Köhler, Carmen; Robitzsch, Alexander; Hartig, Johannes

A bias corrected RMSD item fit statistic. An evaluation and comparison to alternatives

Journal of educational and behavioral statistics 45 (2020) 3, S. 251-273

Quellenangabe/ Reference:
Köhler, Carmen; Robitzsch, Alexander; Hartig, Johannes: A bias corrected RMSD item fit statistic. An evaluation and comparison to alternatives - In: Journal of educational and behavioral statistics 45 (2020) 3, S. 251-273 - URN: urn:nbn:de:0111-pedocs-216818 - DOI: 10.25656/01:21681

https://nbn-resolving.org/urn:nbn:de:0111-pedocs-216818
https://doi.org/10.25656/01:21681

Article

Journal of Educational and Behavioral Statistics
2020, Vol. 45, No. 3, pp. 251–273
DOI: 10.3102/1076998619890566
Article reuse guidelines: sagepub.com/journals-permissions

A Bias-Corrected RMSD Item Fit Statistic: An Evaluation and Comparison to Alternatives

Carmen Köhler
DIPF | Leibniz Institute for Research and Information in Education

Alexander Robitzsch
Leibniz Institute for Science and Mathematics Education (IPN)
Centre for International Student Assessment (ZIB)

Johannes Hartig
DIPF | Leibniz Institute for Research and Information in Education

Testing whether items fit the assumptions of an item response theory model is an important step in evaluating a test. In the literature, numerous item fit statistics exist, many of which show severe limitations. The current study investigates the root mean squared deviation (RMSD) item fit statistic, which is used for evaluating item fit in various large-scale assessment studies. The three research questions of this study are (1) whether the empirical RMSD is an unbiased estimator of the population RMSD; (2) if this is not the case, whether this bias can be corrected; and (3) whether the test statistic provides an adequate significance test to detect misfitting items. Using simulation studies, it was found that the empirical RMSD is not an unbiased estimator of the population RMSD, and nonparametric bootstrapping falls short of entirely eliminating this bias. Using parametric bootstrapping, however, the RMSD can be used as a test statistic that outperforms the other approaches—infit and outfit, $S - X^2$—with respect to both Type I error rate and power. The empirical application showed that parametric bootstrapping of the RMSD results in rather conservative item fit decisions, which suggests more lenient cut-off criteria.

Keywords: *item fit; item response theory; educational measurement; bootstrap*

Applying item response theory (IRT) models to test data in order to draw inferences from the test requires testing whether the model actually fits (American Educational Research Association, American Psychological Association, & National Council for Measurement in Education, 2014). Testing model fit involves several steps, including the calculation of global model fit and local item fit statistics (Hambleton & Han, 2005). Only when the model fits the data can the estimated model parameters be reliably interpreted (Wainer & Thissen, 1987).

Various item fit statistics have been proposed in the literature. The common methods for evaluating item fit can be grouped into two types of general approaches: the chi-square approach and the likelihood-ratio approach (Ames & Penfield, 2015). The former includes, for example, Bock's χ^2 (1972), Yen's Q_1 (1981), Orlando and Thissen's $S - X^2$ (2000), and Wright and Masters's (1982) infit and outfit.[1] The latter involves, for example, McKinley and Mills's G^2 (1985) and Orlando and Thissen's $S - G^2$ (2000). All approaches are based on the computation of residuals between the observed and expected number of correct responses, and it is assumed that, under the null hypothesis, the standardized squared residuals follow a χ^2 distribution. However, there exists no theoretical basis regarding the distribution of the residuals under the null hypothesis of perfect model fit and hence no statistic for testing the null hypothesis. For most statistics, studies found inflated Type I error rates, especially for large sample sizes, as well as a lack of power to detect item misfit (Chon, Lee, & Ansley, 2013; DeMars, 2005; Glas & Suárez Falcón, 2003; Liang, Wells, & Hambleton, 2014; Orlando & Thissen, 2000; Stone & Zhang, 2003). Note that the statistics mentioned so far and the statistics investigated in the remaining article focus on detecting misfit with regard to the functional form assumed by the parametric model. Other possible violations against model assumptions are multidimensionality and local stochastic dependence, which can be tested by statistics specifically designed for this type of model violation (see, e.g., Maydeu-Olivares & Joe, 2005; Reiser, 2008).

The Population Root Mean Squared Deviation (RMSD) Fit Statistic

Another residual-based fit statistic that so far has hardly been investigated is the RMSD, which is implemented in the software *mdltm* (von Davier, 2005). Starting with the 2015 wave, *mdltm* is used for scaling the Program for International Student Assessment (PISA; Organization for Economic Cooperation and Development [OECD], 2016). It has also been used in the Program for the International Assessment of Adult Competencies (PIAAC; Yamamoto, Khorramdel, & von Davier, 2016). The RMSD serves as the criterion for determining both item fit and differential item functioning and is thus of major relevance regarding decisions of model fit with respect to the PISA data. In PISA, the cutoff criterion to identify misfitting items in the cognitive assessment is RMSD > 0.12 (OECD, 2017); in PIAAC, it is RMSD > 0.15.

