Exploring Within-Rater Category Ordering: A Simulation Study Using Adjacent-Categories Mokken Scale Analysis

Educational and Psychological
Measurement

2018, Vol. 78(5) 887–904

© The Author(s) 2017

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164417724841

journals.sagepub.com/home/epm

Stefanie A. Wind¹ and Randall E. Schumacker¹

Abstract

The interpretation of ratings from educational performance assessments assumes that rating scale categories are ordered as expected (i.e., higher ratings correspond to higher levels of judged student achievement). However, this assumption must be verified empirically using measurement models that do not impose ordering constraints on the rating scale category thresholds, such as item response theory models based on adjacent-categories probabilities. This study considers the application of an adjacent-categories formulation of polytomous Mokken scale analysis (ac-MSA) models as a method for evaluating the degree to which rating scale categories are ordered as expected for individual raters in performance assessments. Using simulated data, this study builds on the preliminary application of ac-MSA models to rater-mediated performance assessments, in which a real data analysis suggested that these models can be used to identify disordered rating scale categories. The results suggested that ac-MSA models are sensitive to disordered categories within individual raters. Implications are discussed as they relate to research, theory, and practice for rater-mediated educational performance assessments.

Keywords

Mokken scaling, rating scales, nonparametric IRT, performance assessment, raters

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, Tuscaloosa, AL 35487, USA.

Email: swind@ua.edu

Educational performance assessments often involve human raters who evaluate student performances using a multicategory rating scale. These rating scales are usually paired with a scoring rubric that describes expected characteristics or criteria for each rating scale category, where ratings in higher categories correspond to higher levels of achievement. A meaningful interpretation of ratings from these assessments is based on the assumption that raters use the rating scale categories in the expected order (i.e., higher ratings correspond to higher levels of judged student achievement). However, rating scale category ordering is an empirical question for which evidence must be examined for each rater. To evaluate the alignment between expected and empirical category ordering, it is necessary to use measurement models that do not impose ordering constraints on the rating scale category thresholds, such as item response theory (IRT) models based on adjacent-categories probabilities.

Although issues related to rating scale category ordering have traditionally been considered within the context of parametric IRT (discussed further below), it is also possible to evaluate rating scale category ordering using nonparametric IRT models based on adjacent categories. In general, scholars have recognized the usefulness of nonparametric models for evaluating the psychometric properties of measurement instruments in contexts for which the underlying response processes is complex or not well understood, such as the measurement of affective variables using ordinal rating scales. Recognizing that rater-mediated educational performance assessments also involve complex response processes based on rater judgments of student performances, Wind (2014, 2015, 2016) and colleagues (Wind & Engelhard, 2015; Wind & Patil, 2018) have demonstrated the application of Mokken's (1971) nonparametric IRT models to rater-mediated assessments. When applied to these assessments, Mokken scale analysis (MSA) provides an exploratory approach to evaluating rating activity in terms of fundamental measurement properties, while maintaining an ordinal level of measurement. Accordingly, indices based on MSA provide a method for evaluating the quality of rater judgments in terms of the degree to which they adhere to fundamental measurement properties within a measurement framework supported by ordinal measurement and nonparametric properties.

When considering the application of MSA (Mokken, 1971) to educational assessments, it is important to recognize that the original polytomous MSA models (Molenaar, 1982, 1997) are based on cumulative probabilities, where item step response functions (ISRFs; i.e., category probabilities) are defined as the conditional probability of a response in or above a given category. This cumulative probability formulation has important consequences for interpretation of rating scale category probabilities based on polytomous MSA models across measurement contexts. In particular, the interpretation of cumulative probabilities, which are defined as the probability that a particular examinee receives a rating in or above a given category, is often incongruent with the intended interpretation of rating scales in performance assessments. Furthermore, the cumulative formulation of rating scale probabilities does not result in empirical evidence that thresholds match the intended ordering of the ordinal rating scale categories (Andrich, 2011, 2013, 2015; Linacre, 2002). As

pointed out by Andrich, this diagnostic capability is necessary for the development and improvement of rating scales.

