Exploring Incomplete Rating Designs With Mokken Scale Analysis

Educational and Psychological
Measurement

2018, Vol. 78(2) 319–342

© The Author(s) 2016

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164416675393

journals.sagepub.com/home/epm

Stefanie A. Wind¹ and Yogendra J. Patil¹

Abstract

Recent research has explored the use of models adapted from Mokken scale analysis as a nonparametric approach to evaluating rating quality in educational performance assessments. A potential limiting factor to the widespread use of these techniques is the requirement for complete data, as practical constraints in operational assessment systems often limit the use of complete rating designs. In order to address this challenge, this study explores the use of missing data imputation techniques and their impact on Mokken-based rating quality indicators related to rater monotonicity, rater scalability, and invariant rater ordering. Simulated data and real data from a rater-mediated writing assessment were modified to reflect varying levels of missingness, and four imputation techniques were used to impute missing ratings. Overall, the results indicated that simple imputation techniques based on rater and student means result in generally accurate recovery of rater monotonicity indices and rater scalability coefficients. However, discrepancies between violations of invariant rater ordering in the original and imputed data are somewhat unpredictable across imputation methods. Implications for research and practice are discussed.

Keywords

Mokken scaling, rating quality, missing data, performance assessment

Recent research has demonstrated the application of models based on Mokken scale analysis (MSA; Mokken, 1971) as a probabilistic-nonparametric approach to evaluating rating quality in the context of rater-mediated educational performance assessments (Wind, 2014, 2015, 2017; Wind & Engelhard, 2016). Findings from these

¹The University of Alabama, Tuscaloosa, AL, USA

Corresponding Author:

Stefanie A. Wind, Educational Studies in Psychology, Research Methodology, and Counseling, The University of Alabama, 313C Carmichael Hall, Box 870231, Tuscaloosa, AL 35487, USA.

Email: swind@ua.edu

applications have suggested that adaptations of polytomous MSA models based on adjacent-categories probabilities provide a diagnostic framework in which to explore rating quality at the individual rater level without potentially inappropriate parametric transformations.

As is the case with traditional applications of MSA, the use of MSA to evaluate rating quality requires complete data. In the context of rater-mediated assessments, complete data sets include ratings from every rater on every assessment component (e.g., students, tasks, and rubric domains). Although complete rating designs are interesting theoretically, practical constraints limit their use in operational assessment systems. Accordingly, rating designs are frequently used in which substantial levels of missingness are specified prior to data collection, and systematic links are used to establish connectivity among elements of the assessment system. Previous research on MSA includes the exploration of various data imputation techniques for dichotomous and polytomous items in terms of their impact on Mokken indicators of measurement quality (Sijtsma & van der Ark, 2003; van der Ark & Sijtsma, 2005; van Ginkel & van der Ark, 2005; van Ginkel, van der Ark, & Sijtsma, 2007). In general, this research has focused on the effects of missing data in situations in which missing data is not a systematic component of the data collection design, and often occurs at low rates, such as responses to attitude survey items. Furthermore, previous missing data research in the context of MSA has not included the use of the adjacent-categories formulations of polytomous MSA models proposed for use in the context of rater-mediated assessments. In order to facilitate the use of MSA as a method for evaluating rating quality in practical settings, it is necessary to explore methods for imputing missing ratings and their effects on conclusions related to rating quality.

Purpose

The purpose of this study is to explore the impact of missing data imputation on MSA-based rating quality indicators. Specifically, this study focuses on the following research question:

Research Question 1: What is the effect of missing data imputation techniques and levels of missingness on indicators of rating quality based on adjacent-categories Mokken models?

Using simulated and real data, the influence of imputation methods and levels of missingness is explored across conditions that reflect a range of rater and student sample sizes and rating scales with various numbers of categories.

This study builds on previous research in two main ways. First, it extends previous research on MSA as a probabilistic-nonparametric framework for evaluating rating quality by exploring the application of this technique to data that reflect practical educational assessment settings. Although results from previous studies indicate that models based on MSA provide a useful diagnostic approach to exploring rating

quality, their application is limited by the requirement for complete rating designs. The current study provides an initial exploration of the effects of missing data imputation techniques that can inform the use of MSA-based rating quality indicators when complete ratings are not available.

Second, this study builds on previous research related to missing data imputation methods in the context of MSA, in general, and extends this research through the consideration of data imputation methods for rater-assigned polytomous scores. Although much research has been conducted related to the effects of missing data imputation when the original dichotomous (Mokken, 1971) and polytomous MSA models (Molenaar, 1982, 1997) are applied, the effects of missing data imputation on the adjacent-categories MSA models for rater-mediated assessments (Wind, 2017; discussed further below) has not previously been explored. As noted above, the current application of missing data imputation in the context of MSA is also distinct in that incomplete rating designs generally involve a systematic and a priori specification of missingness that is generally not observed in contexts in which MSA is more often applied, such as affective surveys.

Missing Data in Rater-Mediated Assessments

Research on missing data in the context of educational and psychological measurement typically involves methods for determining patterns of missingness and underlying mechanisms that lead to missingness (Little & Rubin, 2002). Specifically, classification schemes are frequently used to describe various situations in which missing data occur. For example, Little and Rubin (2002) describe three mechanisms that lead to missing data: (1) missing completely at random (MCAR), (2) missing at random (MAR), and (3) missing not at random (MNAR). These mechanisms distinguish patterns of incomplete data based on the relationship between missing values and values of the data, including values of observed scores or values of missing scores. Observations that are MCAR do not depend on values of observed or missing data, observations that are MAR depend on observed values, and observations that are MNAR may depend on the value of the missing observation and/or an unobserved variable (Little & Rubin, 2002; van Ginkel et al., 2007).

In the context of rater-mediated assessment systems, data collection procedures typically specify some level of missingness prior to data collection in order to accommodate practical constraints related to scoring time and other resources. Although these data collection designs vary across settings, the missingness in most large-scale performance assessment systems can be viewed as MCAR because missing observations depend only on the particular scoring design rather than on the values of the missing observations.

Data Collection Designs for Rater-Mediated Assessments

A variety of designs are used to collect data in rater-mediated assessments. When multiple raters score each performance, it is possible to calculate rating quality

indicators such as reliability and agreement (Johnson, Penny, & Gordon, 2009). Furthermore, it is possible to apply parametric item response theory (IRT) models, such as the many-facet Rasch model (Linacre, 1989) to incomplete data sets as long as the data collection design includes sufficient connectivity between facets, such as raters scoring common performances with at least one other rater (Wind, Engelhard, & Wesolowski, 2016).

