Exploring Rating Quality in Rater-Mediated Assessments Using Mokken Scale Analysis

Educational and Psychological
Measurement

2016, Vol. 76(4) 685–706

© The Author(s) 2015

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164415604704

epm.sagepub.com

Stefanie A. Wind¹ and George Engelhard Jr.²

Abstract

Mokken scale analysis is a probabilistic nonparametric approach that offers statistical and graphical tools for evaluating the quality of social science measurement without placing potentially inappropriate restrictions on the structure of a data set. In particular, Mokken scaling provides a useful method for evaluating important measurement properties, such as invariance, in contexts where response processes are not well understood. Because rater-mediated assessments involve complex interactions among many variables, including assessment contexts, student artifacts, rubrics, individual rater characteristics, and others, rater-assigned scores are suitable candidates for Mokken scale analysis. The purposes of this study are to describe a suite of indices that can be used to explore the psychometric quality of data from rater-mediated assessments and to illustrate the substantive interpretation of Mokken-based statistics and displays in this context. Techniques that are commonly used in polytomous applications of Mokken scaling are adapted for use with rater-mediated assessments, with a focus on the substantive interpretation related to individual raters. Overall, the findings suggest that indices of rater monotonicity, rater scalability, and invariant rater ordering based on Mokken scaling provide diagnostic information at the level of individual raters related to the requirements for invariant measurement. These Mokken-based indices serve as an additional suite of diagnostic tools for exploring the quality of data from rater-mediated assessments that can supplement rating quality indices based on parametric models.

Keywords

rater-mediated assessment, Mokken scaling, invariant measurement, nonparametric item response theory

¹The University of Alabama, Tuscaloosa, AL, USA

²The University of Georgia, Athens, GA, USA

Corresponding Author:

Stefanie A. Wind, The University of Alabama, Box 870231, Tuscaloosa, AL 35487, USA.

Email: swind@ua.edu

Mokken scale analysis (MSA; Mokken, 1971) is a nonparametric approach to item response theory (IRT) that has been applied as a technique for evaluating the quality of data obtained from scales in political science, health outcomes, and affective domains (Meijer, Tendeiro, & Wanders, 2015; Sijtsma & Molenaar, 2002). Mokken's scaling models are described as nonparametric because the underlying item response functions (IRFs) are not restricted to specific shapes as long as basic ordering requirements are satisfied. As a result, Mokken models are often viewed as more relaxed, or more general, than parametric IRT models in terms of assumptions about the relationship between the latent variable and the probability for a response (Sijtsma & Hemker, 2000; Sijtsma & Molenaar, 2002). As pointed out by Engelhard (2008), Mokken scaling offers a probabilistic version of Guttman scaling (a deterministic model; Guttman, 1950) that is also closely related to the requirements of invariant measurement as represented in Rasch measurement theory (Rasch, 1960/1980).

Although Mokken's approach to scaling is less strict in terms of model requirements than parametric IRT, the models within this framework are still based on important requirements related to unidimensionality, monotonicity, and invariant ordering. As a result, Mokken scaling offers statistical and graphical tools for evaluating the quality of social science measurement in terms of fundamental measurement properties without placing potentially inappropriate restrictions on the structure of a data set (Meijer et al., 2015; Sijtsma & Meijer, 2007).

Nonparametric item response theory (NIRT) in general, and Mokken scaling in particular, have been recognized as useful tools for evaluating measurement procedures in contexts in which response processes are not well understood, such as personality and affective domains (Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001; Meijer & Baneke, 2004; Meijer et al., 2015; Reise & Waller, 2003). It has not been widely recognized that educational assessments in which raters judge the quality of student work according to a rubric *also* involve response processes that are not well understood—and thus may be suitable candidates for analysis using MSA. In particular, rater-mediated assessments involve complex response processes characterized by interactions among assessment contexts, student artifacts, rubrics, individual rater characteristics, and other variables. Because these complex interactions are not well understood, uncertainty related to the degree to which a rater is able to “translate” their perception of student achievement to a rating scale designed to represent a construct is a persistent theme in research on rater-mediated assessments. Despite a large body of research on many potential contributing factors to rating quality, including the impact of rater background characteristics (Lumley & McNamara, 1995; Pula & Huot, 1993), rater training and feedback procedures (Elder, Knoch, Barkhuizen, & von Randow, 2005; Knoch, 2011; Weigle, 1998), and scoring criteria (Clauser, 2000), advances in these areas are not yet sufficient to fully understand rating processes (Hamp-Lyons, 2007, 2011).

In response to these concerns, a variety of rating quality indicators of rater agreement, errors, systematic biases, and accuracy have been proposed based on parametric IRT models such as the Many-Facet Rasch model (e.g., Congdon &

McQueen, 2000; Engelhard, 2002; Wind & Engelhard, 2012; Wind & Engelhard, 2013; Wolfe, 2009; Wolfe & McVay, 2012). In particular, within the framework of Rasch measurement theory, researchers have emphasized the importance of examining rating quality at the individual rater level, rather than at the overall group level (e.g., as in Generalizability Theory; Shavelson & Webb, 1991), in order to inform rater training, rater monitoring, and the interpretation of ratings (Engelhard, 1994; Linacre, 1996; Lynch & McNamara, 1998; Stahl & Lunz, 1992; Sudweeks, Reeve, & Bradshaw, 2005).

Parametric procedures for evaluating rating quality involve the transformation of ratings to interval-level scales and the imposition of strict requirements related to the shape of the IRF. Because it offers a less-restrictive approach based on basic measurement properties, Mokken scaling is a promising supplementary method for exploring data quality in the context of rater-mediated assessments that may be more appropriate for the ordinal-level ratings that result from complex rater decision-making processes.