Recognizing the potential incongruence between cumulative probabilities and educational assessments, Wind (2016) proposed an alternative formulation of polytomous MSA based on adjacent-categories probabilities (ac-MSA) that has a closer conceptual alignment to educational assessments. Although category ordering was not the primary focus of the original presentation of ac-MSA, an important observation from this study was that ac-MSA provided diagnostic information about category ordering that was not revealed via the original approach to MSA. Because the original analysis was based on real data, the results did not provide insight into the specific conditions in which disordered categories could be identified. Accordingly, the current study is a follow-up analysis to Wind (2016) that includes a more systematic investigation of rating scale category ordering within the framework of ac-MSA. Specifically, a simulation approach is used to experimentally manipulate a variety of characteristics of polytomous ratings in order to examine the scope of conditions in which ac-MSA can successfully identify disordered rating scale categories.

Purpose

The purpose of this study was to explore the degree to which ac-MSA can be used to correctly identify rating scale category disordering (i.e., the sensitivity of ac-MSA to disordered categories) using threshold estimates from the partial credit model (Masters, 1982) as a frame of reference. Two questions guide the analyses:

Research Question 1: How does the sensitivity of ISRFs based on ac-MSA models to disordered categories vary across magnitudes of disordering, the overall proportion of raters with disordered categories, and sample sizes?

Research Question 2: How can graphical displays of ISRFs based on ac-MSA be used to explore within-rater category disordering?

This study contributes to previous research in several ways. First of all, this study extends the original presentation of ac-MSA models (Wind, 2016) to include a more systematic examination of the degree to which this approach can be used to detect disordered rating scale categories within individual raters in the context of rater-mediated educational performance assessments. Specifically, a simulation study was needed to examine the specific conditions within which disordered categories can be identified using this approach. Furthermore, although numerous simulation studies have been conducted with Mokken scaling, none have been published that deal with rating scale category ordering, and only one simulation study has been published related to ac-MSA (Wind & Patil, 2018). This study also contributes to research on rating scale category disordering within the context of IRT in general by presenting a nonparametric approach to exploring category ordering. Although several scholars have discussed methods for evaluating category disordering and the interpretation of

disordered categories based on IRT models (e.g., Adams, Wu, & Wilson, 2012; Andrich, 2013), these discussions have previously been limited to parametric approaches.

Evaluating Rating Scale Category Functioning Using Item Response Theory Models

When rating scales are used in a measurement procedure, it is essential to empirically examine whether or not the categories are functioning as intended. Essentially, evidence of *rating scale category functioning* describes the empirical application of a set of rating scale categories related to polytomous items or rater judgments. In previous research, several scholars have recommended guidelines for evaluating rating scale category functioning based on parametric IRT models (Engelhard & Wind, 2013; Linacre, 2002). Broadly, these guidelines encourage researchers to evaluate whether rating scale categories are ordered in the same direction as the latent variable, whether different categories reflect substantively meaningful differences in terms of the latent variable, and whether responses in each category match the expectations of the measurement model. Similar to model-data fit analyses, the examination of rating scale category functioning can reveal discrepancies between expectations based on the underlying measurement theory and the operational use and interpretation of rating scales.

Recently, there has been some discussion in the IRT literature related to the category ordering component of rating scale functioning. Specifically, several scholars have discussed conceptual issues related to the underlying causes and interpretation of disordered categories in rating scales (Adams et al., 2012; Andrich, 2013, 2015) as well as appropriate modeling procedures when disordered categories are observed (Wetzel & Carstensen, 2014). Although these scholars articulate different opinions regarding the underlying causes, interpretation, and appropriate interpretations and analytic solutions for disordered categories, the current consensus in the literature appears to advocate for empirically examining category ordering when polytomous IRT models are applied to rating scale data. However, these issues have not yet been discussed in the context of nonparametric IRT models, and methods for detecting disordered categories have not been compared across measurement frameworks.

Furthermore, despite these discussions of rating scale category ordering in the psychometric literature, it is interesting to note that issues related to category ordering are not commonly included in methodological or applied research on evaluating the psychometric properties of educational performance assessments (Wind & Peterson, 2018). Instead, most of the discussions of rating quality indices based on IRT models focus on issues related to rater severity/leniency, central tendency and other range restrictions, and model-data fit (Engelhard, 2002; Johnson, Penny, & Gordon, 2009; Myford & Wolfe, 2003, 2004; Wolfe & McVay, 2012).

