

Evaluating the quality of analytic ratings with Mokken scaling

*Stefanie A. Wind*¹

Abstract

Greatly influenced by the work of Rasch (1960/1980), Mokken (1971) presented a nonparametric scaling procedure that is based on the theory of invariant measurement, but draws upon less strict requirements related to the scale of measurement. Because they are theoretically and empirically related to Rasch models, Mokken's nonparametric models have been recognized as a useful exploratory tool for examining data in terms of the basic requirements for invariant measurement before the application of a parametric model. In particular, recent research has explored the use of polytomous versions of Mokken's (1971) nonparametric scaling models as a technique for evaluating the quality of holistic ratings (Wind & Engelhard, in press) and rating scales (Wind, 2014) for performance assessments in terms of the requirements for invariant measurement. The current study continues the extension of Mokken scaling to performance assessments by exploring the degree to which Mokken-based rating quality indices can be used to explore the quality of ratings assigned within domains on an analytic rubric. Using an illustrative analysis, this study demonstrates the use of a generalized rating design to explore the quality of analytic ratings within the framework of Mokken scaling. Findings from the illustrative analysis suggest that a generalized rating design can be used to examine the quality of analytic ratings in terms of the requirements for invariant measurement.

Key words: Mokken scaling; analytic rubric; performance assessment; nonparametric item response theory

¹ *Correspondence concerning this article should be addressed to:* Stefanie A. Wind, PhD, College of Education, The University of Alabama, 313C Carmichael Hall, Box 870231, Tuscaloosa, AL 35487-0231, USA; email: swind@ua.edu

Two major scoring systems are widely employed in large-scale rater-mediated performance assessments: holistic and analytic scoring. Holistic scoring systems involve the use of rubrics whose performance-level descriptors (PLDs) focus on the evaluation of a student's performance as a whole. Although multiple components of a response may be considered, a single score is assigned to reflect the overall quality of the performance. In contrast, analytic scoring systems involve the use of rubrics with PLDs specified for multiple domains on which separate scores are assigned to reflect specific components of a student's performance. For example, an analytic scoring rubric in a large-scale writing assessment might include domains related to style, organization, and conventions. As noted in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), the use of analytic scoring rubrics "provides valuable information for diagnostic purposes and lends itself to evaluating strengths and weaknesses of test takers. However, validation will be required for diagnostic interpretations for particular uses of the separate scores" (p. 79). Across both holistic and analytic scoring systems, the diagnostic value of rater-assigned scores for informing educational decisions depends on the degree to which scoring systems demonstrate useful psychometric properties, including validity, reliability, and fairness.

Current operational methods for monitoring ratings typically include indices of rating quality that are classified within three major categories: 1) rater agreement, 2) rater errors and systematic biases, and 3) rater accuracy (Johnson, Penny, & Gordon, 2009; Murphy & Cleveland, 1991). Indices of rater agreement describe the degree to which raters assign matching (or in some cases, adjacent) scores to the same performance; rater errors and systematic biases are used to describe specific patterns or trends in rating behavior that are believed to contribute to the assignment of scores different from those warranted by a student's performance; rater accuracy is often defined in practice as a match between operational ratings and those established as "true" or "known" ratings by individuals or committees of expert raters (Johnson, Penny, & Gordon, 2009; Murphy & Cleveland, 1991). When indices of rating quality based on these three categories are applied, high levels of agreement, low levels of error and systematic bias, and high levels of accuracy are assumed to reflect high-quality ratings.

Summaries of rating quality indices emphasize the need for increased precision in the description of rating quality in order to use them as evidence to inform the interpretation and use of ratings (Johnson, Penny, & Gordon, 2009; Murphy & Cleveland, 1991). Further, reviews and meta-analyses of rating quality indices based on the categories of rater agreement, error and systematic bias, and accuracy reveal that indices across these categories do not provide consistent information about a set of raters (e.g., Saal, Downey, & Lahey, 1980). As a result, it is not possible to generalize information about rating quality across these three categories; for example, evidence of rater agreement does not imply a lack of halo error or rater accuracy. The lack of alignment across these indices of rater agreement, error, and accuracy is further highlighted by the fact that rating quality within one category is likely to influence interpretation of results from rating quality analyses within another category. For example, because ordinal indices of rater agreement based on correlation coefficients are only sensitive to the consistency of rater rank-ordering of performances, the interpretation of inter-rater reliability estimates based on correlation

coefficients is influenced by rater leniency and severity error (Cronbach, Linn, Brennan, & Haertel 1997; Lane & Stone, 2006; Zhu & Johnson, 2013).

Rater-invariant measurement

When examining the quality of ratings assigned in rater-mediated assessments, several researchers have recognized the usefulness of *invariant measurement* as a framework in which to consider the degree to which rater-assigned scores demonstrate useful measurement properties. In the context of educational achievement tests, invariant measurement is based on two major requirements for the measurement of students and items: 1) Conclusions about students should not depend on which items were used to measure them; and 2) Conclusions about items should not depend on which students were used to calibrate them (Wright & Stone, 1979). Wind and Engelhard (2013) extended the idea of invariant measurement to the context of rater-mediated assessments, such that these two requirements can be re-stated as follows: 1) Conclusions about students should not depend on which rater happened to score their work; and 2) Conclusions about raters should not depend on which students they happened to score. Models based on invariant measurement can be used to provide a coherent set of rating quality indices that describe the degree to which a set of ratings meets the requirements for rater-invariant measurement.

In social science research, invariant measurement is often associated with Rasch measurement theory (Rasch, 1960/1980; e.g., Engelhard, 2013). Rasch measurement theory is a measurement framework that includes a family of parametric latent trait models characterized by their adherence to the requirements for invariant measurement when good model-data fit is observed. Recognizing the necessity of invariance as a prerequisite for measurement in the social sciences, Mokken (1971) proposed a set of scaling models that are theoretically and empirically related to Rasch measurement theory. However, Mokken expressed reservations related to the parametric form of the Rasch model. Specifically, his concerns were related to the tenability of the Rasch model requirement that the relationship between person locations on the latent variable and the probability for a correct response fit the shape of the logistic ogive. Drawing upon these concerns, Mokken proposed a nonparametric scaling procedure that can be viewed as a nonparametric analogue to Rasch measurement theory that shares the same basic underlying model assumptions related to the requirements for invariant measurement (de Koning, Sijtsma, & Hamers, 2002; Engelhard, 2008; Meijer, Sijtsma, & Smid, 1990).

