View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by University of Groningen

University of Groningen

On the Practical Consequences of Misfit in Mokken Scaling

Crisan, Daniela; Tendeiro, Jorge; Meijer, Rob

Published in:
PsyArXiv Preprints

DOI:
10.31234/osf.io/kgc3a

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Early version, also known as pre-print

Publication date:
2020

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):
Crisan, D., Tendeiro, J., & Meijer, R. (Accepted/In press). On the Practical Consequences of Misfit in Mokken Scaling. *PsyArXiv Preprints*. https://doi.org/10.31234/osf.io/kgc3a

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 21-02-2020

On the Practical Consequences of Misfit in Mokken Scaling

Abstract: Mokken scale analysis is a popular method to evaluate the psychometric quality of clinical and personality questionnaires and their individual items. Although many empirical papers report on the extent to which sets of items form Mokken scales, there is less attention for the effect of violations of commonly used rules of thumb. In this study we investigated the practical consequences of retaining or removing items with psychometric properties that do not comply with these rules-of-thumb. Using simulated data, we concluded that items with low scalability had some influence on the reliability of test scores, person ordering and selection, and criterion-related validity estimates. Removing the misfitting items from the scale had, in general, a small effect on the outcomes. Although important outcome variables were fairly robust against scale violations in some conditions, we conclude that researchers should not rely exclusively on algorithms allowing automatic selection of items. In particular, content validity must be taken into account in order to build sensible psychometric instruments.

Keywords: Mokken scale analysis, scale analysis, item response theory, test construction, content validity.

On the Practical Consequences of Misfit in Mokken Scaling

Item response theory (IRT) models are used to evaluate and construct tests and questionnaires, such as, for example, clinical- and personality scales (e.g., Thomas, 2011). A popular IRT approach is Mokken scale analysis (MSA; e.g., Mokken, 1971; Sijtsma & Molenaar, 2002). MSA has been applied in various fields where multi-item scales are used to assess the standing of subjects on a particular characteristic or the latent trait of interest. In recent years, the popularity of MSA has increased. A simple search on Google scholar with the keywords “Mokken Scale Analysis AND scalability” from 2000 through 2019 yielded about 1200 results, including a large set of empirical studies. These studies were conducted in various domains, such as in personality (e.g., Watson, Deary, & Austin, 2007), clinical psychology and health (e.g., Emons, Sijtsma, & Pedersen, 2012), education (e.g., Wind, 2016), and in human resources and marketing (e.g., De Vries, Michielsen, & Van Heck, 2003). Both the useful psychometric properties of MSA and the availability of easy-to-use software (e.g., the R ‘mokken’ package; van der Ark, 2012) explain the popularity of MSA. As we discuss below, within the framework of Mokken scale analysis, there are several procedures that can be used to evaluate the quality of an existing scale or set of items that may form a scale. In practice, however, a set of items may not comply strictly with the assumptions of a Mokken scale and a researcher is then faced with a difficult decision: Include or exclude the offending items (Molenaar, 1997)? The answer to this question is not straightforward. On the one hand, the exclusion of items must be carefully considered because it may compromise construct validity (see the Standards for Educational and Psychological Testing, 2014, for a discussion of the types of validity evidence). On the other hand, it is not well known to what extent the retention of items that violate the premises of a Mokken scale affect important quality criteria.

The present study is aimed at investigating the effects of retaining or removing items that violate common premises in MSA, on several important outcome variables. Our paper therefore offers novel insights over scale construction for practitioners applying MSA, going over and beyond what MSA typically offers. This study is organized as follows. First, we provide some background on Mokken scale analysis. Second, we present the results of a simulation study in which we investigated the effect of model violations on several important outcome variables. Finally, in the discussion section we provide an evaluative and integrated overview of the findings and we discuss main conclusions and limitations.

Mokken Scale Analysis

For analyzing test and questionnaire data, MSA provides much more analytical tools than classical test theory (CTT; Lord & Novick, 1968), while avoiding the statistical complexities of parametric IRT models. One of the most important MSA models is the monotone homogeneity model (MHM). The MHM is based on three assumptions: (a) Unidimensionality: All items predominantly measure a single common latent trait, denoted θ; (b) Monotonicity: The relationship between θ and the probability of scoring in a certain response category or higher is monotonically nondecreasing, and (c) Local independence: An individual's response to an item is not influenced by his/her responses to other items in the same scale. Assumptions (a) through (c) allow the stochastic ordering of persons on the latent trait continuum by means of the sum score, when scales consist of dichotomous items (e.g., Sijtsma & Molenaar, 2002, p. 22). For a discussion on how this property applies to polytomous items, see Hemker, Sijtsma, Molenaar, & Junker (1997) and van der Ark (2005).

In MSA, Loevinger's *H* coefficient (or the scalability coefficient; Mokken, 1971, p. 148-153; Sijtsma & Molenaar, 2002, chap. 4) is a popular measure to evaluate the quality of each item *i* and of sets of items, in relation to the test score distribution. The *H* coefficient can be obtained for pairs of items (H_{ij}), for individual items (H_i), and for the entire scale (H). The H_i is defined as following for dichotomous items (Sijtsma & Molenaar, 2002, pp. 55-58):

$$H_i = \frac{Cov(X_i, R_{-i})}{Cov_{max}(X_i, R_{-i})} = 1 - \frac{\sum_{j \neq i}(P_i - P_{ij})}{\sum_{j>i} P_i \times (1 - P_j) + \sum_{j<i} P_j \times (1 - P_i)}$$

In this formula, X_i denotes individuals' responses to item *i*. P_i and P_j denote the probability of a correct response to - or endorsing - items *i* and *j*, P_{ij} denotes the probability of correct response to- or endorsing both items *i* and *j*, and R_{-i} denotes the vector of restscores (that is, the individuals' sum scores excluding item *i*). The item-pair and scale coefficients can be easily derived from H_i, by removing the summation symbols (for H_{ij}) or adding an additional one (for H) from/to all the terms in the equation above. For polytomous items, the scalability coefficients are based on the same principles as for dichotomous items, but their formulas are more complex, as probabilities are defined at the levels of item steps (Molenaar, 1991; Sijtsma & Molenaar, 2002, p. 123; see also Crisan, van de Pol, & van der Ark, 2016 for a comprehensive explanation of how these can be obtained).