Missing Data Within the Context of Mokken Scaling

Previous research on MSA includes the exploration of a variety of missing data imputation methods. For example, Sijtsma and van der Ark (2003) explored the impact of four methods of missing data imputation on Mokken scalability coefficients. Among the methods they explored, these authors found that the response function (RF) imputation method based on nonparametric regression techniques resulted in the least bias for Mokken scalability coefficients. Van der Ark and Sijtsma (2005) described five methods for missing data imputation and their effects on item clustering solutions and scalability coefficients based on MSA. Although slightly smaller biases were observed based on more complex methods, the authors concluded that a method based on row and column means (two-way [TW] method; described further below) was promising as a result of its computational simplicity.

Most recently, van Ginkel et al. (2007) compared the effects of several imputation methods using simulated data that reflected different missingness mechanisms and levels of missingness on Mokken scalability coefficients and item cluster solutions. Among the imputation methods examined in their study, van Ginkel et al. found that simple imputation methods such as the TW method produced smaller discrepancies in MSA statistics than did more complex multivariate normal imputation methods.

When considering previous MSA missing data research in light of the purpose of the current study, several differences between these applications are important to consider. First, the main purpose of applying MSA to rater-mediated assessments is to consider results for individual raters. Accordingly, overall summaries related to item (or in this case, rater) clusters based on scalability coefficients are not of interest. Furthermore, these methods have been examined in the context of psychological surveys, which generally include lower levels of missingness that is typically not part of the data collection design. As a result, there are often substantial subsets of items and persons with complete results that allow for the use of methods such as the RF imputation technique. In contrast, rating designs often result in higher levels of missingness that does not include large subsets of students or raters with complete data.

Method

Simulated and real data are used to explore the effects of imputation methods and levels of missingness on MSA indices of rating quality. For both types of data, complete rating designs are manipulated to reflect various degrees of missingness, and

Table 1. Simulation Design.

Design factor	Levels
Student sample size	$N_S = 300$ $N_S = 1,000$
Rater sample size	$N_R = 20$ $N_R = 40$
Number of rating scale categories	Three categories (0, 1, 2) Four categories (0, 1, 2, 3) Five categories (0, 1, 2, 3, 4)

Note. One hundred replications were completed within each cell. After ratings were generated within each condition, each simulated data set was subsequently manipulated to reflect varying degrees of missingness, and missing ratings were imputed.

the effects of four missing data imputation methods on MSA indices of rating quality are explored. This section describes the Monte Carlo simulation design, the real data, and the methods used to create missingness and impute missing values. A description of the data analysis techniques based on Mokken scaling follows.

Simulated Data

A Monte Carlo simulation study was used to systematically examine the effects of imputation methods and levels of missingness on MSA indices of rating quality across conditions that reflect data collection designs for operational rater-mediated assessments.

First, complete data sets were specified using a $2 \times 2 \times 3$ factorial design. These complete data sets were specified using combinations of the rater sample sizes, student sample sizes, and number of rating scale categories indicated in Table 1. Within each combination of conditions, holistic ratings were simulated based on the Rasch rating scale model (Andrich, 1978). This model was selected as a result of findings in previous research of theoretical and empirical similarities between MSA and Rasch measurement theory (e.g., Engelhard, 2008; Meijer, Sijtsma, & Smid, 1990; Wind, 2014), as well as the use of IRT models in previous simulation studies of missing data in the context of MSA (e.g., van Ginkel et al., 2007). Rater and student sample sizes, as well as the number of rating scale categories, were selected based on previous simulation studies related to IRT models for rater-mediated assessments (e.g., Marais & Andrich, 2011; Wolfe, Jiao, & Song, 2014; Wolfe & McVay, 2012; Wolfe & Song, 2015).

Generating rater severity parameters were randomly selected within a typical range of rater severity in performance assessments (-3 to $+3$ logits), where each condition included 25% lenient raters (location ≤ -0.50 logits), 25% severe raters (location $\geq +0.50$ logits), and 50% moderate raters (-0.50 logits $<$ location $<$ $+0.50$ logits). Following simulation guidelines within IRT literature (Harwell, Stone,

Hsu, & Kirisci, 1996) and previous missing data research related to MSA (van Ginkel et al., 2007), 100 replications were completed within each cell.

Real Data

In addition to the simulated data, real data from an operational performance assessment were used to explore the effects of missing data imputation on MSA rating quality indicators in an authentic context. The real data explored in this study were collected from an administration of the Georgia High School Writing Test and include scores from 20 raters and 365 students on four domains of writing: Conventions, Organization, Sentence Formation, and Style. All raters scored all students on all domains, such that the ratings formed a complete assessment network (Engelhard, 1997). Ratings were assigned using a four-category rating scale (1 = *low*; 4 = *high*). These ratings were recoded prior to analysis to the following: 0 = *low*; 3 = *high*.

Creating Missing Ratings

For both the simulated and real data, each of the complete sets of ratings was modified to reflect five levels of missingness (10%, 20%, 30%, 40%, or 50%). Missing ratings were selected using random permutation (independent random sampling without replacement). In order to create missing ratings from unobserved raters for the i th student, a random draw of raters was performed sequentially until the last rater. Once the random draw was completed, the predefined percent of missing raters were selected and labeled as missing raters for that particular i th student. A similar procedure was followed for the rest of the students.

Inspection of the final modified data sets within each condition (simulated data) and each domain (real data) indicated links among raters through common students that ranged from at least 10 common students between each rater and at least one other rater to at least 18 common students between each rater and at least one other rater.

Imputing Missing Ratings

For each modified data set, missing values were imputed using four methods that have previously been applied in the context of missing data research related to MSA: (1) *random imputation* (RI), (2) *two-way imputation* (TW), (3) *two-way with normally distributed errors* (TW-E), and (4) *corrected item-mean substitution with normally distributed errors* (CIMS-E).

Random Imputation. The first imputation method used in this study is RI. Using this method, imputation is performed by randomly selecting a single value for student i from rater j (X_{ij}) from the range of observed integer ratings (0, 1, 2, or 3). The main

advantage of this technique is that it requires no calculation, and imputed values remain within the range of possible ratings. However, this method does not take into account information about rater severity or student achievement.