Polytomous Item Response Theory Models for Evaluating Rating Scale Functioning

A variety of polytomous IRT models have been proposed that can be used to explore rating scale functioning (Nering & Ostini, 2010). In general, these polytomous IRT models have been developed as extensions of earlier models for dichotomous responses (Hambleton, van der Linden, & Wells, 2010). A distinguishing feature across polytomous IRT models is the formulation of the $(k - 1)$ thresholds that distinguish between responses in k rating scale categories. These thresholds can be defined using a variety of formulations to represent the probability of a rating in each category. Several scholars have proposed classification schemes for polytomous IRT models based on threshold definitions (Agresti, 2007; Andrich, 2015; Mellenbergh, 1995; Penfield, 2014). Although the models appear similar in form, the interpretation of results varies across parameterizations. It is therefore essential that the selection of a model for the development of a rating scale match the intended category interpretation.

Polytomous Mokken Models. Similar to other IRT models, such as the Rasch model (Rasch, 1960), MSA was originally developed for dichotomous item responses (Mokken, 1971). Building on the original models, Molenaar (1982, 1997) presented the polytomous monotone homogeneity (MH) and double monotonicity (DM) models. The polytomous MH model is based on three requirements: (1) *unidimensionality*—item responses can be explained by a single latent variable; (2) *local independence*—responses to one item are statistically independent from responses to any other item, after controlling for the latent variable; and (3) *monotonicity*—The conditional probability for a rating in category k or higher is non-decreasing over increasing values of the latent variable. The polytomous DM model shares these requirements, with the addition of a fourth requirement: (4) *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. This DM model requirement is evaluated by plotting the *item step response function* (ISRF), or the conditional probability for a rating in each of the $(k-1)$ rating scale categories, across increasing levels of student achievement. Because MSA models are nonparametric, student achievement estimates are calculated using *restscores*, or student total scores minus the item of interest. Students with adjacent *restscores* are combined into *restscore groups* to evaluate model requirements for individual items.

The nonparametric ISRFs for polytomous MSA models can be viewed as analogues to rating scale category thresholds in parametric IRT. Molenaar (1997) described the ISRFs for polytomous MSA models as a set of $(k-1)$ “steps” that reflect the observed rating in category k on item i for person j , calculated as follows using cumulative probabilities:

$$\tau_{ijk} = 1 \text{ when } X_{ij} \geq k, \text{ and } \tau_{ijk} = 0 \text{ otherwise,} \quad (1)$$

where X_{ij} is the observed rating on item i for person j . Because of this definition of τ_{ijk} , Molenaar's polytomous MSA models are classified as cumulative probability models.

Figure 1, Panel B illustrates the cumulative thresholds that characterize the original MSA models. A comparison of these cumulative thresholds with the adjacent-categories thresholds in Panel A highlights the conceptual differences between cumulative and adjacent-categories thresholds. Whereas the adjacent-categories thresholds describe the probability for a rating in a single category at a time, the cumulative thresholds describe the probability for a rating in a category and all subsequent categories in the ordinal rating scale. This distinction leads to different interpretations of thresholds across the two models. Furthermore, the cumulative definition does not facilitate the empirical confirmation of category ordering that characterizes adjacent-categories thresholds; rather, thresholds are ordered within an item (Andrich, 2015; Molenaar, 1997).

Adjacent-Categories Mokken Models. As noted above, Wind (2016) presented an alternative formulation of polytomous MSA models in which the rating scale thresholds are defined using adjacent categories. This approach to MSA is based on an adaptation of the Mokken ISRF in Equation (1) to reflect adjacent-categories thresholds as follows:

$$\tau_{ijk} = 1 \text{ when } X_{ij} = k, \text{ and } \tau_{ijk} = 0 \text{ when } X_{ij} = k - 1, \quad (2)$$

where X_{ij} is the observed score on item i for person j . Based on Equation (2), an adjacent-categories formulation of the MH model (ac-MH model) can be described, where the monotonicity requirement is re-stated as follows: 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 model) has the following nonintersection requirement: The conditional probability for a rating in category k rather than in category $(k-1)$ on item i has the same relative ordering across all values of the latent variable.

Figure 1, Panel C illustrates the adjacent-categories thresholds for polytomous Mokken models. This illustration highlights the distinction between the adjacent-categories MSA models (Wind, 2016) and the original polytomous MSA models (Molenaar, 1982, 1997). The major implication of the adjacent-categories formulation in the context of MSA is the ability to empirically confirm the ordering of rating scale categories to inform the development, revision, and interpretation of rating scales.

