Educational Measurement
ISSUES AND PRACTICE

NCME national council on measurement in education

Educational Measurement: Issues and Practice
xxxx 2017, Vol. 00, No. 0, pp. 1–17

An Instructional Module on Mokken Scale Analysis

Stefanie A. Wind, *University of Alabama*

Mokken scale analysis (MSA) is a probabilistic-nonparametric approach to item response theory (IRT) that can be used to evaluate fundamental measurement properties with less strict assumptions than parametric IRT models. This instructional module provides an introduction to MSA as a probabilistic-nonparametric framework in which to explore measurement quality, with an emphasis on its application in the context of educational assessment. The module describes both dichotomous and polytomous formulations of the MSA model. Examples of the application of MSA to educational assessment are provided using data from a multiple-choice physical science assessment and a rater-mediated writing assessment.

Keywords: Mokken scaling, nonparametric item response theory

Methods based on item response theory (IRT) are frequently used to inform the development, interpretation, and use of educational assessments across a wide range of contexts. These methods are useful because they provide information about the relationship between student locations on a latent variable and the probability for a particular response (i.e., the item response function; IRF). Within the context of educational measurement, the most frequently applied IRT models are based on a parametric formulation, in which a specific algebraic form is specified that defines the shape of the IRF, and transformations are used that result in measures on an equal-interval scale. It is also possible to explore measurement properties using a nonparametric approach to IRT, such as Mokken scale analysis (MSA).

MSA (Mokken, 1971) is a probabilistic-nonparametric approach to IRT that provides a systematic framework for evaluating measurement quality in terms of fundamental measurement properties. Models based on MSA are considered nonparametric because the relationship between the latent variable and the probability for a response (i.e., the IRF) is not required to match a particular shape, as long as basic ordering requirements are met (discussed further below). Mokken (1971) summarized his motivation for developing this approach to item response modeling as follows:

> 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 . . . [and] 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)

In addition to its use as an alternative to parametric IRT, several scholars have recognized the usefulness of MSA in its own right as a method for exploring fundamental measurement properties, including invariant person and item ordering, when an ordinal level of measurement is sufficient to inform decisions based on a measurement procedure. In particular, several authors (Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001; Meijer & Baneke, 2004; Meijer, Tendeiro, & Wanders, 2015; Reise & Waller, 2009) have pointed out that nonparametric IRT in general, and MSA in particular, is an especially useful approach in contexts in which the underlying response processes are not well understood, such as affective variables. Although MSA has been widely applied to a variety of affective domains, the use of this approach in the context of educational assessments is less common. The perspective emphasized throughout this module is that educational achievement tests also involve response processes that are not well understood, and that MSA provides a useful framework for exploring fundamental measurement properties of these assessments, including the degree to which invariant student and item ordering is observed. For example, performance assessments involve complex response processes in which raters are asked to "translate" their perception of student achievement to a rating scale, using a judgmental process that is mediated by many interacting variables, such as student performances, rubrics, rating scales, and individual rater characteristics. Likewise, response processes for multiple-choice (MC) items involve the interaction between students' locations on a latent variable, item formats, assessment contexts, and other student characteristics.

The application of MSA to educational assessments is appropriate when researchers are interested in exploring the degree to which an assessment conforms to fundamental measurement properties, such as during assessment development or revision. Similarly, this approach is useful in contexts in which information about measurement quality and person and item ordering is needed, but sample sizes are not sufficient to achieve stable estimates based on parametric IRT models. In particular, the investigation of MSA models is useful in that evidence of adherence to MSA model requirements provides support for the interpretation of total scores (i.e.,

Stefanie Wind, Department of Educational Research, University of Alabama, Tuscaloosa, AL; swind@ua.edu.

sum scores) to meaningfully order persons and items on the latent variable without any parametric transformations. MSA is also appropriate when the desired conclusions from an assessment procedure do not require an interval-level scale, but can be informed by ordinal information related to students and items. When interval-level measures are needed (e.g., for computerized adaptive testing or certain equating procedures), the procedures illustrated in this module can provide an initial exploratory overview of the degree to which the measurement procedure adheres to basic measurement requirements that can act as a lens through which to explore the data using both numeric results and graphical displays (Sijtsma & Meijer, 2007).

Purpose

The purpose of this instructional module is to provide an introduction to MSA as a probabilistic-nonparametric framework in which to explore measurement quality, with a special emphasis on its application to the context of educational assessment. Several introductions to MSA have been published, including two introductory level books (Sijtsma & Molenaar, 2002; van Schuur, 2011), several book chapters (Meijer et al. 2015; Mokken, 1997; Molenaar, 1997), and a tutorial related to the use of this approach in psychology (Sijtsma & van der Ark, 2017). However, a brief module that presents an accessible and didactic introduction to MSA in the context of educational assessment has not been published. Furthermore, the ITEMS instructional module series has not yet included a presentation of nonparametric techniques for evaluating the quality of educational assessments. By presenting an introduction to MSA in the form of an ITEMS module, I hope to provide a concise and accessible summary of the key features of this approach and its applications in educational assessment that will benefit practitioners, researchers, and students who are interested in learning more about this approach.

The module is organized as follows. First, I present an overview of Mokken's (1971) original dichotomous monotone homogeneity (MH) and double monotonicity (DM) models, followed by a discussion of Molenaar's (1982, 1997) polytomous formulations of the MH and DM models. Second, I present a general procedure for applying MSA to educational assessments. The module concludes with illustrations of the application of the general procedure to educational assessments that highlight the feasibility and usefulness of this approach across multiple types of educational measurement procedures.

Mokken Scaling Models for Dichotomous Responses

In the original presentation of MSA, Mokken (1971) presented two models for evaluating the quality of scales made up of dichotomous items (scored as $X = 0$ or $X = 1$). In order for dichotomous responses to be suitable for MSA, a score of 1 should reflect a higher location on the latent variable (in the context of achievement tests: higher ability) than a score of 0. The first model is the MH model, which is the more general of the two original MSA models. The MH model is based on three underlying assumptions that can be defined as follows: (1) *Monotonicity*: As person locations on the latent variable increase, the probability for correct response ($X = 1$) does not decrease; (2) *Unidimensionality*: Item responses reflect evidence of a single latent variable; and (3) *Local independence*: Responses to an item are not influenced by responses to any other item, after controlling for the latent variable.

Several properties are important to note about the MH model. First, the MH model does not restrict the shape of the IRF beyond the requirement for monotonicity. As a result, IRFs that adhere to the MH model may take on a variety of shapes that do not necessarily match the logistic ogive shape that is typically associated with parametric IRT models. Figure 1(A) shows a pair of IRFs for dichotomous items that meet the assumptions of the MH model. Specifically, the y-axis shows the probability for a correct response (i.e., the IRF), and the x-axis represents the latent variable (θ). Although the IRFs for Item i and Item j are intersecting, they adhere to the MH model requirements because they do not decrease over increasing locations on the latent variable. Second, when data fit the MH model assumptions, the relative ordering of students on the latent variable is invariant across items. Because fit to the MH model assumptions provides evidence for invariant ordering of persons across items, this model can be viewed as analogous to the two-parameter logistic model in parametric IRT.

Mokken's (1971) second model is the DM model, which is a special case of the MH model. The DM model shares the same three assumptions as the MH model, but includes a fourth assumption: (4) *Invariant item ordering (IIO)*: Response functions for individual items do not intersect with response functions for any other item. Under the DM model, IRFs may take on a variety of shapes as long as they do not intersect. Figure 1(B) shows a pair of IRFs for dichotomous items that meet the assumptions of the DM model. In contrast to the plot in Panel A, the pair of items shown in Panel B meets the DM model requirements because the IRFs for Item i and Item k are both monotonic (nondecreasing over the latent variable), and nonintersecting. The important result of this requirement for dichotomous items is that when data fit the DM model assumptions, the items are ordered the same way across students. Because fit to the DM model assumptions provides evidence for invariant ordering of both students and items, this model has been described as an ordinal version of the dichotomous Rasch model, or the one-parameter logistic model in parametric IRT (Engelhard, 2008; Meijer, Sijtsma, & Smid, 1990; van Schuur, 2003).

Mokken Scaling Models for Polytomous Responses

Molenaar (1982, 1997) proposed polytomous formulations of the MH and DM models that have facilitated the widespread application of MSA to a variety of domains in which rating scales are used. The polytomous MH and DM models are based on the same underlying assumptions as Mokken's (1971) dichotomous formulations. However, under the polytomous formulation, the model assumptions are evaluated for each item both at the overall item level and within rating scale categories. Similar to parametric IRT models for polytomous items, the polytomous MH and DM models are based on a set of response functions that describe the probability for a rating in or above a particular rating scale category across values of the latent variable.

In the context of MSA, response functions for rating scale categories (i.e., category response functions) are called *item step response functions* (ISRFs; Molenaar, 1982, 1997). For a rating scale with k categories, a set of $k - 1$ ISRFs are specified

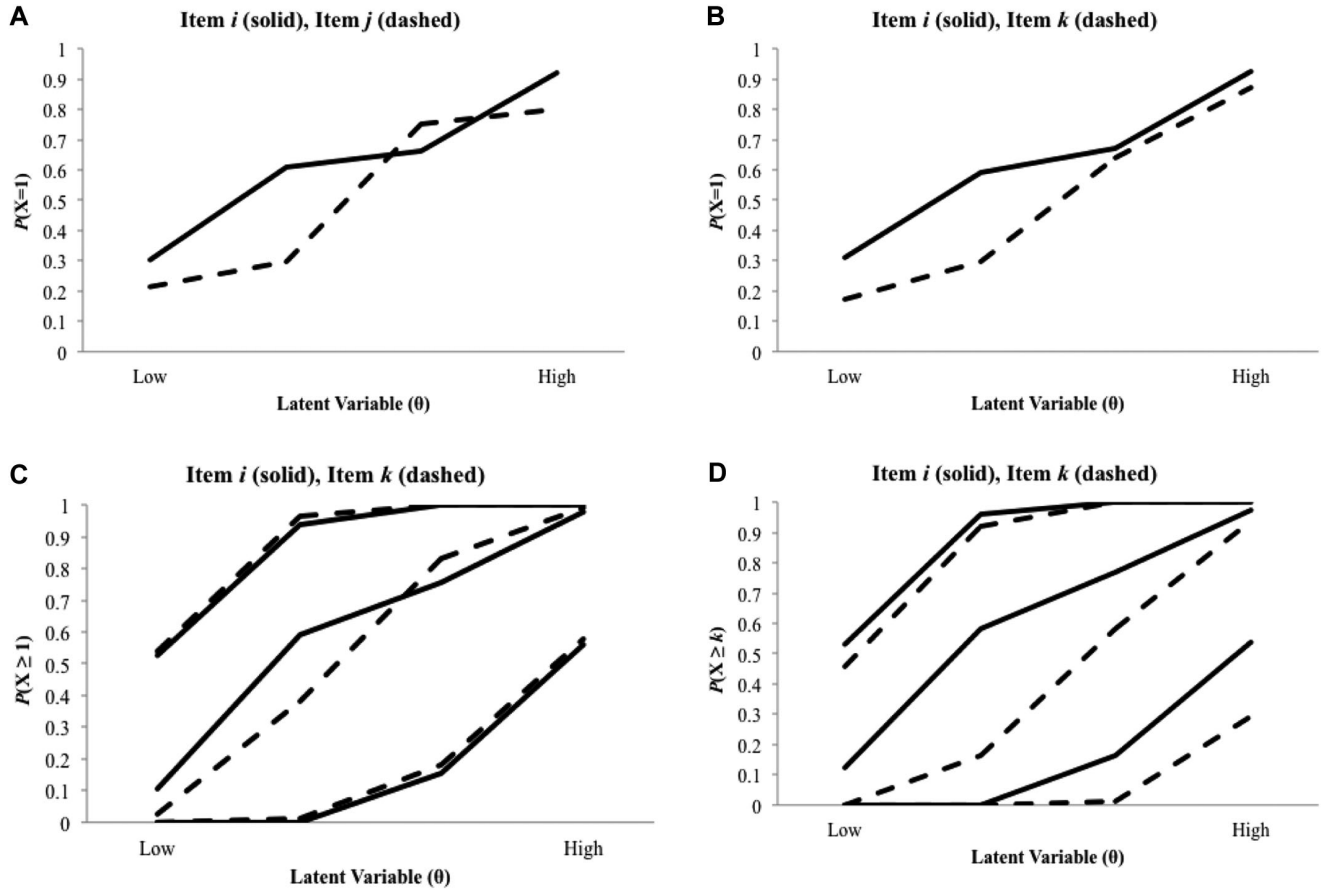

FIGURE 1. Mokken item response functions.

