Article

Investigating the Practical Consequences of Model Misfit in Unidimensional IRT Models

Applied Psychological Measurement
2017, Vol. 41(6) 439–455

DOI: 10.1177/0146621617695522
journals.sagepub.com/home/apm


Daniela R. Crişan[1], Jorge N. Tendeiro[1], and Rob R. Meijer[1]

Abstract

In this article, the *practical* consequences of violations of unidimensionality on selection decisions in the framework of unidimensional item response theory (IRT) models are investigated based on simulated data. The factors manipulated include the severity of violations, the proportion of misfitting items, and test length. The outcomes that were considered are the precision and accuracy of the estimated model parameters, the correlations of estimated ability ($\hat{\theta}$) and number-correct (NC) scores with the true ability (θ), the ranks of the examinees and the overlap between sets of examinees selected based on either θ, $\hat{\theta}$, or NC scores, and the bias in criterion-related validity estimates. Results show that the $\hat{\theta}$ values were unbiased by violations of unidimensionality, but their precision decreased as multidimensionality and the proportion of misfitting items increased; the estimated item parameters were robust to violations of unidimensionality. The correlations between θ, $\hat{\theta}$, and NC scores, the agreement between the three selection criteria, and the accuracy of criterion-related validity estimates are all negatively affected, to some extent, by increasing levels of multidimensionality and the proportion of misfitting items. However, removing the misfitting items only improved the results in the case of severe multidimensionality and large proportion of misfitting items, and deteriorated them otherwise.

Keywords

item response theory, model fit, consequences of model violation

Introduction

Item response theory (IRT; for example, Embretson & Reise, 2000) is a popular psychometric framework for the construction and/or evaluation of tests and questionnaires, and applications range from large-scale educational assessment to small-scale cognitive and personality measures. Although IRT has a number of practical advantages over classical test theory, the price to pay for using IRT models in practice is that inferences made from IRT-based estimates are accurate to the extent that the empirical data meet the sometimes rather restrictive model assumptions and thus the model fits the data. The common assumptions for dichotomously scored data

[1]University of Groningen, The Netherlands

Corresponding Author:
Daniela R. Crişan, Department Psychometrics and Statistics, Faculty of Behavioral and Social Sciences, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands.
Email: d.r.crisan@rug.nl

analyzed using cumulative IRT models are unidimensionality, monotonicity, and local independence (Embretson & Reise, 2000).

In practice, the data rarely, if ever, meet the strict assumptions of the IRT models. Thus, model fit is always a matter of degree (e.g., McDonald, 1981). Therefore, there is a large body of literature that concentrates on developing methods for testing model assumptions and model fit (e.g., Bock, 1972; Haberman, 2009; Orlando & Thissen, 2000; Smith, Schumacker, & Bush, 1998; Stone & Zhang, 2003; Suárez-Falcón & Glas, 2003; Yen, 1981).

When a model does not fit data well enough or when the data violate one or more model assumptions to some degree, practitioners or test constructors are usually advised to use a better fitting model or to remove misfitting items (Sinharay & Haberman, 2014). Item fit is often determined by investigating the differences between the observed and expected proportions of correct item scores, where large residuals indicate misfit. Items that do not fit the model may be removed from the test so that a set of items is obtained that can reasonably be described by the IRT model under consideration. In practice, however, it is not always easy to remove items.

A first complication is that it is often not easy to define what a ''large'' residual should be. Another more practical consideration is that removing items from a test may distort the content validity of the measurement. For example, sometimes items are chosen so that they represent specific content domains that are important for representing the overall construct to be measured. Removing items that do not fit an IRT model may then result in an underrepresentation of the construct that is being measured. A third consideration is that if the test has already been administered, removing badly fitting items could disadvantage the test takers who answered them correctly. Finally, sometimes IRT models are not used for test construction or evaluation but to calibrate the items so that IRT-based methods can be used. Examples can be found in educational research, where IRT is used to link or equate different versions of a test, or in clinical assessment, where IRT is used to conduct IRT-based differential item functioning or IRT-based person–fit analysis (Meijer & Sijtsma, 2001). In these cases, it is decided beforehand which IRT model should be used, and then once implemented, it is often impossible to remove items, change the existing test, or use a different (i.e., better fitting) IRT model. Sometimes, there are even contractual obligations that determine the type of IRT model that is chosen (see Sinharay & Haberman, 2014).

As models give, at best, good approximations of the data, researchers have investigated the effects of model violations on the estimation of item and person parameters. Also, the robustness of the estimated parameters under different model violations has been investigated. The majority of previous studies focused on determining the robustness of different estimation methods against these violations (e.g., Drasgow & Parsons, 1983), and/or on the *statistical* significance of misfit on model parameters or on IRT-based procedures such as test equating (e.g., Dorans & Kingston, 1985; Henning, Hudson, & Turner, 1985). Some studies have explored the robustness of item parameter estimates to violations of the unidimensionality assumption (e.g., Bonifay, Reise, Scheines, & Meijer, 2015; Drasgow & Parsons, 1983; Folk & Green, 1989; Kirisci, Hsu, & Yu, 2001). As Bonifay et al. (2015) noted,

> . . . if a strong general factor exists in the data, then the estimated IRT item parameters are relatively unbiased when fit to a unidimensional measurement model. Accordingly, in applications of unidimensional IRT models, it is common to see reports of ''unidimensional enough'' indexes, such as the relative first-factor strength as assessed by the ratio of the first to second eigenvalues. (p. 505)

Also, indices have been proposed that give an idea about the strength of departure from unidimensionality, such as the DETECT index (Stout, 1987, 1990). DETECT is based on

conditional covariances between items to assess data dimensionality. The idea is that the covariance between two items, conditional on the common latent variable, is nonnegative when both items measure the same secondary dimension and it is negative when they clearly measure different secondary dimensions. Recently, Bonifay et al. (2015) investigated the ability of the DETECT ''essential unidimensionality'' index to predict the bias in parameter estimates that results from misspecifying a unidimensional model when the data are multidimensional.

Although the studies cited above are important, a next logical step is to investigate the impact of model misfit on the *practical decisions* that are being made based on the estimates derived from the model (i.e., the *practical significance* of model misfit), which is a far less studied but important issue (Molenaar, 1997). Practitioners are interested in knowing to what extent the main conclusions of their empirical research are valid under different models and settings—for example, with or without misfitting items, or with or without misfitting item score patterns. Sinharay and Haberman (2014) defined practical significance as ''an assessment of the extent to which the decisions made from the test scores are robust against the misfit of the IRT models'' (p. 23). The assessment of practical significance of misfit involves evaluating the agreement between decisions made based on estimated trait levels derived from misfitting models and the decisions made based on estimated trait levels derived from better fitting models (Sinharay & Haberman, 2014).

Recently, Sinharay and Haberman (2014) investigated the practical significance of model misfit in the context of various operational tests: a proficiency test in English, three tests that measure student progress on academic standards in different subject areas, and a basic skills test. Their study mostly considered the effect of misfit on equating procedures. They found that the one-, two-, and three-parameter logistic models (1PLM, 2PLM, and 3PLM), and the generalized partial credit model (e.g., Embretson & Reise, 2000), did not give a good description of any of the datasets. Moreover, they found severe misfit (i.e., large residuals between observed and expected proportion-correct scores) for a substantial number of items. However, they also found that for several tests that showed severe misfit, the practical significance was small, that is, a *difference that matters* (DTM) index lower than 0.5 (which was the recommended benchmark) and a disagreement of 0.0003% between a poor-fitting and a better fitting model–data combination with regard to pass–fail decisions.