Method

Building on the original presentation of ac-MSA, which was based on a real data analysis (Wind, 2016), this study uses a simulation procedure to explore the degree to which this adaptation of polytomous MSA can be used to detect disordered rating scale categories. The simulation procedure facilitates the systematic exploration of

Figure 1. Conceptual illustration of thresholds.

the sensitivity of ac-MSA to disordered rating scale categories in the presence of varying magnitudes of disordering, proportions of raters with disordering, and sample sizes.

Table 1. Simulation Design Factors.

Design Factor	Conditions
1. Number of rating scale categories	3, 4, 5
2. Number of raters	12, 24
3. Number of students	100, 300, 500
4. Proportion of raters with disordered generating thresholds (based on partial credit model)	0.10, 0.20
5. Logit distance between disordered thresholds	0.4, 1.4, 2.4

Simulation Procedure

Holistic polytomous ratings were generated based on the partial credit model (Masters, 1982) using a Monte Carlo simulation procedure. This model was selected because it is based on adjacent-categories probabilities, which match the probability formulation in ac-MSA models. Accordingly, it is possible to specify disordered rating scale categories for individual raters by specifying disordered generating parameters for the rating scale category thresholds.

To explore the sensitivity of ac-MSA to disordered categories across a variety of conditions, the simulation was based on five design factors: (1) number of rating scale categories (3, 4, or 5 categories), (2) number of raters (12 or 24), (3) number of students (100, 300, 500), (4) proportion of raters with disordered generating thresholds (0.10 or 0.20), and (5) magnitude of disordering on the logit scale (0.4, 1.4, or 2.4 logits) (Table 1). The three magnitudes of disordered thresholds were selected based on Linacre's (2002) recommendation of 1.4 logits as an effective distance between rating scale category thresholds to meaningfully distinguish among categories in terms of the latent variable. For each condition, one threshold was disordered for the specified proportion of raters (0.10 or 0.20), rounded to the nearest integer. One hundred replications were completed for each combination of conditions.

The generating threshold parameters across raters were selected from $\tau \sim (U[-4, 4])$; this range of values reflects the typical range of rater severity in educational performance assessments. Following Linacre (2002), the thresholds for nondisordered raters were spaced 1.4 logits apart. In the disordered conditions, one of the $k-1$ thresholds was disordered using the experimental condition to determine the magnitude. The generating parameters for students were based on $\theta \sim N(0, 1)$.

Data Analysis

A four-step procedure was used to analyze the simulated data. First, ISRFs were calculated for each rater based on adjacent-categories probabilities. Specifically, the data were structured such that each column represented an individual rater, and each row represented an individual student (see Figure 2). Using this structure, raters were treated like polytomous "items" in the ac-MSA analysis. Second, the probabilities

Student	Rater				
	1	2	3	...	<i>J</i>
1	$X_{1,1}$	$X_{1,2}$	$X_{1,3}$	...	$X_{1,J}$
2	$X_{2,1}$	$X_{2,2}$	$X_{2,3}$	...	$X_{2,J}$
3	$X_{3,1}$	$X_{3,2}$	$X_{3,3}$	...	$X_{3,J}$
...	...	...	...	...	...
N	$X_{N,1}$	$X_{N,2}$	$X_{N,3}$	...	$X_{N,J}$

Figure 2. Structure of the data matrix

Note. Each cell entry (X_{nj}) represents the observed rating for student n from rater j .

associated with each rating scale category were used to identify deviations from the expected category ordering based on the ordinal rating scale categories. Specifically, for each pair of adjacent ISRFs within each rater, any instance in which the probability associated with the lower ISRF was higher than the probability associated with the higher ISR was flagged as evidence of category disordering for the rater. Third, the match between the empirical disordering and the disordering specified within the generating parameters was examined by identifying the proportion of replications in which disordering was present for a rater when generating thresholds were disordered for the same rater.

Finally, within-rater category ordering was explored using graphical indices. In addition to statistical summaries of sensitivity to disordered rating scale categories, graphical displays of adjacent-categories ISRFs were considered as they relate to each of the simulation conditions. These displays were examined to explore the degree to which graphical indicators of within-rater rating scale category ordering provide additional insight into rater judgment that can inform the interpretation and use of ratings in educational performance assessments.