that reflect the probability for a response in a given rating scale category or higher across levels of the latent variable. According to Molenaar (1997, p. 370), Mokken ISRFs can be conceptualized as a set of $k-1$ "steps" (τ_{ijk}) that reflect the observed rating in category k on Item i for person j, where $\tau_{ijk}=1$ when $X_{ij}\geq k$, and $\tau_{ijk}=0$ otherwise, where X_{ij} is the observed score on Item i for person j.

Similar to the dichotomous MH model, Molenaar's (1982) polytomous version of the MH model is based on the requirements of unidimensionality, local independence, and monotonicity. However, these requirements are evaluated within rating scale. Accordingly, the monotonicity assumption is restated as: *Monotonicity*: The conditional probability for a rating in category k or higher is nondecreasing over increasing values of the latent variable. Figure 1(C) illustrates a set of ISRFs for a rating scale item with four categories that meets the assumptions of the polytomous MH model. The three lines represent the thresholds for a rating in the second category or higher (highest ISRF), the third category or higher (middle ISRF), and the fourth category or higher (lowest ISRF).

An important difference between the dichotomous and polytomous versions of the MH model should be noted. Specifically, whereas adherence to the MH model requirements for dichotomous items implies that the order of student total scores on the raw score scale reflects their ordering on the latent variable (θ; i.e., stochastic ordering on the latent variable [SOL]), a weaker version of SOL is implied by the polytomous MH model (van der Ark & Bergsma, 2010). *Weak SOL* implies that the total score (X+) can be used to divide a sample into a group of students with high locations on the latent variable and a group of students with low locations on the latent variable, such that students with the highest and lowest θ values can be identified by dividing X+ into two groups.

The polytomous DM model (Molenaar, 1982, 1997) shares the requirements of the polytomous MH model, with the additional assumption of nonintersecting ISRFs: *Nonintersecting ISRFs*: The conditional probability for a rating in category k or higher on Item i has the same relative ordering across all values of the latent variable.

Figure 1(D) illustrates a set of ISRFs for a rating scale item with four categories that meets the assumptions of the polytomous DM model.

Recent research related to the polytomous DM model has highlighted a lack of necessary correspondence between nonintersecting ISRFs and the invariant item-ordering property that characterizes the dichotomous DM model. Specifically, explorations of IIO have highlighted the possibility that nonintersecting response functions for rating scale categories for a pair of items, which suggest fit to the polytomous DM model, do not always correspond to IIO when aggregated to the overall item level (Ligtvoet, van der Ark, Bergsma, & Sijtsma, 2011; Ligtvoet, van der Ark, te Marvelde, & Sijtsma, 2010; Sijtsma & Hemker, 1998; Sijtsma, Meijer, & van der Ark, 2011). This phenomenon is illustrated in Figure 2 for two pairs of polytomous items. For example, Panels A and B illustrate the aggregation phenomenon for Item i and Item j: Panel A shows the ISRFs and Panel B shows the IRFs. As can be seen in Panel A, the ISRFs for the two items do not

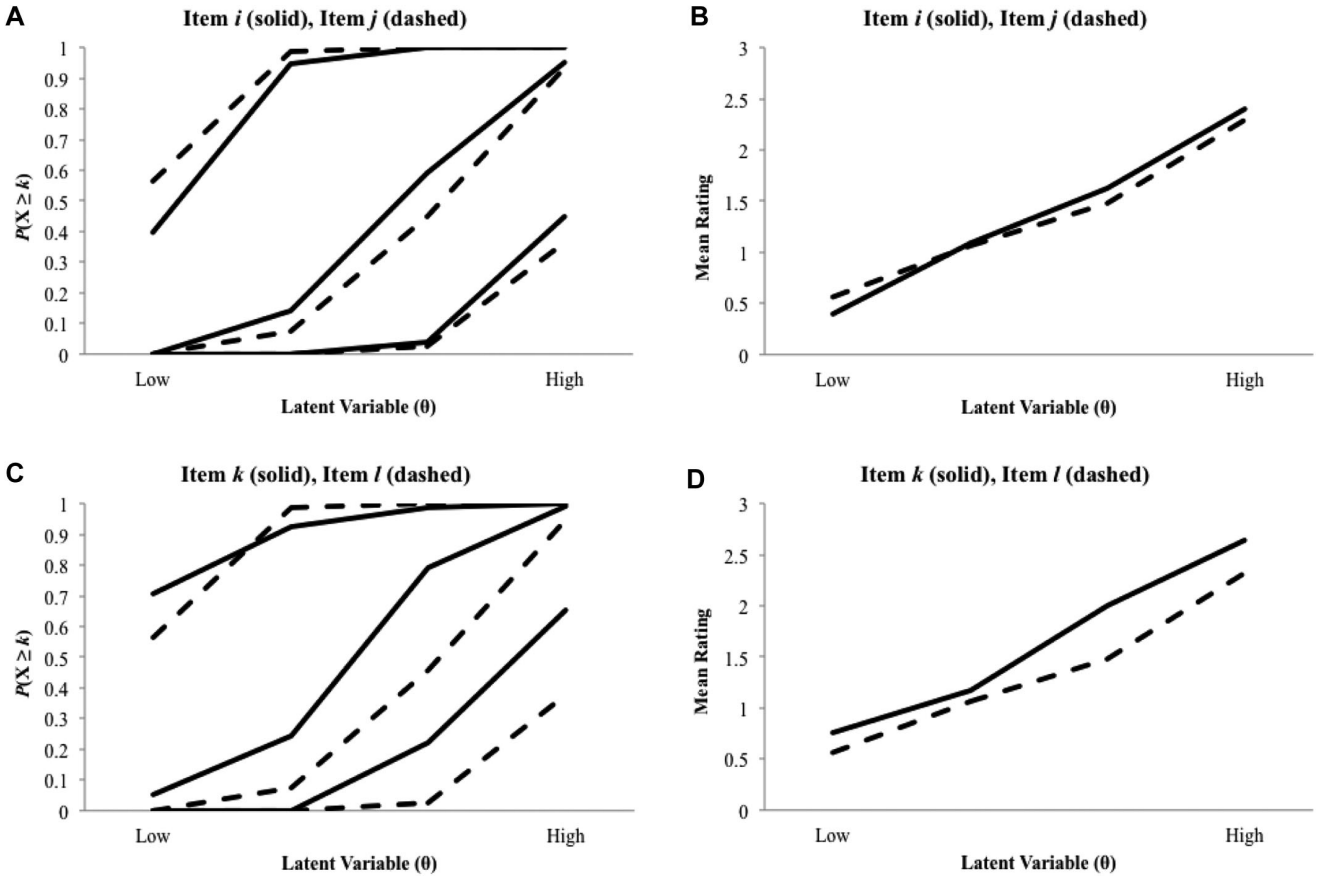

FIGURE 2. Aggregation phenomenon for polytomous items.

intersect—thus meeting the nonintersection requirement of the polytomous DM model. However, the aggregation of the ISRFs in Panel B illustrates intersecting IRFs for the two items, which is a violation of IIO. The opposite result is also true: Figure 2, Panels C and D, illustrates two items for which intersecting ISRFs (Panel C) correspond to nonintersecting IRFs (Panel D).

The recommended solution to this discrepancy between DM and IIO, or *aggregation phenomenon* (Ligtvoet et al., 2011), is to evaluate IIO only as it applies to overall items. Specifically, the combined requirement of invariant ordering at the level of rating scale categories and for the overall item has been described as having "little practical value" and as "unrealistic for many data sets" (Sijtsma et al. 2011, p. 34). Instead, researchers are encouraged to use techniques such as the *manifest IIO* procedure (MIIO) proposed by Ligtvoet et al. (2010, 2011), in which IIO is examined using only overall item scores (using IRFs), while ignoring invariant ordering at the level of rating scale categories. I discuss the MIIO method further in the next section, and demonstrate the approach in the illustrative analyses at the end of the module.

Software for MSA

A variety of graphical displays and statistics are available for evaluating fit to the dichotomous and polytomous MH and DM models. These methods can be implemented automatically using a standalone program: MSP5 (Molenaar & Sijtsma, 2000), as well as in two statistical software programs: (1) the *mokken* package (van der Ark, 2007, 2012) can be used to conduct MSA in *R*; and (2) the commands described by Hardouin, Bonnaud-Antignac, and Sebille (2011) can be used to conduct MSA in the Stata program.

The MSP5 program includes a graphical interface, in which researchers can import data in standard tabular formats (e.g., spreadsheet formats or comma-separated files), and conduct analyses by selecting procedures from dropdown menus. The *mokken* package and Stata commands can also be used to explore data stored in any tabular file format, after it has been read into the software. Although these programs require the use of a command-line interface to conduct analyses, the functions only require a basic understanding of *R* and Stata. Furthermore, the functions are described in detail and examples of procedures for conducting MSA are available in the documentation for both the *mokken* package (van der Ark, 2007, 2012) and the Stata commands (Hardouin et al., 2011). All three programs provide tabular output of numerical results that can be exported to other formats, including comma-separated files, as well as graphical displays that can be exported as images.

Because the MSP5 program is no longer being updated, any developments in MSA research will not be made available in this program. In contrast, the *mokken* package is frequently updated, and includes several advances that are not available in MSP5, such as the MIIO procedure (Ligtvoet et al., 2010), standard errors for scalability coefficients (Kuijpers, van der Ark, & Croon, 2013), and confidence intervals for response function plots. Accordingly, the methods illustrated in this module are illustrated using the *mokken* package.

Evaluating Measurement Quality Using MSA

Figure 3 provides a simplified version of a procedure for applying MSA as a framework for evaluating the measurement properties of an educational assessment that includes four major steps: (1) import the data matrix; (2) analyze assessment opportunities (AOs); (3) interpret results within context; and (4) modify AOs. In this section, I describe each step in Figure 3 theoretically. Then, I illustrate the procedure using data from an educational assessment made up of dichotomously scored MC items, and an educational performance assessment in which students received polytomous ratings on written compositions. These two examples highlight the use of MSA for dichotomous scores and polytomous ratings, respectively. Accordingly, they highlight the wide range of educational assessments that can be explored using MSA.

