00
SD	.64	.68	.03	.31	.33

Note. Raters are presented in logit-scale location order from most severe to least severe.

In the example data set, the raters judged Domain 2 as the most difficult domain. The average rating observed for Domain 2 was 1.62 on the observed score scale, which corresponds to a logit-scale location of .52 logits. In contrast, the raters judged Domain 3 as the easiest domain. The average rating observed for Domain 3 was 2.37 on the observed score scale, which corresponds to a logit-scale location of −.51 logits.

Rating Scale Category Threshold Estimates

In the context of modern measurement models, rating scale category thresholds reflect the difficulty associated with individual rating scale categories. Specifically, a threshold estimate is calculated for each pair of adjacent rating scale categories that reflects the location on the logit-scale at which there is an equal probability that a person will receive a rating in a given category, rather than the category just below it. As a result, $(m - 1)$ thresholds are calculated for a rating scale with m categories.

As noted above, the example data set was analyzed using the MFR-PC model. With the PC model, separate sets of thresholds are estimated for the elements of one or more facets in the model. In the analysis, the model was specified such to estimate separate thresholds for each of the raters. As a result, the analysis produced

one set of rating scale thresholds for each individual rater (a total of 20 sets of rating scale thresholds). Practically, this means that the structure of the rating scale can be separately analyzed for each rater. It is also possible to estimate the rating scale structure separately across the elements within other facets, such as domains/items.

The last 20 columns in Figure 2 show the estimates for the rating scale category thresholds for each rater in the example analysis, where separate columns reflect individual raters. In these columns, horizontal lines indicate the logit-scale locations for each threshold. Each threshold corresponds to the pair of rating scale categories that can be seen just below and above it. Inspection of the rating scale category thresholds indicates that there were differences in the structure of the rating scale for the 20 raters in the example dataset. Specifically, the location of the rating scale category thresholds varied over the 20 raters. Relatedly, there were differences in the length on the logit scale associated with rating scale categories for different raters. Furthermore, one rater (Rater 14) only used categories 0, 1, and 2. Together, these results suggest that the 20 raters interpreted the structure of the rating scale differently. The practical consequence of these differences is that the difficulty associated with each of the rating scale categories was different for different raters. As a result, a rating in a given category from one rater may not have the same interpretation as a rating in the same category from another rater.

Precision of Estimates

The second source of evidence for interpreting the results from the measurement model is indicators of precision for each facet in the model. In the example data, precision indices for person, raters, and domains/items were examined. As shown in Figure 1 and Table 1, there are three main indicators of precision: (1) targeting, (2) standard errors, and (3) separation statistics.

Targeting

The first indicator of measurement precision based on the MFR-PC model is *targeting*. Essentially, targeting refers to the match in location estimates between the elements of each facet in the model. The interpretive question for evidence of targeting is as follows: To what extent are the locations of individual elements of each facet aligned with elements of every other facet? Evidence of close alignment between facet locations indicates that a particular element within a facet (e.g., an individual person) has been measured with enough precision to confidently interpret their logit-scale location. Accordingly, close targeting results in low levels of measurement error. In contrast, large discrepancies between location estimates suggests that the measurement of a given element within a facet may not be precise. Accordingly, a lack of proper targeting results in high levels of measurement error.

It is possible to evaluate targeting using a visual inspection of the variable map shown in Figure 2. Specifically, one can compare the locations of the persons, raters, and domains/items to gauge the degree to which the assessment is targeted. In the example dataset, there appears to be adequate targeting because the distributions of persons, raters, and domains/items are generally overlapping. However, there are some very high scores of persons for which additional measurement with more-difficult

assessment opportunities (e.g., more-severe raters and more-difficult domains/items) would provide more precise estimates. Likewise, there are some very low scores of persons for which additional measurement with easier assessment opportunities (e.g., more-lenient raters and easier domains/items) would provide more precise estimates.

Standard Errors

The second indicator of measurement precision based on the MFR-PC model is *standard errors*. Standard errors (*SE*s) estimate the precision with which individual location estimates were calculated. The interpretive question for evidence of targeting is as follows: How much error is associated with the logit-scale estimate for each element of each facet? SEs are calculated for every element of each facet in the measurement model. Small values of the *SE* indicate that a particular element within a facet (e.g., an individual person) has been measured with enough precision to confidently interpret its logit-scale location. In contrast, large values of *SE* indicate that the measurement of a given element within a facet are not precise.

The illustrative analysis includes calculations of *SE*s for each person, rater, and domain/item. For the persons, *SE*s ranged from .38 logits for the person with the most-precise achievement estimate to .78 logits for the person with the least-precise achievement estimate. For raters, *SE*s ranged from .17 logits for the rater with the least-precise severity estimate to .29 logits for the rater with the most-precise severity estimate. There were differences in the *SE* for person and raters because of the incomplete rating design. Specifically, the random assignment of persons to raters resulted in different degrees of targeting between persons and raters. In contrast, all of the raters rated all of the domains/items, so these elements all had the same *SE*: .08 logits. Furthermore, it is important to note the *SE*s were largest for persons, followed by raters, and lowest for domains/items. This result is expected in incomplete rating designs, because there were many more observations for each domain/item compared to the observations for raters and persons.