Until recently, the use of Mokken scaling had not been explored as a method for monitoring the quality of ratings within the context of performance assessments. Recently, Wind (2014) demonstrated the application of polytomous versions of Mokken's (1971) scaling models to the context of educational performance assessments as a method for evaluating the quality of rating scale category functioning for holistic rating scales. Specifically, Wind (2014) proposed a suite of Mokken-based guidelines to evaluate the overall functioning of rating scales in terms of rating scale category ordering, category information, and model-data fit. The current study continues the application of MSA to the context of performance assessments by proposing and illustrating a suite of indices that can be used to describe and explore the psychometric quality of rating data, and to illustrate the substantive interpretation of Mokken-based statistics and displays for rater-mediated assessments. Whereas Wind (2014) explored the use of Mokken-based statistics and displays to evaluate the operational use of rating scale categories across overall groups of raters, the current study focuses on the use of Mokken scaling as a method for obtaining diagnostic information about rating quality at the level of *individual raters*.

Purpose

The purposes of this study are to describe a suite of indices that can be used to explore the psychometric quality of data from rater-mediated assessments and to illustrate the substantive interpretation of Mokken-based statistics and displays in this context. A case study using data from a large-scale rater-mediated writing assessment is used to illustrate the Mokken-based indices. Techniques that are commonly used in polytomous applications of Mokken scaling are illustrated in the new context, with a focus on the substantive interpretation related to individual raters.

Figure 1. Nonparametric item response function.

Mokken Scale Analysis

The theoretical framework for this study is MSA. MSA can be viewed as a series of steps that focus on identifying items with useful properties in order to construct meaningful scales. Mokken's models describe the probability for a response as a function that is governed by order restrictions, such that the IRF is only required to be nondecreasing in the latent variable (θ). For persons A and B whose latent-variable locations can be ordered such that $(\theta_A < \theta_B)$, the only restriction on the IRF is the requirement that the relative order of the conditional probabilities for a correct or positive response to an item holds across levels of the latent variable (θ):

$$P(X_{i=1}|\theta_A) \leq P(X_{i=1}|\theta_B), \quad (1)$$

where $P(X_{i=1}|\theta_A)$ and $P(X_{i=1}|\theta_B)$ represent the probability that a person at latent variable locations A and B provide a correct or positive response to Item i and Item j , respectively, given their latent variable locations. As long as they are nondecreasing, the IRFs associated with NIRT may take on a variety of shapes. Figure 1 illustrates an IRF that meets the ordering requirement in Equation 1.

Mokken's original scaling procedure includes two models based on the ordering requirement in Equation (1) that describe relationships among items, persons, and latent variables: (a) the Monotone Homogeneity (MH) model and (b) the Double Monotonicity (DM) model. The MH model is based on three assumptions:

- *Monotonicity*—The probability that a person will correctly respond to an item increases as their location on the latent variable increases.

- *Conditional Independence*—Responses to an item are not influenced by responses to any other item, after controlling for the latent variable.
- *Unidimensionality*—Item responses reflect evidence of a single latent variable.

The DM model is based on the same underlying assumptions as the MH model, with one additional assumption:

- *Nonintersecting IRFs*—Item difficulty ordering is consistent across all levels of the latent variable.

For dichotomous items, adherence to the assumption of nonintersecting IRFs results in an *Invariant Item Ordering* (IIO). IIO implies that the ordering of items in terms of difficulty does not depend on which students are used for the comparison. Similarly, IIO implies that the ordering of students does not depend on the particular item by which the students are ordered. With the exception of ties, IIO implies that the expected order of items is identical for each subgroup of persons, and the expected ordering of persons on the latent variable is the same for each item. Ligtoet, van der Ark, te Marvelde, and Sijtsma (2010) described the usefulness of this property for facilitating the “interpretation and comparability of respondents’ measurement results” in several contexts (p. 578), and observed that evidence of IIO is necessary for the interpretation of individual person and item scores. In their words:

Any set of items can be ordered by means of item mean scores, but whether such an ordering also holds for individuals has to be ascertained by means of empirical research. Only when the set of items has an IIO, can their cumulative structure be assumed to be valid at the lower aggregation level. (p. 579)

The IIO property that characterizes Mokken’s DM model is congruent with the requirements for invariant measurement (Engelhard, 2013; Wright & Stone, 1979). In essence, invariant measurement is based on the idea that measures of phenomena of interest must not be affected by irrelevant characteristics of the process used to collect those measures. Invariant measurement can be summarized as adherence to two requirements: (a) *Item-invariant measurement of persons*—The measurement of persons must be independent of the particular items they happen to take; and (b) *Person-invariant calibration of items*—The calibration of items must be independent of the particular persons used to calibrate them (Engelhard, 2013). If invariant measurement is not achieved within a measurement system, persons will appear to possess “more” of the trait being measured on tests that are composed of easier items, and persons will appear to possess “less” of the trait being measured on tests that are composed of harder items. Because of these invariant-ordering properties, the

DM model has been described as a nonparametric or ordinal version of the Rasch model (Meijer, Sijtsma, & Smid, 1990; van Schuur, 2003).

Because it is based on the principles of invariant measurement, MSA provides a coherent measurement framework for examining the quality of social science measurement procedures. Furthermore, because nonparametric models are based on less-strict underlying assumptions regarding the relationship between observed ratings and latent traits, MSA can be used for analyses of ordinal data with less room for improper interpretations that may result from violating requirements with the use of a parametric model, such as the Rasch model.

Method

To demonstrate the use of MSA as a method for evaluating the quality of ratings, a suite of indices of rating quality based on commonly used Mokken-based indices of measurement quality is proposed and illustrated using a data set that has been previously analyzed by several researchers within the context of parametric IRT, with an emphasis on the substantive interpretation in the context of rater-mediated assessments.

Data Source

The illustrative analysis in this study uses data that were previously examined by Andrich (2010), Gyagenda and Engelhard (2009), and Wind and Engelhard (2012). The data come from the Georgia High School Writing Test, and include scores from 365 eighth-grade students whose persuasive essays were rated by 20 operational raters. A rating between 1 and 4 was assigned within four separate domains: Conventions, Organization, Sentence Formation, and Style. All raters scored the entire set of 365 essays, such that the rating design was fully connected (Engelhard, 1997). The ratings were recoded to 0 = *low*; 3 = *high* prior to analysis.