Import the Data Matrix

The first step in the procedure is to import the data matrix into a software program for MSA. The first panel of Figure 3 shows the basic structure of the data matrix that is used as the starting point for MSA. For basic applications of MSA, data that are included in the analysis include student responses to AOs. In this module, the term *AO* is used to describe procedures used to evaluate students in an educational assessment that result in ordinal scores. For example, AOs can be items, such as MC items, or individual raters who score each student in a performance assessment. Each of the rows of the matrix (1 through N) represents a student (n), each of the columns (1 through L) represents an AO (i), and each cell contains the ordinal scored responses for each student to each item (X_{ni}).

Analyze AOs

After the data have been imported into software for MSA, the second step in the procedure for analyzing educational assessments using MSA is to analyze the AOs using indicators of measurement quality based on the MH and DM models. In this module, I focus on three categories of indicators that can be used to evaluate measurement quality: (A) monotonicity, (B) scalability, and (C) invariant ordering. Table 1 summarizes the alignment between these three categories and the Mokken model assumptions described above.

Monotonicity. The first indicator of measurement quality based on the MH model is monotonicity. For dichotomous items, monotonicity suggests that the probability for a correct response is nondecreasing across increasing locations on the latent variable (θ). For polytomous ratings, monotonicity implies that the cumulative probability for a rating in or above each rating scale category [$P(X_i \geq k)$] is nondecreasing across increasing levels of student achievement within a given rating scale item or for a particular rater (discussed further below).

Because the MH model does not facilitate the estimation of latent variable locations for persons (θ), a nonparametric approximation is needed to evaluate the monotonicity

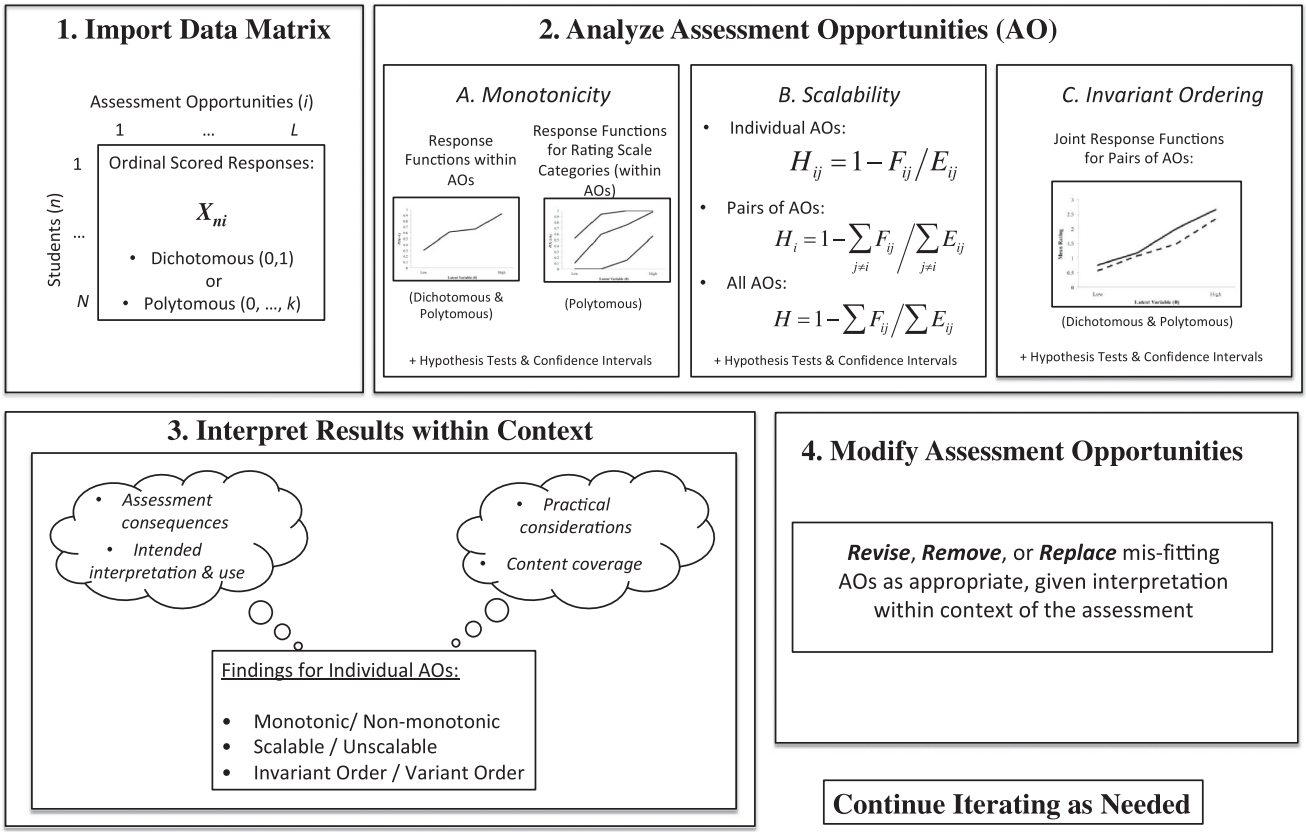

FIGURE 3. Procedure for applying Mokken scale analysis in the context of educational measurement.

Table 1. Alignment Between Mokken Model Assumptions and Indicators of Measurement Quality

Assumptions	Double Monotonicity Model	Monotone Homogeneity Model	Model-Based Indicators
Monotonicity	√	√	(A) Item/Rater monotonicity
Conditional Independence	√	√	(B) Item/Rater scalability coefficients
Unidimensionality	√	√	(A) Item/Rater monotonicity; (B) Item/Rater scalability
Nonintersecting Response Functions	√		(C) Invariant item/rater ordering

assumption. As a result, Mokken analyses are conducted using an estimate based on unweighted sum scores (X_+) for each student across an entire set of items or raters. 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). In order to evaluate model assumptions for an item of interest, Junker (1993) proposed the use of the restscore (R), which is the sum score minus the score on the AO of interest (Hemker, Sijtsma, Molenaar, & Junker, 1997; Junker, 1993; Junker & Sijtsma, 2000; Sijtsma & Molenaar, 2002).

Several methods exist for evaluating monotonicity for nonparametric IRT models in general, including kernel smoothing (e.g., Mazza, Punzo, & McGuire, 2014; Ramsay, 1991). Within the context of MSA, a simple approach to evaluating monotonicity based on restscores is generally used; this method is provided in the *mokken* package for *R* (van der Ark, 2007, 2012). First, restscores are calculated using the total observed score for each student across a set of AOs, minus the scores assigned on the AO of interest. In order to provide additional statistical power, students with adjacent restscores are combined to form restscore groups that reflect the range of latent variable locations. The *mokken* package (van der Ark, 2007, 2012) creates restscore groups automatically based on Molenaar and Sijtsma's (2000) criteria for minimum sample sizes within restscore groups. By default, the minimum sample size within restscore groups is $N/10$ for samples larger than 500; $N/5$ for samples between 200 and 500, and $N/3$ for smaller sample sizes, with a minimum of 50 persons in each group. Along the same lines, van Schuur (2011, p. 52) recommended checking the sample size within restscore groups in order to ensure that a single participant does not make up more than 2% of a restscore group. It is important to note that for monotonicity analyses, restscore groups are calculated separately for each AO of interest. As a result, a different number of restscore groups might be observed across AOs. After restscore groups are calculated, monotonicity is investigated using graphical displays and statistical hypothesis tests.

For dichotomous items, the graphical procedure for evaluating monotonicity involves plotting the probability for a correct response as a function of restscore groups. When MSA is used for dichotomous items, these probabilities are calculated as the proportion of students within a restscore group who earned a score of 1 on the dichotomous AO of interest. Figure 4(A) illustrates this procedure for a dichotomous item. In this figure, student restscores are plotted along the x-axis, and the probability for a correct response is plotted along the y-axis. No violations of monotonicity are observed for Item i because the IRF is nondecreasing over increasing values of restscores.

For polytomous ratings, monotonicity is evaluated for overall AOs and within rating scale categories. First, overall monotonicity is evaluated by examining the average observed ratings across increasing restscore groups. As illustrated in Figure 4(B), nondecreasing average ratings (y-axis) across increasing restscore groups (x-axis) provides evidence of monotonicity for an overall AO. Next, monotonicity can be evaluated within rating scale categories using plots of ISRFs for an item or rater of interest. Figure 4(C) illustrates adherence to the monotonicity assumption because the ISRFs are nondecreasing over increasing restscores. When the graphical approach to investigating monotonicity is used, it is also possible to plot confidence intervals around the estimated response functions as additional evidence regarding the stability of the results. Specifically, the *mokken* package provides optional Wald confidence intervals that can be added to monotonicity plots for both dichotomous and polytomous AOs (van der Ark, 2013).

Statistical hypothesis tests (Z tests; see Molenaar & Sijtsma, 2000, p. 72) for detecting significant violations of monotonicity are available for both dichotomous and polytomous scores. For dichotomous items, a one-sided, one-sample Z test is used to evaluate the null hypothesis that the probability for a correct score is equal across adjacent restscore groups, against the alternative hypothesis that the probability for a correct response is lower in the group with a higher restscore, which would be a violation of monotonicity. A similar Z test is used for polytomous AOs, where the cumulative probability for a rating in category k or higher is compared across adjacent restscore groups. These hypothesis tests for monotonicity can be calculated for individual AOs using the *mokken* package for *R* (van der Ark, 2007, 2012).