Separation Statistics

The third indicator of measurement precision based on the MFR-PC model is *separation statistics*. Rasch model separation statistics are used to evaluate the reliability (i.e., reproducibility) of logit-scale location estimates for each facet included in the model. The interpretive question for evidence of separation is as follows: To what extent are the differences in location estimates meaningful for elements within each facet? Researchers who use Rasch measurement theory often examine two separation statistics: (1) reliability of separation (*Rel*), and (2) a chi-square statistic for separation. Both of these statistics provide information about the spread of individual persons, raters, domains/items, and other facets locations on the logit scale. First, one can calculate the reliability of separation statistic (*Rel*) for each facet as an indicator of the degree to which differences among the individual elements (e.g., each person or each rater) are observed based on the measurement procedure. When there is evidence of acceptable data-to- model fit (described further below), the interpretation of *Rel* for persons is comparable to coefficient alpha. For other facets, *Rel* describes the degree to which individual elements have distinct

logit-scale locations. Second, one can calculate a chi-square separation statistic that describes the degree to which differences among logit-scale locations for elements within a given facet are statistically significant.

The analysis includes calculated separation statistics for each of the facets in Equation 1. The reliability of separation statistic was highest for domains/items: $Rel_{\delta} = .97$, followed by raters ($Rel_{\delta} = .89$), and persons ($Rel_{\theta} = .80$); all of the chi square values were statistically significant ($p < .01$). These results suggest that there were meaningful differences in the logit-scale location estimates for all three facets in the model.

Data-to-Model Fit of Estimates

The third source of evidence for interpreting the results from the measurement model is indicators of data-to-model fit for each facet in the model. When researchers examine data-to-model fit in the context of a Rasch measurement model analysis, they are essentially comparing the observed ratings that were collected during the assessment procedure. Rasch models are quite strict with regard to model requirements. Accordingly, it is expected that there will be discrepancies between the expected ratings and observed ratings. These *residuals* are useful for identifying those elements in an assessment procedure for which there are frequent and substantively noteworthy (i.e., large) discrepancies between the observed and expected ratings. Data-to-model fit analyses involve summarizing residuals related to each facet in the measurement model. In the example data, we examined data-to-model fit indices for persons, raters, and domains/items. As shown in Table 1, there are two main indicators of precision: (1) numeric fit statistics and (2) graphical displays of data-to-model fit.

Numeric Fit Statistics

The first indicator of data-to-model fit based on the MFR-PC model is *numeric fit statistics*. One can use Rasch model fit statistics to summarize the residuals associated with individual elements within each facet of the model. The interpretive question for numeric fit statistics is as follows: To what extent do the estimated locations for each element of each facet match what is expected by the measurement model? Researchers who use Rasch measurement theory often examine two data-to-model statistics: (1) infit and (2) outfit.

The most used form of Rasch infit and outfit statistics are *mean square error* (*MSE*) statistics. Essentially, these statistics are averages of the residuals associated with individual elements of each facet. The difference between the two fit statistics is that infit *MSE* is weighted by statistical information (i.e., the model variance), and outfit *MSE* is not weighted. For raters, Infit *MSE* is defined as follows:

$$\text{Infit } MSE = \frac{\sum_n^N Z_{nmi}^2}{\sum_n^N Q_{nmi}}, \tag{2}$$

where Z is the standardized residual (difference between observed and expected rating) for rater m's rating of person n on domain/item i, and Q is the response variance. Next, Outfit *MSE* is defined as follows:

$$\text{Outfit } MSE = \frac{\sum_{n}^{N} Z_{nmi}^{2}}{N}, \tag{3}$$

where N is the number of persons. One can also calculate infit *MSE* and outfit *MSE* for persons and domains/items using the sum of the residuals associated with a particular person or domain/item, respectively. Outfit *MSE* is not weighted, so it is more sensitive to extreme unexpected ratings, such as a very lenient rater giving a low rating to a high-scoring person. Infit *MSE* is more sensitive to less-extreme unexpected ratings.

The interpretation of values of infit and outfit *MSE* statistics is somewhat contentious in the Rasch measurement literature, with different researchers proposing different critical values that reflect various characteristics of a measurement procedure, such as sample size and the type of data that were collected (e.g., DeAyala, 2009; Smith, 2004; Wolfe, 2013; Wu & Adams, 2013). Nonetheless, researchers who use the Rasch model generally agree that the expected value for infit and outfit *MSE* statistics is around 1.00, where values below 1.00 suggest less variation than expected and values above 1.00 suggest more variation than expected. In this study, infit and outfit are conceptualized as continuous variables, while recognizing the generally accepted range for rating scale data of about .60–1.40 (Bond & Fox, 2015; Engelhard, 2013).