Although rating quality has been examined using indicators based on parametric IRT models such as the Rasch model, the utility of nonparametric IRT models for monitoring rating quality is relatively unexplored. The application of nonparametric IRT models to rater-mediated assessments is promising in light of the fact that “if an IRT model is used for constructing a test, and the measurement of respondents on an *ordinal scale* is sufficient for the application envisaged, parametric models might be unduly restrictive for this purpose,” and the fact that desirable measurement properties such as invariant order-

ing of persons and items can be obtained under nonparametric IRT models (Sijtsma & Molenaar, 2002, p. 15, emphasis in the original).

Recent research includes the application of Mokken scale analysis as a method for evaluating the quality of rater-mediated educational performance assessments (Wind, 2014; Wind & Engelhard, in press). This research included a set of nonparametric statistics and displays based on Mokken scaling to evaluate the overall functioning of rating scales in terms of their orientation with the latent variable, rating scale category precision, and model-data fit. Specifically, Wind (2014) presented adaptations of Mokken's (1971) Monotone Homogeneity model and Double Monotonicity model to the Monotone Homogeneity for Raters (MH-R) and Double Monotonicity for Raters (DM-R) models as a method for exploring rating scale functioning within the context of holistic ratings from a writing assessment. The current study represents a next-step in this program of research, and illustrates the application of Mokken scale analysis to the context of rater-mediated performance assessments with analytic scoring systems.

Purpose

The purpose of this study is to continue the application of Mokken scale analysis to the context of performance assessments by illustrating the use of Mokken-based statistics and displays as a method to evaluate the quality of analytic ratings. Two research questions guide the study:

1. What does Mokken scale analysis reveal about the quality of analytic domain ratings?
2. What does Mokken scale analysis reveal about differences in the quality of ratings assigned to different domains on an analytic rubric?

Mokken scale analysis

Mokken scaling (Mokken, 1971) is a probabilistic-nonparametric approach to scale analysis that consists of a set of exploratory steps for constructing scales with useful measurement properties, such as monotonicity in the latent variable and invariant ordering of persons and items. Models based on Mokken scaling are unidimensional item response theory (IRT) models that describe the probability of an observed response as a function of person and item ordering on a latent variable. Mokken scaling is considered nonparametric because of the lack of restrictions on the shape of the item response functions (IRFs) that describe the relationship between the probability of an observed response and person and item characteristics. Whereas parametric IRT models specify the shape of the IRF using an algebraic form, the functions that underlie Mokken's nonparametric models are only required to adhere to basic ordering requirements (Sijtsma & Molenaar, 2002). Summarizing the rationale for a nonparametric approach to measurement in the social sciences, Mokken (1971) observed:

In vast areas of social research the application of parametric models may often be too far fetched. Their application presupposes a relatively deep insight into the structure of the variable to be measured and the properties of the items by which it can be measured.... Specific parametric assumptions... lead to procedures of inference and estimation that are too pretentious and intricate for the level of information and the precision that can be claimed for the data used in actual measurement. (p. 173)

Based on the idea that the assumptions of parametric IRT models may not be warranted in social science and behavioral applications, nonparametric models have been proposed as an alternative method for examining item response data that can be used as a preliminary step before the application of parametric models. In particular, Mokken's approach to scaling provides an exploratory method for determining the degree to which a set of observations adhere to the requirements for invariant measurement as a preliminary step *before* the application of the Rasch model.

Mokken scaling models

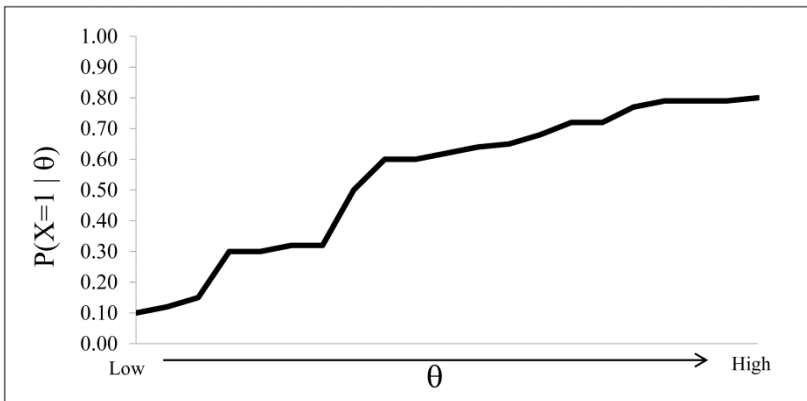
Although Mokken scaling models are considered less strict than the Rasch model in terms of their underlying requirements, they are still based on a set of basic requirements for the relationship between the latent variable (i.e., construct) and the probability for a response. Specifically, Mokken's (1971) scaling models are based on the requirement that the probability for a response is a nondecreasing function of a single latent variable. The ordering requirement states that, for persons A and B who can be ordered in terms of the latent variable such that $(\theta_A < \theta_B)$, the following relationship must hold:

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

where $P(X_{i=1}|\theta_A)$ and $P(X_{i=1}|\theta_B)$ represent the probability that a person A and Person B provide a correct or positive response to Item i . As long as they meet the ordering requirement in Equation 1, Mokken IRFs do not need to match the s -shape of the logistic ogive, as is required for most parametric IRT models, including the Rasch model. Whereas the application of the Rasch model includes checks for adherence to the logistic ogive as evidence that measurement has occurred, models based on Mokken scaling do not specify a particular shape for the IRF, as long as the ordering requirement in Equation 1 is observed. Further, Mokken IRFs may have several points of inflection, and may be constant over several ranges of the latent variable. Figure 1 is an example of an IRF that meets the ordering requirement in Equation 1.

The lack of restrictions on the shape of Mokken IRFs does not facilitate the direct estimation of person locations on a latent variable, similar to person parameter (i.e., theta $[\theta]$) estimates in the Rasch and other parametric models. As a result, it is necessary to use an approximation of the person location parameter to check for adherence to Mokken model assumptions. Specifically, the ordering of persons by their unweighted sum score

Figure 1:
Nonparametric Item Response Function



(X_+) across a set of items has been shown to serve as an estimate of their ordering by θ (Grayson, 1988; Huynh, 1994; Mokken, 1971; Ünlü, 2008). When model assumptions such as Equation 1 are examined for an item of interest, the use of X_+ as a conditional value is limited because this sum score is calculated using the item of interest. As a result, checks for adherence to Mokken scaling assumptions are based on the *restscore* (R), which is calculated as X_+ minus the score on the item of interest (X_i ; Junker, 1993), such that $R = X_+ - X_i$.