Loevinger's H coefficient reflects the accuracy of ordering persons on the θ scale using the sum score as a proxy. If the MHM holds, then the population H values for all item pairs, items, and the entire scale are between 0 and 1 (Sijtsma & Molenaar, 2002, Theorem 4.3). Larger H coefficients are indicative of better quality of the scale ("stronger scales"), whereas values closer to 0 are associated with "weaker scales". A so-called Mokken scale is a unidimensional scale comprised of a set of items with 'large-enough' scalability coefficients, which indicate that the scale is useful for discriminating persons using the sum scores as proxies for their latent θ values. There are some often-used rules of thumb that provide the basis for MSA (Mokken, 1971, p. 185). A Mokken scale is considered a weak scale when $.3 \leq H < .4$, a medium scale when $.4 \leq H < .5$, and a strong scale when $H \geq .5$ (Mokken, 1971; Sijtsma & Molenaar, 2002). A set of items for which $H < .3$ is considered unscalable. Using .3 as a lower bound value for H_i and H is the default option in various software packages, including the R ‘mokken’ package (van der Ark, 2012) and MSP5 (Molenaar & Sijtsma, 2000).

A popular feature of MSA is its item selection tool, known as the automated item selection procedure, AISP (Sijtsma & Molenaar, 2002, chaps. 4 and 5). The AISP assigns items into one or more Mokken (sub-)scales according to some well-defined criteria (see e.g., Meijer, Sijtsma, & Smid, 1990), and identifies items that cannot be assigned to any of the selected Mokken scales (i.e., unscalable items). The unscalable items may not discriminate well between persons and, depending on the researcher’s choice, may be removed from the final scale.

Both the AISP selection tool and the item quality check tool are based on the scalability coefficients. However, it is important to note that a suitable lower bound for the scalability coefficients should ultimately be determined by the user (Mokken, 1971), taking the specific characteristics of the data and the context into account. Although several authors emphasized the importance of not blindly using the rules of thumb (e.g., Rosnow and Rosenthal, 1989, p. 1277, for a general discussion outside Mokken scale analysis), many researchers use the default lower bound offered by existing software when evaluating or constructing scales.

How is Mokken Scale Analysis Used in Practice?

Broadly speaking, there are two types of MSA research approaches: In one approach, MSA is used to evaluate the item- and scale quality when constructing a questionnaire or test (e.g., Ettema, Dröes, De Lange, Mellenberg, & Ribbe., 2007; De Boer, Timmerman, Pijl, & Minnaert, 2012). In the other approach, MSA is used to evaluate an existing instrument (e.g., Bech, Carrozzino, Austin, Møller, & Vassend, 2016; Bielderman et al., 2013; Bouman et al., 2011). Not surprisingly, researchers using MSA in the construction phase tend to remove items more often based on low scalability coefficients and/or the AISP results (e.g., Brenner et al., 2007; De Boer et al., 2012; De Vries et al., 2003) than researchers who evaluate existing instruments. However, researchers seldom use sound theoretical, content, or other psychometric arguments to remove items from a scale.

Researchers evaluating existing scales often simply report that items have low coefficients, but they are typically not in a position to remove items (e.g., Bech et al., 2016; Bielderman et al., 2013; Bouman et al., 2011; Cacciola, Alterman, Habing, & McLellan, 2011, p.12; Emons et al., 2012, p. 349; Ettema et al., 2007). Thus, practical constraints often predetermine researchers' actions, but it is unclear to what extent other variables, such as predictive or criterion validity (Standards for Educational and Psychological Testing, 2014), are affected by the inclusion of items with low scalability. What is, for example, the effect on the predictive validity of the sum scores obtained from a more homogenous scale as compared to a scale that includes lower scalability items? For some general remarks about the relation between homogeneity and predictive validity, and about one of the drawbacks of relying on the *H* coefficient, see the online supplementary materials.

***Practical* Significance**

In this study, we extend the existing literature on the practical use of MSA (see Sijtsma & van der Ark, 2017 and Wind, 2017, for excellent tutorials for practitioners in the fields of psychology and education) by systematically investigating how *practical* outcomes, such as scale reliability and person rank ordering were affected by scores obtained from scales containing items with low scalability coefficients. This study also extends previous literature on the *practical* significance (Sinharay & Haberman, 2014) of the misfit of IRT models (e.g., Crișan, Tendeiro, & Meijer, 2017) by focusing on nonparametric IRT models.

In the remaining of this paper we describe the methodology we used to answer our research questions, we present the findings of our study, and we follow up with some insights for practitioners and researchers regarding scale construction and/or revision.

Method

We conducted a simulation study using the following independent and dependent variables.

Independent Variables

We manipulated the following four factors:

Scale length. We simulated scales consisting of I = 10 and 20 items. These numbers of items are representative for scales often found in practice (e.g., Rupp, 2013, pp. 22-24).

Proportion of items with low H_i values. In the existing literature using simulation studies, the number of misfitting items can vary between 8% and 75% or even 100% (see Rupp, 2013, for a discussion). In the present study, three levels for the proportion of items with $H_i < .30$ were considered: I_{LowH} =.10, .25, and .50. These levels of I_{LowH} operationalized varying proportions of misfitting items in the scale, which we label here as 'small', 'medium', and 'large' proportions, respectively.

Number of response categories. We simulated responses to both dichotomously and polytomously scored items with the number of categories equal to C = 2, 3, and 5. Each data set in a condition was based on one C value only.

Range of H_i values. For the I_{LowH} items, two ranges of item scalability coefficients H_i were considered: R_H = [.1, .2) and [.2, .3). Hemker, Sijtsma, and Molenaar (1995) and Sijtsma and van der Ark (2017) suggested using multiple lower bounds for the H coefficients within the same analysis. They suggested using 12 different lower bounds, ranging from .05 through .55 in steps of .05. However, in order to facilitate the interpretation and to avoid a very large design, we chose the two ranges of item scalability coefficients mentioned above. For all fitting items we set $.3 \leq H_i \leq .7$. We set the upper bound to .7 instead of 1 because few operational scales have H_i values larger than .7.

Design

The simulation was based on a fully crossed design consisting of 2(I) × 3(I_{LowH}) × 3(C) × 2(R_H) = 36 conditions, with 100 replications per condition.