Data Analysis

In this study, adaptations of the polytomous formulations of Mokken's (1971) MH and DM models (Molenaar, 1982, 1997) are adapted in order to arrive at a suite of Mokken-based indicators of data quality for rater-mediated assessments. Similar to parametric polytomous IRT models, polytomous versions of the MH and DM models are based on a conceptualization of rating scale categories as a series of dichotomous "steps," such that $m - 1$ separate response functions are specified for a rating scale that has m unique categories. Within the framework of MSA, item step response functions (ISRFs; i.e., category response functions) are defined as the cumulative probability for a rating within a category [$P(X > k | \theta)$]. Mokken ISRFs are constrained by the same nondecreasing order restriction as the dichotomous IRFs (Equation 1). As in the dichotomous models, the shape of Mokken-based ISRFs is not constrained

Figure 2. Nonparametric category response functions.

Table 1. Alignment Between Mokken Model Assumptions and Mokken Rating Quality Indicators.

Assumptions	Double monotonicity for ratings (DM-R)	Monotone homogeneity for ratings (MH-R)	Model-based indicators
Monotonicity	✓	✓	(A) Rater monotonicity
Conditional independence	✓	✓	(B) Rater scalability coefficients
Unidimensionality	✓	✓	(A) Rater monotonicity; (B) Rater scalability
Nonintersecting response functions	✓		(C) Invariant rater ordering

to a particular form as long as ordering requirements are met. A set of Mokken ISRfS for a four-category rating scale item is illustrated in Figure 2.

In this study, the polytomous MH model is extended to the Monotone Homogeneity for Ratings (MH-R) model, and the polytomous DM model is extended to the Double Monotonicity for Ratings (DM-R) model. The indices presented in this study are based on the typical sequential procedures used to examine fit to the MH and DM models that involve evaluating data quality based on three categories of indices: (A) Monotonicity, (B) Scalability, and (C) Invariant ordering (Meijer et al., 2015; Sijtsma & Molenaar, 2002). Table 1 summarizes the alignment between these three categories and the Mokken model assumptions described above.

In this section, rating quality indices based on the MH-R and DM-R models are presented theoretically. Then, results are presented using the illustrative data set to

illustrate the empirical application of MSA to examining data quality at the individual rater level using MSA. In order to illustrate the use of MSA to examine rating quality, the two Mokken models for ratings (described below) are applied separately to each of the four rubric domains (Conventions, Organization, Sentence Formation, and Style), so that the analyses can be viewed as four applications of the MSA techniques for exploring rating quality for holistic ratings. Wind (2015) presents the use of MSA to examine rating quality for analytic ratings, where multiple domains can be considered together as a single scale. The illustrative analyses were conducted with the *mokken* package for the *R* computer program (*R* Development Core Team, 2015; van der Ark, 2007, 2012).

Monotone Homogeneity for Ratings Model

The MH-R model is an application of the polytomous MH model (Molenaar, 1982, 1997) to the context of polytomous rater-assigned scores. In traditional MSA analyses, indicators of monotonicity are used to check the MH model assumption of monotonicity, and have also been used to check the assumption of unidimensionality (Meijer et al., 2015; Sijtsma & Molenaar, 2002). The next section describes two major indices of data quality for rater-mediated assessments based on the MH-R model: Rater monotonicity and Rater scalability.

Rater Monotonicity. The first indicator of data quality based on the MH-R model is rater monotonicity. For dichotomous items, monotonicity suggests that the probability that a student correctly responds to an item is nondecreasing across increasing locations on the latent variable (θ). In the context of rater-mediated assessments with polytomous ratings, monotonicity implies that the cumulative probability for each rating scale category [$P(X \geq k)$] is nondecreasing across increasing levels of student achievement. A nonparametric approximation of the student ability parameter is obtained by calculating unweighted sum scores (X_+) for each student across an entire set of items. Mokken (1971) demonstrated that an ordering of students according to X_+ serves as an estimate of their ordering according to θ (Molenaar, 1982; van der Ark, 2005). As a type of purification for checking NIRT assumptions for an item of interest, Junker (1993) proposed the use of the restscores (R), which is the sum score minus the score on the item (or in this case, the rating from the rater) of interest.

The *mokken* package (van der Ark, 2007, 2012) offers a simple method for evaluating monotonicity based on restscores. In the context of a rater-mediated assessment, restscores are created using student total scores across a group of raters minus the rating assigned by a rater of interest. Using restscores groups,¹ monotonicity is investigated in two ways. First, monotonicity is examined in terms of average ratings within restscores groups. For each rater, average ratings are plotted as a function of restscores, such that an estimate of a rater response function is constructed. Figure 3 is an example of a diagnostic rater response function plot that can be used to investigate monotonicity at the overall rater level. In this figure, student restscores are plotted

Figure 3. Rater monotonicity at the overall rater level.

along the x -axis, and average ratings on a 4-point rating scale (0 = *low*; 3 = *high*) are plotted along the y -axis. The figure illustrates evidence of monotonicity for a single rater, because average ratings are nondecreasing as restscores increase.

Monotonicity can also be examined at the rating scale category level. Based on the conceptualization of polytomous ratings as a series of dichotomous steps, monotonicity is examined for the $(m - 1)$ meaningful item step response functions by calculating the cumulative probability for a rating in a given category within each restscore group. If a rater demonstrates monotonicity, the cumulative probability for ratings in each category will be nondecreasing as restscores increase. Figure 4 illustrates a diagnostic plot for rater monotonicity at the level of rating scale categories using a rating scale with four categories (0 = *low*; 3 = *high*). Student restscores are plotted along the x -axis, and the y -axis represents the probability for a rating in Category k or higher, given a restscore value [$P(X \geq k | R = r)$]. The highest line represents the probability that a student in a restscore group receives a rating in Category “1” or higher [$P(X \geq 1)$]. Likewise, the second highest line represents the probability for a rating of “2” or higher [$P(X \geq 2)$], and the lowest line represents the probability for a rating of “3” or higher [$P(X \geq 3)$].