Mokken scaling for dichotomous responses

Mokken presented two probabilistic-nonparametric scaling models for dichotomous items: the Monotone Homogeneity (MH) model and the Double Monotonicity (DM) model. The MH model is based on three assumptions about the relationship between item responses and the latent variable: 1) *Unidimensionality* – Item responses reflect evidence of a single latent variable; 2) *Conditional Independence* – Responses to an item are not influenced by responses to any other item, after controlling for the latent variable; and 3) *Monotonicity* – The probability that a person will correctly respond to an item is non-decreasing as their location on the latent variable increases. The DM model is based on the three MH model assumptions, with the addition of a fourth assumption: 4) *Nonintersecting IRFs* – Item difficulty ordering is consistent across all levels of student achievement.

For dichotomous items, the DM model assumption of nonintersecting IRFs results in invariant item ordering (IIO). When IIO is observed, the expected order of items is invariant across students, and the expected ordering of students on the latent variable is invariant across items. Based on these invariance properties, several scholars have noted theoretical similarities between the DM model and the Rasch model, where the DM model

may be viewed as a less-restrictive version of the Rasch model (Engelhard, 2008; Meijer, Sijtsma, & Smid, 1990; van Schuur, 2003). The Rasch model shares the same four requirements as the DM model, but adds an additional requirement related to the use of a parametric transformation to achieve interval-level estimates of items and persons.

Mokken scaling for polytomous responses

Molenaar (1982, 1997) extended Mokken's method for scale analysis to include polytomous versions of the MH model and DM model. The polytomous versions of the MH model and DM model are based on the same underlying assumptions as the original dichotomous formulations, where the assumptions are applied to rating scale categories as well as overall items. Similar to polytomous formulations of the Rasch model, the item step response functions (ISRFs; i.e., category response functions) that underlie polytomous Mokken models are conceptualized as a series of dichotomous "steps" between scale categories, such that $m - 1$ ISRFs are specified for a rating scale that has m categories.

When polytomous responses are examined, monotonicity implies that the probability for a rating in Category k or higher is non-decreasing as a person's location on the latent variable increases. Along the same lines, the polytomous formulation of the DM model maintains the same assumptions of unidimensionality, conditional independence, and monotonicity as the dichotomous DM model. For polytomous data, the DM model assumption related to invariant ordering can be extended to ISRFs as follows: *Non-intersecting ISRFs* imply that the order of cumulative category probabilities is consistent across all levels of student achievement. Recently, Ligtvoet and his colleagues (Ligtvoet, et al., 2010, 2011) pointed out a potential discrepancy between the polytomous DM model and IIO, where adherence to the requirement of nonintersecting ISRFs may in some cases not correspond to invariant ordering of overall IRFs for polytomous items. As a result, he proposed a method called Manifest Invariant Item Ordering (MIIO) for examining IIO at the overall item level. Although nonintersecting ISRFs are interesting theoretically, indices of IIO at the overall item level through MIIO provide more practical diagnostic information about measurement quality for polytomous items than indices of non-intersecting ISRFs (double monotonicity; Sijtsma, Meijer, & van der Ark, 2011).

Mokken scaling for rater-mediated assessments

Recently, Wind (2014) and Wind and Engelhard (in press) applied the polytomous versions of Mokken's MH and DM models to the context of rater-mediated educational assessments. The new application included a set of rating quality indicators based on the Monotone Homogeneity for Ratings (MH-R) model, which is an application of the MH model, and the Double Monotonicity for Ratings (DM-R) model, which is an application of the DM model. The statistics and displays for monitoring rating quality included methods that are typically used to examine measurement quality using the polytomous MH and DM models, with an emphasis on the new substantive interpretations in the

context of rater-mediated assessments with holistic ratings. A detailed exploration of the MH-R and DM-R models and the corresponding Mokken-based rating quality indices for holistic ratings is provided in Wind (2014) and Wind and Engelhard (in press). This study focuses on the interpretation of the statistics and displays as indicators of rating quality for analytic ratings.

Method

In this study, an illustrative data analysis is used to demonstrate the application of Mokken scaling to analytic rating data in order to explore the degree to which ratings demonstrate useful measurement properties. The data used to illustrate application and interpretation of the statistics and displays were recently used to illustrate Mokken-based rating quality indices for holistic ratings (Wind & Engelhard, in press). Andrich (2010) and Wind and Engelhard (2012) also examined these data in the context of parametric IRT. The illustrative data are from the Georgia High School Writing Test, and include scores from 365 eighth-grade students whose persuasive essays were rated by 20 raters. The ratings were assigned using a four-point rating scale (1 = *low* to 4 = *high*) and an analytic rubric with four domains: Conventions, Organization, Sentence Formation, and Style. All 20 raters scored all 365 essays, such that the rating design was fully connected (Eckes, 2011; Engelhard, 1997), and each essay received 20 ratings. The ratings were recoded to 0 = *low*; 3 = *high* for the present analyses. Analyses were conducted using the *mokken* package for the *R* statistical software program (van der Ark, 2007, 2012).

Rating quality indices for analytic ratings

In this study, the applications of polytomous versions of Mokken's (1971) scaling models proposed by Wind and Engelhard (in press) for holistic ratings are extended further as exploratory techniques that can be used to examine the quality of ratings assigned to domains in an analytic rubric. Whereas the original Mokken-based rating quality indices were focused on holistic ratings (Wind & Engelhard, in press), the current study includes the application of these techniques to analytic ratings. For illustrative purposes, analyses are conducted using the ratings assigned by three randomly selected operational raters (Rater A, Rater B, and Rater C) from the illustrative data set. In order to accommodate the analytic rating structure of the data, the ratings assigned by the three raters on the four domains were treated as a set of generalized rating scale items (Wang & Engelhard, 2014) using the layout shown in Table 1. When this generalized view of raters is applied, the ratings assigned by each rater on a particular domain can be seen as a type of rating scale "item," that reflects the combination of a rater and domain to form a unique assessment opportunity. Although this rater-by-domain combination is not an "item" in the traditional sense, the term item will be used throughout the manuscript to describe these assessment opportunities.