Scalability. The next category of measurement quality indices based on the MH model is scalability. Mokken (1971) presented extensions of the scalability coefficients originally proposed by Loevinger (1948) that are used to describe the degree to which individual items, pairs of items, and overall sets of items form a scale that can be used to order persons on a construct. In the context of MSA, scalability coefficients have been described as a method for evaluating a variety of measurement properties, including unidimensionality and local independence (Meijer et al., 2015). However, there is some debate regarding the interpretation of scalability coefficients in dimensionality analyses (e.g., Smits, Timmerman, & Meijer, 2012). Nonetheless, scalability coefficients are used

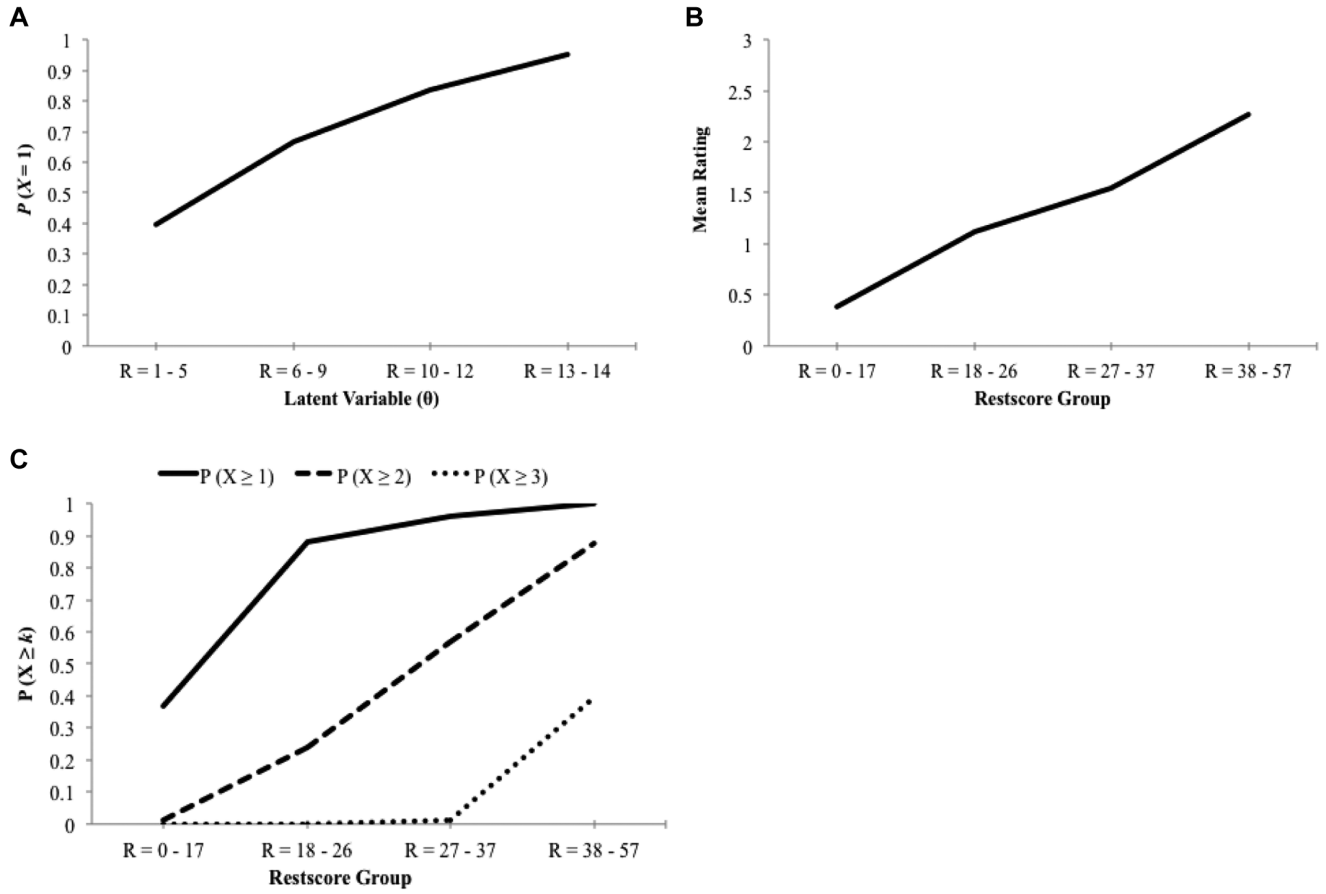

FIGURE 4. Monotonicity plots.

in practice to evaluate adherence to the MH model in the context of MSA.

Indicators of scalability provide a summary of the influence of Guttman errors (described below) on the quality of a measurement procedure, where fewer Guttman errors facilitate the interpretation of person total scores as indicators of person ordering on the construct. Figure 5 illustrates the procedure for detecting Guttman errors for a set of dichotomous items. In both panels, students are ordered from high to low in terms of achievement levels, and items are ordered from easy to difficult. Cell entries with "1" indicate a correct response, and cell entries with "0" indicate an incorrect or negative response. The item responses in Panel A reflect a perfect Guttman pattern, because no combinations of incorrect scores on easier items with correct scores on difficult items are observed. Panel B includes two Guttman errors (indicated with a *).

For both dichotomous and polytomous AOs, scalability coefficients reflect the ratio of observed Guttman errors to the expected frequency of Guttman errors that would be observed based on chance alone within a pair of items:

$$H_{ij} = 1 - \frac{F_{ij}}{E_{ij}}, \tag{1}$$

where F_{ij} is the observed frequency of Guttman errors, and E_{ij} is the expected frequency of Guttman errors. When data fit the MH model, values of scalability coefficients are positive, and range from zero to one, where a value of one indicates no Guttman errors, and values closer to zero indicate many Guttman errors.

In order to illustrate the calculation of H_{ij} for a pair of dichotomous items, Table 2 shows the observed joint frequency of correct and incorrect responses to Item i and Item j for a group of 268 students. Inspection of the marginal frequencies reveals that Item i (total correct = 186) is easier than Item j (total correct = 113). Accordingly, the order of these items in terms of difficulty is $j < i$. Based on this overall ordering, Guttman errors are defined as combinations of an incorrect or negative response on Item i and a correct response on Item j ($X(_{i,j}) = 0,1$). The corresponding error cell is marked, where it can be seen that the observed frequency of Guttman errors is 16. The expected frequency of Guttman errors is calculated based on statistical independence, where the expected frequency within a cell is calculated as: (Row total * Column total)/N. The expected frequency for the error cell in Table 2 is (82 * 113)/268 = 34.57. Using observed and expected frequencies of Guttman errors, scalability is calculated as

$$H_{ij} = 1 - \frac{16}{34.57} = .54. \tag{2}$$

Scalability coefficients are calculated for individual AOs (H_i) using observed and expected Guttman error frequencies for all pairs associated with the item of interest. Likewise, scalability coefficients for an overall set of AOs (H) are calculated using one minus the ratio of observed and expected Guttman error frequencies for all pairs of AOs.

Panel A: Responses Contain No Guttman Errors						
Students		Items				
		Easy	→			*Difficult*
		Item 1	**Item 2**	**Item 3**	**Item 4**	**Item 5**
High	Student 1	1	1	1	1	1
↓	Student 2	1	1	1	1	0
	Student 3	1	1	1	0	0
	Student 4	1	1	0	0	0
Low	Student 5	1	0	0	0	0
Panel B: Responses Contain Two Guttman Errors						
Students		Items				
		Easy	→			*Difficult*
		Item 1	**Item 2**	**Item 3**	**Item 4**	**Item 5**
High	Student 1	0*	1	1	1	1
↓	Student 2	1	1	1	1	0
	Student 3	1	1	1	0	0
	Student 4	1	1	0	0	0
Low	Student 5	1	0	0	0	1*

FIGURE 5. Guttman patterns. *Guttman error.

Table 2. Observed Joint Frequencies of Item *i* and Item *j*

	Item $j = 0$	Item $j = 1$	Total
Item $i = 0$	66	16*	82
Item $i = 1$	89	97	186
Total	155	113	268

In addition to the observed and expected error frequency method described above, scalability coefficients can also be calculated using covariances. This method involves calculating the observed covariance between two AOs, and the covariance that would be obtained if no Guttman errors were observed (maximum covariance). For additional details regarding the calculation of scalability coefficients using the covariance method, see Sijtsma and Molenaar (2002, pp. 52–53).

Molenaar (1991) presented polytomous formulations of the scalability coefficients for individual items, item pairs, and overall groups of items. Using the observed and expected Guttman error method, the polytomous H_{ij} coefficient is calculated as follows. First, the cumulative category probabilities (ISRFs) are used to establish the Guttman pattern for each pair of AOs. For example, the Guttman pattern for a pair including polytomous Item i and polytomous Item j might be defined as $X_i \geq 1, X_i \geq 2, X_j \geq 1, X_j \geq 2, X_i \geq 3, X_j \geq 3$ based on the observed cumulative probabilities for each item. If no Guttman errors were observed, each pair of observed ratings within the pair of items (X_i, X_j) would follow this sequence, such that the expected order with no Guttman errors would be defined as (0,0), (1,0), (2,0), (2,1), (2,2), (3,2), (3,3). Observations in the other cells in the joint frequency table for this pair of items are defined as Guttman errors (see Table 3). Weights are defined for each error cell by calculating the number of errors involved in arriving at the score pattern, based on the Guttman pattern established using the cumulative probabilities (for additional details about weights, see Molenaar & Sijtsma, 2000, pp. 20–22). The (weighted) observed and expected frequencies of Guttman errors are used to calculate H_{ij} using Equation 1.

Interpreting values of scalability coefficients. When data fit the MH model, values of H range from $.00 \leq H \leq 1.00$, where a value of 1.00 indicates a perfect Guttman pattern (no observed Guttman errors). 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. Although these classification criteria frequently appear in empirical applications of MSA, their interpretation and use varies across applications with different purposes. In particular, because scalability coefficients have primarily been considered in psychological measure of affective or developmental variables, these rule-of-thumb values may not hold the same interpretation for educational achievement tests, especially those that involve raters (discussed further below). Furthermore, when scalability coefficients are calculated for persons, several scholars have recommended the use of zero as a lower-bound critical value for evidence of model data (Meijer & Sijtsma, 2001; Sijtsma & Meijer, 1992). As with critical values for any coefficient, the unique measurement context should inform item and person selection criteria.

In addition to critical values, standard errors can also be examined in order to aid in the interpretation of scalability coefficients. Specifically, the *mokken* package for *R* (van der Ark, 2007, 2012) calculates asymptotic standard errors for all three types of scalability coefficients (H, H_i, H_{ij}). Similar to standard errors in other statistical contexts, these values can be used to calculate confidence intervals (Kuijpers et al., 2013) using the general form for a 95% confidence interval: $95\%\text{CI} = H_i \pm 1.96^* se(H_i)$. The *mokken* package also provides significance tests for scalability coefficients in the form of Z tests whose values indicate the degree to which these coefficients are significantly different from zero (for additional details, see Molenaar & Sijtsma, 2000, pp. 59–62; Sijtsma & Molenaar, 2002, p. 40; van der Ark, 2007, 2012). Confidence

Table 3. Joint Frequency Table for Two Polytomous Items

		Item *j*					
		$k = 0$	$k = 1$	$k = 2$	$k = 3$	Marginal Frequency for Item *i*	$P(X_i \geq k)$
Item *i*	$k = 0$	3	0*	0*	0*	3	1.00
	$k = 1$	4	7*	3*	0*	14	.98
	$k = 2$	10	22	34	3*	69	.91
	$k = 3$	9*	17*	40	26	92	.52
	Marginal Frequency for Item *j*	26	46	77	29	178	
	$P(X_j \geq k)$	1.00	.85	.60	.16		

Note. Cells containing Guttman errors are indicated with an asterisk (*). These data are from Ligtvoet (2010, pp. 22–24).

intervals can be used to compare scalability coefficients to other meaningful values, such as zero or Mokken's (1971) lower-bound criteria of .30. Confidence intervals can also be used to compare scalability coefficients across person subgroups of interest. Specifically, separate MSA analyses are needed in order to calculate values of H_i, H_{ij}, or H within each subgroup. Then, confidence intervals can be examined for each AO, AO pair, or overall set of AOs in order to identify differences in scalability that may warrant further investigation.

Scalability and automated item selection. One of the major uses of the scalability coefficient is to select sets of AOs that demonstrate adherence to the assumptions of the MH model. Mokken's (1971) original presentation of his scaling procedures includes a bottom-up method for selecting items that meet the assumptions of the MH model. Computer applications that implement MSA have been developed to include an automated item selection procedure (AISP) that identifies sets of scalable items using H coefficients. Although the AISP procedure has been applied as a technique for evaluating dimensionality and selecting items in affective domains, the use of this procedure for educational assessment items has not been fully explored. As a result, it is not yet clear how automated item selection based on scalability coefficients apply to this context.