In addition to graphical procedures, monotonicity can also be checked using statistical hypothesis tests. Specifically, the null hypothesis that the expected average ratings are equal between two adjacent restscore groups is examined against the alternative hypothesis that the expected average rating is lower in the group with a higher restscore, which is a violation of monotonicity.

Essentially, the evaluation of rater monotonicity involves checking the match between the relative ordering of a set of performances for a rater of interest and the

Figure 4. Rater monotonicity within rating scale categories.

relative ordering of the performances based on the other raters, as defined by the restscore groups. As a result, evidence of monotonicity suggests adherence to the first requirement for rater-invariant measurement: students are ordered the same way across the group of raters. When monotonicity is observed for a group of raters, the interpretation of student achievement ordering is invariant across a group of raters.

Rater Scalability. The second category of Mokken-based indices of data quality is *scalability*. In traditional MSA applications, scalability coefficients are used to examine the impact of Guttman errors on scale functioning. Specifically, Mokken (1971) proposed a set of scalability coefficients based on Loevinger's (1948) H coefficient that can be used to examine scalability for pairs of items (H_{ij}), individual items (H_i), and the overall set of items in a survey or test (H).

For dichotomous items, deviations from a Guttman ordering are identified by determining the overall difficulty ordering of a set of items based on the proportion of students who succeed on an item, and discrepancies with the overall ordering are identified by examining the relative ordering of items within all possible item pairs. After the frequency of Guttman errors is identified, the errors are weighted by the expected cell frequency that would occur given marginal independence. Finally, the ratio of observed errors to the frequency of expected errors is calculated. The item pair scalability coefficient (H_{ij}) is calculated as one minus this observed-to-expected error ratio, and it can be stated as

$$H_{ij} = 1 - \frac{F_{ij}}{E_{ij}}, \quad (2)$$

where H_{ij} is the scalability of the item pair consisting of Item i and Item j , F_{ij} is the frequency of observed Guttman errors between Item i and Item j , and E_{ij} is the expected frequency of Guttman errors between Item i and Item j , given marginal independence of the two items.

When the MH model holds, $0.00 \leq H \leq 1.00$, where a value of 1.00 indicates a Guttman scalogram pattern. Sijtsma and Molenaar (2002) demonstrate a derivation of H coefficients as a ratio of the observed correlation between two items to the highest possible correlation, given the marginal distributions of the two items. Equations for calculating the scalability of individual items and sets of items using the Guttman error method and the covariance method are provided in Sijtsma and Molenaar (2002).

Mokken (1971) proposed a minimum value of $H_i = 0.30$ to identify items that contribute to a meaningful ordering of persons in terms of a latent variable. Most values observed in real data have an approximate range of $0.30 \leq H \leq 0.60$ (Mokken, 1997; Sijtsma, Meijer, & van der Ark, 2011). It is common practice within MSA to apply rule-of-thumb critical values for the H coefficient in order to evaluate the quality of a scale (Mokken, 1971; Molenaar & Sijtsma, 2000). Typically, the following criteria are applied: $H \geq .50$: strong scale; $.40 \leq H < .50$: medium scale; $.30 \leq H < .40$: weak scale. In practice, scalability analyses are used to determine the precision of person ordering on the latent variable by means of the rest score.

Rater Scalability Coefficients. When the MH-R model is applied to polytomous ratings, scalability coefficients describe the degree to which a set of raters can be ordered to form a scale that describes differences among students in terms of a latent variable. For polytomous ratings, deviations from a perfect Guttman ordering are identified by determining the relative difficulty ordering of rating scale categories across pairs of raters. To check scalability for a pair of raters, each of the rating scale category steps are treated like a set of generalized rating scale items, such that the categories form a series of dichotomous item steps made up of the rating scale categories for each of the raters. For example, in the case of a rating scale with four categories (0, 1, 2, 3), the six total category steps for rater i (X_i) and rater j (X_j) might be ordered as follows: $X_i \geq 1$, $X_j \geq 1$, $X_i \geq 2$, $X_j \geq 2$, $X_i \geq 3$, $X_j \geq 3$. An example of a Guttman error would be a student earning a score of 0 on the first category step from Rater i in combination with a score of 1 on the same step from Rater j ($X_i = 0$, $X_j = 1$). After errors are identified, the observed frequency is weighted by the expected frequency that would occur given marginal independence. Finally, the ratio of observed-to-expected errors is calculated. The rater pair scalability coefficient (H_{ij}) is calculated as one minus this observed-to-expected error ratio. In a parallel fashion to item scalability coefficients, scalability coefficients for rater pairs (H_{ij}), individual raters (H_i), and a group of raters (H) can also be calculated using the covariance method (Wind, 2014; Sijtsma & Molenaar, 2002). Sijtsma and Molenaar (2002) should be consulted for a detailed illustration of the calculation of scalability coefficients for pairs of polytomous items.

Essentially, the evaluation of rater scalability involves checking the relative ordering of rating scale categories across pairs of raters. As pointed out by Sijtsma and Molenaar (2002), scalability coefficients describe “whether the items have enough in common for the data to be explained by one underlying latent trait . . . in such a way that ordering the subjects by total score is meaningful” (p. 60). When extended to raters, scalability indices describe the degree to which the total score across raters can be interpreted as an indicator of meaningful student ordering on the latent variable. Evidence of low scalability might suggest idiosyncratic application of a set of rating scale categories, where a rater is not interpreting the rubric in the same way as other raters.

Double Monotonicity for Ratets Model

The next NIRT model used in this study is the DM-R model, which is an adaptation of Mokken’s (1971) DM model for use with polytomous rater-assigned scores. Because of the invariant ordering properties, adherence to the DM-R model requirements suggests that a set of raters meets the second requirement for rater-invariant measurement: raters are ordered the same way across the group of students. When invariant rater ordering is observed, the interpretation of relative rater severity ordering is invariant across the range of student achievement levels—thus facilitating the interpretation of student achievement such that conclusions about relative student ordering do not depend on the particular raters who scored each student.