Data Generation

We generated population item response functions according to two parametric item response theory models: The 2-parameter logistic model (2PLM; e.g., Embretson & Reise, 2000) in the case of dichotomous items, and the graded response model (GRM; Samejima, 1969), in the case of polytomous items. The 2PLM is defined as follows:

$$P(X_i=1|\theta)=\frac{e^{\alpha_i(\theta\text{-}\beta_i)}}{1+e^{\alpha_i(\theta\text{-}\beta_i)}},$$

where X_i denotes the response to item i (coded 0 and 1), a_i denotes the discrimination of item i, β_i denotes the difficulty of item i, and θ denotes the person's level on the latent characteristic (or trait) continuum. Thus, the 2PLM defines the conditional probability of scoring a 1 (typically representing the 'correct' answer) on item i as a function of item and person characteristics. The GRM is a generalization of the 2PLM in case of polytomous items, and is defined as follows:

$$P_{ix}^{*}=\frac{e^{\alpha_i(\theta\text{-}\beta_{ix})}}{1+e^{\alpha_i(\theta\text{-}\beta_{ix})}},$$

where P_{ix}^{*} = $P(X_i \geq x \mid \theta)$, x = 1, …, C, denotes the probability of endorsing at least category x on item i, and β_{ix} denotes the category threshold parameters. By definition, the probability of endorsing the lowest category (x = 0) or higher is 1 and the probability of endorsing category C + 1 or higher is 0. Thus, the GRM defines the probability of scoring at response category x or higher on item i as a function of item and person characteristics. The probability of endorsing response option x is computed as $P(X_i = x \mid \theta) = P_{ix}^{*}$ - $P_{i(x+1)}^{*}$.

The 2PLM or the GRM was used to generate item scores, using discrimination parameters[1] that were constrained to optimize the chances of generating items with H_i in the suitable ranges as required by R_H; see Table 1 for the values used for the true discrimination parameters during the data generation. These values were found after preliminary trial-and-error calibration analyses.

[Insert Table 1 about here]

In Table 1, the column labeled “Misfitting items” denotes the (100 × I_{LowH})% of items with scalability coefficients within the ranges R_H = [.2, .3) and [.1, .2). The column labeled “Fitting items” concerned the remaining items with scalability coefficients in the range [.3, .7]. In all cases, the difficulty/threshold parameters were randomly drawn from the uniform distribution, ensuring that consecutive threshold parameters differed by at least 0.3 units on the latent scale (the GRM requires that the threshold parameters are ordered) and that the items were randomly centered around 0 (thus allowing to generate ‘easy’ and ‘difficult’ items equally likely). This procedure resulted in threshold parameters ranging between approximately -3 and 3. The true θs were randomly drawn from the standard normal distribution. The item parameters together with the θ values defined the item response functions according to the 2PLM/GRM, which represent probabilities of responding in a particular response category. These probabilities were then used to compute the scalability coefficients H_{ij}, H_i, and H (Molenaar, 1991, 1997; see also Crisan et al., 2016). The procedure was repeated for each replication within each simulation condition, until a set of items with (100 × I_{LowH})% of items having scalability coefficients within the range given by R_H was generated.

Finally, for these generated items, item scores for N = 2,000[2] simulees were drawn from multinomial distributions with probabilities given by the 2PLM or the GRM. The

[1] They reflect the strength of the relationship between items and θ, and are in general positively related to H_i.

[2] For part of the design, we ran the simulation with N = 100,000 and we found that this did not affect the results. Hence, N = 2,000 is sufficiently large to yield stable results. The code is available at https://osf.io/vs6f9/.

resulting datasets constituted the *Misfitting datasets*. Subsequently, from each misfitting dataset, we removed the (100 × I_{LowH})% of items with $H_i < .3$, resulting in the *Reduced datasets*. We then computed our dependent variables (listed below) on both the *Misfitting* and the *Reduced* datasets, and we investigated the effect of *DataSet* = "Misfitting", "Reduced" on each outcome.

Dependent Variables

We used the following outcome variables:

1. Scale reliability. Scale reliability was determined as the ratio of true scale score variance to observed scale score variance: $\mathrm{r_{XX}}=\frac{\sigma^2_{\mathrm{True}}}{\sigma^2_{\mathrm{Observed}}}$. The observed scale scores were the sum scores across all items, for the entire sample. The true scale scores were computed as the sum of the expected item scores:

$$\text{True scale score}=\sum_{i=1}^{I}\sum_{k=0}^{C-1} k\times P(X_i=k|\theta)\,.$$

2. *Rank ordering*. We computed Spearman rank correlations between the true and the observed scale scores. The goal was to investigate the differences in the rank ordering of simulees across the simulated conditions. Spearman rank correlations were always computed on the entire sample of simulees.

3. *The Jaccard Index*. We used the Jaccard index (Jaccard, 1912) to compare subsamples of top selected simulees, according to their ordering based on either true scores or observed scores. We focused on subsamples of the highest scoring simulees to mimic decisions based on real selection contexts (e.g., for a job, educational program, or clinical treatment). Four selection ratios were considered: *SR* = 1.0, .80, .50, and .30, thus ranging from high through low selection ratios. The Jaccard index is a measure of overlap between two sets, and is defined as follows:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

The index ranges from 0% (no top selected simulees in common) through 100% (perfect congruence). For each data set we therefore computed four values of the Jaccard index, one for each selection ratio.

4. *Bias in criterion-related validity estimates*. For each dataset, four criterion variables were randomly generated such that they correlated with the true θs at predefined levels (r = .15, .25, .35, and .45; e.g., Dalal & Carter, 2015). The bias in criterion-related validity for each criterion variable was computed as follows:

$$\text{bias} = r(\text{observed scale score, criterion}) - r(\text{true scale score, criterion}).$$

The method was applied to the entire sample (*SR* = 1.0) as well as to the top selected simulees (*SR* = .80, .50, and .30). The goal was to assess the effect of low scalability items on the criterion validity, both for the entire sample and in the subsamples of the top selected candidates. Zero bias indicated that observed scores are as valid as true scores, whereas positive/negative bias indicated that observed scores overpredict/underpredict later outcome variables (in terms of predictive validity, for example).