Invariant ordering. When considering the application of MSA to educational achievement tests, it is important to note that the implications of invariant ordering are somewhat different than the implications in "traditional" MSA analyses. Specifically, a consistent ordering of items or raters across levels of student achievement could be a fairness concern that must be supported with empirical evidence in order to inform the interpretation and use of scores. Further, in educational achievement contexts, the observed order of AO difficulty is generally not compared to some a priori specification of the expected ordering. Instead, evidence is simply needed that the order is consistent for all students in the sample.

Methods for evaluating the invariant ordering assumption in empirical data involve examining response functions for evidence of nonintersection. Similar to the monotonicity assumption, IIO can be evaluated for dichotomous AOs using both graphical and statistical evidence. Several procedures are available for evaluating IIO for dichotomous AOs, including the restscore method, the item-splitting method, proportion matrices (P++/P−− matrices), and the restscore-splitting method (Sijtsma & Molenaar, 2002; van Schuur, 2011). For illustrative purposes, this module focuses on the use of the restscore method to illustrate the evaluation of IIO. Details about additional techniques for examining IIO are provided in Sijtsma and Molenaar (2002).

For dichotomous AOs, the graphical procedure for evaluating IIO based on the restscore method involves plotting the probability for a correct response within a pair of AOs as a function of restscore groups. Figure 6, Panels A and B, illustrates this procedure for two dichotomous items. In this figure, student restscores are plotted along the x-axis, and the probability for a correct response is plotted along the y-axis. It is important to note that restscores are calculated as the total observed score (X_+) minus the score on *both items* within the item pair for which IIO is being evaluated. In Panel A, no violations of IIO are observed for Item i and Item j, because the IRFs have the same relative ordering across the range of restscores. On the other hand, Panel B illustrates two items that do not demonstrate IIO because the relative order of Item i and Item j is reversed for the third restscore group ($R =$ 9–11), compared to the other three restscore groups.

As noted above, recent research on IIO based on the polytomous DM model has highlighted the potential for a discrepancy between nonintersecting ISRFs and IIO, where nonintersection at the level of rating scale categories does not always imply invariant ordering when ISRFs are aggregated to the overall item level. As a result, Ligtvoet et al. (2010, 2011) have encouraged researchers to apply the MIIO method, in which IIO is evaluated for overall items. The graphical procedure for MIIO is illustrated in Figure 6, Panels C and D. In this figure, student restscores are plotted along the x-axis, and the average observed rating is plotted along the y-axis. In Panel C, no violations of IIO are observed for Item i and Item j, because the IRFs have the same relative ordering across the range of restscores. On the other hand, two violations of IIO are observed in Panel D, because the relative order of Item i and item k is reversed for the middle two restscore groups ($R = [17, 25]$ and $R = [26, 36]$, respectively), compared to the first and fourth restscore groups ($R = [0,16]$ and $R = [37, 54]$, respectively). Similar to the graphical procedures for investigating monotonicity, it is also possible to plot Wald confidence intervals around the estimated response functions as additional evidence regarding the stability of the results (van der Ark, 2013).

Statistical hypothesis tests based on the t distribution (t-tests) for detecting significant violations of IIO are available for both dichotomous and polytomous AOs. For a pair of items ordered $i < j$, the null hypothesis that the probability for a correct response is equal across the two items is evaluated against the alternative hypothesis that the item order is reversed ($j < i$), which would be a violation of IIO. A similar test is used for polytomous ratings. For example, if the overall

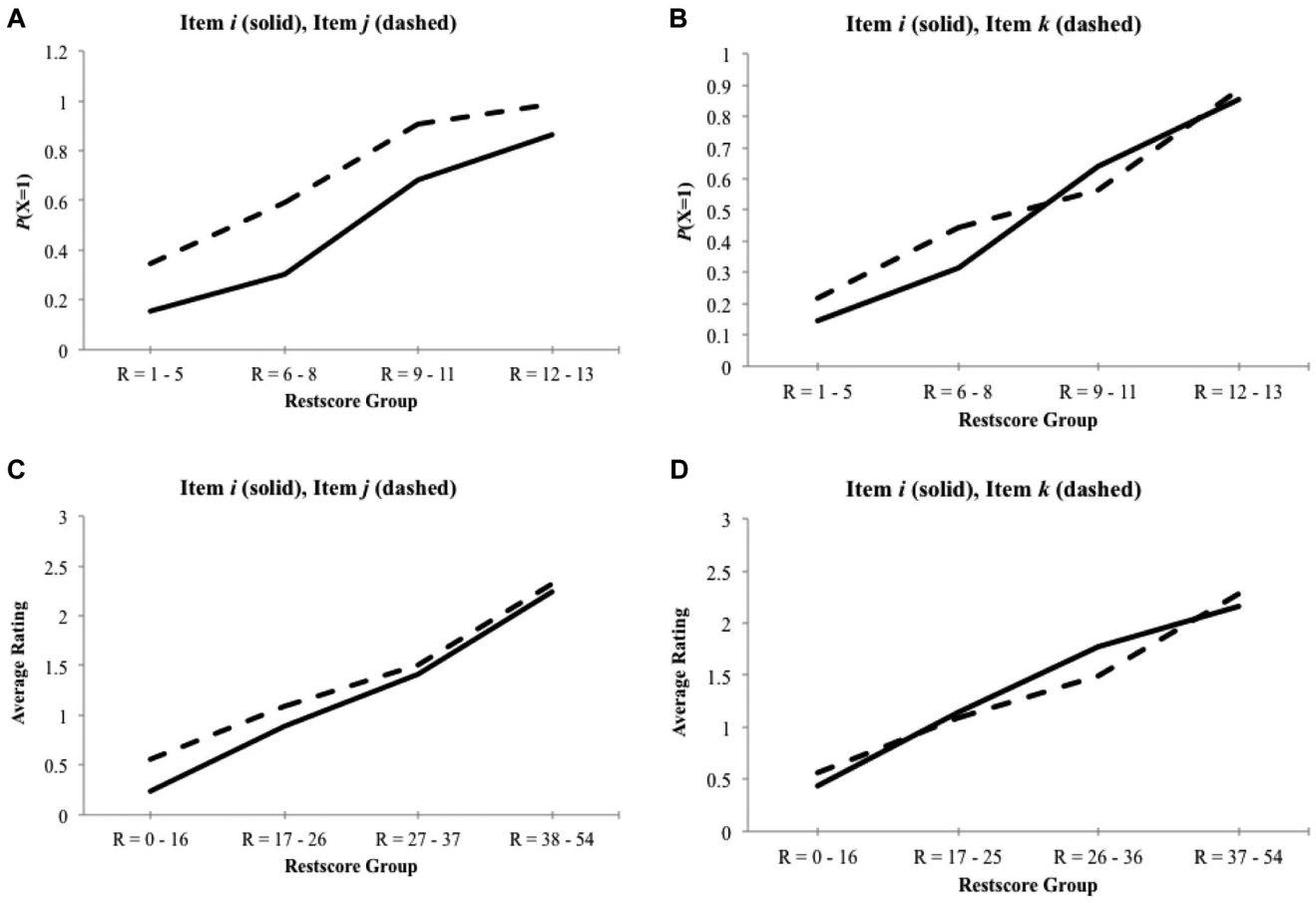

FIGURE 6. Invariant ordering plots.

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 ordering for raters.

Additional Mokken statistics. In addition to the indices of monotonicity, scalability, and invariant ordering presented above, it is important to note that methods based on MSA can also be used to calculate additional statistics related to an overall critical value for model violations ("Crit"; see Molenaar & Sijtsma, 2000) and reliability analyses; both the Crit statistic and Mokken reliability statistics result in single coefficients whose interpretation has not been fully defined or explored in the Mokken scaling literature in general (Meijer et al., 2015), or within the context of Mokken analyses of educational assessments. Rather than using these summary level coefficients to evaluate measurement quality, indicators of monotonicity, users of MSA in the context of educational measurement are encouraged to examine results related to scalability and invariant ordering at the level of individual AOs.

Additionally, person-fit analyses can be conducted within the framework of MSA by transposing the data matrix in Step 1 of Figure 3. For brevity, the analyses in this module are limited to AOs. Additional details regarding the use of MSA for person-fit research are provided in Emons (2008), Meijer, Muijtjens, and van der Vlueten (1996), and Meijer and Sijtsma (2001).

Interpret Results Within Context

The third major step in the analytic procedure outlined in Figure 3 is to interpret the results from the analysis of AOs within the context of the educational assessment. As shown in the figure, findings from the monotonicity, scalability, and invariant-ordering analyses should be considered at the level of individual AOs in terms of the unique context in which the assessment is used. Essentially, the goal of this step is to examine the results in terms of two main considerations: (1) the practical implications of violations of the MSA model requirements in terms of the intended interpretation and use of the assessment and the consequences of the assessment; and (2) opportunities for improving the quality of the AO in subsequent iterations, which depend on practical considerations, such as time and other resources, as well as the role of the AO's content in terms of the alignment with standards or objectives in the blueprint of test content. For example, when violations of IIO are observed related to a particular AO, these results should be considered in terms of the importance of a common item ordering for the interpretation of student scores, given the intended use and consequences of the assessment. Second, the degree to which revisions to the

item are possible in light of practical constraints should be considered, along with the implications of removing or substantially revising the AO on the content coverage within the assessment.

Modify AOs

The fourth step in the analytic procedure shown in Figure 3 is to modify the set of AOs. When it is possible to revise AOs for which violations of monotonicity, scalability, and invariant ordering were observed, these revisions should be informed by best practices for assessment development and revision in educational measurement, such as reviews by expert panels of content experts and cognitive interviews (AERA, APA, & NCME 2014; Lane, Raymond, Haladyna, & Downing, 2011). Following revisions, the assessment will have to be readministered with the revisions before additional analyses can be conducted. In the event that revisions are no longer possible or warranted, the interpretation of results in Step 3 should inform the systematic removal and/or replacement of AOs from the data matrix.

After the data matrix has been updated, additional iterations of the analytic procedure should be conducted until an appropriate set of AOs has been identified. The number of iterations and the degree to which a set of AOs demonstrates properties that are sufficient for operational use should be determined based on the unique context of the educational assessment system.

Illustrations: Applications in Educational Measurement

In the following sections, I illustrate the analytic procedure in Figure 3 using two authentic examples of educational assessments. The two examples were selected in order to demonstrate the application of MSA across two commonly used formats for educational assessments: dichotomously scored MC items, and a writing assessment scored by human raters using polytomous ratings. The procedures for both analyses follow the same general steps, with some differences in the interpretation of results when the dichotomous and polytomous MSA models are applied. The data used in the examples, along with *R* code for the analyses, can be accessed using the link listed at the end of the module.

Illustration 1: MC Assessment Items

In the first illustrative analysis, I demonstrate the application of the four-step analytic procedure shown in Figure 3 to a set of dichotomously scored MC items (N_{items}=15) from a physical science assessment that was administered to a sample of 268 middle school students. The physical science AOs were based on concepts related to physical force, including net force and interpreting force diagrams. Furthermore, an exploratory analysis based on the AISP revealed that the items could be described using a single Mokken scale—thus providing evidence to support the assumption of unidimensionality.