Table 1:
Data Structure: Generalized Ratings

Students	Conventions	Organization	Sentence Formulation	Style
<i>Rater A</i>				
1	r_{CIA}	r_{OIA}	r_{SFIA}	r_{SIA}
...
n	r_{CnA}	r_{OnA}	r_{SFnA}	r_{SnA}
<i>Rater B</i>				
1	r_{CIB}	r_{OIB}	r_{SFIB}	r_{SIB}
...
n	r_{CnB}	r_{OnB}	r_{SFnB}	r_{SnB}
<i>Rater C</i>				
1	r_{CIC}	r_{OIC}	r_{SFIC}	r_{SIC}
...
n	r_{CnC}	r_{OnC}	r_{SFnC}	r_{SnC}

Table 2 presents three categories of Mokken-based statistics and displays that can be used to examine the quality of ratings in the context of rater-mediated educational assessments: A) Scalability; B) Monotone Homogeneity; and C) Invariant ordering. These indices are adapted from the indices presented by Wind (2014).² In this section, the Mokken-based rating quality indices and their interpretations listed in Table 2 are illustrated for analytic ratings using the Georgia writing data.

A. Scalability

The first category of Mokken-based rating quality indices for analytic ratings is *rater scalability*. In the context of Mokken scaling (Mokken, 1971), scalability analysis is a method for examining the degree to which a set of responses approximates a Guttman scale. As a method for examining the impact of error on measurement quality, Mokken (1971) presented an extension of Loevinger’s (1948) homogeneity (H) coefficient to examine scalability within item pairs (H_{ij}), for individual items (H_i), and for the overall set of items in a survey or test (H). Mokken-based scalability coefficients are summary statistics that describe the deviation of an observed data structure from a perfect Guttman (i.e., scalogram) pattern.

² It should be noted that the original set of Mokken-based rating quality indices (Wind, 2014) included indicators of double monotonicity based on the DM-R model, where non-intersection was examined at the level of ISRFs. As noted above, the finding of intersecting ISRFs for polytomous ratings is less diagnostically useful than the finding of intersecting IRFs. The interpretation becomes increasingly complex and less diagnostically useful in the context of analytic ratings. As a result, this study does not include an indicator of rating quality based on non-intersection at the level of ISRFs.

Table 2:
Indicators of Rating Quality based on Mokken Scaling

Category	Rating Quality Indices (Statistics and Plots)	Interpretation: Analytic Ratings
A. Scalability	<ul style="list-style-type: none"> • Individual rater scalability coefficients • Rater pair scalability coefficients • Group rater scalability coefficients 	Can domain ratings be used to distinguish students across achievement levels?
B. Monotone Homogeneity	<ul style="list-style-type: none"> • Monotonicity plots and statistics for overall raters • Monotonicity plots and statistics within rating scale categories 	Is the ordering of students in terms of the latent variable invariant across domain ratings?
C. Invariant Ordering	<ul style="list-style-type: none"> • Manifest invariant ordering statistics and pairwise plots for overall raters 	Is the ordering of domain ratings in terms of the latent variable invariant across students?

Note. This set of Mokken-based rating quality indices is adapted from Wind (2014).

It is simplest to conceptualize Mokken scalability coefficients within the context of pairs of dichotomous items. First, deviations from 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. Then, the relative ordering of items within all possible item pairs is examined in order to check for discrepancies with the overall ordering. The frequency of observed errors are weighted by the expected cell frequency that would occur given marginal independence (i.e., by chance alone). Then, the item-pair scalability coefficient (H_{ij}) is calculated as one minus the ratio of observed-to-expected Guttman errors:

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

where

H_{ij} = Scalability of the item pair consisting of Item i and Item j ,

F_{ij} = Frequency of observed Guttman errors between Item i and Item j , and

E_{ij} = Expected frequency of Guttman errors between Item i and Item j , given marginal independence of the two items.

In addition to item pair scalability coefficients (H_{ij} ; Equation 2), scalability can be calculated for individual items (H_i) and an overall set of items (H). Scalability coefficients can also be calculated using ratio of the observed correlation between two items to the highest possible correlation, given the marginal distributions of the two items (Sijtsma & Molenaar, 2002, pp. 55-58). Equations for both the error frequency method (Equation 2) and the covariance method are given in Sijtsma and Molenaar (2002). Wind (2014) provided a detailed example of the computation of rater scalability coefficients in the context of holistic ratings.

In the case of a perfect Guttman scalogram pattern, scalability coefficients would have a value of $H = 1.00$. When data adhere to the assumptions of the MH model, the lowest possible value for the H coefficient is $H = 0.00$; negative scalability coefficients are theoretically possible, but they indicate lack of adherence to the MH model assumptions. Mokken (1971) proposed rule-of-thumb critical values for the H coefficient that can be used to evaluate the quality of a scale (Mokken, 1971; Molenaar & Sijtsma, 2000): $H \geq .50$: Strong scale; $.40 \leq H < .50$: Medium scale; $.30 \leq H < .40$: Weak scale. In practice, scalability analyses are used to describe the precision of person ordering on the latent variable by means of X_+ . Values of H_{ij} , H_i , and H are frequently used to identify items for potential removal from a scale.

Scalability for analytic ratings. When Mokken scaling is applied to analytic ratings, scalability indices describe the degree to which domain ratings can be ordered to form a scale that describes differences in achievement among a group of students (Wind, 2014). In the context of generalized ratings based on an analytic scoring system, the ordering of items describes the relative ordering of rater severity within domains. Guttman errors are defined as violations of this ordering that are observed in contingency tables for pairs of items. The error frequency is weighted by the expected cell frequency that would occur given marginal independence. Finally, the ratio of observed errors to expected errors is calculated. The scalability coefficient for pairs of items (H_{ij}) is calculated as one minus this observed-to-expected error ratio (Equation 2). Mokken's (1971) criteria for classifying H coefficients provide a potentially useful index of the approximation to a Guttman scale within a set of analytic rating data. High, positive values of H coefficients in the context of analytic ratings suggest that domain ratings can be used to distinguish students in terms of the latent variable.

Results: Georgia writing data. The overall scalability coefficient (H) for the illustrative data is $H = 0.79$, ($SE = 0.02$), which suggests that the set of analytic ratings creates a strong Mokken scale. This finding indicates that the overall ordering of students by domain ratings can be used to distinguish among students in terms of the latent variable. Table 3 includes results from scalability analyses for individual and pairs of items. Results from these analyses indicate that the scalability coefficients for each individual item are quite high ($0.74 \leq H_i \leq 0.83$). This finding suggests that each of the individual items is associated with few Guttman errors, leading to an interpretable ordering of student achievement by the analytic domain ratings. The finding of no negative scalability coefficients among pairs of items further supports the use of student total scores to describe differences in achievement in terms of the latent variable.