For brevity, the focus is on evaluating rater fit in this illustration; however, it is important to examine data-to-model fit for all of the facets in a MFR model analysis (in this case, fit for persons, raters, and domains/items). Table 2 includes numeric fit statistics for each of the raters in the example data set. On average, the infit *MSE* and outfit *MSE* fit statistics were around 1.00 for the raters included in the analysis. However, Rater 2, Rater 10, and Rater 20 have infit *MSE* and outfit *MSE* statistics that are well above the upper critical value of 1.40. These relatively high fit statistics indicate that these three raters gave unexpected ratings frequently. Several researchers describe the rating patterns that correspond to these higher-than-expected fit statistics as "noisy" ratings (e.g., Engelhard, 1994a). On the other hand, although none of the illustrative raters have infit *MSE* and outfit *MSE* statistics below the lower critical value of .60, Rater 17's fit statistics are quite low (infit *MSE* = .61, outfit *MSE* = .62). These values indicate that Rater 17 gave overly consistent ratings. Several researchers describe the rating patterns that correspond to these lower-than-expected fit statistics as "muted" ratings (e.g., Engelhard, 1994a).

Graphical Fit Displays

The second indicator of data-to-model fit based on the MFR-PC model is *graphical fit displays*. Graphical displays of fit to the Rasch model are used to illustrate the direction and magnitude of residuals associated with individual elements within each facet of the model. The interpretive question for graphical fit displays is as follows: What is the pattern of residuals (deviations of observations from model expectation)

associated with a given element of a given facet? Compared to numeric fit statistics, researchers and practitioners use graphical fit displays relatively less frequently. However, these displays are useful for gauging the impact of deviations between empirical observations and model expectations in that they allow researchers to explore data-to-model misfit in terms of the direction and magnitude of residuals. Two particularly useful graphical fit displays for evaluating fit to the Rasch model include plots of standardized residuals and plots of expected and empirical response functions. As with the illustration of numeric fit indices, the focus is on rater fit in the graphical analysis of data-to-model fit. However, it is also possible, and important, to use graphical fit displays to evaluate data-to-model fit for each of the facets included a MFR model analysis.

Standardized residual plots. Figure 3 includes standardized residual plots for three raters who exhibited different levels of data-to-model fit based on the numeric fit statistics. Persons are shown along the *x*-axis, and standardized versions of residuals are shown on the *y*-axis. Because the residuals are standardized, values outside the range of ± 2.00 are generally considered significant deviations from model expectations. Standardized residuals that exceed $+2.00$ indicate that a rater gave a higher-than-expected rating, and standardized residuals that fall below -2.00 indicate that a rater gave a lower-than-expected rating. Values around .00 indicate that a rater's rating matched the model expectations perfectly. Because the MFR-PC model is probabilistic, some variation from model expectations is expected.

The first plot shows standardized residuals for Rater 18, who exhibited acceptable data-to-model fit according to infit *MSE* and outfit *MSE* (infit $MSE = 1.07$, outfit $MSE = 1.09$). The plot indicates that, overall, Rater 18 showed acceptable data-to-model fit, because the residuals are generally between ± 2.00, with one exception. The second plot shows standardized residuals for Rater 20, who exhibited noisy fit according to infit *MSE* and outfit *MSE* (infit $MSE = 1.57$, outfit $MSE = 1.54$). Compared to Rater 18, there is more variation in the standardized residuals for Rater 20, indicating more frequent and more extreme departures from model expectations. Finally, the last plot shows standardized residuals for Rater 17, who exhibited muted fit according to infit *MSE* and outfit *MSE* (infit $MSE = .61$, outfit $MSE = .62$). For this rater, the standardized residuals are less extreme and less varied compared to Rater 18 and Rater 20—indicating slightly less variation than expected by the model.

Expected-and-empirical response functions. Figure 4 includes expected-and-empirical rater response function (RRF) plots for the same three raters as in Figure 3. In these plots, the *x*-axis shows the logit-scale locations for persons, and the *y*-axis shows the rating scale. The dark, solid line shows the expected rater response function, which is a plot of the model-based relationship between persons locations and the expected rating, given perfect data-to-model fit. The dashed line shows the empirical (i.e., observed) relationship between persons locations and a rater's average rating for persons with a given location. Error bands are plotted using thin solid lines to indicate a 95% confidence interval around model expectations.