Two-Way Imputation. The second method is the TW method. In the context of this study, the TW method imputes missing scores using average ratings across students, raters, and the total sample of students and raters. For example, consider a missing score for student i from rater j (X_{ij}). First, the mean rating for i th student is calculated using all nonmissing ratings, and the mean rating for the j th rater is calculated using all nonmissing students. Finally, the grand mean for all nonmissing observations is calculated and the missing value is imputed as

$$TW_{ij} = \bar{X}_i + \bar{X}_j - \bar{X}_{ij}. \quad (1)$$

When Equation (1) is used, it is possible that the imputed value may fall out of the raw score range (in the case of the current data, values between 0 and 3). In this case, the value is rounded to its nearest feasible score.

Two-Way With Normally Distributed Errors. Third, the TW-E imputation method is an adaptation of the TW method described above. In this technique, the standard error (s_e) due to sampling is eliminated by calculating the standard deviation for nonmissing scores. For a given set of observed ratings, where $\#obs$ represents number of observed ratings, the sample error variance (s_e^2) is computed as

$$s_e^2 = \sum_{ij} \sum_{cobs} (X_{ij} - TW_{ij})^2 / (\#obs - 1). \quad (2)$$

Finally, the error ε_{ij} in calculating TW_{ij} using Equation (1) is calculated by randomly drawing samples from the distribution $N(0, s_e^2)$. Therefore, the imputed value is given by

$$TW_{ij}(E) = TW_{ij} + \varepsilon_{ij}. \quad (3)$$

Values of $TW_{ij}(E)$ are rounded to the nearest scores if they are out of range.

Corrected Item Mean Substitution With Normally Distributed Error. The fourth imputation method is the CIMS-E. The CIMS-E imputed value for missing rating X_{ij} is defined as

$$CIMS - E_{ij} = \frac{\bar{X}_i}{\frac{1}{\#obs(i)} \sum_{j \in obs(i)} \bar{X}_j} * \bar{X}_j, \quad (4)$$

where $obs(i)$ is the set of all observed scores for student i . The denominator in Equation (4) is the mean of all the raters who rated student i . In other words, the missing score for rater j and student i is corrected relative to the raters who scored student i .

Data Analysis

After missing ratings were imputed, data analyses were conducted using methods based on the adjacent-categories Mokken models for rater-mediated assessments proposed by Wind (2017). This section provides an overview of the Mokken models that form the basis for the rating quality indices explored in this study, followed by a description of the specific indicators used to evaluate rating quality for individual raters.

Polytomous Mokken Models

The first step in the data analysis procedure for this study was to calculate indicators of rating quality based on MSA using adjacent-categories adaptations of the polytomous monotone homogeneity (MH) and double monotonicity (DM) models (Wind, 2017). These models can be discussed as they relate to Molenaar's (1982, 1997) original formulations of the MH and DM models. Specifically, the original formulation of the MH and DM models is based on a cumulative definition of the item step response function (ISRF; Molenaar, 1997). The monotonicity requirement for this model reflects this formulation:

- *Monotonicity*: The conditional probability for a rating in category k or higher is nondecreasing over increasing values of the latent variable.

Molenaar's (1982, 1997) polytomous DM model also reflects the cumulative formulation; this can be seen in the requirement for nonintersecting ISRFs:

- *Nonintersecting item step response functions*: 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.

Following Ligtoet, van der Ark, Bergsma, and Sijtsma (2011) and Ligtoet, van der Ark, te Marvelde, and Sijtsma (2010), the manifest invariant item ordering (MIIO) technique is used to check the assumption of IIO in polytomous data using average scores on items. First, items are ordered according to overall difficulty based on their average scores. Then, the degree to which this ordering holds across rest-score groups is examined within pairs of items using plots of IRFs and hypothesis tests.

Adjacent-Categories Mokken Models. Mokken ISRFs can also be defined using an adjacent-categories formulation as follows:

$$\begin{aligned}\tau_{ijk} &= 1 && \text{when } X_{ij} = k, \\ \tau_{ijk} &= 0 && \text{when } X_{ij} = k - 1\end{aligned}\tag{5}$$

where X_{ij} is the observed score from rater i for student j . Using this formulation of the ISRF, an adjacent-categories formulation of the MH model (ac-MH) can be described, where the monotonicity requirement is restated as follows:

- *Monotonicity*: The probability for a rating in category k , rather than category $k - 1$, is nondecreasing across the range of the latent variable.

Similarly, the adjacent-categories formulation of the DM model (ac-DM) uses the following nonintersection requirement:

- *Nonintersecting adjacent-categories item step response functions*: The conditional probability for a rating in category k rather than in category $k - 1$ from rater i has the same relative ordering across all values of the latent variable.

Because the MIIO procedure is based on overall items, the MIIO procedure can also be used to evaluate nonintersection for raters using the ac-DM model with no change from the original formulation.

Mokken Rating Quality Indices

This study uses three categories of Mokken-based rating quality indices based on the ac-MH and ac-DM models: rater monotonicity, rater scalability, and invariant rater ordering. A brief overview of these indices is provided below; for additional details, see Wind (2017).

Rater Monotonicity. First, rater monotonicity indices are used to evaluate the degree to which ISRFs for individual raters are nondecreasing across increasing levels of student achievement. When monotonicity is observed, student performances are ordered the same way across raters. These indices are based on the monotonicity requirement for the ac-MH model and include graphical displays and statistical hypothesis tests. In terms of graphical displays, monotonicity can be evaluated for individual raters by examining adjacent-categories ISRFs for evidence of nondecreasing probabilities across increasing levels of student achievement, which would be a violation of monotonicity. In addition, a one-sided, one-sample hypothesis test can be calculated to compare the probability for a rating in a given category across adjacent restscore groups. When the adjacent-categories probability in the lower restscore group exceeds the cumulative probability in the higher group by a value greater than a predefined critical value (usually 0.03; e.g., Mokken, 1971), the violation is examined using a one-sided, one-sample z test.

Rater Scalability. The second category of rating quality indices based on MSA is rater scalability. Within the context of MSA, scalability coefficients are used to evaluate the degree to which Guttman errors affect the overall quality of a scale (Loevinger,

1948; Mokken, 1971). When these coefficients are applied to raters, scalability coefficients describe the impact of Guttman errors on the ordering of ISRFS for individual raters. As in traditional applications of MSA, scalability coefficients for rater-mediated assessments can be calculated for an overall group of raters (H), pairs of raters (H_{ij}), and individual raters (H_i), where values of scalability coefficients describe the degree to which observed total scores across raters provide a meaningful description of student ordering on the latent variable. When evaluating rating quality at the individual rater level, values of scalability coefficients for individual raters (H_i) are of particular interest.