As Sinharay and Haberman (2014) discussed, their study was concerned with the practical significance of misfit on equating procedures. The aim of the present study was to extend the Sinharay and Haberman (2014) study, and to investigate the practical significance of violations of unidimensionality on rank ordering and criterion-related validity estimates in the context of pattern scoring. More specifically, the impact of model misfit and of retaining or removing misfitting items on the rank ordering of simulees and on the bias in criterion-related validity estimates was assessed, as these are important outcomes for applied researchers. Misfit was simulated by inducing violations of the assumption of unidimensionality, which is a common underlying assumption for many IRT models. The validity of IRT applications largely depends on the assumption of unidimensionality (Reise, Morizot, & Hays, 2007). However, as Bonifay et al. (2015) noted, only narrow measures are strictly unidimensional. Often, multidimensionality is caused by diverse item content that is necessary to properly represent a complex construct. The question then is whether, and to what extent, violations of unidimensionality do affect the practical decisions that are made based on the estimated trait levels, and whether removing the items that violate the model with respect to unidimensionality improves the validity of these decisions. Moreover, the authors of this study were interested in whether practical effects associated with model misfit are affected by the selection ratio.

The following research questions were formulated:

Research Question 1 (RQ1): What is the effect of misfit on the estimated latent trait ($\hat{\theta}$) and on the estimated item parameters? The authors focused on the 2PLM (e.g., Embretson & Reise, 2000); hence, the item parameters of interest are the discrimination and difficulty parameters. The 2PLM was chosen as it is a model commonly applied to dichotomous multiple-choice items. The effect of model misfit on the precision and accuracy of item and person parameter estimates was investigated. The authors hypothesized to find evidence in agreement with Bonifay et al. (2015), who found that although some bias in parameter estimates might exist as a consequence of model misspecification, its magnitude is relatively small if a strong general factor exists in the data. Although investigating the effects of misfit on the precision and accuracy of model parameter estimates is not the main focus of this article, it is important to first show that the operationalization of misfit is sensible, so that the practical effects of misfit can be interpreted in relation to these violations. The novelty of this study is brought by RQ2 and RQ3.

Research Question 2 (RQ2): What is the effect of misfit on the rank ordering of persons, in combination with selection ratios? Although it is well known that $\hat{\theta}$ and *NC* scores correlate highly (e.g., Molenaar, 1997), the rank ordering of persons based on the model-fit data outcomes ($\widehat{\theta}_{Fit}, NC_{Fit}$) is expected to outperform the counterpart measures based on the model-misfit data ($\widehat{\theta}_{Misfit}, NC_{Misfit}$). The correlation of $\hat{\theta}$ and *NC* scores with the true θ is expected to decrease as the proportion of misfitting items increases and as the correlation between dimensions decreases. It is unknown from the literature how the trait-level estimates based on the reduced datasets (i.e., datasets from which the misfitting items are removed) would perform in comparison with $\widehat{\theta}_{Misfit}$ and NC_{Misfit}. Moreover, the authors were interested to investigate to what extent the sets of selected examinees coincide across the three scoring settings (model-fit, model-misfit, or misfitting items removed). The authors expected to find similar results across selection ratios, but with larger effect sizes as the selection ratio would decrease.

Research Question 3 (RQ3): What is the effect of misfit on criterion-related validity estimates? The authors hypothesized that the accuracy of estimating criterion-related validity would decrease as the proportion of misfitting items increased. Larger bias was also expected as the correlation between dimensions decreased. The authors had no prior expectation regarding the effect of removing the misfitting items on the bias in criterion-related validity estimates.

Method

Design and Simulation Setup

Independent variables. The following factors were manipulated in this study:

Proportion of misfitting items. Rupp (2013) provided an overview of simulation studies on model fit, and showed that the chosen number of misfitting items varied greatly between simulation studies, with values between 8% (e.g., Armstrong & Shi, 2009a, 2009b) and 75% or even 100% (e.g., Emons, 2008, 2009). Here, three levels were considered: $I_{misfit} = .10, .25, .50$. This is representative of small, medium, and large proportions of misfitting items.

Test length. Two test lengths were used: $I = 25, 40$. Test lengths between 20 and 60 items are typically used in simulation studies. These test lengths are representative of many intelligence and personality questionnaires.

Correlation between dimensions. The responses for the misfitting items were generated from a two-dimensional model (discussed below). Two levels for the correlation between dimensions θ_1 and θ_2 were considered: $r(\theta_1, \theta_2) = .70, .40$. The lower this correlation, the more multidimensional the data are. A correlation of approximately .70 is found between subtests of many educational tests that are considered unidimensional for practical purposes (Drasgow, Levine, & McLaughlin, 1991). A correlation of .40 might be considered too extreme. However, it allows exploring the effects of misfit in the case of severe multidimensionality.

Selection ratio. The selection ratio refers to the proportion of respondents who are selected, for example, for a job or an educational program, based on the test results. When the selection ratio is close to 1, the majority of individuals in the sample are selected. However, when the selection ratio is small, only a small number of individuals are selected. In this study, the following selection ratios (*SR*) were considered: $SR = 1, .80, .50$, and .30. For a given *SR*, the effect of keeping or removing misfitting items on the selected top $(100 \times SR)\%$ of examinees was assessed. Selection ratios of .80, .50, and .30 are representative of high through low selection rates. The proportion of selected top respondents was based on sorting the full sample on the basis of either θ, $\hat{\theta}$, or *NC* scores.

Dependent variables. To investigate the precision and accuracy of the estimated model parameters (RQ1), the mean absolute deviation (*MAD*, given by $\sum_{t=1}^{T} |\omega_t - \widehat{\omega}_t| / T$) and the bias (*BIAS*, given by $\sum_{t=1}^{T} (\omega_t - \widehat{\omega}_t) / T$) for the model parameters were analyzed across conditions, where ω denotes the model parameter under consideration and T denotes the sample size if ω refers to the person parameter, or the test length if ω refers to an item parameter.

To investigate the differences in the *rank ordering* of simulees under the different conditions (RQ2), Spearman's rank correlations between the various ranks were first computed based on θ, $\hat{\theta}$, and *NC* scores across conditions.[1] The Spearman rank correlations were always based on the entire sample of simulees; that is, with $SR = 1$. Second, to compare the *sets* of top selected simulees defined by each *SR* according to the rankings based on θ, $\hat{\theta}$, and *NC* scores, the Jaccard index was computed as a measure of the overlap between pairs of sets. The Jaccard index (Jaccard, 1912) for two sets is defined as the ratio of the cardinals of the intersection set to the union set, ranging from 0 (the two sets do not intersect) through 1 (the sets coincide; see Equation 1).

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}. \tag{1}$$

For $SR = 1$, the Jaccard index is always equal to 1, as all examinees are selected. When one of the two sets of top selected simulees is based on θ, the Jaccard index can be thought of as a measure of sensitivity (when computed in the misfit conditions) or specificity (when computed in the fit or reduced conditions).