Table 3:
Mokken Rating Quality Results

Assessment Opportunity (Generalized Rating)		A. Scalability		B. Monotonicity	C. Invariant Ordering
Domain	Rater	Scalability Coefficient H_i (SE)	Number of Negative Rater Pair Scalability Coefficients	Number of Violations	Number of violations
Conventions	A	0.79 (0.02)	0	0	0
	B	0.83 (0.01)	0	0	1
	C	0.78 (0.02)	0	0	0
Sentence Formation	A	0.80 (0.02)	0	0	1
	B	0.79 (0.02)	0	0	0
	C	0.79 (0.02)	0	0	0
Style	A	0.83 (0.02)	0	0	2
	B	0.82 (0.02)	0	0	0
	C	0.76 (0.02)	0	0	2
Organization	A	0.74 (0.02)	0	0	0
	B	0.79 (0.02)	0	0	1
	C	0.79 (0.02)	0	0	1

Note. These rating quality indices are adapted from the procedures used to examine model-data fit for Mokken's (1971) NIRT models. Indices of Scalability (A) and Monotonicity (B) are based on the MH-R model. Indices of Invariant Ordering (C) are based on the DM-R model (using the Manifest Invariant Rater Ordering technique).

B. Monotonicity

The second category of rating quality indices for analytic ratings based on Mokken scaling is monotonicity. In the context of dichotomous items, monotonicity in the latent variable suggests that the probability that a student correctly responds to an item ($X_i = 1$) is non-decreasing as their location on the latent variable (and, by extension, X_+) increases. An item meets the monotonicity assumption of the MH model when the observed probability for a correct or positive response is non-decreasing as student restscores increase. When polytomous ratings are considered, monotonicity implies that the probability that a student will receive a rating in a higher category (i.e., that the student will "pass" a rating scale category step) is non-decreasing as their latent variable location increases (Wind, 2014).

Monotonicity for analytic ratings. In order to determine whether monotonicity holds in a set of analytic ratings, it is necessary to examine the cumulative probability for each rating scale category [$P(X \geq k)$] for each item across increasing levels of student achievement. This is accomplished by examining the cumulative probability for ratings associated with an item across increasing restscores. In the context of analytic ratings, restscores are calculated as a student's total score across items minus the rating on an item of interest. For example, the restscore for Student n based on the ratings assigned by Rater A on the Conventions domain (r_{CnA}) would be calculated as: $R_n = X_n - r_{CnA}$.

Using restscore groups, there are two graphical methods for examining monotonicity for analytic ratings. First, monotonicity for each item is examined in terms of average ratings within restscore groups. Figure 2, Panel A is an example diagnostic plot from the illustrative data that demonstrates a method for evaluating monotonicity for an overall item. Student restscores are plotted along the x -axis, and average ratings from Rater A on the Conventions domain ($0 = low$; $3 = high$) are plotted along the y -axis. This figure illustrates evidence of monotonicity for Rater A on the Conventions domain because average ratings are non-decreasing as restscores increase. Graphical evidence for monotonicity can also be examined for items at the level of ISRFs, or rating scale categories. Figure 2, Panel B includes an example diagnostic plot for examining monotonicity within rating scale categories based on the illustrative data set. The plot in Panel B illustrates the relationship between the probability for a rating in Category k or higher across increasing levels of student achievement (restscore groups). When polytomous ratings are conceptualized as a series of dichotomous steps, monotonicity is examined for the $m - 1$ meaningful ISRFs by calculating the cumulative probability for a rating in a given category within each restscore group for a particular item. Student restscores are plotted along the x -axis, and the y -axis represents the probability for a rating in Category k or higher from Rater A on the Conventions domain, given a restscore value [$P(X \geq k \mid R = r)$]. 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 [$P(X \geq 3)$]. Monotonicity is implied when the ISRFs for an item are non-decreasing across non-decreasing restscores.

Statistical evidence can also be used to detect violations of monotonicity. In the context of analytic ratings, violations of monotonicity for an item occur when the average rating on an item in two adjacent restscore groups is disordered, such that students in the higher restscore group have a lower average rating on a domain than students in the lower restscore group. The significance of violations of monotonicity for overall items can be examined using a test for the null hypothesis that the expected average ratings are equal between two adjacent restscore groups, against the alternative hypothesis that the expected rating is lower in the group with a higher restscore, which would be a violation of monotonicity.

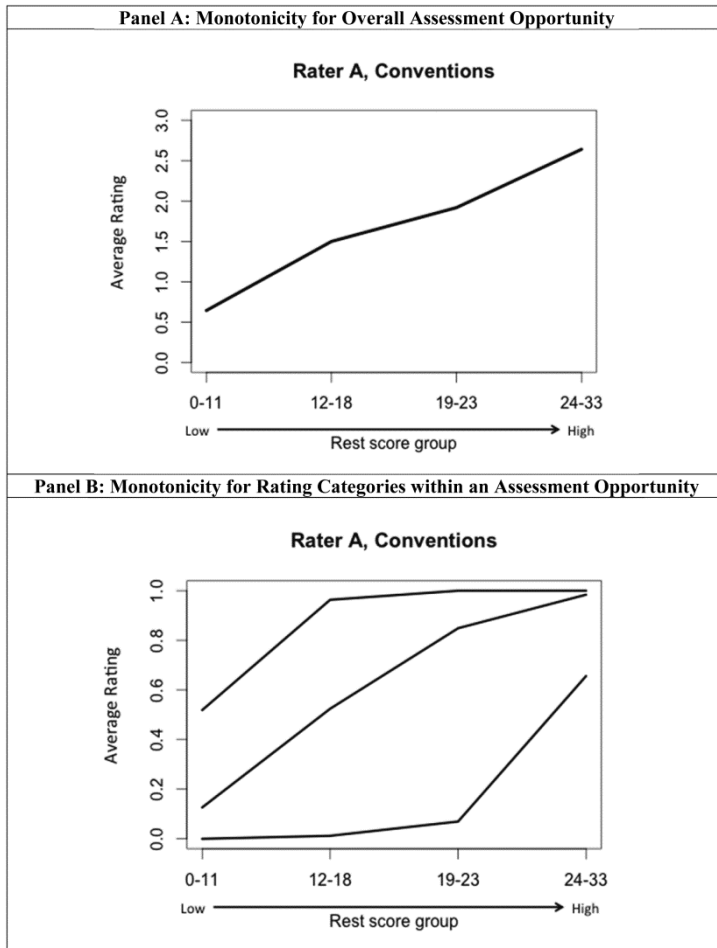


Figure 2:

Diagnostic Plots for Evaluating Monotonicity within Assessment Opportunities (Generalized Ratings)

Results: Georgia writing data. In order to examine monotonicity in the illustrative data, restscores groups specific to each item were calculated for each of the 365 students. The highest rating possible for each item on the recoded rating scale is $X_i = 3$. As a result, the highest possible total score (X_+) for each student across the 12 items is $X_+ = 36$, and the highest possible restscores is $R = 33$, for students with the maximum score [$R = (X_+ - X_i) = (36 - 3) = 33$]. Using restscores groups, monotonicity within items was examined at the overall item level by comparing average ratings associated with an item of interest across

increasing restscore groups through the graphical technique illustrated in Figure 2, Panel A. Next, monotonicity was examined for the 12 items at the level of rating scale categories using plots similar to those shown in Figure 2, Panel B.