Scalability coefficients for raters are calculated in an analogous fashion to scalability coefficients for polytomous items in traditional applications of MSA (Molenaar, 1991), where rating scale category probabilities for raters are used to define the Guttman ordering. However, rater scalability coefficients based on the ac-MH model are calculated using adjacent-categories probabilities to establish the overall rater ordering. Although Mokken (1971) proposed a set of critical values for interpreting scalability coefficients for dichotomous items ($H_i \geq .50$: *strong*; $.40 \leq H_i < .50$: *medium*; $.30 \leq H_i < .40$: *weak*; $H_i < .30$: *unscalable*), the interpretation of these values for polytomous scalability coefficients, in general, and scalability coefficients based on adjacent-categories probabilities, in particular, has not been explored in detail.

Invariant Rater Ordering. The third set of Mokken indicators of rating quality are based on the ac-DM model requirement of invariant ordering. Adherence to this requirement implies that raters can be ordered the same way across students. As noted above, adherence to the DM model is typically evaluated for overall items rather than within rating scale categories (Ligtvoet et al., 2010; Ligtvoet et al., 2011; Sijtsma, Meijer, & van der Ark, 2011). As a result, methods for evaluating invariant ordering based on the ac-DM model are equivalent to those used to evaluate invariant ordering based on the original polytomous formulation of the DM model, with the exception that overall response functions for raters are examined for evidence of nonintersection instead of item response functions. Similarly, statistical tests for violations of invariant rater ordering (IRO) are examined within pairs of items in the traditional manner used in MSA.

Comparison Between Original and Imputed Data Sets

After indices of rater monotonicity, scalability, and IRO were calculated for each of the raters, results were compared between each of the original and imputed data sets. Specifically, discrepancies between the number of violations of monotonicity, scalability coefficients for individual raters (H_i), and the number of violations of IRO observed for the original data set and each of the imputed data sets were considered in terms of imputation techniques and levels of missingness.

Results

In this section, results are presented as they relate to the three Mokken rating quality indicators within the simulated and real data sets: rater monotonicity, rater scalability, and invariant rater ordering.

Rater Monotonicity

Rater monotonicity analyses included examining each of the simulated and real data sets for evidence of nondecreasing adjacent-categories rating scale probabilities across restscore groups. Within both the simulated and real data, monotonicity analyses of the complete ratings revealed no significant violations of monotonicity ($p < .01$) for any of the raters across simulation conditions and the four domains in the real data.

When the imputation methods were applied to the simulated data, no violations of monotonicity were observed across any of the imputed data sets. Similarly, across each of the imputed data sets based on the real data, only one significant violation of monotonicity was observed; this violation was observed for Rater 6 in the Organization domain when the CIMS-E method was used at the 40% level of missingness. Due to the low frequency in observed significant violations of rater monotonicity, no further analyses were used to explore differences in this rating quality indicator related to imputation methods and levels of missingness.

Rater Scalability

Table 2 presents the average individual rater scalability coefficients (H_i) across replications for the complete ratings generated using the simulation. Table 3 includes individual rater scalability coefficients for the 20 raters within the Conventions, Organization, Sentence Formation, and Style domains based on the real data. Examination of these results reveals generally higher scalability coefficients than the simulated data, with the highest average scalability coefficients were observed in the Conventions domain ($M = 0.75$, $SD = 0.03$) and the Style domain ($M = 0.71$, $SD = 0.04$), compared to average scalability for Organization ($M = 0.67$, $SD = 0.07$) and Sentence Formation ($M = 0.51$, $SD = 0.10$).

Discrepancies in Rater Scalability. In order to explore the effects of each of the imputation methods and levels of missingness on rater scalability, individual rater scalability coefficients based on each manipulated data set were compared to those observed in the corresponding complete data set, where positive discrepancies indicate that values of rater scalability coefficients were higher in the original data set ($\text{Discrepancy} = H_{i,\text{Original}} - H_{i,\text{Imputed}}$).

Table 4 includes average values of the discrepancies in rater scalability coefficients between the original and imputed simulated data sets. In both the 300 student conditions (Panel A) and the 1,000 student conditions (Panel B), the largest

Table 2. Mokken Rating Quality Results for Complete Simulated Ratings.

A. Individual rater scalability coefficients					
Number of rating scale categories		300 Students		1,000 Students	
		20 Raters	40 Raters	20 Raters	40 Raters
3	<i>M</i>	0.29	0.29	0.31	0.31
	<i>SD</i>	0.01	0.01	0.01	0.01
4	<i>M</i>	0.27	0.27	0.28	0.28
	<i>SD</i>	0.01	0.01	0.01	0.01
5	<i>M</i>	0.27	0.27	0.27	0.27
	<i>SD</i>	0.01	0.01	0.01	0.01

B. Count of significant violations of invariant rater ordering					
Number of rating scale categories		300 Students		1,000 Students	
		20 Raters	40 Raters	20 Raters	40 Raters
3	<i>M</i>	0.06	0.29	0.17	0.67
	<i>SD</i>	0.25	0.62	0.43	0.94
4	<i>M</i>	0.06	0.22	0.13	0.60
	<i>SD</i>	0.24	0.51	0.27	0.90
5	<i>M</i>	0.05	0.19	0.11	0.57
	<i>SD</i>	0.22	0.53	0.34	0.92

discrepancies were observed when the RI method was used, followed by the CIMS-E method. Across most conditions, the discrepancies based on the TW and TW-E methods differed by 0.01 or were equivalent. With the exception of the four-category rating scale conditions in which 50% of ratings were missing, these values were generally less than ± 0.04 —indicating generally accurate recovery of the original rater scalability coefficients based on the complete data sets.

Similar patterns were observed in the real data. Table 5 includes average values of the discrepancies in rater scalability coefficients between the original and imputed data sets for the real data. Examination of these results suggests that, with the exception of the Sentence Formation domain, the discrepancies between rater scalability coefficients based on the original and imputed data were largest (i.e., highest absolute value) when the RI method was used. The smallest discrepancies (i.e., lowest absolute value) varied across imputation methods for the levels of missingness across domains. For the lower levels of missingness (10% and 20%), the discrepancies in rater scalability were smallest when the CIMS-E method was used to impute missing values. For higher levels of missingness, the discrepancies were smallest when the TW or TW-E methods were used.

Although results for the Sentence Formation domain reflect the general finding that the discrepancies in rater scalability coefficients increased as levels of

Table 3. Mokken Rating Quality Results for Real Data.