Implementation

We implemented the simulation in R (R Development Core Team, 2019). All code is freely available at the Open Science Framework (https://osf.io/vs6f9/)

Results

To investigate the effects of the manipulated variables on the outcomes, we fitted mixed-effects analysis of variance (ANOVA) models to the data, with *DataSet* as a within-subjects factor and the remaining variables as between-subjects factors. In order to ease the interpretation of the results, we plotted most results and we used measures of effect size (η^2 and Cohen's *d*) to determine the strength and practical importance of the effects. Test statistics and their associated *p*-values were not reported in this paper for two reasons. First, the focus of this study is not on *statistical* significance of misfit. Second, due to the very large sample sizes, even small size effects can be statistically significant, which is of little interest. Additionally, we did not report or interpret negligible effects in terms of effect size for parsimony (i.e., $\eta^2 < .01$; Cohen, 1992).

Scale Reliability and Rank Ordering

For score reliability, we obtained an average of 0.87 (*SD* = 0.07). 95% of the estimates of reliability were distributed between 0.71 and 0.96. The ANOVA model with all main effects and two-way interactions explained 91% of the variation in reliability scores. Variation was partly explained by the two-way interactions between I_{LowH} × *DataSet* ($\eta^2 = .02$), and $I \times I_{LowH}$ ($\eta^2 = .02$), and largely explained by the main effects of I ($\eta^2 = .36$), I_{LowH} ($\eta^2 = .26$), C ($\eta^2 = .11$), and *DataSet* ($\eta^2 = .10$). As such, score reliability decreased as I_{LowH} increased, and this effect was stronger for shorter scales of $I = 10$. Removing the misfitting items from the scale led to an increase in score reliability, and this difference in reliability between the data sets increased slightly with I_{LowH} (see Figure 1 for an illustration of these effects).

[Insert Figure 1 About Here]

Elaborating on the effects of I_{LowH} and of removing the misfitting items on score reliability, we found the following: Averaged over I and C, score reliability decreased with .10 (from .91 to .81) in the *DataSet* = "Misfitting" as I_{LowH} increased from 10% to 50%; Removing the misfitting items improved reliability with .02 for I_{LowH} = 10%, .04 for I_{LowH} = 25%, and .06 for I_{LowH} = 50%. For these differences we obtained Cohen's *d* values of 1.70, 1.73, and 1.78 (for I_{LowH} =10%, 25%, and 50% respectively).

Similar conclusions can be drawn for the rank ordering of persons. The average rank correlation over all conditions was 0.93 (SD = 0.04). 95% of the estimated rank correlation coefficients ranged between 0.83 and 0.98. The ANOVA model with all main effects and two-way interaction effects explained 89% of the variability in the Spearman rank correlation values. The findings for person rank ordering were very similar to what we have found for scale reliability. In terms of the values of the Spearman correlation coefficient, as I_{LowH} increased in the *DataSet* = “Misfitting” conditions from 10% to 50%, they decreased, on average, from .95 to .93 and .90 respectively, averaged over *I* and *C*. Removing the misfitting items lead to an improvement in the rank correlation of 0.02, on average. The rank ordering of individuals as determined by their true score was preserved by the observed score, even when 25 – 50 percent of items in a scale had scalability coefficients below .3. Removing those items lead to a small increase in Spearman’s rank correlation.

Regarding score reliability and person rank ordering, our findings show that scale length together with the proportion of MSA-violating items and number of response categories were the main factors affecting these outcomes: Score reliability and rank ordering were negatively affected by the proportion of items violating the Mokken scale quality criteria, especially when shorter scales were used. These outcomes were more robust against violations when longer scales were used. Removing the misfitting items improved scale reliability and person rank ordering to some extent.

Person Classification

Because large rank correlations do not necessarily imply high agreement regarding sets of selected simulees (Bland & Altman, 1986), we also computed the Jaccard index across conditions. For *SR* = 1 the Jaccard index is always 1 (100% overlap), since all simulees in the sample are selected. Figure 2 shows the effect of the manipulated variables on the agreement between sets of selected simulees, for *C* = 2. The effects for the remaining values of *C* were similar and are therefore not shown here.

[Insert Figure 2 About Here]

The degree of overlap between sets of selected simulees was 80.9% averaged over all conditions, with a standard deviation of 0.09. 95% of the values of the Jaccard index were distributed between 0.61 (about 61% overlap) and 0.94 (about 94% overlap). The ANOVA model with all main effects and two-way interactions accounted for 92.7% of the variation in the Jaccard index. The variation was, to a large extent, accounted for by *SR* (η^2 = .66), *I* (η^2 = .10) and I_{LowH} (η^2 = .07), and to some extent by *C* (η^2 = .04), *DataSet* (η^2 = .02), and the interaction between *I* and *SR* (η^2 = .01). All other effects were negligible (η^2 < .01). As such, the overlap between sets of selected simulees increased as scale length and number of response options increased, it decreased as selection rate decreased, and it decreased as the proportion of items with H_i < 0.3 increased. Removing the misfitting items from the scale had a positive effect on the overlap between sets.

Elaborating on the previous findings and focusing on the effects of selection ratio, scale length, proportion of items with H_i < 0.3, and removing the misfitting items, we conclude that the Jaccard index decreased from 0.91, on average, in the conditions with *SR* = .80, to 0.73 in the conditions with *SR* = .30 (Cohen's *d* for this difference was 3.33). Moreover, the Jaccard index value increased from 0.78, on average, when *I* = 10 to 0.84 when *I* = 20 (Cohen's *d* for this difference was 0.68). The Jaccard index decreased, on average, from 0.83 in the conditions where 10% of items had H_i < .3, to 0.76 in the conditions where 50% of items had H_i < .3 (Cohen's *d* = 0.78). Removing the misfitting items resulted in an increase of the Jaccard index to 0.85 (I_{LowH} = 10%; Cohen's *d* = 0.97) and 0.80 (I_{LowH} = 50%; Cohen's *d* = 1.13).Thus, we conclude that person selection is only marginally affected by the proportion of unscalable items or the extent to which the scalability coefficients are deviating from the 0.3 threshold.

Bias in Criterion-Related Validity Estimates

Our results indicated that the bias in criterion validity estimates varied, on average, between -0.05 (*SD* = 0.03; true criterion validity of 0.45) and -0.02 (*SD* = 0.02; true validity of 0.15). The ANOVA model with all main effects and two-way interactions explained between 12.1% and 57.1% of the variance in bias, as true criterion validity increased. Thus, all effects became stronger as true validity increased. The largest effects corresponded to *SR* (η^2 between .04 and .20 across true validity scores), *I* (η^2 between .03 and .15), I_{LowH} (η^2 between .02 and .09), and *C* (η^2 between .01 and .06). There was also an effect of *DataSet* (η^2 between .01 and .03). More specifically, the absolute bias in criterion-related validity estimates increased as *SR* and *I* decreased, as I_{LowH} increased from 10% to 50%, and as *C* decreased. Removing the misfitting items from the scale lead to a very slight reduction in bias. Figure 3 depicts these effects, shown for a validity coefficient of 0.45 and scales consisting of dichotomous items. We further discuss the effects of *SR*, *C*, I_{LowH}, and *DataSet* for the scale characteristics depicted in Figure 3.