The first plot shows the expected-and-empirical RRF for Rater 18. For this rater, there are some notable deviations between the expected RRF and the empirical RRF, particularly for persons with relatively low scores; however, the deviations

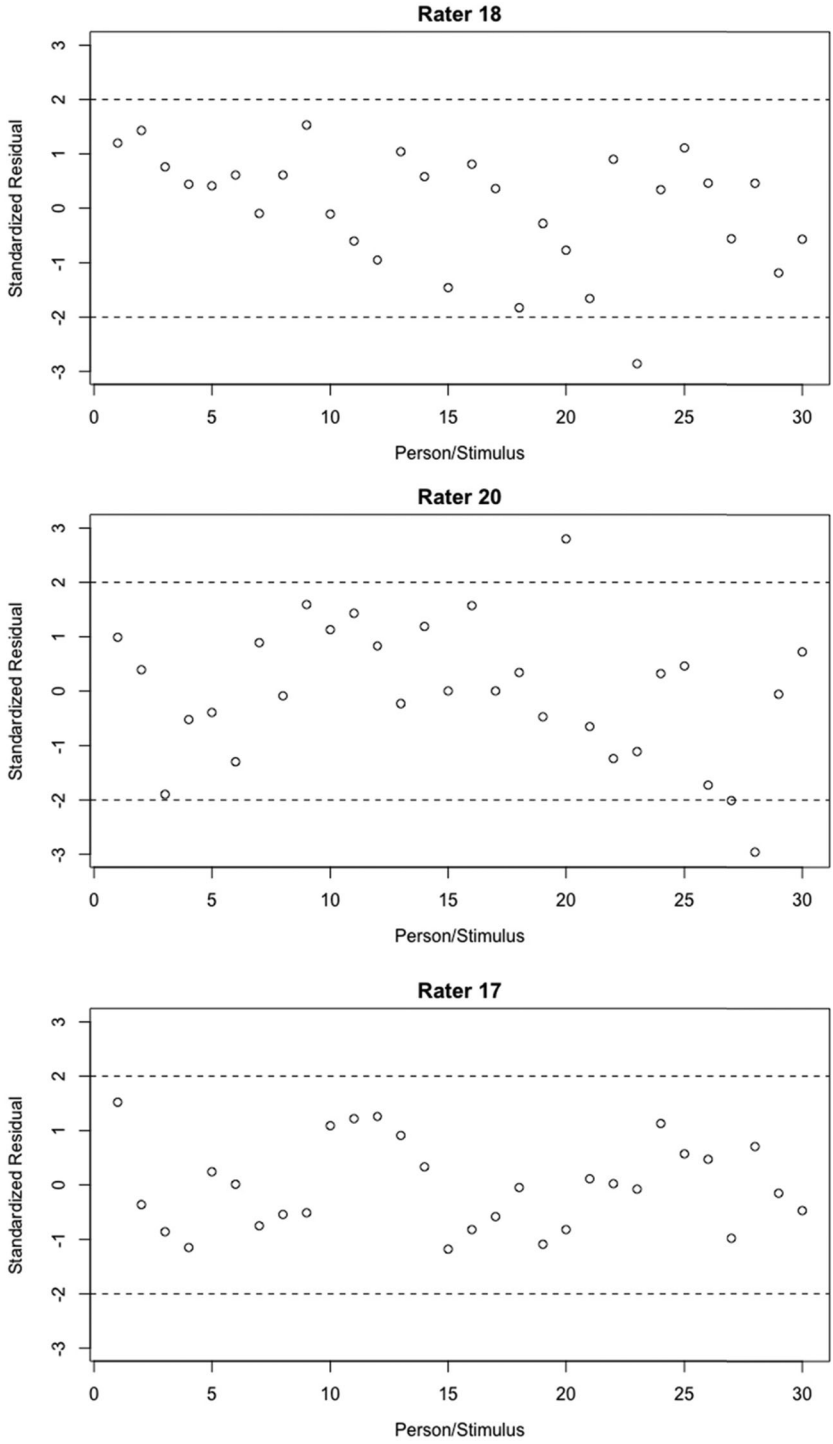

Figure 3. Rater residual plots.