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. For each rating scale category k , the proportion of students in a restscore group who received a rating of $X = k$ on an item is compared to the proportion of students in each of the other restscore groups who received a rating in Category k . Following van der Ark (2013), these comparisons were only performed between restscore groups whose proportion of ratings in Category k is greater than zero. For each item, monotonicity analyses were conducted for each rating scale category ($k = 1, \dots, m$). Violations of monotonicity are identified when the probability that $X = k$ is higher for the group with a lower restscore. Results from hypothesis tests to examine the significance of monotonicity violations for the Georgia writing data are summarized in Table 3, Column B, which indicates that there were no significant violations of monotonicity for the items examined in this study.

C. Invariant ordering

The next Mokken rating quality indicator for analytic ratings is invariant ordering. This indicator is based on the Double Monotonicity for Raters (DM-R) model, which is an adaptation of Mokken's (1971) DM model for use with polytomous rater-assigned scores (Wind, 2014). Methods for examining adherence to the DM model for polytomous data typically consider invariant ordering at the level of ISRFs (double monotonicity) as well as at the level of overall items (invariant item ordering; IIO). As mentioned above, recent research has highlighted a potential mismatch between the polytomous DM model (Molenaar, 1982, 1997) and IIO under certain circumstances. Based on the interpretation of invariant ordering at the overall item level (rather than within rating scale categories), Ligetvoet and his colleagues (2010, 2011) suggested the use of Manifest Invariant Item Ordering (MIIO) to investigate invariant ordering using average ratings on items.

Wind (2014) proposed the use of *Manifest Invariant Rater Ordering* (MIRO) to examine non-intersection across pairs of raters who assigned holistic ratings. As given in Wind (2014), the MIRO procedure consists of two major steps. First, raters are ordered in terms of severity (i.e., difficulty) by their average holistic ratings across the sample of students. Based on the DM-R model, invariant ordering is implied when the following ordering property holds: If the conditional expected ratings assigned by Rater i and Rater j are ordered for a student with restscore R such that $E(X_i) < E(X_j)$, then these two raters are ordered as follows:

$$E(X_i | R) \leq E(X_j | R), \quad (3)$$

for all values r of rest score R .

Invariant ordering for analytic ratings. In order to determine whether invariant ordering holds in a set of analytic ratings, it is necessary to examine the ordering requirement in Equation (3) for pairs of items. This is accomplished by determining the relative or-

dering of rater severity across domains (i.e., ordering the items by difficulty), and comparing average student ratings across pairs of items, given their rest scores. Violations of invariant ordering occur when the ordering of item difficulty shifts across high and low restscore groups. The interpretation of violations of invariant ordering within the context of analytic ratings varies depending on the pair of items included in a comparison. Specifically, the use of the generalized rating structure results in three possible combinations of items in checks for invariant ordering: 1) the same rater assigning ratings on two different domains; 2) two raters assigning ratings on the same domain; and 3) two raters assigning ratings on two domains. The interpretation of crossing response functions depends on the nature of each combination of items.

1. Same rater, different domains. Figure 3, Panel A includes two example diagnostic plots from the illustrative data set for invariant ordering based on the comparison of ratings assigned by the same rater on different domains. The plot on the left shows invariant ordering for two items. The invariant ordering for these two items suggests that Rater B interprets the relative difficulty of the conventions and organization domains in the same way across levels of student achievement. On the other hand, the plot on the right shows a violation of invariant ordering. The crossing response functions suggest that Rater C's severity on the Style and Conventions domains varies across levels of student achievement. Crossing response functions in this context imply that a rater's interpretation of domain difficulty cannot be interpreted in the same way across levels of student achievement.

2. Different raters, same domain. Figure 3, Panel B includes two example diagnostic plots from the illustrative data set for invariant ordering based on the comparison of ratings assigned by different raters on the same domain. The plot on the left shows invariant ordering for the items, where the relative severity of Rater C and Rater B is consistent on the conventions domain across levels of student achievement. On the other hand, the plot on the right shows a violation of invariant ordering, where crossing response functions indicate that the relative severity of Rater B and Rater A on the sentence formation domain varies across levels of student achievement. Crossing response functions in this context imply that rater severity within domains cannot be interpreted in the same way across levels of student achievement.

3. Different raters, different domains. Figure 3, Panel C includes two example diagnostic plots from the illustrative data set for invariant ordering based on the comparison of ratings assigned by different raters on different domains. The plot on the left shows invariant ordering for the items. Non-intersecting response functions for the two items indicate that the relative severity of Rater B and Rater C on the conventions and organization domain can be interpreted in the same way across levels of student achievement. On the other hand, the plot on the right illustrates a violation of invariant ordering for the two items. In the context of this comparison, a violation of invariant ordering implies that the relative severity ordering of Rater A on the conventions domain and Rater C on the organization domain varies across levels of student achievement. Crossing response functions imply that the two raters are not interpreting the relative difficulty of the two domains in the same way across levels of student achievement.