[Insert Figure 3 About Here]

Bias in validity estimates was larger in the top 30% subsample (median of -0.09) compared to the full sample (median of -0.05). Cohen's *d* for this difference was 1.5. In terms of the correlation between predictor and criterion, the absolute difference between the full sample and *SR* = .30 was 0.05, on average. In other words, in the full sample the average estimated validity coefficient was 0.41, while in the *SR* = .30 condition it was 0.36. For scales with 10 dichotomous items the average absolute bias in validity estimates was 0.07, and for scales with 20 items it was 0.04.

Furthermore, the results showed that criterion-related validity was also affected by the proportion the misfitting items. For example, when we wanted to predict the scores on a criterion variable of the top 30% of the simulees using a short scale (top left panel of Figure 3), the difference in bias between I_{LowH} = 10% and I_{LowH} = 50% was 0.03, with Cohen's *d* = 0.67. Thus, a short scale of 10 dichotomous items of which 5 items violated the MSA quality criteria yielded an average criterion validity coefficient of .34. Removing the 50% misfitting items from the scale yielded on average a criterion validity coefficient of .35.

Discussion

In this study, we evaluated the effects of keeping or removing items that are often considered ‘unscalable’ in many empirical MSA studies. Many empirical studies using Mokken scaling either remove items with H_i values smaller than .3 or try to explain why these items should be kept in the scale in spite of them violating this condition. By means of a simulation study, we systematically investigated whether scale reliability, person rank ordering, criterion-related validity estimates, and person classifications were affected by varying levels of incidence of misfitting items (in the MSA sense). Our main results showed that all the outcomes considered were affected, to varying degrees, by some of the manipulated factors (scale length, number of response categories, and proportion of items with low scalability). Removing the misfitting items from the scales had a positive effect on the outcome measures.

Scale score reliability, person rank ordering, and bias of criterion-related validity estimates were most affected by the proportion of items with low scalability. We found a decrease of about .10 in reliability and of about .05 in the Spearman correlation as the proportion of misfitting items increased from 10% to 50%. Removing the misfitting items from the scales led to a slight improvement in reliability and rank correlation (with .04 and .02, respectively). Furthermore, short scales with many misfitting items resulted in an underestimation of the true validity of .11, when predicting the scores on a criterion variable of the top 30% simulees. Removing the misfitting items reduced the bias by .01. Finally, the overlap between sets of selected simulees also decreased with .07, on average, as the proportion of misfitting items increased, and removing the misfitting items improved the overlap with .03. Interestingly, the effect of the range of item scalability coefficients had a negligible effect on the outcomes we studied.

In line with previous findings, scale length, number of response categories, and selection rates also had an effect on the outcome variables (e.g., Crișan et al., 2017; Zijlmans, Tijmstra, van der Ark, & Sijtsma, 2018). The item scalability coefficient is equivalent to a normed item-rest correlation, which, in turn, is used as an index of item-score reliability (e.g., Zijlmans et al., 2018). Therefore, it is not surprising that overall scale reliability decreased as the item scalability coefficients decreased. Moreover, it is well-known that there is a positive relationship between scale length and reliability. This also partly explains our findings regarding the exclusion of misfitting items: Removing the misfitting items from the scales resulted in shorter scales, which had a negative impact on reliability.

Take home message

The take-home message from this study is that, depending on the characteristics of a scale (in terms of length and number of response categories), on the specific use of the scale (e.g., to select a proportion of individuals from the total sample), and on the strength of the relationship between the scale scores and some criterion, the consequences of keeping items that violate the rules-of-thumb often used in MSA item selection can vary in their magnitude. We tentatively conclude the following:

1. The number of items with $H_i < .3$ in a scale has a negative effect on scale reliability, person rank ordering and classification, and on predictive accuracy. The magnitude of this effect varies in terms of variance accounted for, depending on the characteristics and specific uses of the test/scale. In general, (relatively) long scales with several response categories are fairly robust against these violations, especially when they have modest criterion-related validity and they are used with selection ratios above .50.

2. Removing misfitting items from the scale improves practical outcome measures, but the effect is moderate at best. Based on these and previous findings, we do not recommend removing the misfitting items from the scales when there are no other (content) arguments to do so. The relatively small gains in reliability, person selection results, and predictive validity might not outweigh the loss in construct coverage and criterion validity.

3. The distance between the *H* values of the violating items and the .3 threshold had a negligible effect on practical outcomes. So, our results indicate that researchers should not overinterpret H_i differences between .1 and .3

On the one hand these findings are reassuring because, as we discussed above, researchers are often not in a position to simply remove items from a scale (see also Molenaar, 1997). It also discharges the researcher from trying to find opportunistic arguments for keeping an item in the scale with, say, a relatively low *H* value. On the other hand, this is certainly not a plea for lazy test construction. Ideally, when conducting MSA either on existing operational measures or in the scale construction phase, the decision whether to keep or remove items from a scale should be based primarily on *theoretical* considerations and applied researchers should be careful not to use psychometric rules-of-thumb to blindly remove items. In particular, one should not feel obliged to strictly adhere to the discrete qualitative labels of *H* (“weak”, “medium”, and “strong” scale); paraphrasing Rosnow and Rosenthal (1989, p. 1277): “surely, God loves the .29 nearly as much as the .31”. In line with these observations, Sijtsma and van der Ark (2017) recommended that several MSAs should be ran on the data using varying lower bounds for the item scalability coefficients, and the final scale should be chosen such that it satisfies both psychometric *and* theoretical considerations.

On a more general note, one should keep in mind that items can exhibit other kinds of misfit apart from low scalability, such as violations of invariant item ordering or of local independence. Thus, adequate scalability does not mean that items are free from other potential model violations.