Rater	A. Rater scalability coefficient (H_i)				B. Count of significant violations of invariant rater ordering			
	C	O	SF	S	C	O	SF	S
1	0.76	0.71	0.36	0.75	3	0	6	0
2	0.71	0.69	0.55	0.72	1	0	2	0
3	0.75	0.69	0.47	0.73	3	0	2	0
4	0.77	0.70	0.38	0.64	5	3	1	0
5	0.77	0.70	0.61	0.65	0	1	1	0
6	0.70	0.41	0.50	0.67	0	0	2	0
7	0.79	0.66	0.61	0.72	5	2	1	1
8	0.79	0.73	0.56	0.79	1	0	1	0
9	0.75	0.65	0.58	0.68	1	0	1	0
10	0.78	0.67	0.61	0.73	3	0	0	0
11	0.77	0.71	0.48	0.73	4	2	2	0
12	0.78	0.67	0.52	0.72	0	1	2	0
13	0.73	0.68	0.40	0.70	9	1	10	0
14	0.73	0.77	0.37	0.71	7	1	1	0
15	0.73	0.63	0.56	0.72	5	1	2	3
16	0.78	0.68	0.62	0.74	0	0	1	2
17	0.74	0.66	0.60	0.63	1	0	3	0
18	0.67	0.66	0.31	0.72	2	1	8	0
19	0.76	0.69	0.60	0.74	4	1	2	1
20	0.71	0.60	0.54	0.67	0	2	0	1
M	0.75	0.67	0.51	0.71	2.70	0.80	2.40	0.40
SD	0.03	0.07	0.10	0.04	2.58	0.89	2.60	0.82

Note. Domains are abbreviated as follows: C = Conventions; O = Organization; SF = Sentence Formation; S = Style. No violations of monotonicity were observed in the original (complete) data set in any of the four domains.

missingness increased, the results in this domain did not match the patterns observed in the other three domains regarding the imputation techniques. Specifically, discrepancies in rater scalability coefficients in the Sentence Formation domain were generally smaller based on the RI technique compared to the other imputation techniques across all levels of missingness. This finding is interesting in light of the finding of more variable scalability coefficients based on the complete data for this domain ($SD = 0.10$) compared to the other domains in the real data ($SD \leq 0.07$), and compared to the variability of scalability coefficients in the simulated data sets ($SD = 0.01$). Although rater scalability coefficients are calculated for each individual rater, the underlying category ordering used to define Guttman errors is based on the overall group of raters minus the raters of interest. Accordingly, the finding that the RI technique was best at reproducing the original scalability coefficients may reflect the fact that there was more variation among raters in the original data for this domain, and the use of imputation techniques based on mean values may have introduced additional consistency across raters that was not observed in the original data.

Table 4. Average Discrepancies in Rater Scalability Coefficients: Simulated Data.

Panel A: 300 Student conditions																
N raters = 20																
Imputation method	3 Categories					4 Categories					5 Categories					
	% Missing					% Missing					% Missing					
	10	20	30	40	50	10	20	30	40	50	10	20	30	40	50	
Random imputation	0.09	0.14	0.17	0.09	0.19	0.08	0.12	0.15	0.08	0.17	0.08	0.12	0.15	0.08	0.17	0.17
Two-way	0.00	0.00	-0.01	0.00	-0.04	0.00	-0.01	-0.02	0.00	-0.07	0.01	0.02	0.01	0.01	-0.02	-0.02
Two-way error	0.00	0.00	-0.01	0.00	-0.04	0.00	0.00	-0.01	0.00	-0.06	0.02	0.03	0.03	0.02	0.01	0.01
CIMS-E	-0.02	-0.04	-0.06	-0.02	-0.13	-0.03	-0.06	-0.10	-0.03	-0.20	-0.03	-0.06	-0.11	-0.03	-0.22	-0.22
N raters = 40																
Imputation method	3 Categories					4 Categories					5 Categories					
	% Missing					% Missing					% Missing					
	10	20	30	40	50	10	20	30	40	50	10	20	30	40	50	
Random imputation	0.09	0.15	0.19	0.09	0.23	0.08	0.14	0.18	0.08	0.21	0.08	0.13	0.17	0.08	0.21	0.21
Two-way	0.01	0.01	0.02	0.01	0.01	0.00	-0.01	-0.02	0.00	-0.06	0.01	0.01	0.01	0.01	-0.01	-0.01
Two-way error	0.01	0.01	0.01	0.01	0.01	0.00	-0.01	-0.02	0.00	-0.05	0.01	0.01	0.01	0.01	0.01	0.00
CIMS-E	-0.02	-0.03	-0.04	-0.02	-0.10	-0.03	-0.05	-0.08	-0.03	-0.18	-0.03	-0.06	-0.09	-0.03	-0.19	-0.19

(continued)

Table 4. (continued)

Panel B: 1,000 Student conditions																
N raters = 20																
Imputation method	3 Categories					4 Categories					5 Categories					
	% Missing					% Missing					% Missing					
	10	20	30	40	50	10	20	30	40	50	10	20	30	40	50	
Random imputation	0.09	0.14	0.17	0.08	0.20	0.08	0.12	0.15	0.08	0.17	0.08	0.13	0.16	0.08	0.18	
Two-way	0.00	0.00	-0.01	0.00	-0.04	0.00	-0.01	-0.03	0.00	-0.08	0.01	0.02	0.01	0.01	-0.01	
Two-way error	0.00	0.00	-0.01	0.00	-0.04	0.00	-0.01	-0.02	0.00	-0.07	0.02	0.03	0.03	0.02	0.01	
CIMS-E	-0.02	-0.03	-0.05	-0.02	-0.12	-0.03	-0.06	-0.09	-0.03	-0.20	-0.03	-0.06	-0.10	-0.03	-0.21	
N raters = 40																
Imputation method	3 Categories					4 Categories					5 Categories					
	% Missing					% Missing					% Missing					
	10	20	30	40	50	10	20	30	40	50	10	20	30	40	50	
Random imputation	0.10	0.16	0.20	0.10	0.24	0.08	0.14	0.17	0.08	0.21	0.07	0.13	0.17	0.07	0.21	
Two-way	0.01	0.01	0.02	0.01	0.02	0.00	-0.01	-0.03	0.00	-0.06	0.01	0.01	0.01	0.01	0.00	
Two-way error	0.01	0.01	0.02	0.01	0.02	0.00	-0.01	-0.02	0.00	-0.06	0.01	0.02	0.01	0.01	0.01	
CIMS-E	-0.01	-0.02	-0.04	-0.01	-0.09	-0.02	-0.05	-0.08	-0.02	-0.17	-0.03	-0.05	-0.09	-0.03	-0.19	

Table 5. Average Discrepancies in Rater Scalability Coefficients: Real Data.