Results

Overall Sensitivity to Category Disordering

The sensitivity of ac-MSA to within-rater category disordering is summarized in Tables 2 through 4 for the simulation conditions based on rating scales with three, four, and five categories. Overall, the results suggested that ac-MSA ISRFs are quite sensitive to disordered categories within raters across all simulation conditions. For the three-category conditions (Table 2), complete agreement between the partial credit thresholds and the ac-MSA ISRFs was observed across replications with two minor exceptions. The first deviation from complete agreement (sensitivity = 0.94) was observed for the condition based on 12 raters and 100 students, where 20% of the raters had a disordered generating threshold by 0.40 logits. The second deviation from complete agreement (sensitivity = 0.94) was also observed in conjunction with the 100-student sample size and disordered generating threshold magnitude of 0.40 logits, but within the condition based on 24 raters, where 10% of the raters had a disordered generating threshold.

Table 2. Sensitivity of Adjacent-Categories Mokken Ordering and Partial Credit Model Generating Thresholds for Detecting Disordered Raters (Three Category Conditions).

Rater sample size	Proportion of raters with disordering	Distance between disordered thresholds	Proportion consistency between generating thresholds and Mokken ordering for disordered raters across 100 replications		
			N = 100	N = 300	N = 500
12	0.10	0.40	1.00	1.00	1.00
		1.40	1.00	1.00	1.00
		2.40	1.00	1.00	1.00
	0.20	0.40	0.94	1.00	1.00
		1.40	1.00	1.00	1.00
		2.40	1.00	1.00	1.00
	0.10	0.40	0.99	1.00	1.00
		1.40	1.00	1.00	1.00
		2.40	1.00	1.00	1.00
24	0.20	0.40	1.00	1.00	1.00
		1.40	1.00	1.00	1.00
		2.40	1.00	1.00	1.00
	0.10	0.40	1.00	1.00	1.00
		1.40	1.00	1.00	1.00
		2.40	1.00	1.00	1.00

Similarly, results for the four-category conditions (Table 3) indicate complete agreement across replications, except for the conditions where $N_{\text{students}} = 100$ and the distance between disordered thresholds was 0.40 logits. For these conditions, the sensitivity proportions ranged from 0.88 for the condition based on 12 raters and a proportion of within-rater disordering of 0.10, to 0.96 for the condition based on 24 raters and a proportion of within-rater disordering of 0.10.

Finally, results from the five-category conditions (Table 4) indicate complete agreement across all conditions except for minor discrepancies in the conditions where $N_{\text{students}} = 100$ and the distance between disordered thresholds was 0.40 logits. Within these conditions, the sensitivity of the ac-MSA ISRFs to disordered thresholds ranged from 0.95 for both conditions based on 24 raters to 0.98 for the conditions based on 12 raters.

Graphical Displays

To further explore the use of ac-MSA to evaluate within-rater category ordering, graphical displays of adjacent-categories ISRFs were examined across the simulation conditions. Specifically, graphical displays of ISRFs were created for 20 randomly sampled replications within each simulation condition. Within the randomly selected replications, ISRFs were examined for raters specified as disordered based on generated threshold parameters. These plots were used to consider the degree to which

Table 3. Sensitivity of Adjacent-Categories Mokken Ordering and Partial Credit Model Generating Thresholds for Detecting Disordered Raters (Four Category Conditions).

Rater sample size	Proportion of raters with disordering	Distance between disordered thresholds	Proportion consistency between generating thresholds and Mokken ordering for disordered raters across 100 replications		
			N = 100	N = 300	N = 500
12	0.10	0.40	0.88	1.00	1.00
		1.40	1.00	1.00	1.00
		2.40	1.00	1.00	1.00
	0.20	0.40	0.93	0.99	1.00
		1.40	1.00	1.00	1.00
		2.40	1.00	1.00	1.00
	0.10	0.40	0.96	1.00	1.00
		1.40	1.00	1.00	1.00
		2.40	1.00	1.00	1.00
24	0.20	0.40	0.95	1.00	1.00
		1.40	1.00	1.00	1.00
		2.40	1.00	1.00	1.00

Table 4. Sensitivity of Adjacent-Categories Mokken Ordering and Partial Credit Model Generating Thresholds for Detecting Disordered Raters (Five Category Conditions).