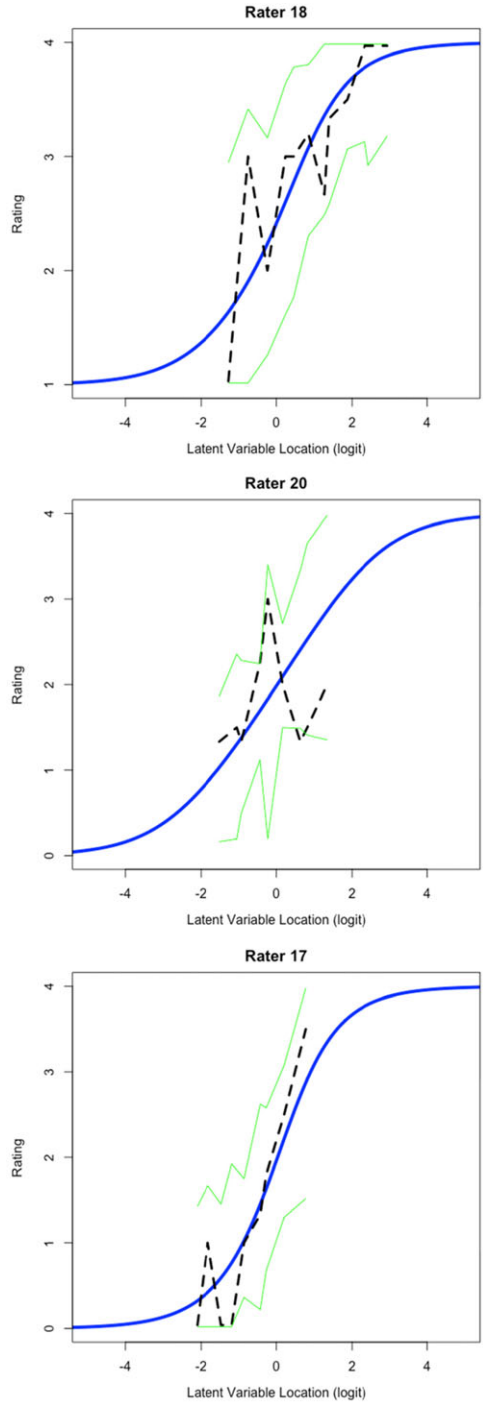

Figure 4. Expected and empirical rater response functions. (Color figure can be viewed at wileyonlinelibrary.com)

are within the error bands. Next, the expected-and-empirical RRF for Rater 20 shows more-frequent and more-extreme deviations between the expected RRF and the empirical RRF. For this rater, the overall shape of the empirical RRF is quite different from the expected RRF. Furthermore, the plot indicates that Rater 20 gave higher-than-expected ratings for persons with relatively low scores (positive residuals between around -2 and .5 logits on the x-axis), and lower-than expected ratings for persons with relatively high scores (negative residuals for persons with judged achievement greater than about 1.00 logits). Finally, Figure 4 includes the expected-and-empirical RRF plot for Rater 17. For this rater, the empirical RRF follows the general shape of the expected RRF, and there are very few notable deviations between the two response functions—indicating close alignment between observed and expected ratings.

Model 3: Interaction Model

The third model in this framework is an interaction model. Interaction model analyses allow researchers and practitioners to investigate systematic, construct-irrelevant variability with the context of an evaluative system. This type of analysis allows researchers to empirically investigate the degree to which the logit-scale locations of elements within a given facet, such as raters, are invariant over the elements of another facet, such as domains/items or persons subgroups.

The procedure for conducting an interaction analysis in the context of Rasch measurement theory involves adding an interaction parameter to the measurement model (Equation 1) that reflects a combination of facets between which an interaction is of interest. Then, an omnibus test is used to evaluate whether the overall locations of the elements within one facet are consistent over the elements of the second facet. If a significant interaction effect is observed, one can then use differential facet functioning (DFF) analyses to identify the individual elements within each of the facets for which a lack of invariance occurred.

Using the same structure as Table 1, Table 3 includes a summary of several outcomes from an interaction model analysis that are relevant in the context of a rater-mediated assessment. A brief illustration of an MFR-PC interaction model analysis is presented in the following section that includes each of the sources of evidence included in Table 3.

Illustration

In this section, the illustration from the presentation of Model 2: Measurement Model is continued. Specifically, the same simulated data set is used to investigate the interaction between rater severity and the judged difficulty of domains/items using the indices summarized in Table 3. Evidence of a statistically significant interaction between rater severity and domain/item difficulty indicates that there is not a consistent interpretation of the difficulty of domains/items over all of the raters—in other words, the difficulty of domains/items depends on the rater who scores a person. Put another way, when there is evidence of a statistically significant interaction between rater severity and domain/item difficulty, rater severity ordering changes depending on the domain/item.

Table 3
Interaction Model Evidence and Interpretive Questions

Evidence Category	Source of Evidence	Interpretive Question
Interaction parameter	Omnibus test for interaction effect	Is there evidence that the elements of one facet are ordered the same way over elements of a second facet?
Differential facet functioning	*Z* statistics for pairwise comparisons	To what extent are there differences in the calibrations of individual elements of one facet when they are calculated separately using individual elements of another facet?
	Graphical displays of interaction effects	What is the direction and magnitude of differences between observed locations and expected locations for a pair of elements?

The following interaction model demonstrates the interaction between rater severity and domain/item difficulty:

$$\ln\left[\frac{P_{nmi(x=k)}}{P_{nmi(x=k-1)}}\right] = (\theta_n - \lambda_m - \delta_i - \tau_{mk}) - \lambda_m\delta_i, \tag{4}$$

where all of the terms in Equation 4 are defined as they were in Equation 1, and $\lambda_m\delta_i$ is the interaction between rater severity and judged domain/item difficulty. The interaction model analyses was conducted using the *FACETS* software (Linacre, 2015).

Interaction Parameter

The first source of evidence for interpreting the results from the interaction model is an omnibus test for the interaction effect. As shown in Table 3, this omnibus test is used to address the following interpretive question: Is there evidence that the elements of one facet are ordered the same way over elements of a second facet? In the example data, the omnibus test for the interaction between rater severity and judged domain/item difficulty is examined. Accordingly, the interpretive question for the illustrative analysis can be restated as follows: Is there evidence that the individual raters are ordered the same way over each of the domains/items?