Imputation method	Conventions					Organization				
	% Missing					% Missing				
	10	20	30	40	50	10	20	30	40	50
Random imputation	0.03	0.07	0.10	0.10	0.12	0.02	0.06	0.08	0.08	0.08
Two-way	−0.02	−0.03	−0.05	−0.08	−0.10	0.05	0.02	0.02	0.00	−0.02
Two-way error	−0.02	−0.03	−0.06	−0.07	−0.10	0.01	−0.01	0.03	−0.03	−0.04
CIMS-E	−0.01	−0.03	−0.05	−0.08	−0.11	−0.02	−0.02	−0.07	−0.10	−0.15

Imputation method	Sentence Formation					Style				
	% Missing					% Missing				
	10	20	30	40	50	10	20	30	40	50
Random imputation	−0.01	0.03	0.02	0.02	−0.05	0.10	0.08	0.12	0.13	0.14
Two-way	−0.03	−0.11	−0.11	−0.22	−0.15	−0.02	−0.03	−0.05	−0.08	−0.08
Two-way error	−0.06	−0.08	−0.08	−0.25	−0.22	−0.02	−0.03	−0.05	−0.07	−0.07
CIMS-E	−0.04	−0.10	−0.13	−0.13	−0.27	−0.01	−0.03	−0.06	−0.07	−0.13

Invariant Rater Ordering

Similar to the rater scalability coefficients, counts of significant violations of IRO observed within each manipulated data set were compared to those observed in the complete data set, where positive discrepancies indicate that more violations of IRO were observed in the imputed data set than in the complete data ($\text{Discrepancy} = \text{Count}_{\text{Original}} - \text{Count}_{\text{Imputed}}$).

Table 6 includes average discrepancies in counts of significant violations of IRO based on the simulated data. Specifically, the discrepancies between the original and imputed data sets are quite small across the simulation and imputation conditions, and the range of discrepancies does not appear to show a clear pattern related to imputation techniques. As was observed in the real data, the average discrepancies appear to increase across increasing levels of missingness.

Table 7 shows a similar pattern of average discrepancies in significant violations of IRO based on the real data. Examination of the average discrepancies in each domain reveals that, with the exception of some of the three-category rating scale conditions in which 50% of the ratings were missing, there were generally small differences (<0.50) in the frequency of significant violations of IRO between the original and imputed data sets. Furthermore, although the average discrepancies appear to increase as the level of missingness increases, they do not appear to show a clear pattern related to the imputation techniques across the four domains.

Summary and Conclusions

The purpose of this study was to explore the impact of missing data imputation on MSA-based rating quality indicators. Specifically, adjacent-categories formulations

Table 6. Average Discrepancies in Counts of Significant Violations of Invariant Rater Ordering: Simulated Data.

Panel A: 300 Student conditions																
N raters = 20																
3 Categories					4 Categories					5 Categories						
% Missing					% Missing					% Missing						
Imputation method	10	20	30	40	50	10	20	30	40	50	10	20	30	40	50	
Random imputation	-0.04	-0.05	-0.11	-0.04	-0.12	-0.01	-0.04	-0.08	-0.01	-0.12	-0.04	-0.05	-0.09	-0.04	-0.17	
Two-way	0.01	-0.03	-0.07	0.01	-1.79	-0.01	-0.02	0.00	-0.01	-0.05	0.00	-0.01	-0.04	0.00	-0.02	
Two-way error	-0.03	-0.04	-0.06	-0.03	-1.77	0.00	-0.01	-0.04	0.00	-0.07	-0.01	-0.03	-0.02	-0.01	-0.04	
CIMS-E	-0.01	-0.03	-0.04	-0.01	-0.04	-0.01	-0.01	0.00	-0.01	-0.01	0.02	0.01	0.00	0.02	0.03	
N raters = 40																
3 Categories					4 Categories					5 Categories						
% Missing					% Missing					% Missing						
Imputation method	10	20	30	40	50	10	20	30	40	50	10	20	30	40	50	
Random imputation	-0.23	-0.30	-0.36	-0.23	-0.45	-0.21	-0.27	-0.34	-0.21	-0.39	-0.17	-0.27	-0.31	-0.17	-0.42	
Two-way	-0.19	-0.16	-0.17	-0.19	-0.31	-0.15	-0.24	-0.25	-0.15	-0.41	-0.15	-0.14	-0.17	-0.15	-0.20	
Two-way error	-0.21	-0.22	-0.20	-0.21	-0.29	-0.13	-0.18	-0.27	-0.13	-0.44	-0.17	-0.16	-0.15	-0.17	-0.14	
CIMS-E	-0.20	-0.18	-0.22	-0.20	-0.19	-0.15	-0.15	-0.15	-0.15	-0.08	-0.11	-0.09	-0.07	-0.11	-0.02	

(continued)

Table 6. (continued)

Panel B: 1,000 Student conditions															
N raters = 20															
3 Categories						4 Categories					5 Categories				
% Missing						% Missing					% Missing				
Imputation method	10	20	30	40	50	10	20	30	40	50	10	20	30	40	50
Random imputation	-0.06	-0.11	-0.18	-0.06	-0.28	-0.11	-0.11	-0.15	-0.11	-0.30	0.02	-0.01	-0.01	0.02	-0.07
Two-way	-0.04	-0.14	-1.08	-0.04	-5.53	-0.03	0.01	-0.04	-0.03	-0.07	0.08	0.07	0.06	0.08	0.04
Two-way error	-0.04	-0.15	-1.01	-0.04	-5.38	0.01	0.01	-0.05	0.01	-0.06	0.06	0.04	0.05	0.06	0.05
CIMS-E	-0.01	-0.06	-0.01	-0.01	0.05	0.00	0.02	-0.01	0.00	0.10	0.10	0.07	0.08	0.10	0.10
N raters = 40															
3 Categories						4 Categories					5 Categories				
% Missing						% Missing					% Missing				
Imputation method	10	20	30	40	50	10	20	30	40	50	10	20	30	40	50
Random imputation	-0.10	-0.29	-0.34	-0.10	-0.58	-0.18	-0.32	-0.40	-0.18	-0.79	-0.19	-0.41	-0.55	-0.19	-1.03
Two-way	-0.11	-0.75	-3.88	-0.11	-11.50	0.02	-0.04	-0.14	0.02	-0.33	0.03	-0.10	-0.05	0.03	-0.05
Two-way error	-0.04	-0.15	-1.01	-0.04	-5.38	0.01	-0.10	-0.22	0.01	-0.13	0.02	-0.06	-0.06	0.02	-0.11
CIMS-E	-0.02	-0.02	-0.08	-0.02	-0.04	0.09	0.03	0.04	0.09	0.19	0.38	0.45	0.48	0.38	0.54

Table 7. Average Discrepancies in Counts of Significant Violations of Invariant Rater Ordering: Real Data.