The assessment was developed in conjunction with a semester-long experimental science curriculum for middle school students in the United States, and the assessment data explored in this illustration were collected following the initial implementation of the curriculum. The major purpose of the first administration of the assessment was to explore its psychometric properties, with the goal of revising the assessment such that it could be used to inform instructional materials and teacher-training procedures related to the science curriculum. Because the desired information from the physical science assessment was related to the overall quality of the assessment procedure in terms of fundamental measurement properties, MSA was an appropriate analytic tool for this assessment context. Additional details about the assessment data are provided in Wind (2016).

Import the Data Matrix

In the context of the physical science assessment, the MC items made up the AOs that were represented in the columns of the data matrix. After the student responses to the MC items were collected, they were scored dichotomously, such that a score of zero indicated an incorrect response, and a score of one indicated a correct response. For the purpose of this illustration, the matrix of dichotomously scored responses to the MC science items was imported into the *mokken* package (van der Ark, 2007, 2012) for analysis.

Analyze AOs

Next, the AOs were evaluated using the three categories of measurement quality indicators described above: (A) monotonicity; (B) scalability; and (C) invariant ordering.

Monotonicity. Results for the physical science MC items revealed no violations of monotonicity. However, inspection of the IRFs used to evaluate monotonicity revealed differences in the slope and shape of IRFs across the 15 items, with some items demonstrating sharper areas of discrimination for various achievement levels. Figure 7 illustrates the range of shapes of IRFs that were observed among the MC items. Because no violations of monotonicity were observed for the MC items, these findings suggest that the relative ordering of students in terms of the construct is consistent across the MC items. In other words, this finding suggests that the relative conclusions about the ordering of the middle school students in terms of physical science knowledge are consistent across all 15 items.

Scalability. The overall scalability coefficient for the 15 physical science items is $H = .41$ ($SE = .02$). The 95% confidence interval for this estimate ($H \pm 1.96{*}SE(H) = [.37, .45]$) ranges from a weak Mokken scale to a medium Mokken scale. Individual item scalability coefficients (H_i) and their corresponding standard errors are given in Table 4. Examination of these coefficients reveals that the scalability of each of the MC items was above Mokken's (1971) minimum value of $H_i = .30$, and ranged from $H_i = .32$ ($SE = .04$) for Item 1, which was the least scalable item, to $H_i = .55$ ($SE = .05$) for Item 12, which was the most scalable item. No negative scalability coefficients were observed for the MC items. Overall, these results suggest that some Guttman errors were observed related to each of the assessment items examined in this study, but that each of the MC items contributes to a meaningful overall ordering of students in terms of physical science knowledge.

Invariant ordering. The final step in the Mokken analysis of the physical science MC items involves evaluating the dichotomous DM model requirement for IIO. Results from IIO analyses of the MC items using the restscore method are

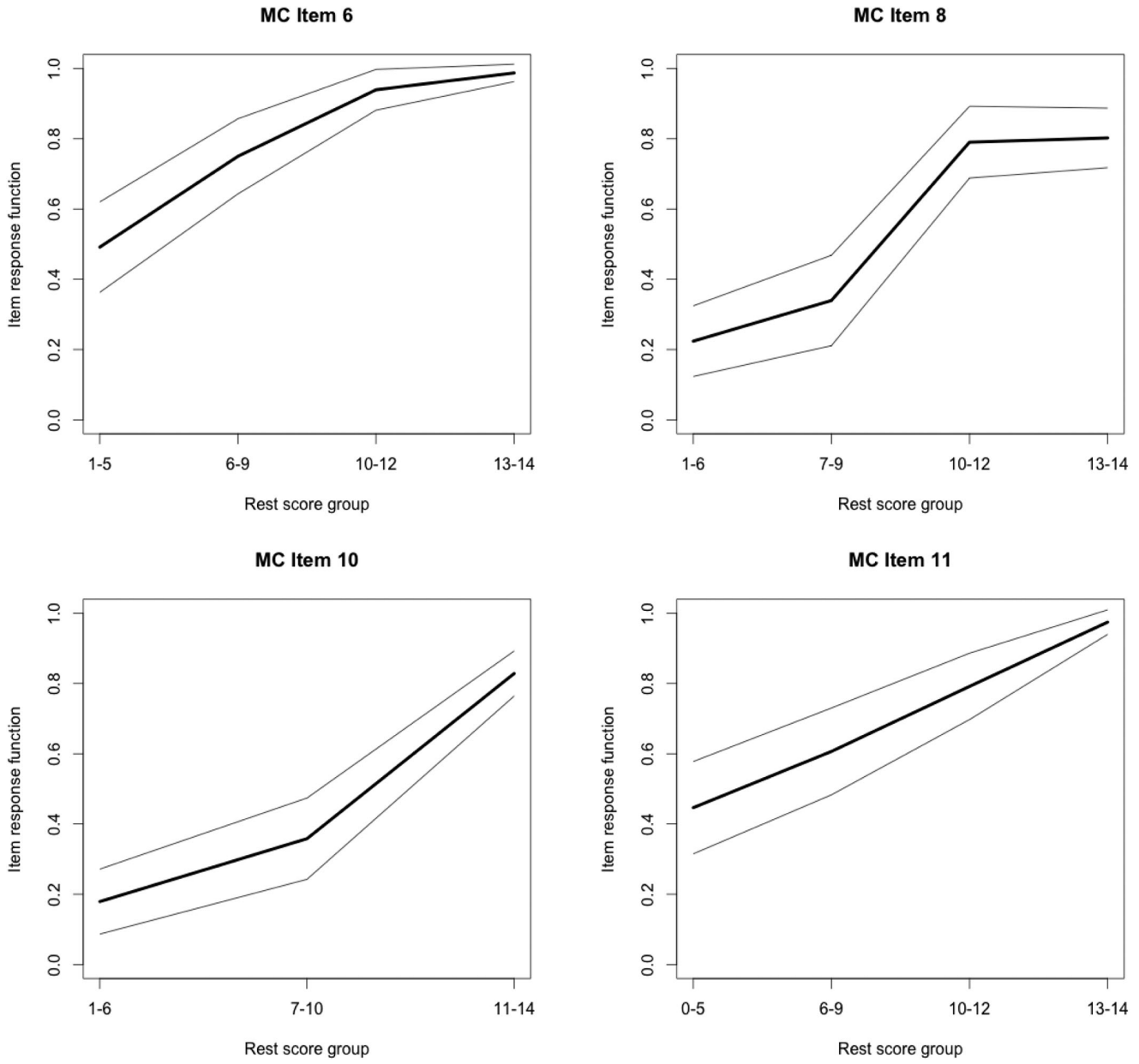

FIGURE 7. Selected monotonicity plots for physical science items. *Note*. In each plot, the *x*-axis shows levels of student achievement based on restscores, and the *y*-axis shows the probability for a correct response. The thick line shows the relationship between student achievement and the probability for a correct response (the item response function), and the thin lines show the confidence interval around the response function.

presented in Table 4. For each item, the frequency of observed and significant violations of IIO is presented. Overall, these results suggest that there were very few significant violations of IIO among the physical science assessment items. However, one significant violation was observed for Item 3 and Item 10, and two significant violations were observed for item pairs involving Item 8. The implication of the finding of crossing IRFs for these MC items is that the difficulty of items 3, 8, and 10 is not invariant across the range of student achievement. In other words, this finding suggests that students with the same total scores might have important differences in their understanding of the physical science concepts assessed by items 3, 8, and 10. As a result, these results do not support a consistent interpretation of the difficulty order of physical science items across the range of student achievement.

Interpret Results Within Context

Because the physical science assessment was a low-stakes classroom assessment administered during the first implementation of an experimental curriculum, the observed variations in item scalability and violations of invariant ordering did not pose serious threats to the interpretation of student achievement in terms of decisions about individual students. Instead, these violations were used to identify MC items for further investigation and revision. Specifically, Items 3, 8, and 10 were identified as potential candidates for revision.

Modify AOs

Based on the interpretation of results from the analysis of the MC items, the fourth step involved revising the MC items

Table 4. Multiple-Choice Item Results

	Item Scalability		Invariant Item Ordering
Item	H_i	*SE*	Count of Significant Violations
1	.32	.04	0
2	.43	.03	0
3	.40	.05	1
4	.36	.04	0
5	.41	.04	0
6	.42	.04	0
7	.43	.03	0
8	.42	.05	2
9	.44	.03	0
10	.49	.05	1
11	.33	.04	0
12	.55	.05	0
13	.44	.03	0
14	.42	.04	0
15	.35	.04	0

identified in Step 3, and administering the assessment again during the next implementation of the curriculum in order to obtain new data for analysis. Results from the subsequent analysis revealed adequate fit to the MH and DM models for the purposes of the assessment.

Illustration 2: Rater-Mediated Assessment

In the second illustrative analysis, I illustrate the use of MSA to evaluate measurement quality when polytomous ratings are used. The key distinction between the analyses based on polytomous MSA models from the analyses based on dichotomous MSA models used in Illustration 1 is related to the evaluation of AOs at the level of rating scale categories in addition to the evaluation of these model assumptions for overall AOs.

The illustrative analysis demonstrates the application of Molenaar's (1982, 1997) polytomous MSA models to an assessment in which human raters evaluated student essays in terms of writing achievement. Specifically, 365 students composed essays that were rated by a group of 20 raters in the context of a rater training program prior to operational scoring. Each rater scored each student's composition in terms of four domains: Conventions, Organization, Sentence Formation, and Style. Ratings were assigned using a four-category rating scale ($0 = low; 3 = high$). The illustrative analysis presented here focuses on the Style ratings. Because the desired information to be obtained from the rating procedure was related to the overall quality of the ratings provided by each rater in terms of fundamental measurement properties, MSA was an appropriate analytic tool for this assessment context. Additional details about these data, including results from the other three domains, are provided in Wind and Engelhard (2015).

When considering the results from MSA analyses of assessment data based on a rater-mediated performance assessment, it is important to note that the interpretation of the results is not necessarily consistent with the interpretation of MSA indices based on more traditional applications of this approach. In particular, rules of thumb for evaluating scalability coefficients, along with statistics for statistical tests of monotonicity and invariant ordering, have not been thoroughly explored in the context of rater-mediated assessments.

Import the Data Matrix

In order to apply the MSA models to the writing assessment data, raters were treated as AOs. Specifically, the example assessment data were structured with individual students as rows and raters as columns. The cells in the matrix included each rater's polytomous rating of each student's essay. This approach is similar to the methods used to apply parametric polytomous IRT models to ratings, including the Rasch model (e.g., Engelhard, 1994; Wolfe & McVay, 2012). When the polytomous MH and DM models are applied to raters, indicators of measurement quality describe the degree to which individual raters demonstrate useful measurement properties related to monotonicity, scalability, and invariant ordering.

Analyze AOs

Next, the AOs were evaluated using indicators of: (A) monotonicity; (B) scalability; and (C) invariant ordering.

