Article

Practical Significance of Item Misfit in Educational Assessments

Applied Psychological Measurement
2017, Vol. 41(5) 388–400

DOI: 10.1177/0146621617692978
journals.sagepub.com/home/apm
SAGE

Carmen Köhler[1] and Johannes Hartig[1]

Abstract

Testing item fit is an important step when calibrating and analyzing item response theory (IRT)-based tests, as model fit is a necessary prerequisite for drawing valid inferences from estimated parameters. In the literature, numerous item fit statistics exist, sometimes resulting in contradictory conclusions regarding which items should be excluded from the test. Recently, researchers argue to shift the focus from statistical item fit analyses to evaluating practical consequences of item misfit. This article introduces a method to quantify potential bias of relationship estimates (e.g., correlation coefficients) due to misfitting items. The potential deviation informs about whether item misfit is practically significant for outcomes of substantial analyses. The method is demonstrated using data from an educational test.

Keywords

model fit, item fit, practical significance, item response theory, educational measurement

To draw valid inferences from an item response theory (IRT) model, the fit of the model needs to be assessed and evaluated (Embretson & Reise, 2000). Model misfit indicates that one or several model assumptions are violated. In unidimensional IRT, these assumptions include local stochastic independence between item responses and assumptions resulting from restrictions of parameters of the item characteristic curves (ICCs), such as setting all discrimination parameters equal to 1. In case of model misfit, the estimated ability and item parameters might be biased and cannot be interpreted reliably (Wainer & Thissen, 1987; Yen, 1981). Testing model fit is thus considered an important step when calibrating and analyzing IRT-based tests, as is documented in Standard 4.10 of the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council for Measurement in Education, 2014).

Since no model perfectly fits any given data set, model misfit will always be present to some degree (Box & Draper, 1987). The important question researchers frequently find themselves confronted with revolves around how much misfit is acceptable. Swaminathan, Hambleton, and Rogers (2006) identify two main steps for assessing model fit: (a) Testing underlying assumptions, and (b) comparing predictions of the model with observed values. On a purely statistical

[1]German Institute for International Educational Research (DIPF), Frankfurt, Germany

Corresponding Author:
Carmen Köhler, German Institute for International Educational Research (DIPF), Schloßstraße 29, Frankfurt 60486, Germany.
Email: carmen.koehler@dipf.de

level, numerous tools for evaluating model fit exist (see, for example, Ames & Penfield, 2015; Swaminathan et al., 2006). In educational assessments, commonly applied methods include differential item functioning (DIF) analyses to evaluate item parameter invariance across groups, testing for unidimensionality, comparing different scaling models, assessing reliability, and scrutinizing item fit indices (see, for example, Organisation for Economic Co-Operation and Development [OECD], 2012; Pohl & Carstensen, 2012). Practitioners often apply heuristics or rules of thumb in order to evaluate the significance of any deviations from the expected outcomes. With regard to item fit, such rules of thumb encompass the evaluation of item fit plots and cutoff scores. The consequences of item misfit oftentimes involve the collapsing of categories of polytomous items, changing the phrasing of the item, or removing the item from the test and/or the empirical analysis altogether.

In tests constructed under IRT, the strict model assumptions typically lead to at least some items to be identified as misfitting. In some instances, even large percentages of items show bad model fit. Having to remove items due to misfit is undesirable for test developers in several aspects: For one, item development costs time and money; another important aspect concerns the sufficient representation of the construct that is to be measured. Tests are often developed according to a specific theory, and the generated item pool is supposed to cover certain aspects of a construct. Each of these aspects (i.e., subdomains) are typically assessed via a limited number of items. If several items measuring a specific subdomain are removed because of model misfit, this aspect can no longer be appropriately assessed if the number of items in the respective subdomain is insufficiently large. All in all, retaining items in the test is commonly desired.

Considering that item misfit is not necessarily relevant with regard to the test outcome, the practice of removing items seems somewhat rash—especially in light of the ongoing debate about the validity of many item fit statistics (see, for example, Ames & Penfield, 2015; Orlando & Thissen, 2000; Swaminathan et al., 2006). The criticism mostly targets the validity of the derived cutoff scores. Recent work by Hambleton and Han (2005), Molenaar (1997), as well as Sinharay (2005) emphasizes the importance of looking beyond statistical significance of item fit and focusing more on its practical significance. The assessment of model fit should be viewed as a multifaceted process that also comprises an examination of the consequences of model misfit (Hambleton & Han, 2005; Sinharay, 2005; Sinharay & Haberman, 2014). Practical consequences pertain to the purpose of the test and the implications from the assessment. In high-stakes assessments, for example, tests might function as a selection criterion for admission into a certain program or educational institution. Sinharay and Haberman (2014) investigated data from three educational tests that were used to derive cut scores, categorizing students according to their competence levels: It was demonstrated that although item misfit was prevalent in all data sets, their practical significance was minor: In two out of the three examples, the removal of items resulted in negligible changes regarding the categorization of students. The authors propose that the decision of whether misfit is practically significant should be based on the change in test outcomes, and conclude that the removal of items is unnecessary if it has no practical relevance.

Note that Sinharay and Haberman (2014) focused on high-stakes testing, in which the accuracy of individual scores is of major importance. A study by van Rijn, Sinharay, Haberman, and Johnson (2016) investigated practical significance of item misfit in the area of low-stakes educational assessments. By *low-stakes*, the authors refer to tests where the assessment outcome has no immediate individual consequences for the examinee. In low-stakes educational assessments such as the Programme for International Student Assessment (PISA) or the National Assessment of Educational Progress (NAEP), most analyses revolve around relationships between competence and other variables or competence comparisons between groups. Van Rijn et al. estimated subgroup means and the percent of examinees at different ability levels to

investigate practical significance of item misfit. The outcomes of these estimates were compared when misfitting items were kept in the measurement model versus when they were excluded from the model. Like Sinharay and Haberman (2014), they found that item misfit hardly impacted the outcome.