Imputation method	Conventions					Organization				
	% Missing					% Missing				
	10	20	30	40	50	10	20	30	40	50
Random imputation	1.20	1.30	0.70	1.00	2.20	0.00	0.30	−0.90	0.20	−0.60
Two-way	0.40	0.60	1.80	2.00	2.00	−0.10	−0.10	0.20	0.50	0.20
Two-way error	0.60	1.00	1.20	2.00	2.40	0.30	0.50	−0.30	0.50	0.70
CIMS-E	0.40	1.30	1.70	2.00	2.00	−0.20	0.30	0.40	0.10	0.50

Imputation method	Sentence Formation					Style				
	% Missing					% Missing				
	10	20	30	40	50	10	20	30	40	50
Random imputation	0.40	−0.30	1.00	−0.50	1.30	−0.80	0.10	0.00	−0.70	0.20
Two-way	−0.30	0.90	1.10	2.00	1.80	−0.40	0.20	−0.30	0.10	0.20
Two-way error	−0.20	0.60	0.60	1.80	2.10	0.00	0.20	0.30	0.10	0.40
CIMS-E	−0.40	0.60	1.30	1.50	2.00	−0.20	0.20	−0.10	0.30	0.40

of the MH and DM models were used to calculate indicators of rater monotonicity, scalability, and invariant ordering. Simulated data and real data from a large-scale rater-mediated writing assessment were manipulated to reflect varying levels of missingness, and data were imputed using four methods that have previously been explored in missing data research related to MSA. The effects of levels of missingness and imputation methods were considered across simulation conditions and domains in the real data.

Research Question 1: What is the effect of missing data imputation techniques and levels of missingness on indicators of rating quality based on adjacent-categories Mokken models?

The guiding research question for this study focused on the effects of missing data imputation techniques on indicators of rating quality based on ac-MSA. First, rater monotonicity analyses for the imputed data sets resulted in nearly perfect agreement with those based on the complete data across the simulated data sets and across all four domains in the real data. This finding suggests that indicators of rater monotonicity appear robust to imputation at the individual rater level when the ac-MH model of assumption is met in the original data.

In terms of rater scalability, the results indicated that discrepancies between the original and imputed data sets tended to increase across increasing levels of

missingness. This pattern was observed across the three-, four-, and five-category rating scale and sample size conditions in the simulated data, as well as across the domains in the real data. Furthermore, the results indicated generally close correspondence to the original scalability coefficients when imputation techniques besides RI were used. In general, the TW and TW-E methods appeared to produce the smallest average discrepancies between the original and imputed data sets when more than 30% of the ratings were missing, and the CIMS-E method appeared to produce the smallest average discrepancies for lower levels of missingness.

Finally, the effects of missing data imputation on IRO were explored by comparing the observed frequency of significant violations of IRO in the original simulated and real data sets to the data sets with imputed ratings. Similar to the scalability results, the discrepancies in significant violations IRO tended to increase as the level of missing ratings increased across simulation conditions and the four domains in the real data. However, the values of discrepancies related to IRO were not systematic across the imputation techniques within the simulation conditions and across all four domains in the real data.

Discussion

The results from this study provide initial insight into the consequences of various imputation methods that is essential for the widespread use of MSA to evaluate rating quality. In particular, this study provides additional insight into the use of adjacent-categories Mokken models (Wind, 2017) to explore the quality of rating data when fully crossed ratings are not available. The finding that the application of missing data imputation techniques to impute missing polytomous ratings resulted in generally small discrepancies in rater scalability coefficients based on adjacent categories suggests that these techniques may facilitate the use of ac-MSA models in the presence of incomplete ratings when rater scalability is of interest. Furthermore, the finding that the TW and TW-E techniques resulted in fairly accurate imputation of missing ratings reflects the results of previous MSA missing data research (Sijtsma & van der Ark, 2003; van Ginkel et al., 2007) and suggests that relatively simple missing data imputation techniques may be effective in this context.

Although the results related to rater monotonicity and IRO were somewhat inconclusive, the general finding of small discrepancies between the original and imputed data sets suggests that the missing data imputation techniques explored in this study may be applicable to missing ratings when these indices are of interest. Although these results provide an initial step toward the application of ac-MSA to rater-mediated assessments when complete ratings are not available, additional research is needed in order to more fully understand the effects of these techniques across a variety of conditions that characterize data collection designs in operational rater-mediated assessments (discussed further below).

It is important to note that the methods presented in this study are not intended to replace existing parametric techniques that are well suited for use in the context of

incomplete rating designs, such as the many-facet Rasch model (Linacre, 1989), but rather as an additional tool that can be used to explore rating quality within the framework of invariant measurement from a nonparametric perspective. Whereas the treatment of missing data in parametric approaches is generally based on estimation techniques that rely on connectivity within a data set, the general approach in MSA is to impute missing values. Previous studies related to the use of MSA for evaluating rating quality have demonstrated the diagnostic value of this approach in conjunction with other methods, such as the MFR model (e.g., Wind, 2014, 2017; Wind & Engelhard, 2016). The current study builds on these initial findings to consider the use of MSA when complete ratings are not available. Future research could include the exploration of alternative methods for handling missing data in the context of MSA that do not rely on imputation techniques.

Implications

This study has several implications in terms of previous MSA research related to missing data in general. First, the finding of comparable performance in recovering rater scalability coefficients across the TW and TW-E methods, along with the superior performance of these methods relative to RI, and, in some cases, CIMS-E, reflects findings described in previous research (e.g., van der Ark & Sijsma, 2005; van Ginkel et al., 2007). Previous research on missing data in the context of MSA has not included examinations of the influence of imputation techniques on indicators of monotonicity or invariant ordering; as a result, it is not possible to compare the current results to previous findings related to these indices.

Second, it is important to note that the context of the current study differs in several significant ways from previous explorations of missing data imputation in the context of MSA. As noted above, levels of missingness in rater-mediated assessments are naturally higher than missingness observed in traditional applications of MSA, such as surveys and questionnaires. Accordingly, the current study included higher levels of missingness than were explored in previous investigations of missing data imputation in the context of MSA that resulted in larger discrepancies between the original and imputed data sets than had been described in previous research.