Monotonicity. Evidence of monotonicity in the context of a rater-mediated assessment suggests that a set of student performances have the same relative ordering across a group of raters, such that the interpretation of relative student achievement does not depend on the particular rater who scored a student's performance. In order to evaluate monotonicity for the 20 operational raters in the example data set, overall rater response functions were examined for evidence of nondecreasing average ratings across increasing restscores. Further, ISRFs for individual raters were examined in order to evaluate the monotonicity assumption within rating scale categories. Neither the graphical nor the statistical techniques revealed violations of monotonicity for any of the 20 raters in the Style domain. Inspection of IRFs and ISRFs for individual raters revealed differences in the slope and shape of IRFs and ISRFs across the 20 raters, with some raters demonstrating sharper areas of discrimination and overall difficulty across levels of student achievement. Figure 8 illustrates some differences in the shape of IRFs that were observed among the raters. However, the finding of adherence to the MH model requirement of monotonicity suggests that the relative conclusions about the ordering of the middle school students in terms of writing achievement are consistent across all 20 raters.

Scalability. Low values of rater scalability for individual raters are of particular interest as an indicator of rating quality, because they indicate frequent Guttman errors that might suggest idiosyncratic use of a set of rating scale categories. The overall scalability coefficient for the group of 20 raters on the Style domain indicated a strong Mokken scale ($H = .77$; $SE = .01$). Values of individual rater scalability coefficients are presented in Table 5. These results indicate that some small differences exist in the relative frequency of Guttman errors across the group of raters. The highest scalability coefficient was observed for Rater 8 ($H_i = .82, SE = .02$), and the lowest scalability coefficient was observed for Rater 6 and Rater 20 ($H_i = .74$, $SE = .02$). Further, results indicated no negative rater pair scalability coefficients among the 20 raters. Taken together, these results indicate that student total scores across the group of raters can be interpreted as a meaningful indicator of student ordering in terms of the construct. In other words, each of the raters contributed

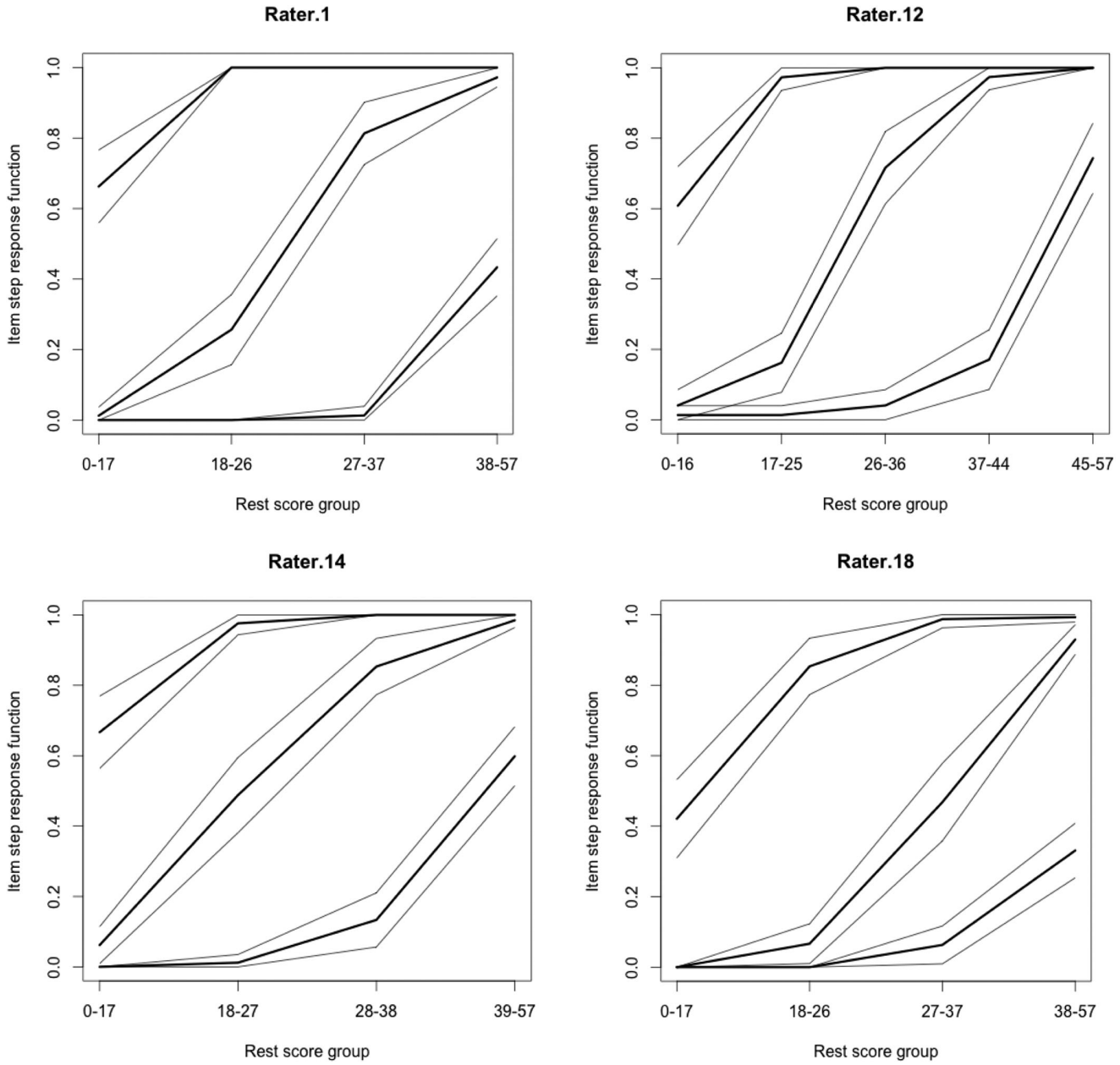

FIGURE 8. Selected monotonicity plots for raters. *Note*. In each plot, the *x*-axis shows levels of student achievement based on restscores, and the *y*-axis shows the probability for a correct response. The thick line shows the relationship between student achievement and the probability for a rating in a particular category (the item step response function), and the thin lines show the confidence interval around the response functions.

to a meaningful order of the students in terms of writing achievement.

Invariant ordering. Evidence for invariant rater ordering (IRO) suggests that it is possible to interpret the relative ordering of rater severity in the same way across levels of student achievement. As I noted above, evidence of IRO is a fairness issue that has implications for the interpretation of rater-assigned scores. Specifically, evidence of IRO suggests that rater severity can be interpreted the same way for all students, and conclusions about the relative ordering of students do not depend on the particular rater or raters who happened to score their performance. Conversely, violations of IRO suggest that the interpretation of student achievement ordering varies across raters, such that conclusions about individual students depend on the raters who scored their performances.

Following Ligtvoet et al. (2010, 2011), I conducted IRO analyses for overall raters, rather than within rating scale categories. For each rater, Table 5 includes the frequency of significant violations of IRO. Violations of IRO were observed most frequently for Rater 15 (3 significant violations), followed by Rater 16 (2 violations). One significant violation of IRO was observed for Raters 7, 19, and 20. All of the other raters demonstrated IRO. This finding suggests that Raters 7, 19, and 20 may have a different interpretation of student writing achievement than the other raters. As a result, these results do not support a consistent interpretation of rater severity across the range of student achievement.

Table 5. Rater-Mediated Assessment Results

	Rater Scalability		Invariant Rater Ordering
Rater	H_i	*SE*	Count of Significant Violations
1	.77	.02	0
2	.76	.02	0
3	.78	.02	0
4	.77	.02	0
5	.76	.02	0
6	.74	.02	0
7	.78	.02	1
8	.82	.02	0
9	.78	.02	0
10	.78	.02	0
11	.78	.02	0
12	.78	.02	0
13	.76	.02	0
14	.77	.02	0
15	.78	.02	3
16	.80	.02	2
17	.75	.02	0
18	.76	.02	0
19	.78	.02	1
20	.74	.02	1

Interpret Results Within Context

Because the writing assessment ratings were collected during a rater training procedure, the observed variations in rater scalability and violations of invariant ordering did not pose serious threats to the interpretation of student achievement in terms of decisions about individual students. Instead, these violations were used to identify raters for further investigation and potential retraining. Specifically, Raters 7, 19, and 20 could be identified as potential candidates for revision.

Modify AOs

Based on the interpretation of results from the analysis of the raters, additional research and rater training procedures could be targeted toward the three raters who demonstrated misfit to the DM model in order to more fully understand their idiosyncratic rating patterns and, if necessary, administer additional training. Alternatively, if retraining is not possible, it may be necessary to remove these three raters from the pool of operational raters.

Summary

In this module, I provided an introduction to MSA as a probabilistic-nonparametric approach to exploring the quality of measurement procedures in the social sciences. I emphasized the usefulness of MSA for exploring measurement quality in the context of educational measurement using examples from a MC science assessment and a rater-mediated writing assessment. Further, I presented three categories of Mokken-based indicators and displays as a systematic technique for evaluating fundamental measurement properties for dichotomous and polytomous AOs: (A) monotonicity, (B) scalability, and (C) invariant ordering.

Although a variety of sophisticated parametric techniques exist that can be applied to the types of educational assessment data discussed in this module, the Mokken approach illustrated here should be recognized as an additional tool that can be used to explore fundamental measurement properties. In particular, the procedures illustrated in this module can be used to explore measurement quality in contexts in which an ordinal scale of measurement is sufficient to inform decisions based on assessment results, when sample sizes are not sufficient for the application of parametric models, and as an exploratory technique used in combination with parametric procedures. A major benefit of this approach is that, when there is evidence of adherence to the MH model, it is possible to interpret total scores as an indicator of student ordering on the latent variable. Likewise, evidence of adherence to the DM model supports a common interpretation of AO ordering across the range of student achievement.

Another major benefit is the diagnostic nature of the graphical displays, which provide researchers with a valuable tool for examining the underlying properties of items or raters that go beyond the summary-level item or rater fit statistics that are often explored in the context of parametric IRT. Summarizing this perspective, Meijer et al. (2015) observed: "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" (p. 89). Furthermore, when statistics and displays based on MSA are examined in combination with parametric techniques, the nonparametric indices have frequently been shown to reveal additional diagnostic information regarding the location of model-data misfit that are not evident based on parametric models (Sijtsma & Meijer, 2007; Wind, 2014, 2016)—thus, providing a more complete summary of the data than what would be obtained based on a single approach.

As noted above, it is essential that researchers consider the unique characteristics of the assessment context when interpreting results from an application of MSA to educational achievement test data. Because these procedures were originally developed for use with affective measures, the interpretation of traditionally used rules of thumb and critical values may not translate directly to educational tests—especially when raters are involved. Additional research is needed in order to provide additional insight into the interpretation of these rules of thumb and critical values for educational achievement tests in general, including rater-mediated assessments.

The current illustrative analyses should encourage researchers to consider the use of nonparametric procedures based on Mokken scaling as a systematic approach for evaluating the quality of educational assessment, particularly when invariant measurement is of interest.