Note that in both studies on practical significance—Sinharay and Haberman (2014) and van Rijn et al. (2016)—the investigation regarding practical significance of item misfit was conducted separately for each of their empirical examples. Testing practical significance for each outcome of interest can be a quite demanding and cost consuming task. Up to date, no general approach exists to evaluate practical significance of misfitting items in educational tests, and no software program reports influences of misfitting items on important outcome variables. In this article, a method to assess consequences of keeping misfitting items in a low-stakes achievement test is proposed. The focus of this study lies on tests that are primarily used to compare competences across groups or to analyze relationships between competences and other variables. In most instances, these relationships are investigated through the analysis of variance components, for example, ANOVA, regression analysis, or correlation coefficients. Results from such analyses allow evaluating the size and significance of the relationship between variables. The correlation coefficient—and the according R-squared—is especially relevant for evaluating whether the relationship between two variables is substantial. The authors argue that if the correlation coefficient significantly changes due to misfitting items in the model, item misfit is practically significant. The authors consider a change in the correlation coefficient as significant when inferences on substantial research questions are altered, for example, if the estimated size of the relationship between ability and a covariate is distorted by including misfitting items in the measurement model so that an actually existing medium size relationship decreases to a low size relationship.

A general approach—applicable to any competence test—is offered to evaluate potential bias in the correlation coefficient when misfitting items are kept in the analysis. Note that the reader can choose which item fit statistics to use, and the debate on the most appropriate item fit statistic is disregarded in the current article. This study's approach is basically an additional aspect in the process of evaluating model fit, and picks up after the researcher has decided—based on statistical item fit analyses and closer inspection of the items—which items might potentially be removed from the test. The approach can be used even when covariates of interest have not been assessed yet, for example, in trial administrations of new tests.

The next section describes this study's approach in detail for the Rasch model (Rasch, 1960), followed by a short description of its generalization to the two-parameter logistic (2PL) model (Birnbaum, 1968). A small simulation that illustrates which factors influence the potential change of the correlation coefficient is subsequently provided. The authors then give an empirical example, demonstrating the effectiveness of the approach in evaluating practical significance of item misfit. Note that an R code was developed for easy implementation of this study's method (see the Online Appendix).

Method

Approach for Rasch Model

To evaluate whether the exclusion of several items has an impact on any analysis of substantive interest, the researcher could simply compare the parameters of interest from the model where misfitting items are included in the measurement model for ability and the model where misfittings items are removed from the measurement model for ability. For example, if the substantive research question concerned the relationship between ability in mathematics and interest in

mathematics, the correlation (or standardized regression coefficient) between ability and interest in mathematics could be calculated (a) using only fitting items in the measurement model for ability, so that the latent ability variable, θ^F, is based on the response indicators of all fitting items, x_{ij}, where i indexes the items from $i = 1, \ldots, I_x$, and j indexes the persons from $j = 1, \ldots, N$. Another option (b) for obtaining the correlation coefficient between ability and interest in mathematics is to include the total number of items, that is, the fitting items, x_{ij}, as well as the misfitting items, m_{ij}, in the measurement model for ability, θ^T. The index for the misfitting items runs from $i = 1, \ldots, I_m$. The difference between the two obtained correlation coefficients indicate how much the parameter of interest, that is, the correlation coefficient, is influenced by the presence of misfitting items.

As this procedure has very limited generalizability and would require a new interpretation of practical significance of item misfit for each research question, a mathematical approach that allows establishing the potential bias in the correlation coefficient for all possible values of $r(\theta^F, Z)$, and therefore all possible covariates, Z, is proposed. The fact that the approach of determining practical significance is applicable irrespective of the observation of Z is especially convenient for pretest situations, which sometimes lack the assessment of the covariates of interest, as they are often only assessed in subsequent main studies.

The underlying idea of this study's approach is based on the decomposition of variance components. The inclusion of misfitting items affects the variance of the latent variable as well as its correlation with the covariate. The potential change in the correlation coefficient is limited to a certain range, however, which can be mathematically computed. The minimum and maximum change depends on the amount of additional variance that is induced by the misfitting items, on the amount of misfitting items relative to the fitting items, and on the strength of the relationship between the latent variable and the covariate. Let θ^M denote the latent ability variable with item indicators m_{ij} of the misfitting items. Given the covariance between θ^F and θ^M, the standard deviation of both variables, σ_{θ^F} and σ_{θ^M}, and the standard deviation of the covariate Z, σ_z, it is possible to calculate the minimum and maximum change of the correlation coefficient if the item indicators m_{ij} are included in measuring ability. For means of identification, the number of misfitting items needs to exceed 2 under the Rasch model and 3 under the 2PL model; apart from these conditions, this study's approach is generalizable to any amount of misfitting items. Note that in most settings, the assumption that the misfitting items measure a single latent dimension, θ^M, will probably not hold. However, making this assumption means that the inclusion of the misfitting items can have the maximum impact on the correlation with Z. Thus, the assumption of all m_{ij} measuring a single dimension was used as as a worst case scenario to calculate the possible range of changes in $r(\theta^F, Z)$. If responses to m_{ij} are multidimensional, or if all m_{ij} are uncorrelated even, the impact on $r(\theta^F, Z)$ will be smaller.