Invariant Rater Ordering. The third indicator of data quality is based on the DM model assumptions of invariant ordering. Ligetvoet et al. (2010) and Ligetvoet, van der Ark, Bergsma, and Sijtsma (2011) observed that the interpretation of IIO for polytomous items was more meaningful at the overall item level than within rating scale categories (Sijtsma et al., 2011). As a method for investigating this DM model requirement, Ligetvoet et al. proposed a method called Manifest Invariant Item Ordering (MIIO) to check the assumption of IIO in polytomous data using average scores on items. The MIIO method is used to investigate nonintersection at the overall item level through a combination of statistical and graphical techniques.

Extended to the context of rater-mediated assessments, Ligetvoet et al.’s (2010, 2011) method will be referred to as *Manifest Invariant Rater Ordering* (MIRO) in this study. The MIRO procedure for examining nonintersection involves two major steps. First, the raters are ordered in terms of severity (i.e., difficulty) by their mean ratings across the entire group of students. Violations of MIRO are apparent when rater severity ordering shifts across high and low restscores groups within pairs of raters. If a violation is observed, a hypothesis test is used to determine whether or not the reversal of rater severity is significant.

Figure 5 illustrates the graphical technique for examining MIRO with polytomous rating data. In Panel A, the severity ordering for the rater pair cannot be interpreted consistently across the observed score scale. On the other hand, in Panel B, Rater *j*

Figure 5. Diagnostic plots for manifest invariant rater ordering.

(dashed line) is more severe (lower expected ratings) than Rater i (solid line) for all restscore groups; in other words, the relative severity of the two raters is consistent for all levels of student achievement. Using the *mokken* package (van der Ark, 2007, 2012), hypothesis tests can be used to determine the significance of intersections. For example, if the overall average ratings from Rater i and Rater j can be ordered such that $\bar{X}_i < \bar{X}_j$, a violation of this ordering is observed for a particular restscore group r when this ordering is reversed, such that $(\bar{X}_i|R=r) > (\bar{X}_j|R=r)$. The significance of this violation can be examined by testing the null hypothesis that the conditional mean ratings for the two raters are equal, $(\bar{X}_i|R=r) = (\bar{X}_j|R=r)$, against the alternative hypothesis of the reversed severity ordering, which is a violation of invariant rater ordering.

Evidence of invariant rater ordering is essential to the interpretation of ratings in performance assessments. A consistent ordering of raters across the range of student achievement is a fairness issue that must be empirically verified in order to inform the interpretation and use of rater-assigned scores. The importance of invariant rater ordering in this context stands in contrast to other domains in which MSA is traditionally applied, where “many symptoms go together and there is no real ordering in item severity” (Meijer et al., 2015). Furthermore, the interpretation of invariant ordering for raters is different from traditional MSA applications in that the observed rater ordering is not compared to an a priori specification of the expected severity ranking; rather, it is simply essential that the severity ordering of raters remain consistent for all students such that conclusions about student achievement do not depend on the “luck of the rater draw.”

Results

In this section, results from analyses with the illustrative data are presented related to the three major categories of data quality indicators for rater-mediated assessments: (A) Rater monotonicity, (B) Rater scalability, and (C) Invariant rater ordering.

Rater Monotonicity

In order to illustrate the first indicator of rating quality based on the MH-R model, monotonicity was examined for the 20 raters who scored the Georgia High School Writing Test. First, restscore groups specific to each rater were calculated for each of the 365 students. The highest possible rating from each rater is $X = 3$, so the highest possible total score (X_+) within each domain for each student across the 20 raters is $X_+ = 60$. Thus, the highest possible restscore is $R = 57$, for students with the maximum score [$R = (X_+ - X_i) = (60 - 3) = 57$]. Using these restscore groups, rater monotonicity was examined at the overall rater level and within rating scale categories.

Examination of the monotonicity plots indicated differences in the shape and location of response functions, suggesting differences in discrimination and severity/leniency across the 20 raters for all four domains. However, none of the raters

violated the MH-R model requirement for monotonicity at the overall rater level or within rating scale categories. In addition to graphical displays of monotonicity, the *mokken* package (van der Ark, 2007, 2012) was used to examine the statistical significance of violations of monotonicity. Results from statistical tests for monotonicity for the Georgia writing data are summarized in Table 2, Column A, which indicates that there were no significant violations of monotonicity for the Georgia raters within any of the four domains. Overall, adherence to the monotonicity assumption for the Georgia writing data suggests that the 20 raters share a consistent interpretation of relative student ordering across achievement levels within the domains; that is, the ordering of students is invariant across the 20 operational raters.

Rater Scalability

Next, the H coefficient for overall rater scalability was calculated for the 20 raters. Based on Mokken's (1971) critical values for the overall scalability coefficient, this group of raters appears to form a strong Mokken scale within each domain (Conventions: $H = 0.81$, $SE = 0.01$; Organization: $H = 0.78$, $SE = 0.01$; Sentence Formation: $H = 0.81$, $SE = 0.01$; Style: $H = 0.77$; $SE = 0.01$). Values of individual rater scalability coefficients are presented in Table 2, Column B; these values were calculated using separate applications of the MH-R model to the four domains. As can be seen in the table, there are differences in the relative frequency of Guttman errors across the group of raters within each domain. Across all four domains, the highest rater scalability coefficient is observed for Rater 8 in the Sentence Formation domain: $H = 0.86$, $SE = 0.02$. The lowest rater scalability coefficient is observed for Raters 2 and 9 in the Organization domain, and Raters 6 and 20 in the Style domain: $H = 0.74$, $SE = 0.02$; nonetheless, this scalability coefficient suggests that the ratings assigned by these raters form a strong scale.