Rater sample size	Proportion of raters with disordering	Distance between disordered thresholds	Proportion consistency between generating thresholds and Mokken ordering for disordered raters across 100 replications		
			N = 100	N = 300	N = 500
12	0.10	0.40	0.98	1.00	1.00
		1.40	1.00	1.00	1.00
		2.40	1.00	1.00	1.00
	0.20	0.40	0.98	1.00	1.00
		1.40	1.00	1.00	1.00
		2.40	1.00	1.00	1.00
	0.10	0.40	0.95	1.00	1.00
		1.40	1.00	1.00	1.00
		2.40	1.00	1.00	1.00
24	0.20	0.40	0.95	1.00	1.00
		1.40	1.00	1.00	1.00
		2.40	1.00	1.00	1.00

graphical evidence of within-rater disordering based on ac-MSA can be used to inform interpretations of rating activity.

Figure 3 includes ISRFs for a subset of raters from the 20 randomly selected replications who were specified to have disordered rating scale categories. For the sake of brevity, the plots in Figure 3 are limited to the simulation conditions based on a three-category rating scale with 24 raters, of whom 20% had disordered generating thresholds; however, the patterns in Figure 3 reflect the general pattern observed across both rater sample sizes as well as the four-category and five-category conditions.

Each plot in Figure 3 includes ISRFs for an individual rater who was specified as disordered in the generating rating scale threshold parameters for the simulation. The x -axis displays student *restscore* groups, which are ordered from low to high. The number of *restscore* groups varies across raters, because *restscores* are calculated specific to each rater to construct ISRFs. The y -axis reflects the adjacent-categories probability for a rating in the higher rating scale category within each pair of adjacent categories ($P[x = k] / P[x = k - 1]$). Because Figure 3 is based on the three-category rating scale conditions, each plot includes two ISRFs, which reflect the probability that students within a *restscore* group receive a rating in category 1 rather than category 0 (solid line with circle plotting symbols) and the probability that students within a *restscore* group receive a rating in Category 2 rather than Category 1 (dashed line with triangle plotting symbols).

It is interesting to note that the rating scale categories were disordered across all *restscore* groups for the conditions based on disordered thresholds with a magnitude of 1.4 and 2.4 logits. In contrast, for the conditions based on disordered thresholds with a magnitude of 0.40 logits, the ISRFs were disordered within a subset of *restscore* groups.

Inspection of the plots in Figure 3 highlights the diagnostic value of ac-MSA ISRFs for exploring rating activity in general as well as in terms of category ordering. Specifically, visual inspection of the category probabilities allows researchers to identify the specific *restscore* groups (if any) in which a rater demonstrated category disordering. Furthermore, these plots also highlight differences in the magnitude of disordering for individual raters across *restscore* groups. For example, although the ISRFs are disordered in the second *restscore* group for the rater in the condition based on 100 students and a disordering magnitude of 0.4 logits, the overall magnitude of this disordering is small. Similarly, although the ISRFs are disordered across all the *restscore* groups for each of the remaining illustrated raters, the plots reveal that the magnitude of disordering generally varies across *restscore* groups. In other words, these plots suggest that the rater's interpretation of the difficulty of rating scale categories varies across achievement levels.

In addition to category ordering, the plots in Figure 3 suggest that, although these raters demonstrated category disordering, the ISRFs were generally nondecreasing across increasing *restscore* groups—suggesting adherence to the ac-MH model. This

finding suggests that rating scale category ordering should be examined in addition to other rating quality indices, including rater monotonicity.

Summary and Conclusions

The purpose of this study was to consider the degree to which ac-MSA is sensitive to within-rater category disordering in the context of rater-mediated performance assessments. Using simulated data, adjacent-categories ISRFs were examined in terms of their sensitivity to disordered generating rating scale category thresholds based on the partial credit model (Masters, 1982). Using simulated data, the sensitivity of ac-MSA to disordered thresholds was evaluated by comparing the proportion of replications in which raters who were specified to have disordered thresholds based on the partial credit model to the proportion of raters who were observed as demonstrating disordered rating scale categories based on ISRFs. Visual displays of ac-MSA ISRFs were also examined to consider the diagnostic value of these graphical indices of category ordering in terms of rating quality. Overall, the results suggested that ISRFs based on ac-MSA were quite sensitive to within-rater category disordering. Furthermore, graphical displays of ISRFs for raters with disordered categories highlighted several characteristics of raters' use of the rating scale categories that provided additional insight into rater judgments that can be used to guide rater training procedures and scoring materials. In this section, the results are summarized as they relate to the two guiding research questions. A discussion of these results, including their implications for research, theory, and practice, follows.