The minimum and maximum correlations, $r_{\min(\theta^T, Z)}$ and $r_{\max(\theta^T, Z)}$, occur in the extreme cases in which the residual of Z—after conditioning Z on θ^F—perfectly explains the residual of θ^M—after conditioning θ^M on θ^F. That is, $r(\theta^M, Z\theta^F) = \pm 1$, or rather, the partial correlation $pr(\theta^M, Z) \pm 1$. The direction of the partial correlation $pr(\theta^M, Z)$ can be contrary to the correlation $r(\theta^F, Z)$, since the misfitting items might measure something completely different from what the fitting items measure. They might measure a different ability dimension or have a low discrimination: For whatever reason they were flagged misfitting, the additional variance they bring to the ability variable, θ^T, might differently depend on Z than the ability measured by only the fitting items, θ^F. That is to say, the correlation between θ^M and Z might differ from the correlation between Z and θ^F, and this correlation has certain predictable limits. Given $r_{\min(\theta^T, Z)}$ and $r_{\max(\theta^T, Z)}$, the minimum and maximum change in $r(\theta^T, Z)$ can be calculated. In the following, the authors first describe how to compute $r_{\min(\theta^M, Z)}$ and $r_{\max(\theta^M, Z)}$. The computation of $r_{\min}$ and $r_{\max}$ is described thereafter.

Estimating $r_{\min(\theta^M,Z)}$ and $r_{\max(\theta^M,Z)}$. To calculate the minimum and maximum change of $r(\theta^T, Z)$, $r_{\min(\theta^M,Z)}$ and $r_{\max(\theta^M,Z)}$ first need to be computed. Because established rules for such computations already exist for manifest variables, they are provided first. They are easily transferable to the latent variable context, considering that the latent variable is typically measuring a single, unidimensional trait and each person can be assigned their respective trait level.

For the manifest context, let Y^F represent θ^F, Y^M represent θ^M, and Y^T represent θ^T. Given the three manifest variables Y^F, Y^M, and Z with known correlations between $r(Y^F, Z)$ and $r(Y^F, Y^M)$, the minimum and maximum $r(Y^M, Z)$, and, subsequently, the minimum and maximum $r(Y^T, Z)$, can be computed under the condition that the partial correlation between Y^M and Z is constricted to the range between -1 and 1, that is,

$$-1 \leq pr_{Y^M Z} = \frac{sr_{Y^M Z}}{\sqrt{1 - r^2_{Y^F Z}}} \leq 1, \tag{1}$$

where $sr_{Y^M Z}$ is the semi partial correlation between Y^M and Z,

$$sr_{Y^M Z} = \frac{r_{Y^M Z} - r_{Y^M Y^F} r_{Y^F Z}}{\sqrt{1 - r^2_{Y^M Y^F}}}. \tag{2}$$

Equation 1 can be solved for $r_{Y^M Z}$, with $pr_{Y^M Z} = -1$ and $pr_{Y^M Z} = 1$, respectively. Thus, the minimum and maximum correlation between Y^M and Z is given by

$$r_{\min(Y^M,Z)} = -\left(\sqrt{1 - r^2_{Y^M Y^F}}\right)\sqrt{1 - r^2_{Y^F Z}} + r^2_{Y^M Y^F} r^2_{Y^F Z}, \tag{3}$$

$$r_{\max(Y^M,Z)} = \left(\sqrt{1 - r^2_{Y^M Y^F}}\right)\sqrt{1 - r^2_{Y^F Z}} + r^2_{Y^M Y^F} r^2_{Y^F Z}. \tag{4}$$

Estimating $r_{\min(\theta^T,Z)}$ and $r_{\max(\theta^T,Z)}$. As this study's main interest does not lie in the minimum and maximum correlation between Z and θ_M but between Z and θ^T, the next step involves calculating the minimum and maximum correlation between Z and Y^T. To obtain this minimum and maximum correlation, the minimum and maximum covariance between Z and Y^T were calculated first. Note that the items measuring θ^T are the items measuring θ^F plus the items measuring θ^M. In the manifest case, Y^T can be expressed as a function of Y^F and Y^M, such that $Y^T = Y^F + Y^M$. The covariance between an aggregated variable and a third variable can generally be calculated as

$$\operatorname{cov}(X + Y, Z) = \operatorname{cov}(X, Z) + \operatorname{cov}(Y, Z). \tag{5}$$

The variance of the aggregated variable is given by

$$\operatorname{var}(X + Y) = \operatorname{var}(X) + \operatorname{var}(Y) + 2\operatorname{cov}(X, Y). \tag{6}$$

Keep in mind, however, that the latent variables θ^M and θ^F differently contribute to θ^T, since the number of items varies between the two latent variables. To put Y^M and Y^F on the same scale, weight w is applied, which is calculated by dividing the number of misfitting items m_{ij} by the number of fitting items x_{ij}, to the standard deviation of Y^M.

$$SD\left(Y^M\right)_{\text{adj}} = SD\left(Y^M\right)w = SD\left(Y^M\right)\frac{\text{Number of misfitting items}}{\text{Number of fitting items}}. \tag{7}$$