Limitations and Future Research

This study has the following limitations: (a) The data generation algorithm of the simulation study was based on a trial-and-error process to sample items with scalability coefficients within the desired range. A more refined method to generate the data could have improved the efficiency of our algorithm; (b) In this study we only considered either dichotomous or polytomous items with a fixed number of response categories (i.e., either 3 or 5) per replication. It is of interest to consider mixed-format test data in future studies; (c) The practical outcomes we considered here are by no means exhaustive or equally relevant in all situations. Depending on the type of data and the application purpose, other outcomes might also be relevant. Therefore, this type of research can be extended to other outcomes of interest. Moreover, other types of scalability (e.g., person scalability) could have important practical consequences. These aspects should be addressed in future research.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bech, P., Carrozzino, D., Austin, S. F., Møller, S. B., & Vassend, O. (2016). Measuring euthymia within the Neuroticism Scale from the NEO Personality Inventory. A Mokken analysis of the Norwegian general population study for scalability. *Journal of Affective Disorders*, *193,* 99-102. doi:10.1016/j.jad.2015.12.039

Bielderman, A., Van der Schans, C., Van Lieshout, M.-R. J., De Greef, M. H. G., Boersma, F., Krijnen, W. P., & Steverink, N. (2013). Multidimensional structure of the Groningen Frailty Indicator in community-dwelling older people. *BMC Geriatrics*, *13*. doi:10.1186/1471-2318-13-86

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, *327*, 307–310. doi:10.1016/S0140-6736(86)90837-8

Bouman, A. J. E., Ettema, T. P., Wetzels, R. B., Van Beek, A. P. A., De lange, J., & Dröes, R. M. (2011). Evaluation of QUALIDEM: a dementia-specific quality of life instrument for persons with dementia in residential settings; Scalability and reliability of subscales in four Dutch field surveys. *International Journal of Geriatric psychiatry*, *26*, 711-722. doi:10.1002/gps.2585

Brenner, K., Schmitz, N., Pawliuk, N., Fathalli, F., Joober, R., Ciampi, A., & King, S. (2007). Validation of the Enlgish and French versions of the Community Assessment of Psychic Experiences (CAPE) with a Montreal community sample. *Schizophrenia Research*, *95,* 86-95. doi:10.1016/j.schres.2007.06.017

Cacciola, J. S., Alterman, A. I., Habing, B., & McLellan, A. T. (2011). Recent status scores for version 6 of the Addiction Severity Index (ASI-6). *Addiction*, *106*(9), 1588-1602. doi:10.1111/j.1360-0443.2011.03482.x

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155 -159. doi:10.1037/0033-2909.112.1.155

Crișan, D. R., Tendeiro, J. N., & Meijer, R. R. (2017). Investigating the Practical Consequences of Model Misfit in Unidimensional IRT Models. *Applied Psychological Measurement*, *41*, 439-455. doi:10.1177/0146621617695522

Crisan, D. R., Van de Pol, J. E., & van der Ark, L. A. (2016). Scalability coefficients for two-level polytomous item scores: An introduction and an application. In L. A. van der Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas, & M. Wiberg (Eds.), *Quantitative Psychology Research: The 80th Annual Meeting of the Psychometric Society*, Beijing, China, 2015 (pp. 139-153). New York, NY: Springer. doi:10.1007/978-3-319-38759-8_11

Dalal, D. K., & Carter, N. T. (2015). Consequences of ignoring ideal point items for applied decisions and criterion-related validity estimates. *Journal of Business and Psychology*, *30*, 483–498. doi:10.1007/s10869-014-9377-2

De Boer, A., Timmerman, M., Pijl, S. J., & Minnaert, A. (2012). The psychometric evaluation of a questionnaire to measure attitudes towards inclusive education. *Eur. J. Psychol. Educ.*, *27*, 573-589. doi:10.1007/s10212-011-0096-z

De Vries, J., Michielsen, H. J., & Van Heck, G. L. (2003). Assessment of fatigue among working people: comparisons of six questionnaires. *Occupational and Environmental Medicine*, *60*, Supplement 1, i10-i15.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Eribaum Associates, Inc.

Emons, W. H. M., Sijtsma, K., & Pedersen, S. S. (2012). Dimensionality of the Hospital Anxiety and Depression Scale (HADS) in cardiac patients. *Assessment, 19,* 337-353. doi:10.1177/1073191110384951

Ettema, T. P., Dröes, R.-M., De lange, J., Mellenberg, G., J., & Ribbe, M. W. (2007). QUALIDEM: Development and evaluation of a Dementia Specific Quality of Life Instrument. Scalability, reliability, and internal structure. *International Journal of Geriatric Psychiatry, 22,* 549-556. doi:10.1002/gps.1713

Hemker, B. T., Sijtsma, K., & Molenaar, I. W. (1995). Selection of unidimensional scales from a multidimensional item bank in the polytomous Mokken IRT model. *Applied Psychological Measurement, 19*, 337–352. doi:10.1177/014662169501900404

Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika, 62,* 331-347. doi:10.1007/BF02294555

Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New Phytologist*, *11*, 37–50. doi:10.1111/j.1469-8137.1912.tb05611.x

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading MA: Addison-Welsley Publishing Company.

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, The Netherlands: Mouton.

Molenaar, I. W. (1991). A weighted Loevinger H-coefficient extending Mokken scaling to multicategory items. *Kwantitatieve Methoden*, *12*(37), 97–117.

Molenaar, I. W. (1997). Nonparametric models for polytomous responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 369–380). New York, NY: Springer.

Molenaar, I. W., & Sijtsma, K. (2000). *MSP5 for Windows. A program for Mokken scale analysis for polytomous items*. Groningen, The Netherlands: iecProGAMMA.

R Development Core Team (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, *44*, 1276-1284.

Rupp, A. A. (2013). A systematic review of the methodology for person fit research in item response theory: Lessons about generalizability of inferences from the design of simulation studies. *Psychological Test and Assessment Modeling*, *55*(1), 3–38.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, No. 17.