In the example data set, the omnibus test for the interaction between raters and domains/items was statistically significant: $\chi^2(60) = 104.7$, $p < .01$. This result suggests that, overall, the individual raters did not have a consistent interpretation of the difficulty of ordering of the domains/items in the assessment context.

Table 4
Z-Statistics for Raters With Statistically Significant Pairwise Comparisons

Rater	Domain	Observed Average Rating	Expected Average Rating	*Z* Statistic*
2	2	24.00	15.11	3.13
	3	15.00	22.95	−2.84
10	1	29.00	18.74	3.09
	3	12.00	23.79	−3.60
20	2	21.00	15.01	2.38
	3	15.00	21.68	−2.58

Note. * Statistically significant at $p < .05$, degrees of freedom = 9.

Differential Facet Functioning

Given the statistically significant interaction parameter, additional analyses are needed in order to identify individual raters whose judgments of the domain/item difficulties are significantly different from the overall calibration of the domains/items. As shown in Table 3, researchers can use DFF analyses to examine interaction effects at the level of individual elements within facets. Specifically, one can use pairwise comparisons and graphical displays to identify pairs of elements of two facets for which there is evidence of a lack of invariance. In the example dataset, these analyses will include pairwise combinations of individual raters with individual domains/items. As the raters were the objects of investigation, a differential rater functioning (DRF) analysis was conducted.

Z-Statistics for Pairwise Comparisons

The first source of evidence of DRF is *Z*-statistics for the pairwise comparisons of the elements of one facet with the elements of a second facet. As shown in Table 3, one can use these pairwise comparisons to address the following interpretive question: To what extent are there differences in the calibrations of individual elements of one facet when they are calculated separately using individual elements of another facet? In the illustrative analysis, this question can be restated as follows: To what extent are there differences in the difficulty of each of the domains/items when they are calculated separately from each of the raters?

Table 4 includes *Z*-statistics for the pairwise comparisons of individual raters and individual domains/items. For brevity, only the comparisons with statistically significant results are included. For each comparison, positive *Z*-statistics indicate that a rater scored persons higher than expected on a given domain, and negative *Z*-statistics indicate that a rater scored persons lower than expected on a given domain. As shown in Table 4, there were three raters with statistically significant pairwise interaction statistics: Rater 2, Rater 10, and Rater 20. Specifically, Rater 2 gave higher-than-expected ratings on Domain 2 ($Z = 3.13$, $p < .05$) and gave lower-than-expected ratings on Domain 3 ($Z = -2.84$, $p < 0.05$). Likewise, Rater 10 gave higher-than-expected ratings on Domain 1 ($Z = 3.09$, $p < 0.05$) and lower-than-expected ratings on Domain 3 ($Z = -3.60$, $p < .05$). Finally,

Rater 20 gave higher-than-expected ratings on Domain 2 ($Z = 2.38$, $p < .05$) and lower-than-expected ratings on Domain 3 ($Z = -2.58$, $p < .05$).

Graphical Displays of Interaction Effects

The second source of evidence of DRF is graphical displays of interaction effects. These graphical displays include essentially the same information as the *Z*-statistics. However, similar to the graphical displays of data-to-model fit, these displays help researchers examine the direction and magnitude of interactions between facets. As shown in Table 3, one can use graphical displays of interaction effects to address the following interpretive question: What is the direction and magnitude of differences between observed locations and expected locations for a pair of elements?

Figure 5 includes a graphical display of the results from the DFF analysis. Specifically, Figure 5 includes separate plots for each of the three domains in which the pairwise *Z*-statistic is shown for each of the 20 raters specific to the individual domain. The results in Figure 5 correspond to the pairwise statistics in Table 4. Specifically, positive values indicate that a rater gave higher-than-expected ratings on a particular domain, and negative values indicate that a rater gave lower-than-expected ratings on a particular domain. Asterisk-plotting symbols are used to indicate statistically significant *Z*-statistics, and circle-plotting symbols to indicate non-significant *Z*-statistics. As indicated in Table 4, there are three raters with statistically significant pairwise interaction statistics: Rater 2, Rater 10, and Rater 20. Inspection of the graphical displays in Figure 5 highlights the magnitude and direction of the higher-than-expected or lower-than-expected ratings in each domain.

Outcome Considerations

In many cases of educational research, one important purpose is to make inferences that generalize to a particular population. In educational research, researchers often use various classes of inferential *statistics* to make these generalizations. The framework discussed in this article is concerned with *measurement*, and should be conducted as an a priori investigation into the validity of the raters, items, and persons under investigation. The purpose of conducting the data transformations and data investigations above is to gain better insight into the collected data, and more importantly, to collect both construct validity (e.g., item-centered arguments for the quality of data collected) and predictive validity (e.g., person-centered arguments for the quality of data collected) evidence to support the interpretation and use of the assessment data.