The minimum and maximum $\text{cov}(Y^{\text{T}}, Z)$ (see Equation 5) is then given by

$$\text{cov}_{\min(Y^{\text{T}},Z)} = r_{\min(Y^{\text{M}},Z)} SD\left(Y^{\text{M}}\right)_{\text{adj}} SD(Z) + r\left(Y^{\text{F}}, Z\right) SD\left(Y^{\text{F}}\right) SD(Z), \tag{8}$$

$$\text{cov}_{\max(Y^{\text{T}},Z)} = r_{\min(Y^{\text{M}},Z)} SD\left(Y^{\text{M}}\right)_{\text{adj}} SD(Z) + r\left(Y^{\text{F}}, Z\right) SD\left(Y^{\text{F}}\right) SD(Z), \tag{9}$$

where the product on the left of the plus sign constitutes $\text{cov}_{\min(Y^{\text{M}},Z)}$ and $\text{cov}_{\max(Y^{\text{M}},Z)}$, respectively; the product on the right of the plus sign equals $\text{cov}(Y^{\text{F}}, Z)$.
The minimum and maximum correlation between θ^{T} and Z can then be computed as

$$r_{\min\left(\theta^{\text{T}},Z\right)} = \frac{\text{cov}_{\min(Y^{\text{T}},Z)}}{SD(Y^{\text{T}})SD(Z)}, \tag{10}$$

$$r_{\max\left(\theta^{\text{T}},Z\right)} = \frac{\text{cov}_{\min(Y^{\text{T}},Z)}}{SD(Y^{\text{T}})SD(Z)}, \tag{11}$$

where the standard deviation of Y^{T} is given by

$$SD\left(Y^{\text{T}}\right) = \sqrt{\text{var}(Y^{\text{T}})_{adj}} = \sqrt{SD(Y^{\text{T}})^2_{adj} + SD(Y^{\text{F}})^2 + 2w\text{cov}(Y^{\text{F}}, Y^{\text{M}})}. \tag{12}$$

In sum, all calculations can be realized given $\text{cov}(\theta^{\text{F}}, \theta^{\text{M}})$, $\text{cov}(\theta^{\text{F}}, Z)$, $SD(\theta^{\text{M}})$, $SD(\theta^{\text{F}})$, and $SD(Z)$. Furthermore, letting Z be a standardized variable with $SD(Z) = 1$, only $cov(\theta^{\text{F}}, \theta^{\text{M}})$, $SD(\theta^{\text{M}})$, and $SD(\theta^{\text{F}})$ need to be estimated in order to establish $r_{\min(\theta^{\text{T}},Z)}$ and $r_{\max(\theta^{\text{T}},Z)}$ for all possible values of $\text{cov}(\theta^{\text{F}}, Z)$. Thus, the only model to estimate is a two-dimensional Rasch model, where θ^{F} and θ^{M} constitute the first and second dimension, respectively (see Figure 1).

The likelihood equation for the between-item multidimensional Rasch model is given by

$$L = \prod_{j=1}^{N} \prod_{i=1}^{I_x} p\left(x_{ij} | \theta_j^{\text{F}}, \beta_i\right) \prod_{j=1}^{N} \prod_{i=1}^{I_m} p\left(m_{ij} | \theta_j^{\text{M}}, \delta_i\right), \tag{13}$$

where β_i and δ_i are the item difficulty parameters for the fitting and misfitting items, respectively.

Take, for example, a test where five out of 20 items show misfit according to an arbitrary item fit index. The two-dimensional Rasch model would include the 15 fitting items measuring the first dimension, θ^{F}, and the five misfitting items measuring the second dimension, θ^{M}. Using the estimated parameters $\text{cov}(\theta^{\text{F}}, \theta^{\text{M}})$, $SD(\theta^{\text{M}})$, and $SD(\theta^{\text{F}})$, $r_{\min(\theta^{\text{T}},Z)}$ and $r_{\max(\theta^{\text{T}},Z)}$ can be calculated for each possible $r(\theta^{\text{F}}, Z)$, thus supplying the boundaries of the minimum and maximum change in the parameter of interest—that is, the correlation coefficient—if the five items that were formerly identified as misfitting are included. The sizes for $r_{\min(\theta^{\text{T}},Z)}$ and $r_{\max(\theta^{\text{T}},Z)}$ are good indicators as to how much the standardized regression coefficient potentially changes due to the inclusion of the misfitting items.

Approach for 2PL Model

The previously described method for obtaining the minimum and maximum change in the correlation coefficient when misfitting items are included in the model is easily transferable to 2PL models. Instead of estimating a between-item multidimensional Rasch model, a between-item multidimensional model that allows for varying item discrimination parameters should be used (see, for example, Adams & Wu, 2007; Muraki, 1992). Such a model is identified by fixing the

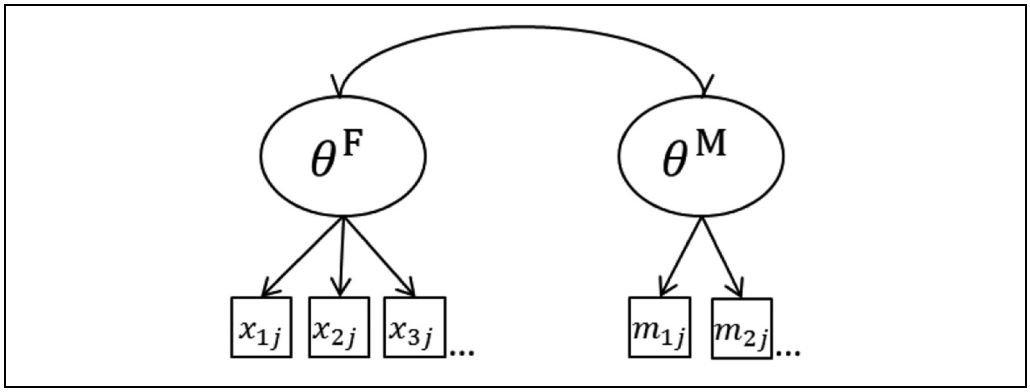

Figure 1. Between-item multidimensional IRT model with item indicators from fitting items, x_{ij}, loading on θ^F, and item indicators from misfitting items, m_{ij}, loading on θ^M.
Note. IRT = item response theory.