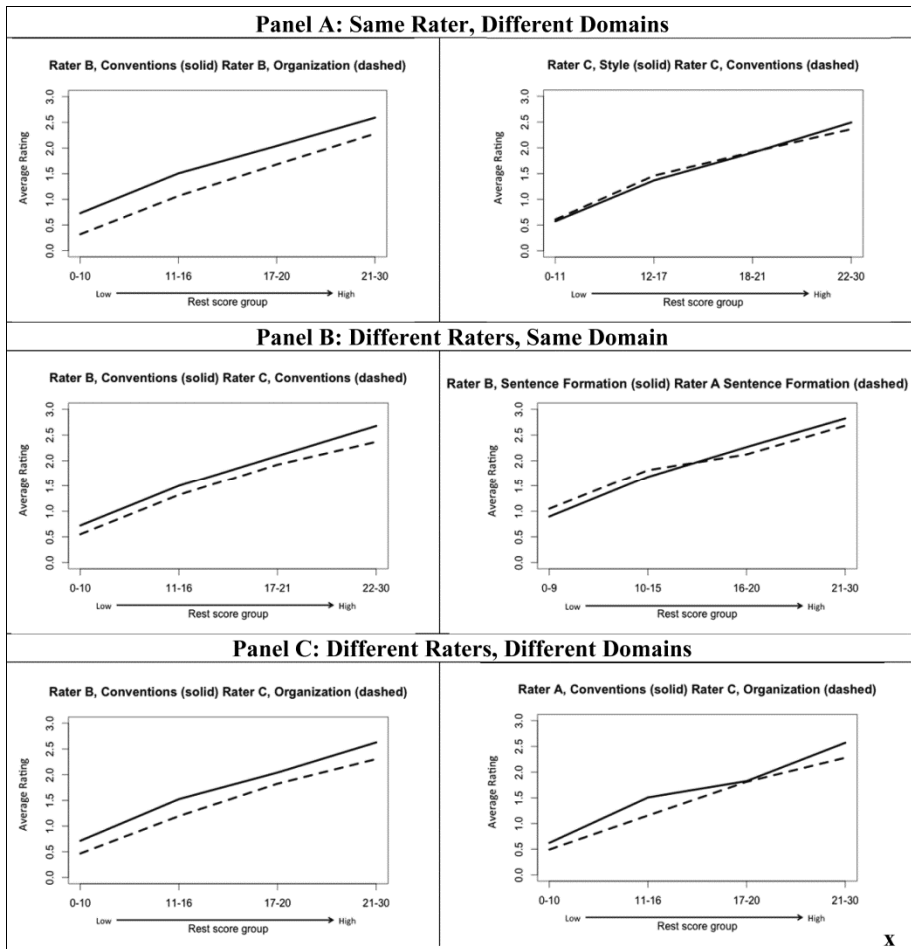


Figure 3:

Diagnostic Plots for Evaluating Invariant Ordering across Assessment Opportunities (Generalized Ratings)

Violations of invariant ordering for analytic ratings. Statistical hypothesis tests can be used to determine the significance of intersections. For example, if the overall average ratings from Item i and Item 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 items 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 ordering. If a violation is observed, a hypothesis test is used to determine whether or not the reversal of rater severity is significant. Intersecting response functions over values of restscores imply violations of invariant ordering.

Results: Georgia writing data. In this study, graphical and statistical evidence was used to evaluate invariant ordering for the 12 items using the restscore method illustrated above. Table 3, Column C summarizes results from the invariant ordering analyses for the Georgia writing data. These results suggest that there are significant violations of invariant ordering for several items. The most significant violations of invariant ordering are observed for the style ratings from Rater C and the style ratings from Rater A. For both of these items, violations of invariant ordering suggest that comparisons with the other items in this sample resulted in significant disordering of item difficulty (i.e., intersecting response functions). It is important to note that this finding that a particular item was involved in a violation of invariant ordering does not necessarily imply low-quality ratings associated with a particular rater on a given domain. Because pairwise comparisons are used to investigate invariant ordering, the interpretation is ambiguous with regard to which item within a pair contributes to a model violation.

Conclusions

When applied to analytic scoring systems, indicators of rating quality provide a method for evaluating the diagnostic quality of ratings assigned in separate domains in order to inform the interpretation and use of scores from rater-mediated assessments. The major contribution of this study was the use of a generalized rating design for investigating rating quality from the perspective of Mokken scaling. Although previous research on Mokken scaling has demonstrated the use of Mokken's (1971) nonparametric scaling models for examining polytomous ratings in a variety of contexts, these models have not been applied to the context of analytic ratings in which each performance is evaluated in terms of multiple domains. Overall, results from this study suggest that a generalized rating design can be used to explore rating quality within the framework of Mokken scale analysis (Mokken, 1971). In this section, conclusions are drawn for the two guiding research questions for this study. A discussion of directions for future research follows.

What does Mokken scale analysis reveal about the quality of analytic domain ratings?

The first research question for this study focused on the use of Mokken scale analysis (Mokken, 1971) as a method for evaluating the quality of analytic ratings. This study illustrated the application of Mokken-based statistics and displays that are traditionally used within Mokken scale analysis to the context of an analytic scoring system. Building upon the Mokken-based rating quality indices for holistic ratings proposed by Wind (2014), this study presented and illustrated the use of Mokken scale analysis to explore

major aspects that reflect the psychometric quality of analytic ratings: A) Scalability; B) Monotonicity; and C) Invariant Ordering. The rating quality indices presented in the current study differ from those presented by Wind (2014) and Wind and Engelhard (in press) in the substantive interpretation of each index. Whereas the original Mokken-based rating quality indices were used to evaluate rater scalability, monotonicity, and invariance, the indices presented in the current study were presented as a method for examining these psychometric properties at the level of analytic ratings.

The Mokken-based rating quality indices described in Table 2 are based on a view of analytic ratings as a type of generalized ratings, where each rater-by-domain combination is treated as a unique item. The rating quality indices are used in an exploratory fashion to determine the degree to which a set of items form a strong Mokken scale (scalability) can be used to order students on the latent variable (monotonicity) and can be used to describe the ordering of rater severity within domains (invariant ordering). As pointed out by Wind (2014), these Mokken-based rating quality indices should not be viewed as a method for rejecting or discarding raters based on statistical or graphical criteria alone; the unique purpose of an assessment system should determine levels of acceptable adherence to the Mokken model assumptions. The rating quality indices presented here should be viewed as a set of exploratory tools that can be used to examine the psychometric properties of analytic ratings. Overall, the illustrative data analyses suggested that Mokken scale analysis provides useful information for examining the requirements of invariant measurement in analytic rating data.

What does Mokken scale analysis reveal about differences in the quality of ratings assigned to different domains in an analytic rubric?