To answer RQ3, the bias in criterion-related validity estimates was computed as the difference between the sample estimated validity and the population validity. Similar to Dalal and Carter (2015), the authors of the present study simulated, for each person, scores on a criterion variable that correlated $r = .15, r = .25, r = .35$, or $r = .45$ with θ. These values represent the population criterion-related validities.

Model-fit items. Dichotomous item scores (0 = *incorrect*; 1 = *correct*) were generated according to the 2PLM. All datasets were based on sample sizes equal to $N = 2,000$. True item and θ

parameters were drawn for each condition represented by a combination of the levels of all independent variables. The true item discrimination parameters $\alpha_i (i = 1, \ldots, I)$ were randomly drawn from the uniform distribution in the interval $(0.5, 2.0)$, and the true difficulty parameters β_i were randomly drawn from the standard normal distribution bounded between $\beta_i = -2.0$ and $\beta_i = +2.0$. Each simulee's θ_n value was randomly drawn from a standard normal distribution. This configuration of model and parameters is in line with similar types of simulation studies.

Model-misfit items. Violations of unidimensionality were generated using a two-dimensional model for a proportion of I_{misfit} randomly selected items. The following model based on Yu and Nandakumar (2001, Equation 7) was used:

$$P_i(\theta_{1n}, \theta_{2n}) = \frac{1}{1 + e^{-\alpha_{1i}(\theta_{1n} - \beta_i) - \alpha_{2i}(\theta_{2n} - \beta_i)}}. \quad (2)$$

Each pair $(\theta_{1n}, \theta_{2n})$ was randomly drawn from a uniform distribution in the interval $(\theta_n - 1.15, \theta_n + 1.15)$ or the interval $(\theta_n - 2.15, \theta_n + 2.15)$, depending on whether the desired correlation between dimensions was around $r = .7$ or $r = .4$. The intervals from which the pairs $(\theta_{1n}, \theta_{2n})$ were drawn were obtained by trial and error: Preliminary analyses showed that this sampling procedure generated pairs of θ values that correlated around the desired values. The discrimination parameters α_{1i} and α_{2i} in Equation 2 were equal to $\alpha_{1i} = \alpha_i \sin(\gamma_i)$ and $\alpha_{2i} = \alpha_i \cos(\gamma_i)$, where α_i is the discrimination parameter for item i that was generated for the model-fitting data situation, and γ_i is an angle randomly drawn from the uniform distribution in the interval $(0, \pi/2)$. As a consequence, two underlying correlated latent variables were used to generate the item scores, where each latent variable was partly contributing to the probability of correctly answering the items.

Model-fit checks. Some nonparametric model-fit checks (Sijtsma & Molenaar, 2002) were performed. In particular, violations of manifest monotonicity (Sijtsma & Molenaar, 2002) and unidimensionality were investigated. Manifest monotonicity is similar to the usual IRT latent monotonicity property but, instead of conditioning on the latent trait θ, one conditions on the observable total or rest score. It has been shown that, for dichotomous items, latent monotonicity implies manifest monotonicity (Junker & Sijtsma, 2000). Thus, violations of the latter imply violations of the former. Violations of unidimensionality were checked using the DETECT procedure (Stout, 1987, 1990). The DETECT value was computed based on one partitioning of the items into two disjoint clusters: Model-fitting items and model-misfitting items. The confirmatory approach of this study therefore sought for multidimensionality induced by the model assumption violation that was manipulated. The DETECT values of Roussos and Ozbek (2006) were used as reference: $0.2 < \text{DETECT} < 0.4$: weak multidimensionality; $0.4 < \text{DETECT} < 1.0$: moderate to large multidimensionality; $\text{DETECT} > 1.0$: strong multidimensionality (see, however, Bonifay et al., 2015 for a discussion of these benchmarks). Possible values for the DETECT index range between $-\infty$ and $+\infty$. The DETECT index was computed for both the model-fit and the model-misfit data. Furthermore, to assess item fit, the authors of this study computed the adjusted chi-square to degrees of freedom ratios for item singles, pairs, and triples (χ^2/df; Drasgow, Levine, Tsien, Williams, & Mead, 1995). Adjusted χ^2/df ratios above 3 are considered to be indicative of substantial misfit (Stark, Chernyshenko, Drasgow, & Williams, 2006). For the model-fit data, which were stochastically generated from the 2PLM, no model-fit issues were expected. The authors decided to perform these checks to serve as benchmarks for the similar model-fit outcomes for data displaying model-fit problems.

Design and implementation. A fully crossed design consisting of 3 (I_{misfit})×2(I)×2($r(\theta_1, \theta_2)$) = 12 conditions, with 100 replications per condition, was used. To test the adequacy of the chosen number of replications, the asymptotic Monte Carlo errors (MCEs; Koehler, Brown, & Haneuse, 2009) for each outcome were estimated across all experimental conditions. The MCEs for all outcomes were always smaller than 0.02, which was deemed acceptable for the purpose of this study. The simulation study was implemented in R (R Development Core Team, 2016). The R package *mirt* (Chalmers, 2012) was used to fit the 2PLM to each dataset. The function ''check.monotonicity'' from the *mokken* R package (Van der Ark, 2007, 2012) was used to check manifest monotonicity. The DETECT program, which was used to compute the DETECT index, comes with the DIM-Pack software (Version 1.0; Measured Progress, 2016). The adjusted χ^2 degrees of freedom ratios were computed using an *R* implementation of Stark's (2001) MODFIT program (the code is available upon request from the authors of this article).

Results

To test the hypotheses and to answer the research questions of this study, several ANOVAs were performed in two stages: First, the authors included in the models all the independent variables with their main effects, two-way interactions, and, where the plots suggested it, three-way interactions. Then, if some effects were nonsignificant, they fitted a second set of models in which they only retained the significant main effects and interactions. In reporting and interpreting the effects, the authors will focus on those for which $\eta^2 \geq .02$ which corresponds to a small effect according to Cohen (1992).

Model-Fit Checks

The monotonicity assumption was not affected by the manipulation of unidimensionality. A significant difference was not found between the proportion of violations of monotonicity in the fit and in the misfit data, $\chi^2(1) = .018, p = .894$, as in both conditions about 10% of the generated samples contained at least one significant violation of monotonicity.

Regarding unidimensionality, the DETECT procedure was sensitive to the operationalization of model misfit. As shown in Figure 1, the average DETECT values increased as I_{misfit} increased and as $r(\theta_1, \theta_2)$ decreased. About 38% of the variation in the DETECT values was caused by the differences between the fit and misfit conditions, 16% was caused by the proportion of misfitting items, and 10% was caused by the variation in $r(\theta_1, \theta_2)$. In the misfit condition, the average DETECT value ranged between 0.013 ($SD = 0.03$, for $I = 25, I_{misfit} = .10$, and $r(\theta_1, \theta_2) = .70$) and 0.69 ($SD = 0.09$, for $I = 25, I_{misfit} = .50$, and $r(\theta_1, \theta_2) = .40$).

The χ^2/df values also suggest that the operationalization of misfit was sensible. The average proportion of χ^2/df larger than 3 was very close to 0 for the model-fit data, in all conditions, for item singles, doubles, and triples. The χ^2/df statistic for item singles was insensitive to the manipulation of dimensionality, as all values were smaller than 3, on average, in all conditions. However, the proportion of χ^2/df for item doubles and triples that were larger than 3 increased in the misfit condition, as the proportion of misfitting items increased and as $r(\theta_1, \theta_2)$ decreased from .70 to .40. In the reduced condition (i.e., in datasets from which the misfitting items have been removed), the results were very similar to the model-fit condition.