variance of both latent dimensions to 1. The relevant parameters for calculating $r_{\min(\theta^T, Z)}$ and $r_{\max(\theta^T, Z)}$ are thus the discrimination parameters of all items and $\text{cov}(\theta^F, \theta^M)$. The discrimination parameter estimates are necessary to obtain the weight w. Compared with the Rasch model, where each item equally contributes to measuring the latent variable, the contribution from each item in the 2PL model is determined by its discrimination parameter. In the step of combining Y^F and Y^M, the relation of how much each item contributes to measuring θ^T is contained by adjusting $\text{SD}(Y^M)$ (see Equation 7). The weight w for the adjustment with regard to the 2PL model is computed by dividing the sum of all discrimination parameters of items m_{ij}, measuring θ^M, by the sum of all discrimination parameters of items x_{ij}, measuring θ^F. Besides the different computation of w, the computation of $r_{\min(\theta^T, Z)}$ and $r_{\max(\theta^T, Z)}$ is equal for the Rasch and the 2PL model.

Simulation

The potential change in $r(\theta^F, Z)$ depends on $r(\theta^F, Z)$ itself, the amount of misfitting items relative to the fitting items, and on $r(\theta^F, \theta^M)$. To illustrate this, four examples varying (a) the amount of misfitting items (few vs. many) and (b) the size of the correlation between θ^F and θ^M (low vs. high) were simulated. Four data sets with $N = 1{,}000$ each were generated. In the first two data sets, the number of items per dimension were $I_x = 20$ and $I_m = 4$, thus presenting an example with only few misfitting items; in the last two data sets, the number of items were $I_x = 20$ and $I_m = 20$, thus presenting an example where many items show misfit. Data Sets 1 and 3 were simulated under no correlation between θ^F and θ^M; Data Sets 2 and 4 were simulated under $r(\theta^F, \theta^M) = .8$. The R function was subsequently used (see the Online Appendix) to calculate $r_{\min(\theta^T, Z)}$ and $r_{\max(\theta^T, Z)}$ under the Rasch model for 11 equally spaced values between $r(\theta^F, Z) = -1.0$ and $r(\theta^F, Z) = 1$.

The results of these analyses are displayed in Figure 2. The possible change in $r(\theta^F, Z)$ is highest for $r(\theta^F, Z) = 0$, and decreases as the correlation between θ^F and Z increases. A comparison between the top and the bottom row of Figure 2 shows that the possible change in $r(\theta^F, Z)$ is greater for higher amounts of misfitting items. The comparison between the first and the second column of Figure 2 illustrates that the possible change in $r(\theta^F, Z)$ decreases as the correlation between θ^F and θ^M increases. However, this effect is much weaker in the condition with

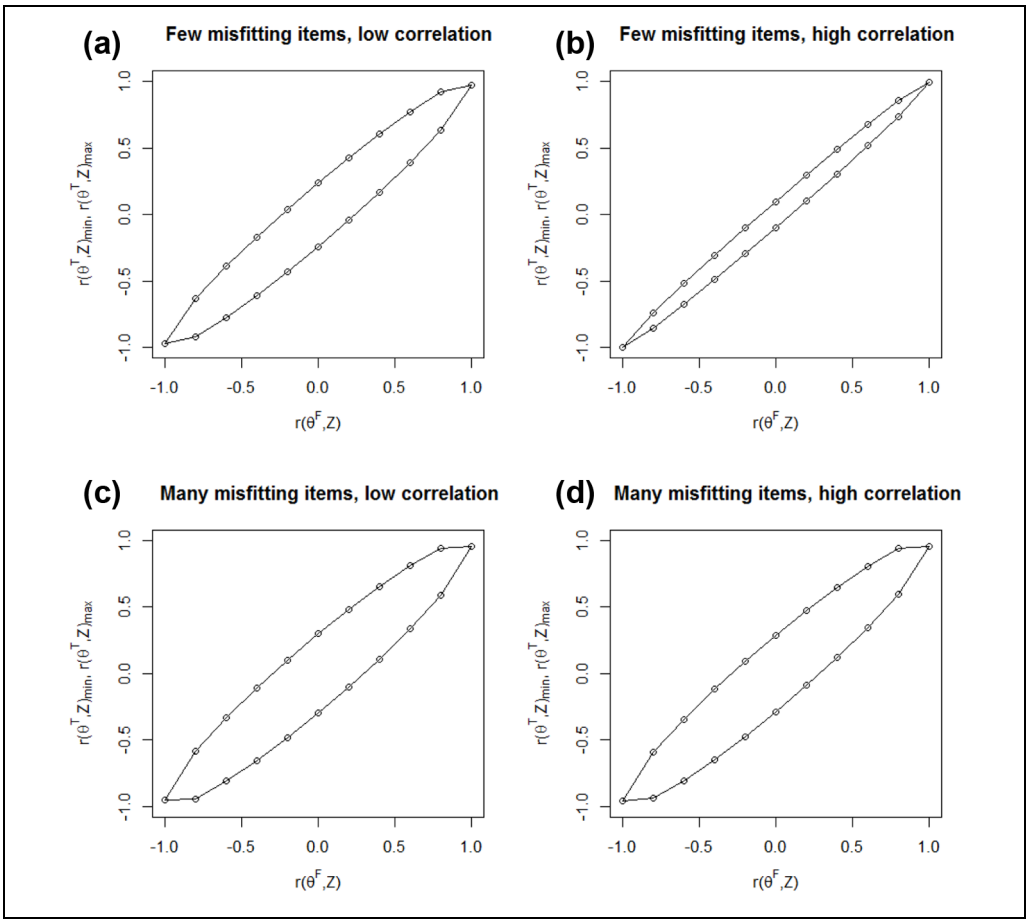

Figure 2. Minimum and maximum change in the regression coefficient when misfitting items are included in the measurement model for varying levels of the amount of misfitting items and the size of correlation between the latent variable containing only fitting items, θ^F, and the latent variable containing the misfitting items, θ^M.