It is important to note that under most statistical operations using rating-scale data, linearity of the data is an assumption. However, in many cases when statistical procedures are conducted on observed, observed data as collected in Model I (the observation model), the assumption of linearity is violated as observed score data is ordinal in nature. The models presented here take the ordinal, observed score data and construct linear, objective measures of known precision (i.e., standard error estimations for all elements within all facets) and quality (i.e., fit index estimations for all elements within all facets) in a manner that is free from the level of persons estimates, rater severity estimates, and item difficulty estimates.

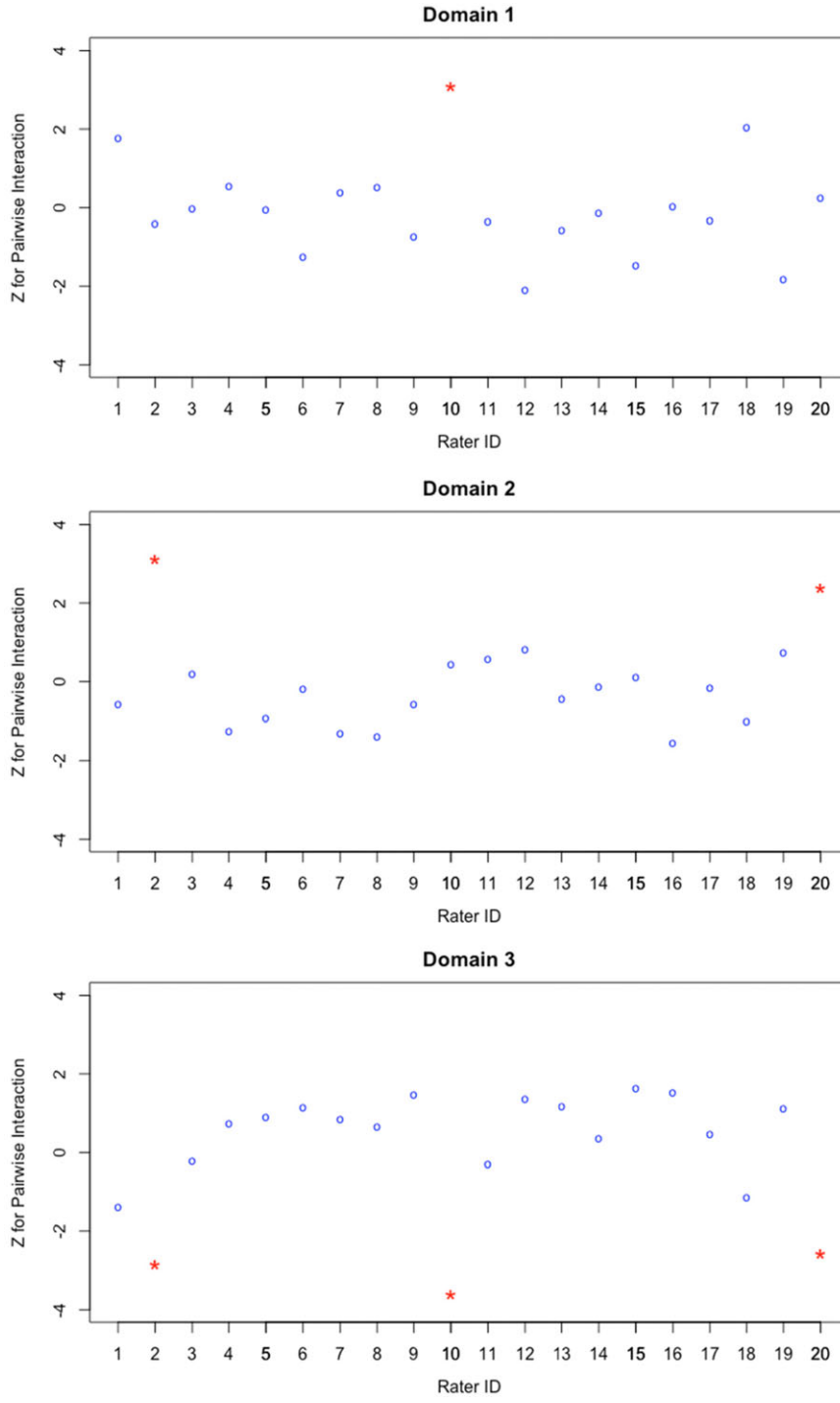

Figure 5. Plots of pairwise interaction statistics. (Color figure can be viewed at wileyonlinelibrary.com)

In the case that facets and elements demonstrate good model fit for the expectations of the models and do not demonstrate evidence of DFF, it is then appropriate to make inferential generalizations. As one example, the constructed measures can be used as dependent variables within the context of generalized linear models for examining the degree to which there are differences in group means (e.g., analysis of variance) or predicting scores based upon linear relationships (i.e., regression).

In the case that evidence of DFF, or more specifically, DRF exists, it should prompt the researcher to reexamine the construct, as operationally defined by the domains/items, and qualitatively investigate for substantive meaning based upon the instances of statistically significant DFF/DRF. Fundamentally, when DFF/DRF exists, a different construct is mapped to the measurement system. In these instances, an understanding of the new construct may provide new insights into the construct being defined by the researcher.