Sinharay, S., & Haberman, S. J. (2014). How often is the misfit of item response theory models practically significant? *Educational Measurement: Issues and Practice*, *33*(1), 23–35. doi:10.1111/emip.12024

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

Sijtsma, K., & van der Ark, L.A. (2017). A tutorial on how to do a Mokken scale analysis on your test and questionnaire data. *British Journal of Mathematical and Statistical Psychology*, *70*, 137-158. doi:10.1111/bmsp.12078

Thomas, M. L. (2011). The value of item response theory in clinical assessment: A review. *Assessment, 18*, 291-307. doi:10.1177/1073191110374797

van der Ark, L. A. (2005). Stochastic ordering of the latent trait by the sum score under various polytomous IRT models. *Psychometrika, 70*, 283-304. doi:10.1007/s11336-000-0862-3

van der Ark, L. A. (2012). New developments in Mokken scale analysis in R. *Journal of Statistical Software*, *48*(5), 1–27. doi:10.18637/jss.v048.i05

Watson, R., Deary, I., & Austin, E. (2007). Are personality trait items reliably more or less “difficult”? Mokken scaling of the NEO-FFI. *Personality and Individual Differences, 43*, 1460-1469. doi:10.1016/j.paid.2007.04.023

Wind, S. (2016). Examining the psychometric quality of multiple-choice assessment items using Mokken scale analysis. *Journal of Applied Measurement*, *17*(2), 142-165.

Wind, S. (2017). An instructional module on Mokken scale analysis. *Educational Measurement: Issues and Practice*, *36*, 50-66. doi: 10.1111/emip.12153

Zijlmans, E. A. O., Tijmstra, J., van der Ark, L. A., & Sijtsma, K. (2018). Item-score reliability in empirical-data sets and its relationship with other item indices. *Educational and Psychological Measurement, 78*, 998-1020. doi:10.1177/0013164417728358

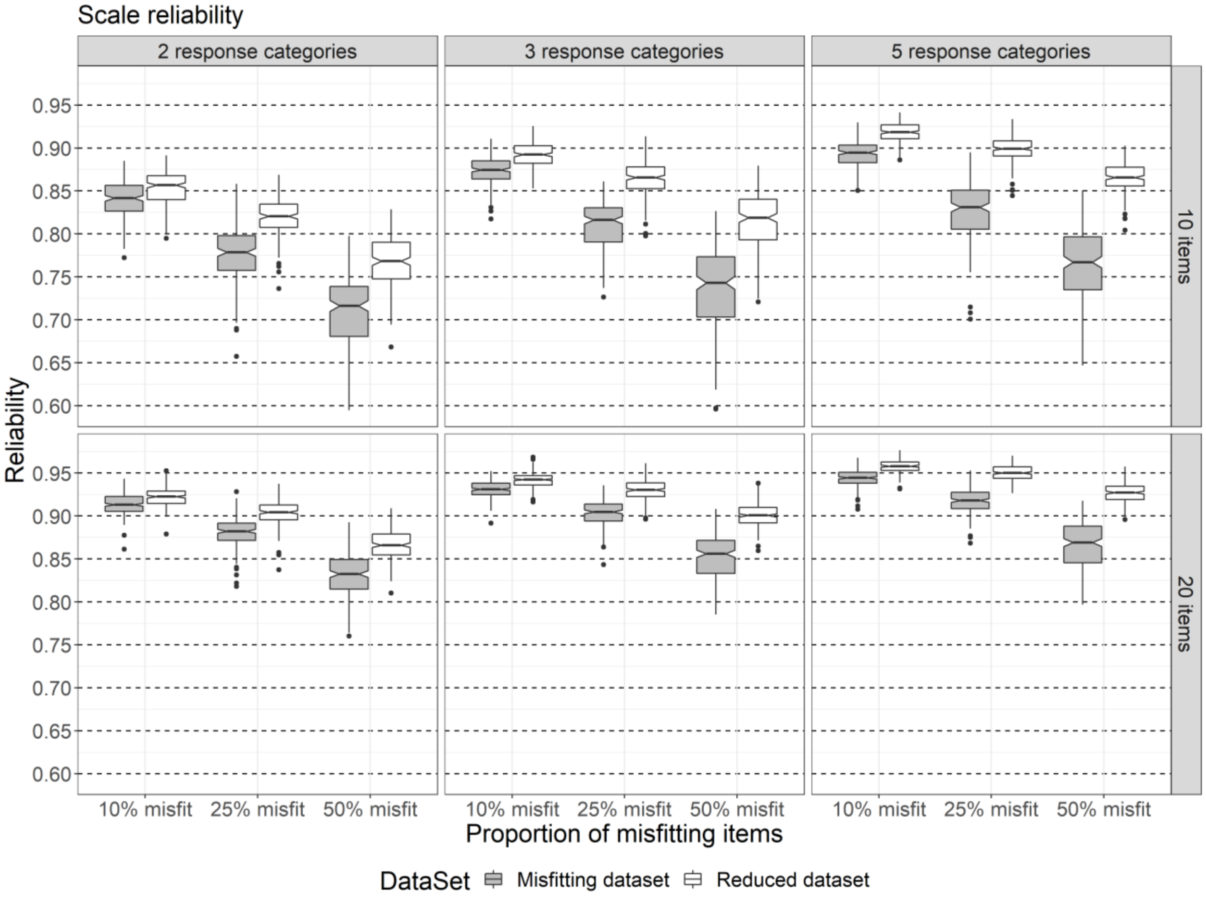

Figure 1. The distribution of reliability scores across the levels of *I*, *C*, and I_{LowH}, over all levels of R_H.