many misfitting items. This indicates that the size of the correlation becomes a less important factor when the amount of misfitting items is large.

Data Example

Method

Data from a pilot study were used to illustrate the applicability of the proposed method. The study was developed to assess different competence areas of German and English as a foreign language of ninth graders in Germany (DESI-Konsortium, 2008). The present data example, the subdomain German Communication and Argumentation, consisted of 28 dichotomously and polytomously scored items. The sample size in the pilot study comprised $N = 529$ students.

In a first step, a Rasch model including all items was estimated and the weighted mean square (WMNSQ) item fit indices were calculated using the package TAM (Kiefer, Robitzsch, & Wu, 2014) in the open source software R (R Core Team, 2016).[1] The WMNSQ is a residual based item fit statistic; Wu, Adams, Wilson, and Haldane (2007) developed it based on the Infit (see Wright & Masters, 1982). A value of 1.15 was chosen as critical for item misfit, and all items with a WMNSQ >1.15 were considered misfitting.[2] The two-dimensional Rasch model was subsequently estimated with the fitting items loading on θ^F and the misfitting items loading on θ^M. After estimating the weight w and choosing 11 equally spaced values between $r(\theta^F, Z) = -1$ and $r(\theta^F, Z) = 1$, w, $r(\theta^F, Z)$, and the estimated parameters $\mathrm{cov}(\theta^F, \theta^M)$, $SD(\theta^M)$, and $SD(\theta^F)$ were inserted into Equations 3, 4, and 7 to 12. The resulting $r_{\min(\theta^T, Z)}$ and $r_{\max(\theta^T, Z)}$ inform about the minimum and maximum change of the 11 arbitrarily chosen values for $r(\theta^F, Z)$. In a last step, Mplus 7.4 (Muthén & Muthén, 2012) was used to calculate the standardized regression coefficient $r(\theta^F, Z)$ when regressing ability on 10 covariates from the data sample, estimated $r_{\min(\theta^T, Z)}$ and $r_{\max(\theta^T, Z)}$ for these 10 regression coefficients, and finally estimated the respective $r(\theta^T, Z)$ to examine whether the standardized regression coefficient when including the missing values in the measurement model actually lay within the predicted intervals. Four dichotomous covariates (*gender, possession of own room, any books read within the last quarter, currently reading a book*), three ordinal covariates (*books for Christmas, asking about an unfamiliar word, parental expectations for graduate degree*), which were rescaled into dichotomous variables by collapsing the first two and final two categories, respectively, as well as three continuous covariates (*learning for German, attitude toward reading, reading for fun*) were investigated. To obtain the scale scores for the continuous covariates, the ordinal items measuring the respective construct by calculating their mean were combined. Examinees with a missing value on any of the 10 covariates were excluded from the analysis, which resulted in $N = 396$ students.[3]

Results

Under the Rasch model, the WMNSQ of four out of the 28 items measuring German Communication and Argumentation exceeded 1.15. A two-dimensional between-item model with all 24 fitting items loading on θ^F, and the four misfitting items loading on θ^M was estimated (see Figure 1 and Equation 13). The two dimensions correlated at $r(\theta^F, \theta^M) = .027$, with a covariance of $cov(\theta^F, \theta^M) = 0.003$, and standard deviations $SD(\theta^F) = 0.829$ and $SD(\theta^M) = 0.126$. Based on these estimates, the minimum and maximum $r(\theta^T, Z)$ for each $r(\theta^F, Z) = -1.0, -0.8, \ldots, 1$ were calculated. For this purpose, Z was assumed to be standardized, with $SD(Z) = 1$. According to Equations 3 and 4, the minimum and maximum correlation between Y^M and Z, with $r(\theta^F, \theta^M)^2 = .027^2 = .001$, and $r(\theta^F, Z)^2 = -1.0^2, -0.8^2, \ldots, 1^2$ were first computed. According to Equations 7 to 12, the adjusted standard deviation $SD(Y^M)_{adj}$, the minimum and maximum covariances between Y^M and Z, $SD(Y^T)$, and finally the minimum and maximum correlation between the latent ability variable measured by all items, θ^T, and the covariates Z were calculated. The results are displayed in Figure 3. The figure also shows the standardized regression coefficient when regressing ability on the 10 covariates for both the model excluding (x axis) and the model including the misfitting items (y axis). As is evident from the figure, the estimated coefficients $r(\theta^T, Z)$ lay within the computed minimum and maximum boundaries $r_{\min(\theta^T, Z)}$ and $r_{\max(\theta^T, Z)}$.