Effect of Misfit on Model Parameters

Although investigating the effects of misfit on the precision and accuracy of model parameter estimates (RQ1) is not a major objective of this article, it is important to first show that the

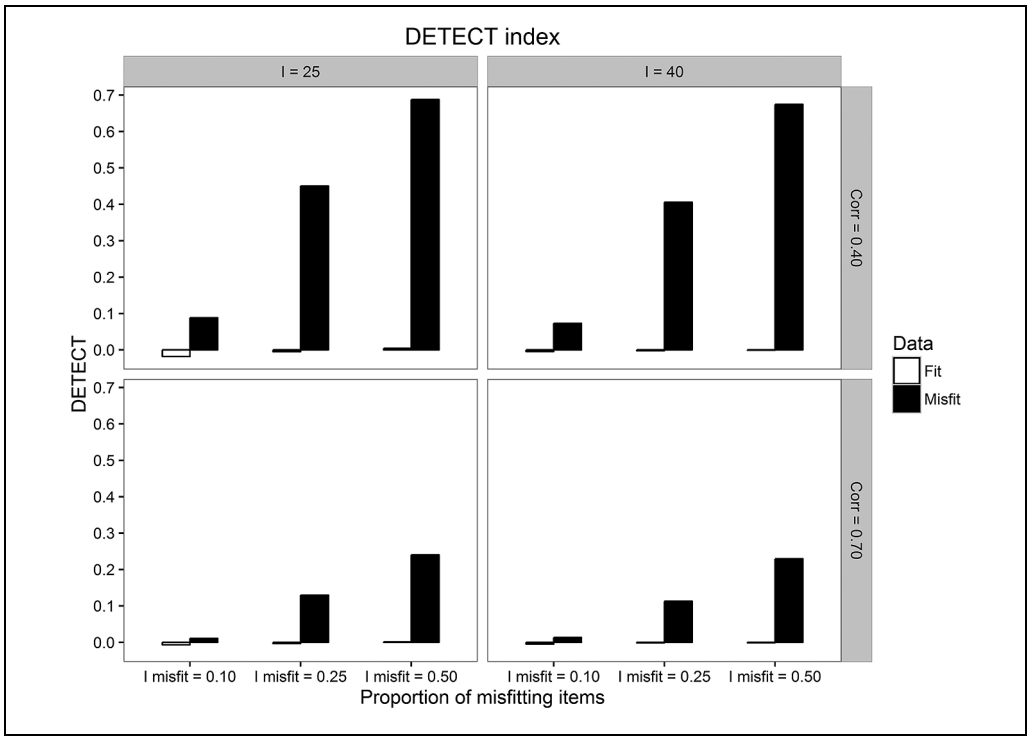

Figure 1. Average DETECT index of departure from unidimensionality for both the fit and the misfit data, across proportions of misfitting items, test lengths, and correlations between dimensions.

operationalization of misfit actually affected these estimates, so that the *practical* effects of misfit can be interpreted in relation to these violations. Therefore, the authors of this study first discuss the effects of violations of unidimensionality on person and item parameter estimates in terms of *MAD* and *BIAS*.

Effect of misfit on $\hat{\theta}$. The authors of this study analyzed the effect of violations of unidimensionality on the precision (*MAD*) and accuracy (*BIAS*) of $\hat{\theta}$. The analyses confirmed their expectations.

Figure 2 shows that the *MAD* of the $\hat{\theta}$ values for the misfit condition increased as the proportion of misfitting items increased and as multidimensionality became stronger; that is, $r(\theta_1, \theta_2)$ decreased. Also, the *MAD* decreased as test length increased. The ANOVA showed large main effects of Dataset (fit/misfit conditions), $F(1, 2384) = 8{,}723.13, p < .001, \eta^2 = .20, I,\ F(1, 2384) = 7{,}526.54,\ p < .001,\ \eta^2 = .17,\ I_{misfit},\ F(2, 2384) = 3{,}687.34,\ p < .001,\ \eta^2 = .17,$ and $r(\theta_1, \theta_2)$, $F(1, 2384) = 3{,}106.51, p < .001, \eta^2 = .07$. A small, but significant, three-way interaction effect of *Dataset*, I_{misfit}, and $r(\theta_1, \theta_2)$ was also found, $F(2, 2384) = 1{,}058.62, p < .001, \eta^2 = .05$.

Subsequent analyses showed that, as expected, introducing violations of unidimensionality deteriorated the precision of $\hat{\theta}$ as compared with the fit condition in terms of *MAD*; precision decreased even more as I_{misfit} and multidimensionality increased. Thus, when $r(\theta_1, \theta_2) = .70$, the effect size *d* of the difference between the *MAD* of $\hat{\theta}_{Fit}$ and the *MAD* of $\hat{\theta}_{Misfit}$ increased from very small in the case of 10% misfitting items to very large in the case of 50% misfitting items. When $r(\theta_1, \theta_2) = .40$, the effects were stronger, and they increased faster. Overall, increasing the number of items improved the precision of $\hat{\theta}$.

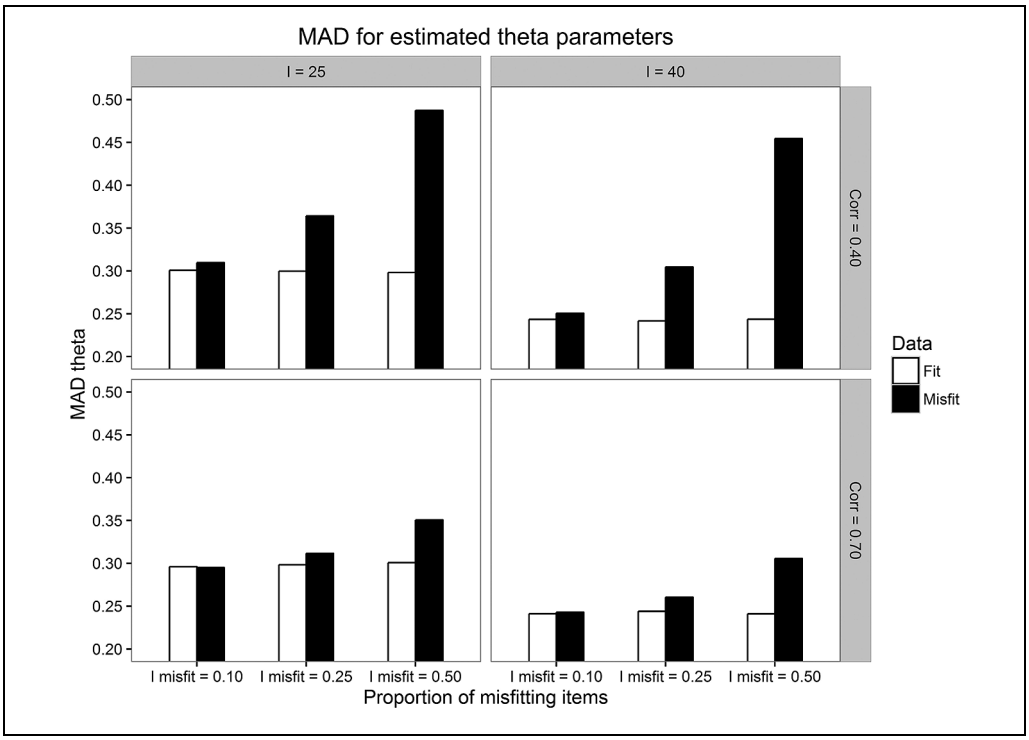

Figure 2. Average *MAD* for $\hat{\theta}$ for both the fit and the misfit data across proportions of misfitting items, test lengths, and correlations between dimensions.
Note. MAD = mean absolute deviation.