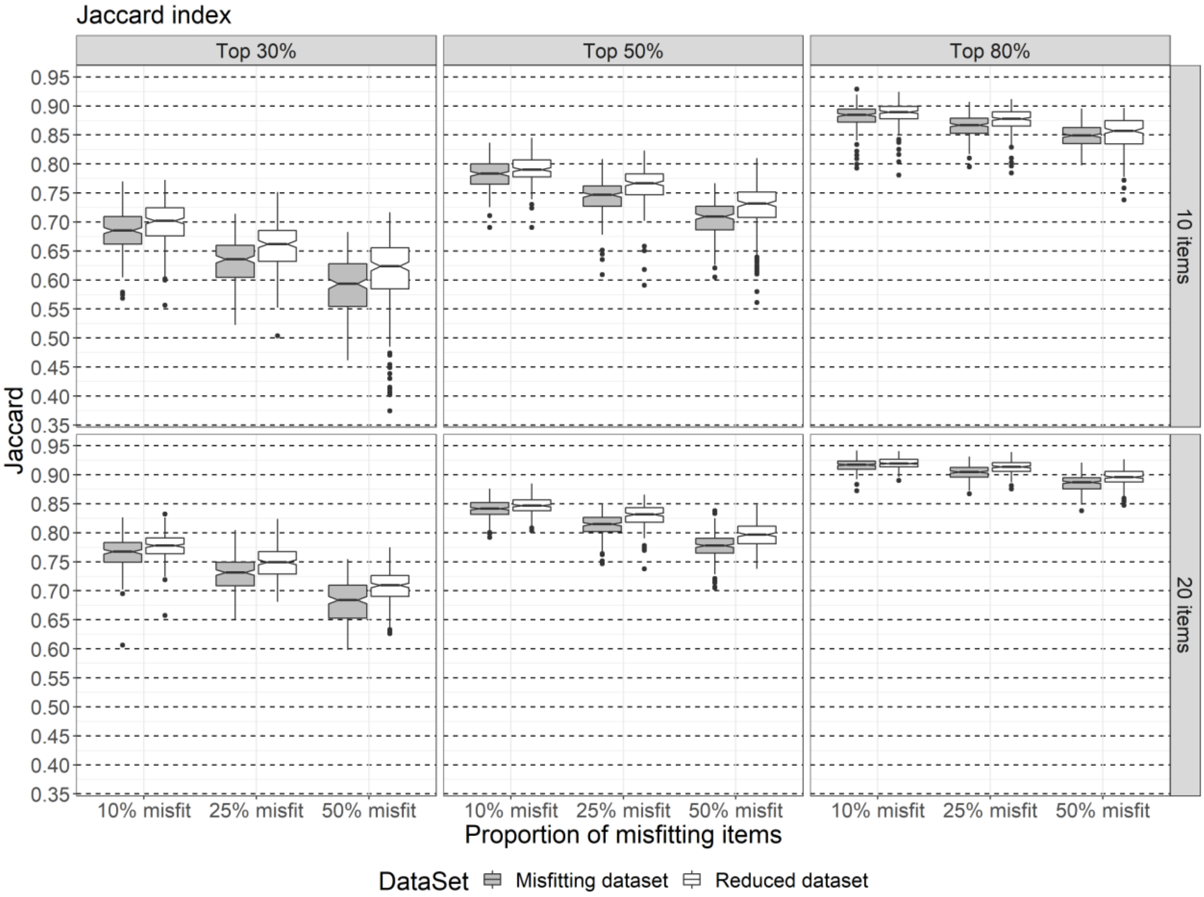

Figure 2. The distributions of the Jaccard index as a function of I_{LowH}, *DataSet*, *SR*, and *I*, when *C* = 2.

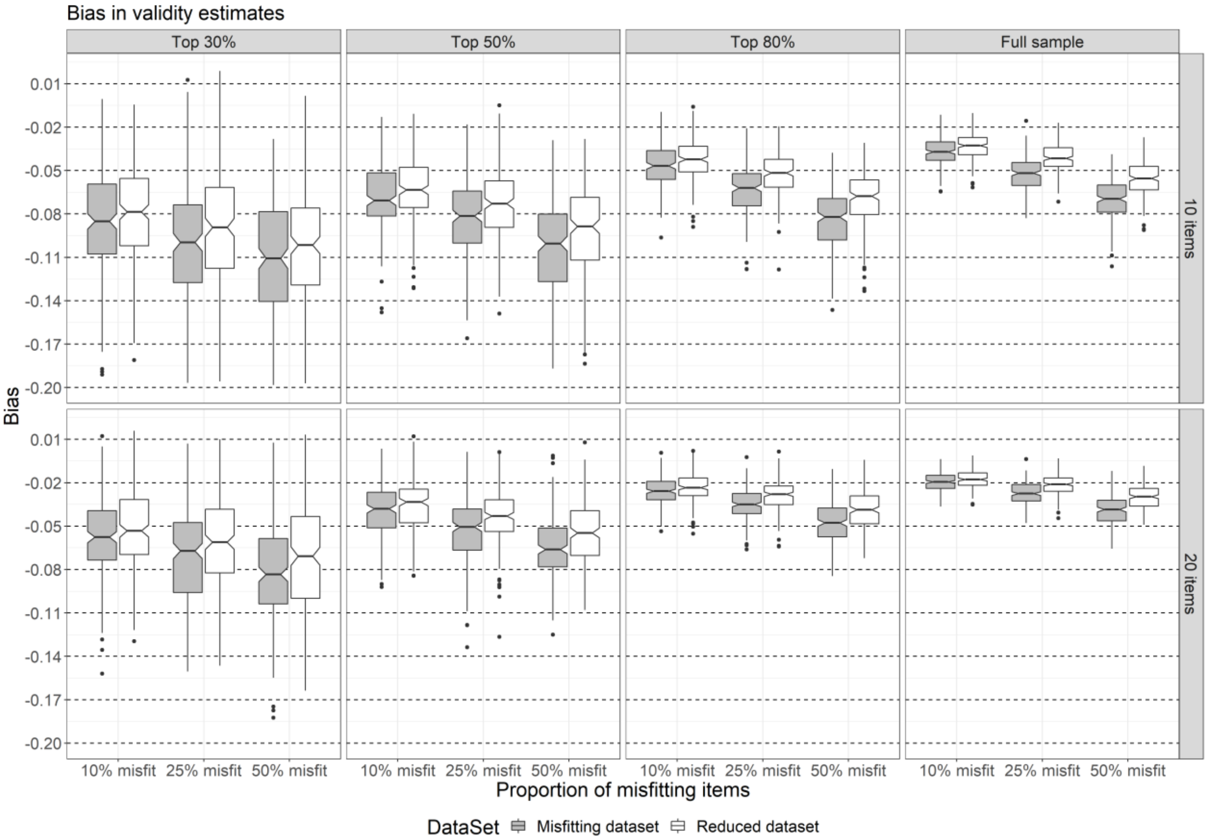

Figure 3. Bias in criterion-related validity estimates across I_{LowH}, *DataSet*, *I*, and *SR*, for scales with dichotomous items (C = 2) and true validity coefficient equal to .45

Table 1

Ranges of Discrimination Parameters Used For Data Generation

R_H	α_i	
	Fitting items	Misfitting items
$.10 \leq H_i < .20$	*U*(2.30, 2.70)	*U*(0.35, 0.75)
$.20 \leq H_i < .30$	*U*(2.30, 2.70)	*U*(0.50, 0.90)

Note: The discrimination parameters were randomly generated from a uniform distribution *U* bounded by the values in parentheses.