The first research question asked: *How does the sensitivity of ISRFs based on ac-MSA models to disordered categories vary across magnitudes of disordering, the overall proportion of raters with disordered categories, and sample sizes?* The results from this study indicated that ISRFs based on ac-MSA were highly sensitive ($sensitivity \geq 0.99$) to within-rater rating scale category disordering for all conditions in which the magnitude of disordering in the generating thresholds was either 1.40 or 2.40 logits. Slightly lower sensitivity was observed for the conditions based on disordered generating thresholds whose magnitude was 0.40, particularly within the conditions based on $N_{\text{students}} = 100$ ($0.88 \leq sensitivity \leq 0.98$); for the larger student sample sizes, sensitivity for these conditions was comparable to the conditions based on 1.40 or 2.40 logit disordering.

The second research question asked: *How can graphical displays of ISRFs based on ac-MSA be used to explore within-rater category disordering?* In addition to statistical indices of disordering, the results from this study suggested that graphical displays of ISRFs based on ac-MSA also reveal within-rater rating scale category disordering. Furthermore, these graphical displays provide additional insight into rating scale category disordering, including the level or levels of student achievement for which rating scale categories are disordered as well as the magnitude of disordering for levels of student achievement. Thus, these graphical displays can be used to

explore idiosyncratic rating patterns related to category disordering in detail for individual raters.

In addition to the utility of plots of ISRFs for exploring within-rater category ordering based on ac-MSA, it is interesting to consider the diagnostic utility of these plots in contrast with other approaches to evaluating rating scale category ordering. For example, indicators of rating scale category disordering based on polytomous Rasch models include numeric estimates of threshold locations on a linear scale that can be compared to evaluate the degree to which categories are ordered as expected. Furthermore, when the partial credit formulation of the Rasch model is used, threshold location estimates can be calculated for individual raters to evaluate within-rater category ordering. However, these threshold estimates do not provide insight into the specific subgroups of students for which rating scale categories are ordered and disordered. Although graphical displays of rating scale category probabilities based on polytomous Rasch models can be examined across student achievement levels that highlight a variety of issues related to rating scale category functioning, these displays do not necessarily correspond to numeric indices of disordering based on threshold estimates (Andrich, 2015).

Discussion

This study has several implications for research and practice related to rater-mediated assessments. In terms of research, this study continues the exploration of ac-MSA as a nonparametric tool for exploring rater activity and rating quality. This simulation study offered additional insight into the properties of these models and corroborates the initial findings based on real data (Wind, 2016). In particular, the simulation approach included in the current study reflects a broader range of conditions than the characteristics of the data used in the original presentation of ac-MSA. Furthermore, this study includes the first simulation analysis related to Mokken scaling in which disordered categories have been included as a simulation condition. The findings from the current study suggest that ac-MSA provides an additional technique for exploring category disordering in polytomous IRT models in general. Whereas previous research on category disordering has been limited to parametric IRT models, the ac-MSA approach offers a nonparametric perspective on category ordering that contributes to the ongoing discussions related to this topic (Adams et al., 2012; Andrich, 2013, 2015; Wetzel & Carstensen, 2014).

In terms of practice, the finding that ac-MSA ISRFs are sensitive to disordered categories has implications for rater training and monitoring procedures for rater-mediated performance assessments. Ac-MSA offers a diagnostic tool for evaluating rating quality that provides insight into rating scale category use at the individual rater level within a nonparametric framework. Accordingly, this approach could be applied in practical performance assessment settings as a method for empirically investigating raters' interpretation of rating scale categories in terms of ordering. Unless data imputation techniques are applied (Wind & Patil, 2018), ac-MSA

requires complete ratings (i.e., no missing data). Thus, the application of this approach to practical settings in educational assessment is most promising in settings based on small samples for which complete ratings are obtained, such as during rater training or in subsets of performances used to monitor raters in an ongoing fashion.

Limitations

Several limitations are important to note. First, although the simulation design was intended to reflect a wide range of performance assessment settings, operational performance assessments may differ from the characteristics of the simulated data in important ways that limit the generalizability of the current findings to these assessments. For example, the sample sizes that were included in the simulation design were relatively small. These samples were intended to reflect assessment contexts in which ac-MSA would most likely be appropriate, such as rater training sessions prior to operational scoring. Additional research is needed that considers the sensitivity of ac-MSA to within-rater category disordering when larger sample sizes of students and raters are involved. Recognizing the practical constraints associated with obtaining complete ratings in the presence of large-scale performance assessments, this research should also consider the implications of missing data and data imputation techniques (e.g., Wind & Patil, 2018) for identifying within-rater category ordering from the perspective of ac-MSA.