The second research question for this study focused on the use of Mokken scaling to identify differences in the quality of ratings assigned to different domains in an analytic scoring rubric. This research question was explored through the application of Mokken-based statistics and displays to generalized ratings in four domains. When examined at the level of rater-by-domain combinations, scalability indices were used to examine the degree to which a set of generalized ratings could be used as a scale to meaningfully order students and raters in terms of a single latent variable. Essentially, the scalability indicators can be seen as a method for identifying individual rater-by-domain combinations that did not scale well with other items in a set of generalized ratings. Monotonicity indices at the level of items describe the degree to which individual raters share a common understanding of student achievement within domains. Indicators of invariant ordering for items provide useful diagnostic information about individual rater interpretation of student achievement across domains, and across pairs of raters within and across domains. Findings of violations of invariant ordering for a particular rater-by-domain combination could be used to guide rater training or remediation in order to ensure a consistent interpretation of a scoring rubric across a group of raters.

Taken together, the results from the illustrative analyses did not indicate major differences in rater scalability or monotonicity across domains. Some differences in invariant

ordering were observed across domains, with more frequent significant violations of invariant ordering observed for ratings assigned to the Style domain than the other domains. However, these results should be considered in light of the fact that the illustrative analyses were conducted using only three randomly selected raters, and the use of different raters might have resulted in different findings. The key idea in the illustrative analysis was to demonstrate the application of Mokken-based statistics and displays to generalized ratings that can be extended to other empirical data and assessment contexts in order to evaluate rating quality within domains.

Discussion

Greatly influenced by the work of Rasch (1960/1980), Mokken's theory and procedure for scale analysis is characterized by an emphasis on an underlying theory and parsimony when evaluating the usefulness of a measurement model. Traditional applications of the Rasch model to rater-mediated assessments include the use of person and item fit statistics based on residuals that describe the degree to which observed responses match the expectations of the model. However, these fit statistics may not be sufficient to fully examine measurement quality in terms of basic measurement properties that are necessary for invariant measurement, including monotonicity, scalability, and invariant ordering. The nonparametric Mokken models for analytic ratings presented in this study can be used as an initial step that can be used to explore the underlying requirements for invariant measurement that must be met before the Rasch model is applied – a step that is often overlooked in Rasch measurement and other modern measurement research. In essence, this research reflects a call to the Rasch measurement community to examine what is really happening “under the hood,” while still maintaining strict requirements for defining measurement.

This study can be viewed as part of a larger program of research on the application of Mokken scaling to the context of rater-mediated educational assessments. Additional research is needed in order to more fully understand the diagnostic utility of Mokken-based rating quality indicators for informing the interpretation and use of rater-assigned scores. Specifically, research is needed that considers the application of Mokken scale analysis to scoring systems with incomplete rating designs and partial-credit scoring. Future research should also consider the application of Mokken-based item selection techniques (e.g., Straat, van der Ark, & Sijtsma, 2013) to the context of rater-mediated assessments. In addition, the use of Mokken scale analysis for examining rating quality in content areas other than writing assessment is warranted.

Additional research on Mokken scaling in general is also needed that examines the degree to which Mokken-based indices of measurement quality detect similar deviations from model assumptions to traditional indices of measurement quality for IRT models, including local dependencies, differential item functioning, and differential person functioning within the context of various types of educational assessments, including rater-mediated and multiple-choice assessments. Additional research should also build upon nonparametric techniques for examining person fit (e.g., Meijer & Sijtsma, 2001;

Tendeiro & Meijer, 2014) to consider the degree to which Mokken-based indices of person fit align with parametric indices, such as those provided by the Rasch model.

References

- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (AERA, APA, NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: Author.
- 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 in Copenhagen, Denmark.
- de Koning, E. Sijtsma, K., & Hamers, J. H. M. (2002). Comparing of four IRT models when analyzing two tests for inductive reasoning. *Applied Psychological Measurement, 26*(3), 302-320.
- Eckes, T. (2011). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Frankfurt am Main: Peter Lang.
- Engelhard, G., Jr. (1997). Constructing rater and task banks for performance assessments. *Journal of Outcome Measurement, 1*(1), 19-33.
- Engelhard, G., Jr. (2008). Historical perspectives on invariant measurement: Guttman, Rasch, and Mokken. *Measurement: Interdisciplinary Research & Perspective, 6*(3), 155-189.
- Engelhard, G., Jr. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. New York: Routledge.
- Grayson, D. A. (1988). Two-group classification in latent trait theory: Scores with monotone likelihood ratio. *Psychometrika, 53*(3), 383-392.
- 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*(3), 225-246.
- Huynh, H. (1994). A new proof for monotone likelihood ratio for the sum of independent Bernoulli random variables. *Psychometrika, 59*(1), 77-79.
- Junker, B. W. (1993). Conditional association, essential independence and monotone unidimensional item response models. *The Annals of Statistics, 21*, 1359-1378.
- 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*(4), 578-595.
- 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.
- Meijer, R. R. & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25*(2), 107-135.
- 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*(3), 283-298.

- Murphy, K. R., & Cleveland, J. N. (1991). *Performance appraisal: An organizational perspective*. Allyn & Bacon.
- Mokken, R. J. (1971). *A theory and procedure of scale analysis*. The Hague: Mouton/Berlin: De Gruyter.
- 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: Springer.
- R Development Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org/>.
- Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research. (Expanded edition, Chicago: University of Chicago Press, 1980).
- 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.
- Straat, J. H., van der Ark, L. A., & Sijtsma, K. (2013). Comparing optimization algorithms for item selection in Mokken scale analysis. *Journal of Classification*, 30, 75-99.
- Tendeiro, J. N. & Meijer, R. R. (2014). Detection of invalid test scores: The usefulness of simple nonparametric statistics. *Journal of Educational Measurement*, 51(3), 239-259.
- Ünlü, A. (2008). A note on monotone likelihood ratio of the total score variable in unidimensional item response theory. *British Journal of Mathematical and Statistical Psychology*, 61(1), 179-187.
- van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software*, 20(11), 1-19. URL <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), 1-27. URL <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(2), 139-163.
- Wang, J. & Engelhard, G., Jr. (2014). A pairwise algorithm in R for rater-mediated assessments. *Rasch Measurement Transactions*, 28(1), 1457-1459.
- Wind, S. A. (2014). Guidelines for rating scales based on Rasch measurement theory and Mokken scaling. *Journal of Applied Measurement*, 15(2), 100-133.
- Wind, S. A. & Engelhard G., Jr. (2013). How invariant and accurate are domain ratings in writing assessment? *Assessing Writing*, 18(4), 278-299.
- Wind, S. A. & Engelhard, G., Jr. (in press). Exploring rating quality in rater-mediated assessments using Mokken scaling. *Educational and Psychological Measurement*.