Note that the minimum and maximum possible change in the standardized regression coefficient when the misfitting items were included was rather small. The possible discrepancy between $r(\theta^T, Z)$ and $r(\theta^F, Z)$ was greatest for no correlation and small to medium sized correlations between the explanatory variable and the latent ability. Overall, the potential changes of

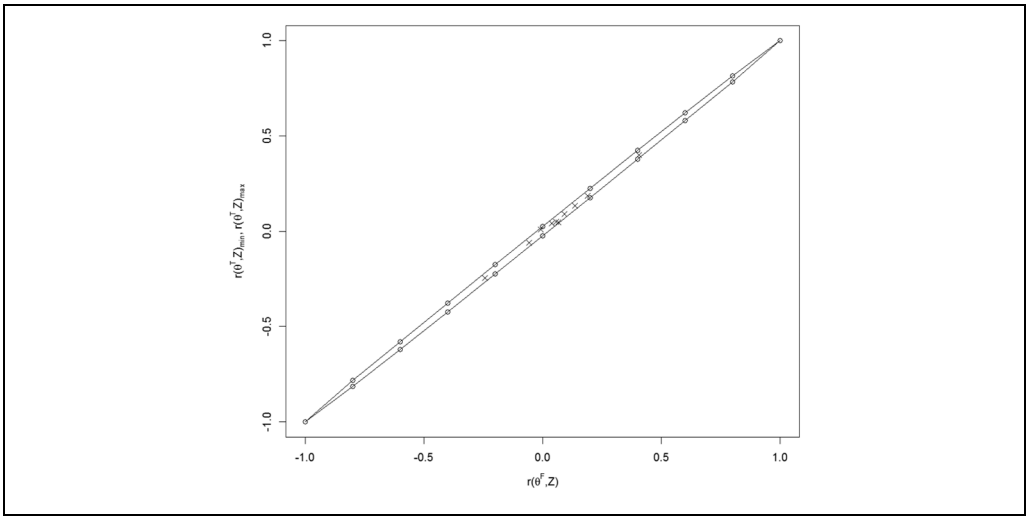

Figure 3. Change in regression coefficient when misfitting items are included in the measurement model for German communication and argumentation.

Note. The points represent the minimum and maximum change of $r(\theta^F, Z)$. The lines connect the 11 calculated $r_{\min(\theta^T, Z)}$ (bottom line) and $r_{\max(\theta^T, Z)}$ (top line), respectively. The crosses mark the estimated $r(\theta^F, Z)$—on the *x* axis—and $r(\theta^T, Z)$—on the *y* axis—when regressing competence on the ten covariates.

the standardized regression coefficient were quite small and the inclusion of the misfitting items seems rather negligible: Inferences drawn from regressing ability in German Communication and Argumentation on covariates remain the same when misfitting items are included in the latent regression model.

Discussion

The goal of this article was to introduce a method that assists in determining practical significance of item misfit in educational low-stakes large-scale assessments. Testing the substantial consequences of item misfit should be an integral part of assessing model fit (Hambleton & Han, 2005). The proposed method is based on basic mathematical principles regarding correlations, and can be applied routinely to any sort of test that involves analyses of relationships. Thus far, hardly any approaches assessing practical significance of item misfit existed.

Compared with the approach by van Rijn et al. (2016), who specifically compared relevant outcomes when misfitting items were either included in the measurement model or not, a major advantage of the proposed method lies in its generalizability. In the R function (see the Online Appendix), only the item response data, the misfitting items, and the IRT model need to be specified, and it returns the minimum and maximum potential change in the correlation coefficient for possible correlations between -1 and 1. These potential changes apply to any variable that might be of interest. This is especially valuable for trial administrations of tests (i.e., pretests). In the pretest, relevant covariates are not always part of the assessment. Using the proposed method, potential consequences of misfitting items on contextual analyses can be evaluated nevertheless. It also works in large-scale assessments with a multi-matrix data sampling approach, since parameters of an IRT model are well approximated even if some items are missing by design. This also holds for the proposed approach, which is based on a multidimensional latent regression IRT model. As long as the misfitting items from different booklets

measure the same trait (e.g., reading literacy), they can be grouped together and the influence of the misfitting items when regressing reading literacy on a covariate can be computed. Furthermore, the approach is generalizable to any scenario where researchers are interested in a potential change of the strength or direction of the relationship between two variables when the scope of the items measuring the latent variable changes. The potential change pertains to the parameter estimate *given* the respective latent measurement model the researcher chose to answer his research question with. Keep in mind that the calculated minimum and maximum values should be considered the worst cases, meaning that the (misfitting) items added to measuring the construct of interest differently relate to the covariate than the rest of the items. In cases where the items add no additional information to measuring the latent variable and only produce measurement error, their consequence on the estimated relationship with a covariate is limited.

The gain of retaining items that were pronounced misfitting depends on the purpose of the test. Certainly, an item that only produces irrelevant noise in the data, that has been translated improperly, or that simply lies outside the examinees' ability range should be altered or removed from the test. The authors neither promote that investigating why items have a poor item fit becomes unnecessary, nor do they intend for their method to replace any of the existing methods. In some situations, however, the reason for item misfit is unclear and the test developers might be reluctant to delete an item for reasons of construct representation. The proposed method allows examining how the misfit influences relevant outcomes, thus giving an additional option of evaluating item and model fit. Another criterion for evaluating whether the items can be kept in the measurement model is to compare the reliability of the test when misfitting items are included in the test or not. A decrease in reliability after the removal of misfitting items could be regarded as a reason to keep them despite of the misfit.