Regarding the accuracy of $\hat{\theta}$, there was hardly any impact on the *BIAS* index when unidimensionality was violated, regardless of the proportion of misfitting items or of the degree of multidimensionality.

Effect of misfit on item parameters. Concerning the precision and accuracy of item parameter estimates, the results support the conclusion of Bonifay et al. (2015) that the effect of misfit on parameter estimates is small if a strong general factor underlies the data. The precision of both the discrimination and difficulty parameters decreased when violation of unidimensionality was induced and as I_{misfit} increased (results not shown). Thus, in the case of α_i, Cohen's d for the difference between the fit and the misfit conditions increased from 1.55 to 6.49 as I_{misfit} increased from .10 to .50 and $r(\theta_1, \theta_2) = .70$. Respectively, the average *MAD* increased from 0.09 in the misfit condition with 10% misfitting items to 0.26 in the misfit condition with 50% misfitting items. For $r(\theta_1, \theta_2) = .40$, the pattern was similar, but the effects were somewhat stronger. For β_i, Cohen's d for the difference between the fit and the misfit conditions increased from 3.04 for $I_{misfit} = .10$ to 6.50 for $I_{misfit} = .50$. The average *MAD* for the β_i estimates increased from 0.16 in the misfit condition with 10% misfitting items to 0.54 in the misfit condition with 50% misfitting items.

Regarding the accuracy of the item parameter estimates, the effect of misfit on the accuracy of β_i was negligible (results not tabulated): The average bias ranged between -0.01 and 0.01 across conditions. For α_i, however, the effect of violation of unidimensionality was stronger, and the bias increased with I_{misfit}. Thus, Cohen's d for the difference between the fit and the

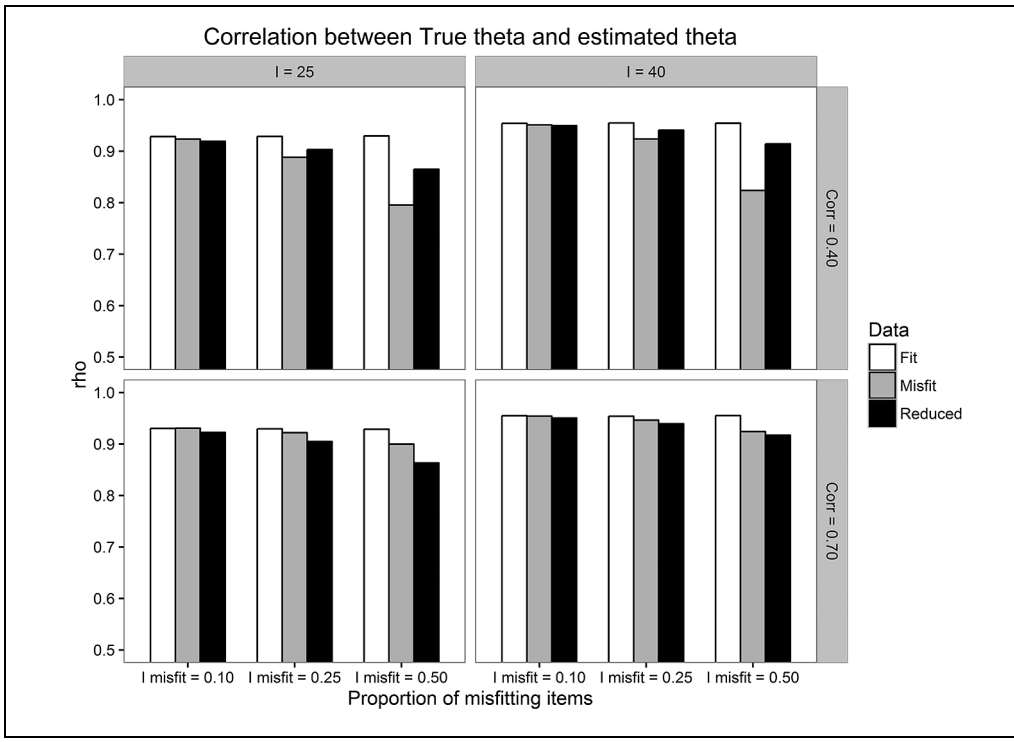

Figure 3. Average Spearman rank correlation between θ and $\hat{\theta}$ computed over the fit, the misfit, and the reduced datasets, across proportions of misfitting items, test lengths, and correlations between dimensions.

misfit conditions increased from 0.15 for I_{misfit} = .10 (the average *BIAS* for the misfit condition was −0.006) to 4.26 for I_{misfit} = .50 (the average *BIAS* for the misfit condition was −0.19).

Effect of Misfit on Rank Ordering of Persons

The effect of violations of unidimensionality on rank ordering and selection decisions (RQ2), and on criterion-related validity (RQ3, findings presented in the next section) represents the main focus of this study.

Correlations between θ, $\hat{\theta}$, and NC scores. The Spearman correlations between θ, $\hat{\theta}$, and *NC* scores (RQ2) were analyzed. On one hand, as expected, very high correlations were found between $\hat{\theta}$ and *NC* scores (all greater than .95), which were not affected by violations of unidimensionality, regardless of their severity. Also, removing the misfitting items did not have any impact. On the other hand, it was found that misfit affected the correlation between θ and $\hat{\theta}$, and between θ and *NC* scores. The results for these two correlations were very similar; thus, for simplicity, only the results concerning the rank correlation between θ and $\hat{\theta}$ are now presented (Figure 3). The conclusions can be extended to the rank correlation between θ and *NC* scores.

As can be seen in Figure 3, the analyses of the present study showed that violations of unidimensionality (misfit condition) had some decreasing effect on the magnitude of the correlation between θ and $\hat{\theta}$; only when multidimensionality was extreme (i.e., $r(\theta_1, \theta_2)$ = .40) and the percentage of misfitting items was 25 or 50, removing the misfitting items (reduced condition) led

to a higher rank correlation. Overall, increasing the test length had a positive effect on the correlation. ANOVA showed a significant main effect of test length, $F(1, 3576) = 6{,}686.25$, $p < .001$, $\eta^2 = .17$, and a small but significant three-way interaction between *Dataset*, I_{misfit}, and $r(\theta_1, \theta_2)$, $F(4, 3576) = 698.28$, $p < .001$, $\eta^2 = .07$, which is discussed next. In the condition with $r(\theta_1, \theta_2) = .70$, the rank correlation remained high, on average, ranging between .86 and .95. This variation in the average correlation was explained by the interaction between *Dataset* and I_{misfit}: Violations of unidimensionality caused a medium to large decrease in the average correlation only when there were 25% or 50% misfitting items, but removing these items actually reduced the correlation even more, for all values of I_{misfit}. For the extreme condition with $r(\theta_1, \theta_2) = .40$, the effects were stronger, and removing the misfitting items led to higher rank correlations.