Data for Illustrative Analyses

The data used in the illustrative analyses, along with R code for implementing the analyses, are available at this link: https://gitlab.com/stefaniewind/Mokken_ITEMS/tree/master

Self-Test

1. Describe the major differences between the monotone homogeneity and double monotonicity models in terms of their underlying requirements.
2. Describe how the monotone homogeneity and double monotonicity model assumptions relate to the

requirements for invariant measurement: (a) persons are ordered the same way across items; (b) items are ordered the same way across persons.
3. A researcher discovers nonintersecting ISRFs within a pair of polytomous items. Does this imply fit to the double monotonicity model? Why or why not?
4. Calculate the item pair scalability coefficient for the dichotomous item pair (i, j) using the following joint frequency table:

	Item $j = 0$	Item $j = 1$	Total
Item $i = 0$	41	41	82
Item $i = 1$	27	159	186
Total	68	200	268

5. Describe violations of monotonicity for dichotomous and polytomous assessment opportunities.
6. Describe the relationship between the number of Guttman errors associated with an assessment opportunity and scalability for that assessment opportunity.
7. Using the data and R code for the illustrative analyses, calculate numeric and graphical indicators of monotonicity, scalability, and invariant ordering. Check your work against the results shown in the module.

Answers to Self-Test

1. The monotone homogeneity (MH) and double monotonicity (DM) models share three common underlying assumptions: monotonicity, unidimensionality, and local independence. The DM model is more restrictive in that it also includes a fourth assumption: invariant item ordering.
2. When data fit the MH model, there is evidence that persons are ordered the same way across items (a). When data fit the DM model, there is evidence that items are ordered the same way across persons (b).
3. No. As pointed out by Ligtvoet et al. (2010, 2011), there is an *aggregation phenomenon* that is sometimes observed for polytomous items, where nonintersecting ISRFs do not always aggregate to nonintersecting IRFs, and vice versa. Accordingly, additional checks are necessary in order to ensure that the aggregation phenomenon does not affect the conclusion of adherence to the DM model.
4. $H_i = .428$. This value can be calculated using the observed and expected ratio method as follows:
- Inspection of the marginal frequencies for the two items reveals that Item i (total correct = 186) is more difficult than Item j (total correct = 200), so the item difficulty ordering is $i < j$.
- Guttman errors are defined as combinations of a correct response to Item i and an incorrect response to Item j; the cell $(X_{(ij)} = 1{,}0)$. The frequency of observed Guttman errors is 27.
- The expected frequency of Guttman errors is: [(Row total * Column total)/N] = [(186 * 68)/268)] = 47.19.
- Scalability for the item pair is calculated as: $H_{ij} = 1 - \frac{27}{47.19} = .428$.
5. For assessment opportunities with dichotomous scores, a violation of monotonicity occurs when the probability for a correct response is higher among students with lower restscores (i.e., lower achievement) than it is among students with higher restscores. For assessment opportunities with polytomous scores, a violation of monotonicity occurs when the probability for a rating in category k or higher is higher among students with lower restscores than it is among students with higher restscores. When the assumption of monotonicity is violated, students are not ordered the same way across assessment opportunities.
6. As the number of Guttman errors increases, the value of the scalability coefficient decreases.
7. Please compare your results for Illustration 1 to Table 4 and Figure 7, and your results for Illustration 2 to Table 5 and Figure 8.

List of Selected Introductory Texts on Mokken Scaling

Conceptual overviews:

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory* (Vol. 5). Thousand Oaks, CA: Sage.

Sijtsma, K., Meijer, R. R., & van der Ark, L. A. (2011). Mokken scale analysis as time goes by: An update for scaling practitioners. *Personality and Individual Differences*, *50*(1), 31–37. https://doi.org/10.1016/j.paid.2010.08.016

van Schuur, W. H. (2011). *Ordinal item response theory: Mokken scale analysis*. Los Angeles, CA: Sage.

Applications:

Sijtsma, K., Emons, W. H., Bouwmeester, S., Nyklíček, I., & Roorda, L. D. (2008). Nonparametric IRT analysis of quality-of-life scales and its application to the World Health Organization Quality-of-Life scale (WHOQOL-Bref). *Quality of Life Research*, *17*(2), 275–290.

van Schuur, W. H., & Vis, J. C. P. M. (2000). What Dutch parliamentary journalists know about politics. *Acta Politica*, *35*, 196–227.

Watson, R., Deary, I. J., & Shipley, B. (2008). A hierarchy of distress: Mokken scaling of the GHQ-30. *Psychological Medicine*, *38*, 575–580.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Chernyshenko, O. S., Stark, S., Chan, K.-Y., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research*, *36*, 523–562. https://doi.org/10.1207/S15327906MBR3604_03

Emons, W. H. M. (2008). Nonparametric person-fit analysis of polytomous item scores. *Applied Psychological Measurement*, *32*, 224–247. https://doi.org/10.1177/0146621607302479

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*, 93–112. https://doi.org/10.2307/1435170

Engelhard, G. (2008). Historical perspectives on invariant measurement: Guttman, Rasch, and Mokken. *Measurement*, *6*(3), 155–189. https://doi.org/10.1080/15366360802197792

Hardouin, J.-B., Bonnaud-Antignac, A., & Sebille, V. (2011). Nonparametric item response theory using Stata. *Stata Journal*, *11*(1), 30–51.

Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1997). Stochastic ordering using the latent trait and the sum

score in polytomous IRT models. *Psychometrika*, *62*, 331–347. https://doi.org/10.1007/BF02294555
Junker, B. W. (1993). Conditional association, essential independence and monotone unidimensional item response models. *Annals of Statistics*, *21*, 1359–1378.
Junker, B. W., & Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement*, *24*, 65–81. https://doi.org/10.1177/01466216000241004
Kuijpers, R. E., van der Ark, L. A., & Croon, M. A. (2013). Standard errors and confidence intervals for scalability coefficients in Mokken scale analysis using marginal models. *Sociological Methodology*, *43*(1), 42–69. https://doi.org/10.1177/0081175013481958
Lane, S., Raymond, M., Haladyna, T. M., & Downing, S. M. (2011). *Handbook of test development*. New York, NY: Routledge.
Ligtvoet, R. (2010). *Essays on invariant item ordering*. Tilburg University, Tilburg, Netherlands. Retrieved from https://pure.uvt.nl/portal/en/publications/essays-on-invariant-item-ordering(872d8f6b-778e-49f7-8fdb-c74237a70ffc).html
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. https://doi.org/10.1177/0013164409355697
Loevinger, J. (1948). The technic of homogeneous tests compared with some aspects of scale analysis and factor analysis. *Psychological Bulletin*, *45*, 507–529.
Mazza, A., Punzo, A., & McGuire, B. (2014). KernSmoothIRT: An R package for kernel smoothing in item response theory. *Journal of Statistical Software*, *58*(6). https://doi.org/10.18637/jss.v058.i06
Meijer, R. R., & Baneke, J. J. (2004). Analyzing psychopathology items: A case for nonparametric item response theory modeling. *Psychological Methods*, *9*, 354–368. https://doi.org/10.1037/1082-989X.9.3.354
Meijer, R. R., Muijtjens, A. M. M., & van der Vlueten, C. P. M. (1996). Nonparametric person-fit research: Some theoretical issues an empirical example. *Applied Measurement in Education*, *9*(1), 77–89.
Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*, 107–135. https://doi.org/10.1177/01466210122031957
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. https://doi.org/10.1177/014662169001400306
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. Revick (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 85–110). New York, NY: Routledge.
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.
Mokken, R. J. (1971). *A theory and procedure of scale analysis*. The Hague, The Netherlands: Mouton.
Molenaar, I. W. (1982). Mokken scaling revisited. *Kwantitative Methoden*, *3*(8), 145–164.
Molenaar, I. W. (1991). A weighted Loevinger H-coefficient extending Mokken scaling to multicategory items. *Kwantitative Methoden*, *37*(12), 97–117.
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, W., & Sijtsma, K. (2000). MSP5 for Windows User's Manual. Psychometrics and Statistics: Iec ProGAMMA. Retrieved October 26, 2015, from https://www.rug.nl/research/portal/en/publications/msp5-for-windows-users-manual(2581d067-1df4-4888-8299-7def6f1159a5).html
Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, *56*, 611–630. https://doi.org/10.1007/BF02294494
Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, *5*, 27–48. https://doi.org/10.1146/annurev.clinpsy.032408.153553
Sijtsma, K., & Hemker, B. T. (1998). Nonparametric polytomous IRT models for invariant item ordering, with results for parametric models. *Psychometrika*, *63*, 183–200. https://doi.org/10.1007/BF02294774
Sijtsma, K., & Meijer, R. R. (1992). A method for investigating the intersection of item response functions in Mokken's nonparametric IRT Model. *Applied Psychological Measurement*, *16*, 149–157. https://doi.org/10.1177/014662169201600204
Sijtsma, K., & Meijer, R. R. (2007). Nonparametric item response theory and special topics. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Psychometrics* (Vol. 26, pp. 719–747). Amsterdam, The Netherlands: Elsevier.
Sijtsma, K., Meijer, R. R., & van der Ark, L. A. (2011). Mokken scale analysis as time goes by: An update for scaling practitioners. *Personality and Individual Differences*, *50*(1), 31–37. https://doi.org/10.1016/j.paid.2010.08.016
Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory* (Vol. 5). Thousand Oaks, CA: Sage.
Sijtsma, K., & van der Ark, L. A. (2017). A tutorial on how to do a Mokken scale analysis on your test and questionnaire data. *British Journal of Mathematical and Statistical Psychology*, *70*(1), 137–158. https://doi.org/10.1111/bmsp.12078
Smits, I. A. M., Timmerman, M. E., & Meijer, R. R. (2012). Exploratory Mokken scale analysis as a dimensionality assessment tool: Why scalability does not imply unidimensionality. *Applied Psychological Measurement*, *36*, 516–539. https://doi.org/10.1177/0146621612451050
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. https://doi.org/10.1007/s11336-000-0862-3
van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software*, *20*(11), 1–19.
van der Ark, L. A. (2012). New developments in Mokken scale analysis in R. *Journal of Statistical Software*, *48*(5), 1–27.
van der Ark, L. A. (2013). Visualizing uncertainty of estimated response functions in nonparametric item response theory. In R. E. Millsap, L. A. van der Ark, D. M. Bolt, & C. M. Woods (Eds.), *New developments in quantitative psychology* (pp. 59–68). New York, NY: Springer. https://doi.org/10.1007/978-1-4614-9348-8_5
van der Ark, L. A., & Bergsma, W. P. (2010). A note on stochastic ordering of the latent trait using the sum of polytomous item scores. *Psychometrika*, *75*, 272–279. https://doi.org/10.1007/s11336-010-9147-7
van Schuur, W. H. (2003). Mokken scale analysis: Between the Guttman scale and parametric item response theory. *Political Analysis*, *11*(2), 139–163.
van Schuur, W. H. (2011). *Ordinal item response theory: Mokken scale analysis*. Los Angeles, CA: Sage.
Wind, S. A. (2014). Examining rating scales using Rasch and Mokken models for rater-mediated assessments. *Journal of Applied Measurement*, *15*(2), 100–132.
Wind, S. A. (2016). Examining the psychometric quality of multiple-choice assessment items using Mokken scale analysis. *Journal of Applied Measurement*, *17*(2), 142–165.
Wind, S. A., & Engelhard, G. (2015). Exploring rating quality in rater-mediated assessments using Mokken scaling. *Educational and Psychological Measurement*, *76*(4), 685–706.
Wolfe, E. W., & McVay, A. (2012). Application of latent trait models to identifying substantively interesting raters. *Educational Measurement: Issues and Practice*, *31*(3), 31–37. https://doi.org/10.1111/j.1745-3992.2012.00241.x