Procedures for rater scalability analysis also include an examination of the scalability of rater pairs using the H_{ij} coefficient. H_{ij} is the normed covariance between ratings assigned by two raters, and positive values suggest adherence to the assumptions of the MH-R model within a pair of raters. Results from the rater pair scalability analysis revealed that there were no negative rater pair scalability coefficients among the group of raters who scored the Georgia High School Writing Test within the four domains. This finding of fit to the MH-R model within each domain further suggests that student total scores can be interpreted as a meaningful indicator of student ordering on the latent variable.

Invariant Rater Ordering

Next, the MIRO procedure was used to examine the DM-R model requirement of invariant rater ordering within each possible rater pair. Table 2, Column C, summarizes results from the MIRO analyses, and indicates that there were violations of MIRO for several raters when the DM-R model was examined separately for the four

Table 2. Mokken Rating Quality Results.

Rating Quality Indicators									
A. Rater Monotonicity		B. Rater Scalability				C. Invariant Rater Ordering			
Rater	Number of significant violations*	Rater scalability coefficient H_i (SE)				Number of significant violations			
		C	O	SF	St	C	O	SF	St
1	0	0.81 (0.01)	0.80 (0.02)	0.84 (0.01)	0.77 (0.02)	3	0	6	0
2	0	0.78 (0.02)	0.74 (0.02)	0.78 (0.02)	0.76 (0.02)	1	0	2	0
3	0	0.82 (0.01)	0.79 (0.02)	0.81 (0.02)	0.78 (0.02)	3	0	2	0
4	0	0.81 (0.02)	0.81 (0.01)	0.78 (0.02)	0.77 (0.02)	5	3	1	0
5	0	0.83 (0.02)	0.79 (0.02)	0.81 (0.02)	0.76 (0.02)	0	1	1	0
6	0	0.76 (0.02)	0.80 (0.02)	0.78 (0.02)	0.74 (0.02)	0	0	2	0
7	0	0.83 (0.01)	0.79 (0.02)	0.82 (0.02)	0.78 (0.02)	5	2	1	1
8	0	0.83 (0.01)	0.83 (0.01)	0.86 (0.01)	0.82 (0.02)	1	0	1	0
9	0	0.82 (0.02)	0.74 (0.02)	0.82 (0.02)	0.78 (0.02)	1	0	1	0
10	0	0.82 (0.01)	0.76 (0.02)	0.83 (0.01)	0.78 (0.02)	3	0	0	0
11	0	0.81 (0.02)	0.79 (0.02)	0.81 (0.02)	0.78 (0.02)	4	2	2	0
12	0	0.83 (0.01)	0.78 (0.02)	0.80 (0.02)	0.78 (0.02)	0	1	2	0
13	0	0.84 (0.01)	0.78 (0.02)	0.85 (0.01)	0.76 (0.02)	9	1	10	0
14	0	0.79 (0.02)	0.83 (0.02)	0.81 (0.02)	0.77 (0.02)	7	1	1	0
15	0	0.78 (0.02)	0.76 (0.02)	0.81 (0.02)	0.78 (0.02)	5	1	2	3
16	0	0.83 (0.01)	0.78 (0.02)	0.82 (0.01)	0.80 (0.02)	0	0	1	2
17	0	0.77 (0.02)	0.75 (0.02)	0.81 (0.02)	0.75 (0.02)	1	0	3	0
18	0	0.75 (0.02)	0.77 (0.02)	0.77 (0.02)	0.76 (0.02)	2	1	8	0
19	0	0.81 (0.02)	0.78 (0.02)	0.83 (0.01)	0.78 (0.02)	4	1	2	1
20	0	0.78 (0.02)	0.75 (0.02)	0.80 (0.02)	0.74 (0.02)	0	2	0	1

Note. The domain labels are as follows: C = Conventions; O = Organization; SF = Sentence formation; St = Style.

*No violations of monotonicity were observed in any of the four domains.

domains. Violations of invariant rater ordering were most frequently observed for Rater 13 on the Sentence Formation domain, whose comparisons with the other raters within this domain resulted in 10 significant intersecting rater response functions, followed by Rater 13 on the Conventions domains whose comparisons resulted in 9 significant intersections. Intersecting response functions suggest that the interpretation of rater severity ordering is *not* invariant across the range of the latent variable for all of the raters who scored the Georgia High School Writing assessment within each of the domain areas. As a result, it is not possible to interpret the relative ordering of rater severity in the same way across levels of student achievement. Further investigation, such as qualitative analyses, may shed light on potential causes for the lack of invariant rater ordering for these raters.

Summary and Discussion

In this study, Mokken's (1971) nonparametric models were applied to rater-assigned polytomous scores as a method for evaluating the quality of ratings. In order to illustrate the use of MSA to explore rating quality, indicators based on the MH-R and DM-R models were examined separately within four domains on the Georgia High School Writing test. Major findings from the study indicated that indices of rating quality based on MSA can provide diagnostic information about the quality of ratings in terms of the requirements for rater-invariant measurement. Although results were presented for all four domains, it is important to note that the analyses illustrated here involved treating each domain as a separate Mokken scale in order to demonstrate the generalizability of the Mokken-based indices for evaluating the quality of multiple sets of holistic ratings. The use of MSA to examine the quality of analytic ratings that include multiple domains as a single Mokken scale is explored in Wind (2015).

The nonparametric approach to measurement that characterizes Mokken's procedure for scale analysis is desirable in settings where relations among variables are difficult to define, such as rater perceptions of student achievement. Among the major motivations for the application of NIRT models in the social sciences is the lack of confidence in the assumption that transforming ordinal observations (such as ratings) to an interval scale is an appropriate way to describe a latent construct, which may or may not possess these interval-level properties. In other words, there is a distinction between data analyses with transformed observations assumed to reflect a construct and the actual properties of the construct. Furthermore, as pointed out by Cliff and Keats (2003), the desired conclusions to be drawn from these investigations are usually ordinal in nature, and often do not require the interval-level metric that is achieved through the application of parametric models. Mokken's (1971) approach to scaling provides a promising method for examining the degree to which a set of observations adhere to important aspects of measurement, such as monotonicity and nonintersection, without imposing potentially inappropriate assumptions on the level of measurement.