Furthermore, the nature of rating designs in operational performance assessment systems limit the applicability of several imputation techniques that depend on complete responses for at least some subset of responses. In particular, it was not possible to explore the RF imputation method that has been explored in other missing data research related to MSA (e.g., Sijsma & van der Ark, 2003; van Ginkel et al., 2007), because this method requires complete observations for a subset of respondents. Because complete subsets of ratings are generally not included in data collection designs for operational rater-mediated performance assessments, these designs were not included in the current study.

Limitations and Directions for Further Research

When considering the results from this study, several limitations and corresponding directions for further research are important to note. First, the design of the simulation study only included two sample sizes for students ($N = 300$ and $N = 1,000$) and two sample sizes for raters ($N = 20$ and $N = 40$). This design was intended to reflect operational rater-mediated assessment systems in general, as well as assessment systems in which the use of MSA may be desirable, where relatively small sample sizes are included. Additional simulation-based research should explore the effects missing data imputation techniques on ac-MSA rating quality indicators across a larger range of student and rater sample sizes.

Along the same lines, the simulation design did not include the explicit manipulation of characteristics related to rating quality in terms of rater monotonicity, scalability, and invariant ordering. Because the rating quality indicators based on ac-MSA that were examined in this study are relatively new, much additional research is needed related to the sensitivity of these indicators in general, as well as in the context of missing data. Specifically, additional simulation studies in which violations of rater monotonicity, Guttman errors, and violations of invariant rater ordering were specified would potentially provide more definitive insight into the effects of missing data imputation techniques on these rating quality indicators. In particular, the finding of less-clear patterns across imputation methods within the Sentence Formation domain in the real data suggests that the effectiveness of imputation techniques may vary across conditions that were not included in the simulation study.

Finally, the rating designs explored in both the simulated and real data sets were based on incomplete but connected data collection designs that reflect those used in many operational rater-mediated assessment systems (Johnson et al., 2009). However, additional rating designs are available, such as those based on disconnected designs, anchor raters, and anchor students (Eckes, 2015; Engelhard, 1997). Additional explorations of missing data in the context of ac-MSA models for rater-mediated assessments should consider the effects of imputation techniques across these designs, as well as additional incomplete designs in which the strength of connectivity is systematically manipulated.

Authors' Note

A previous version of this article was presented at the 10th meeting of the International Test Commission in Vancouver, British Columbia, Canada, July 2016.

Declaration of Conflicting Interests

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

Funding

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

References

- Andrich, D. A. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561-573. doi:10.1007/BF02293814
- Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (2nd ed.). Frankfurt am Main, Germany: Peter Lang.
- Engelhard, G. (1997). Constructing rater and task banks for performance assessments. *Journal of Outcome Measurement*, 1(1), 19-33.
- Engelhard, G. (2008). Historical perspectives on invariant measurement: Guttman, Rasch, and Mokken. *Measurement*, 6(3), 155-189. doi:10.1080/15366360802197792
- Harwell, M., Stone, C. A., Hsu, T.-C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement*, 20, 101-125. doi:10.1177/014662169602000201
- Johnson, R. L., Penny, J. A., & Gordon, B. (2009). *Assessing performance: Designing, scoring, and validating performance tasks*. New York: Guilford Press.
- 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. doi:10.1177/0013164409355697
- Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago, IL: MESA Press.
- Little, R. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: Wiley.
- Loevinger, J. (1948). The technic of homogeneous tests compared with some aspects of scale analysis and factor analysis. *Psychological Bulletin*, 45, 507-529.
- Marais, I., & Andrich, D. A. (2011). Diagnosing a common rater halo effect using the polytomous Rasch model. *Journal of Applied Measurement*, 12, 194-211.
- 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. doi:10.1177/014662169001400306
- Mokken, R. J. (1971). *A theory and procedure of scale analysis*. The Hague, Netherlands: De Gruyter.
- 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 multcategory items. *Kwantitative Methoden*, 37(12), 97-117.
- Molenaar, I. W. (1997). Nonparametric models for polytomous responses. In W. J. van der Linden & Hambleton, R. K. (Eds.), *Handbook of modern item response theory* (pp. 369-380). New York, NY: Springer.
- 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, 31-37. doi:10.1016/j.paid.2010.08.016

- Sijtsma, K., & van der Ark, L. A. (2003). Investigation and treatment of missing item scores in test and questionnaire data. *Multivariate Behavioral Research*, 38, 505-528. doi: 10.1207/s15327906mbr3804_4
- van der Ark, L. A., & Sijtsma, K. (2005). The effect of missing data imputation on Mokken scale analysis. In L. A. van der Ark, M. A. Croon, & K. Sijtsma (Eds.), *New developments in categorical data analysis for the social and behavioral sciences* (pp. 147-166). Mahwah, NJ: Lawrence Erlbaum.
- van Ginkel, J. R., & van der Ark, L. A. (2005). SPSS syntax for missing value imputation in test and questionnaire data. *Applied Psychological Measurement*, 29, 152-153. doi: 10.1177/0146621603260688
- van Ginkel, J. R., van der Ark, L. A., & Sijtsma, K. (2007). Multiple imputation of item scores in test and questionnaire data, and influence on psychometric results. *Multivariate Behavioral Research*, 42, 387-414. doi:10.1080/00273170701360803
- Wind, S. A. (2014). Examining rating scales using Rasch and Mokken models for rater-mediated assessments. *Journal of Applied Measurement*, 15, 100-132.
- Wind, S. A. (2015). Evaluating the quality of analytic ratings with Mokken scaling. *Psychological Test and Assessment Modeling*, 3, 423-444.
- Wind, S. A. (2017). Adjacent-categories Mokken models for rater-mediated assessments. *Educational and Psychological Measurement*, 77, 330-350. doi:10.1177/0013164416643826
- Wind, S. A., Engelhard, G., Jr., & Wesolowski, B. (2016). Exploring the effects of rater linking designs and rater fit on achievement estimates within the context of music performance assessments. *Educational Assessment*, 21, 278-299. doi: 10.1080/10627197.2016.1236676
- Wind, S. A., & Engelhard, G. (2016). Exploring rating quality in rater-mediated assessments using Mokken scaling. *Educational and Psychological Measurement*, 76, 685-706.
- Wolfe, E. W., Jiao, H., & Song, T. (2014). A family of rater accuracy models. *Journal of Applied Measurement*, 16, 153-160.
- 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. doi:10.1111/j.1745-3992.2012.00241.x
- Wolfe, E. W., & Song, T. (2015). Comparison of models and indices for detecting rater centrality. *Journal of Applied Measurement*, 16, 228-241.