The Jaccard index. It is well known that large rank correlations do not necessarily imply high agreement with respect to the sets of selected examinees (Bland & Altman, 1986). To check to what extent the sets of top selected examinees coincided, the Jaccard index was used across conditions. The results regarding the agreement between θ and $\hat{\theta}$ were very similar to the results regarding the agreement between θ and *NC* scores; thus, only the results for the former case are discussed in detail.

The analyses of this study showed that the disagreement between θ and $\hat{\theta}$ regarding the top selected simulees was larger when multidimensionality was severe. Moreover, removing the misfitting items only improved the agreement if their proportion in the test was large (50%) and multidimensionality was severe. Figure 4 clearly shows that there is a strong main effect of selection ratio on the Jaccard index between θ and $\hat{\theta}$; that is, overall, the proportion of overlap between sets of top selected examinees decreased greatly with selection ratio. In fact, the results showed that 89% of all the variance in the Jaccard index between θ and $\hat{\theta}$ can be attributed to the different selection ratios. However, there seems to be some interesting effects within selection ratios; for example, there is a three-way interaction effect between *Dataset*, I_{misfit}, and $r(\theta_1, \theta_2)$, which is further interpreted below in the case when $SR = .80$. The other levels of *SR* led to very similar effects in terms of strength and direction.

The three-way interaction between *Dataset*, I_{misfit}, and $r(\theta_1, \theta_2)$ was small, but significant, $F(4, 3579) = 194.12$, $p < .001$, $\eta^2 = .04$. When $r(\theta_1, \theta_2) = .70$ and $I_{misfit} = .10$, keeping or removing the misfitting items had a trivial effect on the agreement between θ and $\hat{\theta}$, which was around 91% for all conditions given by *Dataset* (i.e., in the model-fit datasets, in the datasets where misfit was induced, and in the datasets from which the misfitting items were removed). When $I_{misfit} = .25$ or .50, larger differences between conditions were found: Keeping the misfitting items in the test reduced the overlap between sets, especially in the latter case where a very large effect size was found, but removing them led to an even lower agreement. The average Jaccard index ranged between 0.88 for the reduced condition with 50% misfitting items and 0.91 for the misfit condition with 10% misfitting items.

As can be seen in Figure 4, when $r(\theta_1, \theta_2) = .40$, the pattern of results is different for the conditions with 25% or 50% misfitting items, where violations of unidimensionality reduced the overlap between sets to a very large extent, and removing the misfitting items led to a medium to very large improvement of the agreement between sets of top selected examinees. Here, the average Jaccard index ranged between .85 in the misfit condition with 50% misfitting items and .90 in the reduced condition with 25% misfitting items.

The effects were slightly stronger, and the overlap between sets of top selected examinees decreased as the selection ratio became smaller (not tabulated). Thus, when $SR = .30$, the overlap between sets of examinees selected based on θ and $\hat{\theta}$ ranged between 56% (for the condition with $r(\theta_1, \theta_2) = .40$ and $I_{misfit} = .50$) and 74% (for the fit and misfit conditions, with $I_{misfit} = .10$

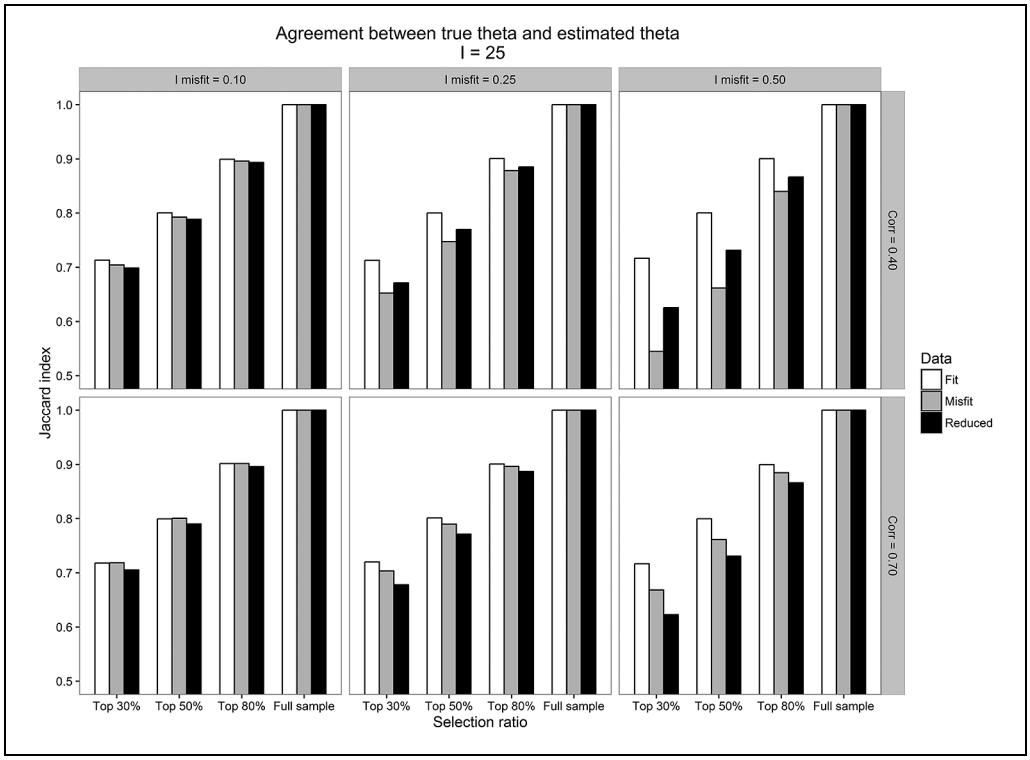

Figure 4. Average Jaccard index of agreement between θ and $\hat{\theta}$ computed over the fit, the misfit, and the reduced datasets, across proportions of misfitting items, test lengths, and correlations between dimensions.

and $r(\theta_1, \theta_2) = .70$). Similar percentages were found for the overlap between sets of examinees selected based on θ and *NC* scores. The overlap between sets of examinees selected based on θ and *NC* scores was larger than 80% in all experimental conditions.

Effect of Misfit on Criterion-Related Validity Estimates

For the bias in criterion-related validity estimates, the effects were tested on several population validity values, and it was found that the pattern of results was the same in all cases. Interestingly, the absolute value of *BIAS* increased overall as the population validity increased from .15 to .45. This is consistent with the findings of Dalal and Carter (2015). To avoid redundancy, more detailed results were provided here for the case when the population validity is .45. The conclusions can be generalized to the other validity values.