This study demonstrated that statistics and displays based on Mokken scaling can be used to evaluate the quality of rater-assigned scores in terms of a variety of

desirable properties. As pointed out by Meijer et al. (2015), the use of graphical techniques, such as those that characterize MSA, is essential for understanding the underlying measurement properties in a data set and informing substantive interpretation of data quality over and above statistical summaries of model-data fit. They explain:

There seems to be a great reluctance by especially trained psychometricians to use graphs. We often see fit statistics and large tables full of numbers that certainly do not provide more information than graphs (see also Wainer, 2005, for a more thorough presentation of this topic). (p. 89)

In keeping with the theme of the benefits of Mokken-based indices of data quality for rater-mediated assessments, it is important to note that the rating quality indices presented in this study are not intended to serve as criteria for rejecting scaling models or discarding raters who demonstrate model violations. This study does not propose critical values for determining the severity of violations of monotonicity, double monotonicity, or invariant ordering for raters besides the commonly used criteria for scalability coefficients (Mokken, 1971). Although the “Crit” statistic proposed by Molenaar and Sijtsma (2000) provides an index of the overall severity of violations of Mokken model assumptions that could technically be applied to the Mokken-based rating quality indices, the interpretation of this statistic has not been fully defined or explored in the Mokken scaling literature (Meijer et al., 2015). As a result, it is also unclear whether the interpretation of the critical values holds in the application of Mokken models to rater-mediated assessments.

Rather than focusing on model rejection or critical values, the Mokken-based rating quality indices presented in this study should be viewed as a series of diagnostic indicators that provide statistical and graphical evidence about the degree to which observed ratings match the expectations of models with useful measurement properties. The unique purpose of an assessment and scoring system should determine the relative importance of each indicator, and the degree to which violations of MH-R and DM-R model assumptions warrant further attention.

Implications

The overall goal in developing indicators of rating quality is to provide a method for evaluating the quality of rater-mediated assessments that can inform score interpretation and use for large-scale rater-mediated assessments. Although additional research is needed in order to more fully understand the utility of nonparametric indicators of data quality for rater-mediated assessments, the illustrations and substantive interpretations of Mokken-based statistics and displays presented here suggest that Mokken scaling indices can be used as a preliminary step to investigate important measurement properties, such as invariance, in rating data before a parametric model is applied. As pointed out by Meijer et al. (2015), nonparametric approaches to evaluating data quality, such as MSA, should not be viewed as a replacement for parametric models. Rather, “nonparametric approaches are excellent tools to decide whether

parametric models are justified. Moreover, given the often no-so-easy-to-interpret fit statistics for parametric models, nonparametric tools provide a nice extension of the parametric toolkit to IRT modeling” (p. 107) with the possibility of developing and supporting parametric models in the next iteration of assessment development.

Future Research

Future research is needed in order to develop a more complete understanding of the application of Mokken scaling to rater-mediated educational assessments. Specifically, the modification of MH-R and DM-R models when incomplete rating designs are utilized should be explored. Another issue is the modeling of items with different numbers of rating scale categories. Research on these issues is essential for operational applications of NIRT-based methods to monitor rating quality. Research is also needed that considers the practical utility of the Mokken-based rating quality indices. Specifically, a complete understanding of the implications for this study requires investigation of the degree to which nonparametric indices provide diagnostic information that can be used to inform rater training, monitoring rater performance during a scoring session, and other issues related to operational scoring in large-scale assessment settings.

Authors' Note

A previous version of this article was presented at the annual meeting of the American Educational Research Association in Philadelphia, PA, April 2014.

Declaration of Conflicting Interests

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

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Note

1. The current practice for creating restscore groups that is implemented in the *mokken* package (van der Ark, 2007, 2012) combines students with adjacent restscores into restscore groups following the criteria for the minimum sample size within each group proposed by Molenaar and Sijtsma (2000). Specifically, the default number of restscore groups is $N/10$ for $N > 500$; $N/5$ for N between 200 and 500; and $N/3$ for smaller sample sizes, with a minimum of 50 persons in each group. Van Schuur (2003) also discussed this issue, and stated that the groups should be sufficiently large such that a single participant does not make up more than 2% of a restscore group.

References

- Andrich, D. A. (2010). *The detection of a structural halo when multiple criteria have the same generic categories for rating*. Paper presented at the international conference on Rasch Measurement, Copenhagen, Denmark.
- Chernyshenko, O. S., Stark, S., Chan, K., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research*, 36, 523-562.
- Clauser, B. E. (2000). Recurrent issues and recent advances in scoring performance assessments. *Applied Psychological Measurement*, 24, 310-324.
- Cliff, N., & Keats, J. A. (2003). *Ordinal measurement in the behavioral sciences*. Mahwah, NJ: Lawrence Erlbaum.
- Congdon, P. J., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement*, 37, 163-178.
- Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2005). Individual feedback to enhance rater training: Does it work? *Language Assessment Quarterly*, 2, 175-196.
- Engelhard, G., Jr. (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., Jr. (1997). Constructing rater and task banks for performance assessments. *Journal of Outcome Measurement*, 1(1), 19-33.
- Engelhard, G., Jr. (2002). Monitoring raters in performance assessments. In G. Tindal & T. Haladyna (Eds.), *Large-scale assessment programs for ALL students: Development, implementation, and analysis* (pp. 261-287). Mahwah, NJ: Erlbaum.
- Engelhard, G., Jr. (2008). Historical perspectives on invariant measurement: Guttman, Rasch, and Mokken. *Measurement: Interdisciplinary Research & Perspective*, 6, 155-189.
- Engelhard, G., Jr. (2013). *Invariant Measurement: Using Rasch Models in the Social, Behavioral, and Health Sciences*. New York: Routledge.
- Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, & S. A. Clausen (Eds.), *Measurement and prediction* (Vol. IV, pp. 60-90). Princeton, NJ: Princeton University Press.
- Gyagenda, I. S., & Engelhard, G. (2009). Using classical and modern measurement theories to explore rater, domain, and gender influences on student writing ability. *Journal of Applied Measurement*, 10, 225-246.
- Hamp-Lyons, L. (2007). Worrying about rating. *Assessing Writing*, 12(1), 1-9.
- Hamp-Lyons, L. (2011). Writing assessment: Shifting issues, new tools, enduring questions. *Assessing Writing*, 16(1), 3-5.
- Junker, B. W. (1993). Conditional association, essential independence and monotone unidimensional item response models. *Annals of Statistics*, 21, 1359-1378.
- Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior: A longitudinal study. *Language Testing*, 28, 179-200.
- Ligtvoet, R., van der Ark, L. A., Bergsma, W. P., & Sijtsma, K. (2011). Polytomous latent scales for the investigation of the ordering of items. *Psychometrika*, 76, 200-216.
- Ligtvoet, R., van der Ark, L. A., te Marvelde, J. M., & Sijtsma, K. (2010). Investigating an invariant item ordering for polytomously scored items. *Educational and Psychological Measurement*, 70, 578-595.
- Linacre, J. M. (1996). Generalizability theory and many-facet Rasch measurement. In G. Engelhard & M. Wilson (Eds.), *Objective measurement: Theory into practice* (pp. 85-98). Norwood, NJ: Ablex.