In some instances when there is evidence of unacceptable data-to-model fit, it may be warranted to remove misfit elements; however, this is a qualitative decision based upon extremes in the data (Linacre, 2010), flexibility in the range of fit criteria based upon the intended use and/or context (Wright & Linacre, 1994), or whether the research determines that the data accurately depicts a useful representation of the data structure (Adams & Wright, 1994). In these cases, it is important for researchers to re-examine their initial theory of the construct relevant factors included in the model: Do the domains/items appropriately reflect the construct being investigated? Are the demographics, geographics, or psychographics of the sample pool of raters appropriate for the study? Is the pool of persons most appropriate or representative of the construct being investigated?

Conclusion

The models presented enhance conceptualization of educational research processes in the context of rater-mediated assessments. It is recommended that the considerations of measurement in the research process be more integrated as part of validation processes. In particular, when raters are used to facilitate the data collection process, it is important to consider variability associated with them as a construct-relevant factor that can be controlled for, or directly investigated, when necessary and appropriate. In order for the considerations provided in this framework to be effective, implementation into research must be guided by careful theoretical analyses throughout the data collection processes and grounded in substantive interpretation. This framework contains explicit knowledge integration and extraction throughout the research process. We hope the future implementation of this framework better supports the exploration and knowledge gained of research processes mediated by raters.

References

Adams, R., & Wright, B. D. (1994). When does misfit make a difference? In M. Wilson (Ed.), *Objective measurement: Theory into practice II* (pp. 244–270). Norwood, NJ: Ablex.

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

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York, NY: Routledge.

Braun, H. I., Jackson, D. N., & Wiley, D. E. (2002). *The role of constructs in psychological and educational measurement*. New York, NY: Routledge.

Brunswik, E. (1952). *The conceptual framework of psychology*. Chicago, IL: University of Chicago Press.

DeAyala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: The Guilford Press.

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

Engelhard G., Jr. (1994b). Historical views of the concept of invariance in measurement theory. In M. Wilson (Ed.), *Objective measurement: Theory into practice: Volume 2* (pp. 73–99). Norwood, NJ: Ablex.

Engelhard G., Jr. (1997). Constructing rater and task banks for performance assessments. *Journal of Outcome Measurement*, *1*, 19–33.

Engelhard, G., Jr. (2008). Historical perspectives on invariant measurement: Guttman, Rasch, and Mokken. *Measurement: Interdisciplinary Research and Perspectives*, *6*(3), 155–189.

Engelhard, G., Jr. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. New York, NY: Routledge.

Hammond, K. R. (1955). Probabilistic functioning and the clinical method. *Psychological Review*, *62*, 255–262.

Kane, M. (1992). An argument-based approach to validation. *Psychological Bulletin*, *112*, 527–535.

Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*, 319–342.

Kane, M., Crooks, T. J., & Cohen, A. S. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, *18*(2), 5–17.

Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin*, *134*, 404–426.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago, IL: MESA Press.

Linacre, J. M. (2000). Comparing and choosing between "partial credit models" (PCM) and rating scale models (RSM). *Rasch Measurement Transactions*, *14*(3), 768.

Linacre, J. M. (2010). When to stop removing items and persons in Rasch analysis? *Rasch Measurement Transactions*, *23*(4), 1241.

Linacre, J. M. (2015). *Facets Rasch measurement*. Chicago, IL: Winsteps.com.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*, 635–694.

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

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan

Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests* (Expanded edition, 1980). Chicago, IL: University of Chicago Press.

Smith, R. M. (2004). Fit analysis in latent trait models. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 73–92). Maple Grove, MN: JAM Press.

Wind, S. A., Engelhard, G. E., & Wesolowski, B. C. (2016). Exploring the effects of rating designs and rater fit on achievement estimates within the context of music performance assessment. *Educational Assessment*, *21*(4), 278–299.

Wolfe, E. W. (2013). A bootstrap approach to evaluating person and item fit to the Rasch model. *Journal of Applied Measurement*, *14*(1), 1–9.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370.

Wu, M., & Adams, R. J. (2013). Properties of Rasch residual fit statistics. *Journal of Applied Measurement*, *14*, 339–355.

Zieky, M. J. (2016). Developing fair tests. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 81–99). New York, NY: Routledge.

Authors

BRIAN C. WESOLOWSKI is Associate Professor of Music Education at the University of Georgia, Hugh Hodgson School of Music, 250 River Road, Athens, GA 30602; bwes@uga.edu. His primary research interests include rater behavior, scale development, and policy of educational assessment in music.

STEFANIE A. WIND is Assistant Professor of Educational Measurement at the University of Alabama, Box 270831, Tuscaloosa, AL 35487; stefanie.wind@ua.edu. Her primary research interests include the exploration of methodological issues in the field of educational measurement, with emphases on methods related to rater-mediated assessments, rating scales, Rasch models and item response theory models, and nonparametric item response theory, as well as applications of these methods to substantive areas related to education.