In Figure 5, the average *BIAS* in validity estimates is depicted across conditions. A very interesting result was that when multidimensionality was not extreme (i.e., $r(\theta_1, \theta_2) = .70$), it was better to keep the misfitting items in the test even when their proportion was large. Overall, the effect of violation of unidimensionality on the accuracy of criterion-related validity estimates is small: The largest mean value of *BIAS* equals -0.088 and was found for the misfit condition, with $I_{misfit} = .50$, $r(\theta_1, \theta_2) = .40$, and $I = 25$. Although not very large in magnitude, the bias in criterion-related validity estimates is affected by whether misfitting items are kept in the test and by the proportion of misfitting items, and these effects differ when $r(\theta_1, \theta_2) = .70$ or

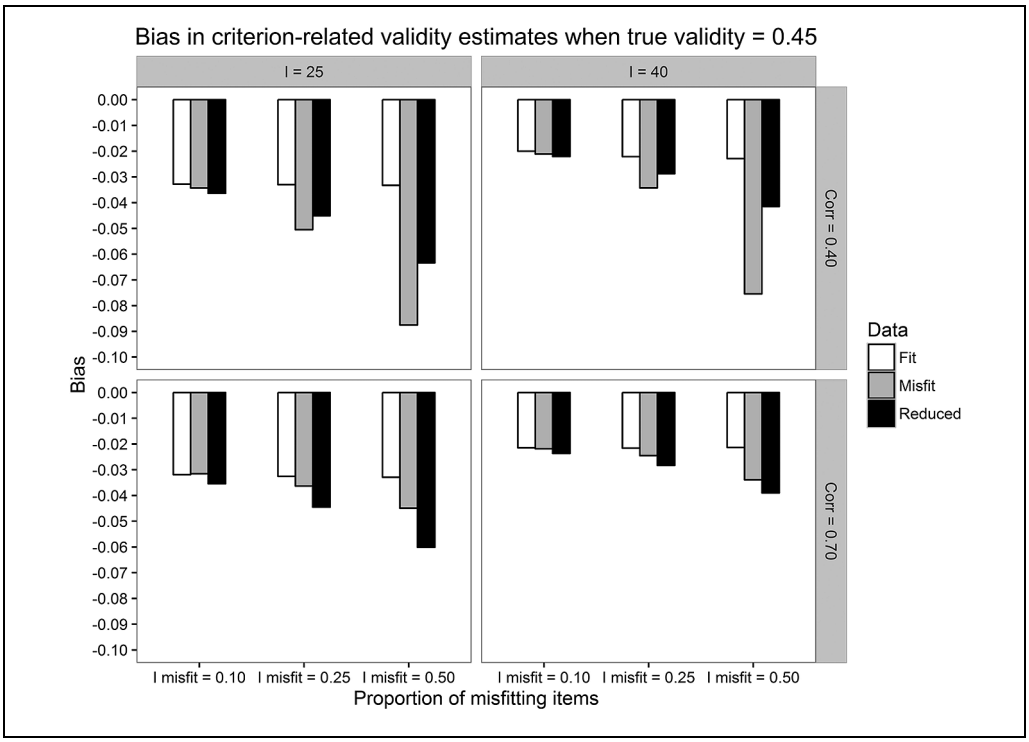

Figure 5. Average *BIAS* in criterion-related validity estimates computed over the fit, the misfit, and the reduced datasets, across proportions of misfitting items, test lengths, and correlations between dimensions.

.40. That is, there is a significant, albeit small three-way interaction effect between *Dataset*, I_{misfit}, and $r(\theta_1, \theta_2)$, $F(4, 3576) = 179.97$, $p < .001$, $\eta^2 = .05$. The interaction between I_{misfit} and *Dataset* was further analyzed separately at both levels of $r(\theta_1, \theta_2)$.

When $r(\theta_1, \theta_2) = .70$, bias in estimating criterion-related validity did not increase with violations of unidimensionality, but removing the misfitting items from the test resulted in a small increase in bias. When $I_{misfit} = .25$ or .50, bias increased due to violation of unidimensionality to a medium to very large extent, but removing the items from the test actually increased the bias even more.

When $r(\theta_1, \theta_2) = .40$ and $I_{misfit} = .10$, the situation was similar to the condition with $r(\theta_1, \theta_2) = .70$. For $I_{misfit} = .25$ or .50, however, the effect of removing the misfitting items was reversed. More specifically, in these cases removing the misfitting items actually led to a medium to strong decrease of the bias in criterion-related validity estimates.

Discussion

Model-fit assessment is an important step when fitting an IRT model to a set of item responses. The validity of the conclusions derived from an estimated IRT model decreases as the severity of violations of model assumptions in the data increases. As, in practice, there is always some degree of misfit between the data and the model, researchers and practitioners are constantly faced with having to choose between essentially three strategies, all of which might have

important drawbacks: (a) ignore the misfit, which might affect the accuracy of model parameter estimates; (b) remove the items that seem to cause the misfit and reassess the model fit. This might be problematic for reasons discussed in the ''Introduction'' (e.g., underrepresentation of the construct or impossibility to remove items); or (c) use a better fitting model. This strategy might also be problematic because, for example, the additional complexities might pose new estimation problems. It would be very useful to know which strategy would be more appropriate in which situation, based on the *practical* consequences of model misfit (i.e., the robustness of conclusions that are made based on a poorly fitting model). In cases where removing items is unfeasible, practitioners could profit from knowing how severe the problems may be when all items are taken into consideration.

In this study, the authors only focused on model misfit caused by violations of the assumption of unidimensionality in the context of dichotomously scored, educational tests. Three main conclusions can be drawn from the results of this study based on the three stated research questions: First, regarding the precision and accuracy of the model parameters, the precision of $\hat{\theta}$ decreased with increasing severity of multidimensionality and with increasing proportion of misfitting items (to a small extent), but $\hat{\theta}$ was essentially unbiased by violations of unidimensionality. The item parameters seemed to be robust against violations of unidimensionality. Precision decreased with increasing severity of multidimensionality and with increasing proportion of misfitting items, for both the discrimination and difficulty parameters, but the magnitude of this effect was small. The accuracy of the discrimination parameter also decreased with increase of both multidimensionality and the proportion of misfitting items. Again, the amount of overestimation was small.

Second, regarding the effect of misfit on the rank ordering of examinees according to either θ, $\hat{\theta}$, or *NC* scores, similar patterns were found for the Spearman rank correlations and for the Jaccard index of overlap between sets of top selected examinees. More specifically, when multidimensionality was not extreme, both the rank correlation and the Jaccard index decreased with increasing proportion of misfitting items, but they remained high. Removing the misfitting items slightly lowered both the correlations and the overlap between sets, especially for shorter tests with many misfitting items. Regarding the overlap between pairs of sets, it was found that it quickly decreased as the selection ratio decreased.

Third, the effect of misfit on the accuracy of criterion-related validity estimates was rather small, and it became even smaller as the population validity decreased. Nevertheless, the patterns of effects were similar to those found in this study for the rank ordering of examinees: If multidimensionality was not severe, removing the misfitting items actually decreased the accuracy of criterion-related validity estimates.

Practical Implications

Perhaps the most important message derived from this study is that, with respect to violations of unidimensionality, practical decisions seem to be only affected by model misfit to a small extent. Moreover, perhaps surprisingly, removing the misfitting items had in general a negative effect on practical decisions. One explanation for these findings might be that removing the misfitting items decreases the reliability of the test scores, which, in turn, might lead to poorer selection decisions and predictive validity. Although in the practice of large-scale educational testing misfitting items are more often replaced or revised rather than removed, in small-scale testing the practice of removing the misfitting items from the test is encountered more often (e.g., Bolt, Deng, & Lee, 2014; Sinharay, Haberman, & Jia, 2011; Sinharay & Haberman, 2014; Sondergeld & Johnson, 2014). Therefore, practitioners and researchers should be very careful when removing misfitting items from a test, whenever this possibility exists. Both the content

validity and the psychometric quality of a test as a whole may be influenced by removing items, and this may have an effect on important outcome measures. This message is particularly comforting in settings in which item removal is not an option, as one may now better gauge the consequences of retaining the misfitting items.