- Loevinger, J. (1948). The technique of homogeneous tests compared with some aspects of "scale analysis" and factor analysis. *Psychological Bulletin*, 45, 507-530.
- Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, 12, 54-71.
- Lynch, B. K., & McNamara, T. F. (1998). Using G theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, 15, 158-180.
- Meijer, R. R., & Baneke, J. J. (2004). Analyzing psychopathology items: A case for nonparametric item response theory modeling. *Psychological Methods*, 9, 354-368.
- Meijer, R. R., Sijtsma, K., & Smid, N. G. (1990). Theoretical and empirical comparison of the Mokken and the Rasch approach to IRT. *Applied Psychological Measurement*, 14, 283-298.
- Meijer, R. R., Tendeiro, J. N., & Wanders, R. B. K. (2015). The use of nonparametric item response theory to explore data quality. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 85-110). New York, NY: Routledge.
- Mokken, R. J. (1971). *A theory and procedure of scale analysis*. The Hague: Mouton/Berlin: De Gruyter.
- Mokken, R. J. (1997). Nonparametric models for dichotomous responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 351-367). New York, NY: Springer.
- Molenaar, I. W. (1982). Mokken scaling revisited. *Kwantitative Methoden*, 3(8), 145-164.
- Molenaar, I. W. (1997). Nonparametric models for polytomous responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 369-380). New York, NY: Springer.
- Molenaar, I. W., & Sijtsma, K. (2000). MPS5 for Windows: A program for Mokken scale analysis for polytomous items (Version 5.0) [Computer software]. Groningen, Netherlands: ProGAMMA.
- Pula, J. J., & Huot, B. A. (1993). A model of background influences on holistic raters. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 237-265). Cresskill, NJ: Hampton Press.
- R Development Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <http://www.R-project.org/>
- Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research. (Expanded edition, Chicago: University of Chicago Press, 1980).
- Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, 5, 27-48.
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Thousand Oaks, CA: Sage.
- Sijtsma, K., & Hemker, B. T. (2000). A taxonomy of IRT models for ordering persons and items using simple sum scores. *Journal of Educational and Behavioral Statistics*, 25, 391-415.
- Sijtsma, K., & Meijer, R. R. (2007). Nonparametric item response theory and special topics. In C. R. Rao & S. Sinharay (Eds.), *Psychometrics: Handbook of statistics* (Vol. 26, pp. 719-747). Amsterdam, Netherlands: Elsevier.

- Sijtsma, K., Meijer, R. R., & van der Ark, L. A. (2011). Mokken scale analysis as time goes by: An update for scaling procedures. *Personality and Individual Differences, 50*, 31-37.
- Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory* (Vol. 5). Thousand Oaks: Sage.
- Stahl, J. A., & Lunz, M. E. (1992, May). *A comparison of generalizability theory and multifacet Rasch measurement*. Paper presented at the Midwest Objective Measurement Seminar, Chicago, IL.
- Sudweeks, R. R., Reeve, S., & Bradshaw, W. S. (2005). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing, 9*, 239-261.
- van der Ark, L. A. (2005). Stochastic ordering of the latent trait by the sum score under various polytomous IRT models. *Psychometrika, 70*, 283-304.
- van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software, 20*(11). Retrieved from <http://www.jstatsoft.org/v20/i11/>
- van der Ark, L. A. (2012). New developments in Mokken scale analysis in R. *Journal of Statistical Software, 48*(5). Retrieved from <http://www.jstatsoft.org/v48/i05/>
- van Schuur, W. H. (2003). Mokken scale analysis: Between the Guttman scale and parametric item response theory. *Political Analysis, 11*, 139-163.
- Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*, 263-287.
- Wind, S. A. (2014). Guidelines for rating scales based on Rasch measurement theory and Mokken scaling. *Journal of Applied Measurement, 15*, 100-133.
- Wind, S. A. (2015). Evaluating the quality of analytic ratings with Mokken scaling. *Psychological Test and Assessment Modeling, 57*(3), 423-444.
- Wind, S. A., & Engelhard, G. (2012). Examining rating quality in writing assessment: Rater agreement, error, and accuracy. *Journal of Applied Measurement, 13*, 321-335.
- Wind, S. A., & Engelhard, G., Jr. (2013). How invariant and accurate are domain ratings in writing assessment? *Assessing Writing, 18*, 278-299.
- Wolfe, E. W. (2009). Item and rater analysis of constructed response items via the multi-faceted Rasch model. *Journal of Applied Measurement, 10*, 335-347.
- Wolfe, E. W., & McVay, A. (2012). Application of latent trait models to identifying substantively interesting raters. *Educational Measurement: Issues and Practice, 31*, 31-37.
- Wright, B. D., & Stone, M. (1979). *Best test design: Rasch measurement*. Chicago, IL: MESA Press.