An important finding from the empirical example is the robustness of the correlation coefficient against violations of model fit. Not only was the potential change of the standardized regression coefficient rather small, but the actual change when misfitting items were included in the model was even lower. Certainly, these results are restricted to the presented data example. As the simulated data examples show, the potential change might be greater for tests with relatively larger amounts of misfitting items and more dissimilarity between misfitting and fitting items. In cases where the potential change is large, keep in mind that the potential change should be considered the worst case scenario and that the actual parameter change lies somewhere in between the calculated boundaries. Therefore, large potential changes do not necessarily mean that the misfit is practically significant for all possible research questions, but makes it more likely. If a researcher wants to know the actual practical significance regarding a specific research question, the relevant outcome parameters need to be compared when the misfitting items are included in the measurement model or not.

So far, the presented approach is only applicable for bivariate analyses in which the explaining variable is either continuous or binary. This limits statements regarding multiple group comparisons, which are typically relevant in large-scale assessments such as NAEP or PISA. Furthermore, it would be interesting to apply the approach to more complex research questions which require, for example, multilevel models, latent multiple regression models, or multidimensional models. For other study designs such as computer adaptive testing (CAT), item fit might play a more crucial role. Practical significance for CAT goes beyond changes in the parameters that measure relationships, as the reliability and validity of a single item or several items play a role in which items will be presented to the individual. Thus, item misfit needs to be evaluated in terms of changes in item presentation when misfitting items remain in the test, and whether this change has an effect on crucial outcomes.

Finally, note that the presented method is not restricted to the area of item fit. It is a general method that allows estimating the minimum and maximum possible change in a correlation

coefficient (or standardized regression coefficient) if some of the items are kept in the measurement model. It can thus be applied in any scenario where decisions on dropping items from a test have to be made.

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.

Supplemental Material

The online appendix are available at http://journals.sagepub.com/doi/suppl/10.1177/0146621617692978.

Notes

1. The same analyses were conducted using the two-parameter logistic (2PL) model, calculating the weight according to the estimated discrimination parameters. As the results hardly differed, only the analyses with the Rasch model were reported here.
2. The chosen cutoff score is arbitrary. The literature gives no definite answer on an adequate cutoff score for the weighted mean square (WMNSQ; see, for example, Linacre, 2003; Wu, 1997).
3. Certainly, listwise deletion is one of the least favorable methods for dealing with missing values. As our study refrains from any substantive claims and the data purely serve to demonstrate the proposed method, the simplest missing data approach was used.

References

Adams, R., & Wu, M. (2007). The mixed-coefficients multinomial logit model: A generalized for of the Rasch model. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 57-75). New York, NY: Springer.

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

Ames, A. J., & Penfield, R. D. (2015). An NCME instructional module on item-fit statistics for item response theory models. *Educational Measurement: Issues and Practice*, *34*, 39-48. doi: 10.1111/emip.12067

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395-479). Reading, MA: Addison-Wesley.

Box, G. E. P., & Draper, N. R. (1987). *Empirical model-building and response surfaces*. New York, NY: Wiley.

DESI-Konsortium. (Ed.). (2008). *Unterricht und Kompetenzerwerb in Deutsch und Englisch: Ergebnisse der DESI-Studie* [Learning and instruction of German and English. Results from the DESI study]. Weinheim, Germany: Beltz.

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

Hambleton, R. K., & Han, N. (2005). Assessing the fit of IRT models to educational and psychological test data: A five step plan and several graphical displays. In W. R. Lenderking & D. Revicki (Eds.), *Advances in health outcomes research methods, measurement, statistical analysis, and clinical applications* (pp. 57-78). Washington, DC: Degnon Associates.

Kiefer, T., Robitzsch, A., & Wu, M. (2014). *TAM: Test analysis modules* (R package Version 1.5-2). Retrieved from https://cran.r-project.org/web/packages/TAM/

Linacre, J. M. (2003). Rasch power analysis: Size vs. significance: Infit and outfit mean-square and standardized chi-Square fit statistic. *Rasch Measurement Transactions*, *17*, 918. Retrieved from http://rasch.org/rmt/rmt171n.htm

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

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159-176.

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Author.

Organisation for Economic Co-Operation and Development (2012). *PISA 2009 technical report*. Paris, France: OECD Publishing.

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

Pohl, S., & Carstensen, C. H. (2012). *NEPS technical report: Scaling the data of the competence tests* (NEPS Working Paper No. 14). Bamberg, Germany: Nationales Bildungspanel, Otto-Friedrich–Universität. Retrieved from https://www.neps-data.de/Portals/0/Working%20Papers/WP_XIV.pdf

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research (Expanded edition published by University of Chicago Press, 1980).

Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a Bayesian approach. *Journal of Educational Measurement*, *42*, 375-394.

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

Swaminathan, H., Hambleton, R. K., & Rogers, H. J. (2006). Assessing the fit of item response theory models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Psychometrics* (Vol. *26*. pp. 683-718.). Amsterdam: Elsevier.

van Rijn, P. W., Sinharay, S., Haberman, S. J., & Johnson, M. S. (2016). Assessment of fit of item response theory models used in large-scale educational survey assessments. *Large-scale Assessments in Education*, *4*, Article 10. doi:10.1186/s40536-016-0025-3

Wainer, H., & Thissen, D. (1987). Estimating ability with the wrong model. *Journal of Educational Statistics*, *12*, 339-368.

Wright, B. D., & Masters, G. N. (1982). *Rating scale Analysis: Rasch measurement*. Chicago: MESA Press.

Wu, M. L. (1997). *The development and application of a fit test for use with marginal maximum likelihood estimation and generalized item response models*. Unpublished Masters Dissertation, University of Melbourne.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *Australian Council for Educational Research (ACER) ConQuest Version 2.0: Generalised item response modeling software*. Victoria, Austalia: ACER Press.

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