On a more general note, the aim of constructing and using psychological tests is not to obtain unidimensional measures but measures that are theoretically and practically *useful*. As Gustafsson and Åberg-Bengtsson (2010) discussed, the strict requirement of unidimensionality may have negative effects on the interpretability and usefulness of the resulting measures. In fact, there are many widely used instruments that do not meet the unidimensionality requirement but are highly useful for theoretical, diagnostic, and predictive purposes (e.g., intelligence test batteries such as the Wechsler series). As such, a narrow instrument measuring a homogeneous construct may have some desirable psychometric properties but may be of limited usefulness for some practical purposes.

Limitations and Future Research

In this study, the following limitations are discussed: (a) the authors only considered dichotomously scored items; investigating these effects on polytomous items would also be of great practical value; (b) they only considered misfit due to violations of unidimensionality; the practical significance of violations of other assumptions and/or their interactions could also be insightful; (c) regarding practical decisions, the authors focused on various outcomes in educational settings. One can expect that, in context other than educational assessment, different conclusions about the impact of model misfit on practical decisions may arise. All these limitations ought to be addressed in future research.

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.

Note

1. Kendall's coefficient was also considered. Results showed negligible differences between both approaches, and therefore, only the results based on Spearman's coefficients were discussed.

References

Armstrong, R. D., & Shi, M. (2009a). A parametric cumulative sum statistic for person fit. *Applied Psychological Measurement*, *33*, 391-410. doi:10.1177/0146621609331961

Armstrong, R. D., & Shi, M. (2009b). Model-free CUSUM methods for person fit. *Journal of Educational Measurement*, *46*, 408-428. doi:10.1111/j.1745-3984.2009.00090.x

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, *327*, 307-310.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29-51. doi:10.1007/BF02291411

Bolt, D. M., Deng, S., & Lee, S. (2014). IRT model misspecification and measurement of growth in vertical scaling. *Journal of Educational Measurement*, *51*, 141-162. doi:10.1111/jedm.12039

Bonifay, W. E., Reise, S. P., Scheines, R., & Meijer, R. R. (2015). When are multidimensional data unidimensional enough for structural equation modeling? An evaluation of the DETECT multidimensionality index. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*, 504-516. doi:10.1080/10705511.2014.938596

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*, 1-29. doi:10.18637/jss.v048.i06

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155-159.

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

Dorans, N. J., & Kingston, N. M. (1985). The effects of violations of unidimensionality on the estimation of item and ability parameters and on item response theory equating of the GRE verbal scale. *Journal of Educational Measurement*, *22*, 249-262. doi:10.1111/j.1745-3984.1985.tb01062.x

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1991). Appropriateness measurement for some multidimensional test batteries. *Applied Psychological Measurement*, *15*, 171-191. doi:10.1177/014662169101500207

Drasgow, F., Levine, M. V., Tsien, S., Williams, B., & Mead, A. D. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement*, *19*, 143-166. doi:10.1177/014662169501900203

Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement*, *7*, 189-199. doi:10.1177/014662168300700207

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

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

Emons, W. H. M. (2009). Detection and diagnosis of person misfit from patterns of summed polytomous item scores. *Applied Psychological Measurement*, *33*, 599-619. doi:10.1177/0146621609334378

Folk, V. G., & Green, B. F. (1989). Adaptive estimation when the unidimensionality assumption of IRT is violated. *Applied Psychological Measurement*, *13*, 373-390. doi:10.1177/014662168901300404

Gustafsson, J.-E., & Åberg-Bengtsson, L. (2010). Unidimensionality and interpretability of psychological instruments. In S. E. Embretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches* (pp. 97-121). American Psychological Association.

Haberman, S. J. (2009). *Use of generalized residuals to examine goodness of fit of item response models* (ETS Research Report Series No. RR-09-15). Princeton, NJ: Educational Testing Service.

Henning, G., Hudson, T., & Turner, J. (1985). Item response theory and the assumption of unidimensionality for language tests. *Language Testing*, *2*, 141-154. doi:10.1177/026553228500200203

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

Junker, B. W., & Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement*, *24*, 65-81. doi:10.1177/01466216000241004

Kirisci, L., Hsu, T.-c., & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement*, *25*, 146-162. doi:10.1177/01466210122031975

Koehler, E., Brown, E., & Haneuse, S. J.-P. (2009). On the assessment of Monte Carlo error in simulation-based statistical analyses. *The American Statistician*, *63*, 155-162. doi:10.1198/tast.2009.0030

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology*, *34*, 100-117. doi:10.1111/j.2044-8317.1981.tb00621.x

Measured Progress. (2016). *DIM-pack*. Available from http://psychometrictools.measuredprogress.org

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*, 107-135. doi:10.1177/01466210122031957

Molenaar, I. W. (1997). Lenient or strict application of IRT with an eye on practical consequences. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 38-49). New York, NY: Waxmann.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*, 50-64. doi:10.1177/01466216000241003

R Development Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Available from http://www.R-project.org

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research, 16*(Suppl. 1), 19-31. doi:10.1007/s11136-007-9183-7

Roussos, L. A., & Ozbek, O. Y. (2006). Formulation of the DETECT population parameter and evaluation of DETECT estimator bias. *Journal of Educational Measurement*, *43*, 215-243. doi:10.1111/j.1745-3984.2006.00014.x

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*, 3-38.

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

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

Sinharay, S., Haberman, S. J., & Jia, H. (2011). *Fit of item response theory models: A survey of data from several operational tests* (ETS Research Report Series No. RR-11-29). Princeton, NJ: Educational Testing Service.

Smith, R. M., Schumacker, R. E., & Bush, J. M. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement*, *2*, 66-78.

Sondergeld, T. A., & Johnson, C. C. (2014). Using Rasch measurement for the development and use of affective assessments in science education research. *Science Education*, *98*, 581-613. doi:10.1002/sce.21118

Stark, S. (2001). MODFIT: A computer program for model-data fit. Unpublished manuscript, University of Illinois, Urbana-Champaign.

Stark, S., Chernyshenko, O. S., Drasgow, F., & Williams, B. A. (2006). Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology*, *91*, 25-39. doi:10.1037/0021-9010.91.1.25

Stone, C. A., & Zhang, B. (2003). Assessing goodness of fit of item response theory models: A comparison of traditional and alternative procedures. *Journal of Educational Measurement*, *40*, 331-352. doi:10.1111/j.1745-3984.2003.tb01150.x

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*, 589-617. doi:10.1007/BF02294821

Stout, W. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika*, *55*, 293-325. doi:10.1007/BF02295289

Suárez-Falcón, J. C., & Glas, C. A. W. (2003). Evaluation of global testing procedures for item fit to the Rasch model. *British Journal of Mathematical and Statistical Psychology*, *56*, 127-143. doi:10.1348/000711003321645395

Van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software*, *20*, 1-19. doi:10.18637/jss.v020.i11

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

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, *5*, 245-262. doi:10.1177/014662168100500212

Yu, F., & Nandakumar, R. (2001). Poly-DETECT for quantifying the degree of multidimensionality of item response data. *Journal of Educational Measurement*, *38*, 99-120. doi:10.1111/j.1745-3984.2001.tb01118.x