Finally, it is important to note that the analyses in this study focused primarily on exploring the sensitivity of ac-MSA to within-rater category disordering. Additional research is needed that includes systematic investigations of the degree to which ac-MSA can also be used to correctly identify non-disordered rating scale categories (i.e., the specificity of the approach). This research will shed light on the overall alignment between rating scale category disordering across parametric IRT models and ac-MSA.

Authors' Note

A previous version of this article was presented at the annual meeting of the National Council on Measurement in Education in San Antonio, Texas in April 2017.

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

- Adams, R. J., Wu, M. L., & Wilson, M. (2012). The Rasch rating model and the disordered threshold controversy. *Educational and Psychological Measurement*, 72, 547-573. <https://doi.org/10.1177/0013164411432166>
- Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). Hoboken, NJ: Wiley-Interscience.
- Andrich, D. A. (2011). Rating scales and Rasch measurement. *Expert Review of Pharmacoeconomics & Outcomes Research*, 11, 571-585. doi:10.1586/erp.11.59
- Andrich, D. A. (2013). An expanded derivation of the threshold structure of the polytomous Rasch Model that dispels any "threshold disorder controversy." *Educational and Psychological Measurement*, 73, 78-124. doi:10.1177/0013164412450877
- Andrich, D. A. (2015). The problem with the step metaphor for polytomous models for ordinal assessments. *Educational Measurement: Issues and Practice*, 34(2), 8-14. doi:10.1111/emip.12074
- Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal & G. Haladyna (Eds.), *Large-scale assessment programs for ALL students: Development, implementation, and analysis* (pp. 261-287). Mahwah, NJ: Lawrence Erlbaum.
- Engelhard, G., & Wind, S. A. (2013). *Rating quality studies using Rasch measurement theory* (Research Report No. 2013-3). New York, NY: College Board.
- Hambleton, R. K., van der Linden, W. J., & Wells, C. S. (2010). IRT models for the analysis of polytomously scored data: Brief and selected history of model building advances. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 21-42). New York, NY: Routledge.
- Johnson, R. L., Penny, J. A., & Gordon, B. (2009). *Assessing performance: Designing, scoring, and validating performance tasks*. New York, NY: Guilford Press.
- Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, 3, 85-106.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174. doi:10.1007/BF02296272
- Mellenbergh, G. J. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement*, 19, 91-100. doi:10.1177/014662169501900110
- Mokken, R. J. (1971). *A theory and procedure of scale analysis*. The Hague, the Netherlands: De Gruyter.
- Molenaar, I. W. (1982). Mokken scaling revisited. *Kwantitative Methoden*, 3, 145-164.
- Molenaar, I. W. (1997). Nonparametric models for polytomous responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 369-380). New York, NY: Springer.
- Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, 4, 386-422.
- Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, 5, 189-227.
- Nering, M. L., & Ostini, R. (Eds.). (2010). *Handbook of polytomous item response theory models*. New York, NY: Routledge.
- Penfield, R. D. (2014). An NCME instructional module on polytomous item response theory models. *Educational Measurement: Issues and Practice*, 33(1), 36-48. doi:10.1111/emip.12023

- Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests* (Expanded edition, 1980. Chicago, IL: University of Chicago Press). Copenhagen, Denmark: Danish Institute for Educational Research.
- Wetzel, E., & Carstensen, C. H. (2014). Reversed thresholds in partial credit models: A reason for collapsing categories? *Assessment*, 21, 765-774. doi:10.1177/1073191114530775
- 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. (2016). Adjacent-categories Mokken models for rater-mediated assessments. *Educational and Psychological Measurement*, 77, 330-350. doi:10.1177/0013164416643826
- Wind, S. A., & Engelhard, G. (2015). Exploring rating quality in rater-mediated assessments using Mokken scale analysis. *Educational and Psychological Measurement*, 76, 685-706. doi:10.1177/0013164415604704
- Wind, S. A., & Patil, Y. J. (2018). Exploring incomplete rating designs with Mokken scale analysis. *Educational and Psychological Measurement*, 78, 319-342. doi:10.1177/0013164416675393
- Wind, S. A., & Peterson, M. E. (2018). A systematic review of methods for evaluating rating quality in language assessment. *Language Testing*, 35, 161-192. doi:10.1177/0265532216686999
- 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