In the following, it is assumed that item responses X_i, with $i = 1, \ldots, I$, follow a unidimensional item response model with item response functions (IRFs) $P_i(\theta) = P(X_i = 1 \,|\theta)$ and a standard normally distributed latent ability θ. Under local stochastic independence,

$$P(\mathbf{X} = \mathbf{x}) = P(X_1 = x_1, \ldots, X_I = x_I) = \int \prod_{i=1}^{I} \left(P_i(\theta)^{x_i} (1 - P_i(\theta))^{1-x_i} \right) w(\theta) d\theta, \quad (1)$$

where $w(\theta)$ denotes the density function of the standard normal distribution.

In most applications, parametrically modeled IRFs $P_i^*(\theta;\boldsymbol{\gamma}_i)$ are imposed (e.g., one-parameter logistic [1PL] or two-parameter logistic [2PL] IRFs; Embretson & Reise, 2000) to estimate a unidimensional item response model where a vector of item parameters $\boldsymbol{\gamma}_i$ is estimated for all items $i = 1, \ldots, I$. This parametric item response model will be typically misspecified to at least some extent. As a consequence, the true IRF $P_i(\theta)$ deviates from the estimated IRF $P_i^*(\theta;\boldsymbol{\gamma}_i)$ even in case of infinitely large samples. The IRFs $P_i^*(\theta;\boldsymbol{\gamma}_i)$ can be interpreted as quasi maximum likelihood estimates of the parametric item response model (White, 1982). The discrepancy between the two IRFs is quantified in the population RMSD statistic for item i:

$$\mathrm{RMSD}_i = \sqrt{\int \left[P_i(\theta) - P_i^*(\theta;\boldsymbol{\gamma}_i)\right]^2 w(\theta)d\theta}. \tag{2}$$

Note that in this definition of the population RMSD, the two IRFs $P_i(\theta)$ and $P_i^*(\theta;\boldsymbol{\gamma}_i)$ are unknown. Integration of the latent variable θ is accomplished by numerical integration based on a finite grid of θ values with quadrature nodes θ_t. The discrete version of the RMSD for an item i is thus defined as

$$\mathrm{RMSD}_i = \sqrt{\sum_t \left[P_i(\theta_t) - P_i^*(\theta_t;\boldsymbol{\gamma}_i)\right]^2 w_t}. \tag{3}$$

The weights w_t correspond to normalized values of the normal density at the quadrature point θ_t. They serve as a discrete prior distribution in order to calculate the area between the two IRFs as the sum of the differences between the observed and expected probability of success at each quadrature node. The two IRFs $P_i(\theta_t)$ and $P_i^*(\theta_t;\boldsymbol{\gamma}_i)$ need to be substituted by estimated functions based on the observed data, which are described in the next section.

The Estimated RMSD Fit Statistic

Basically, the RMSD measures the distance between a true and a fitted IRF. However, the true IRF $P_i(\theta_t)$ and the parametrically modelled IRF $P_i^*(\theta_t;\boldsymbol{\gamma}_i)$ need to be estimated—denoted as $\hat{P}_i(\theta_t)$ and $\hat{P}_i^*(\theta_t;\boldsymbol{\gamma}_i)$—at nodes θ_t. For computing the RMSD, estimated nonparametric IRFs $\hat{P}_i(\theta_t)$ are based on individual posterior probabilities $h_p^*(\theta_t) = P(\theta_t|\mathbf{x}_p)$ at the prespecified nodes, where p indexes the persons. These posterior distributions make use of the fitted parametric item response model and, up to a constant, can be calculated as

$$h_p^*(\theta_t) = P(\theta_t|x_{p1,}, \ldots, x_{pI}) \propto \prod_{i=1}^{I} \left(\hat{P}_i^*(\theta_t;\boldsymbol{\gamma}_i)^{x_{pi}}\left(1 - \hat{P}_i^*(\theta_t;\boldsymbol{\gamma}_i)\right)^{1-x_{pi}}\right). \tag{4}$$

The true IRF at a particular node θ_t (see also Sueiro & Abad, 2011) is estimated by

$$\hat{P}_i(\theta_t) = \frac{\sum_p x_{pi} h_p^*(\theta_t|\mathbf{x}_p)}{\sum_p h_p^*(\theta_t|\mathbf{x}_p)}. \tag{5}$$

Based on these estimated IRFs, the RMSD statistic is calculated as

$$\widehat{\text{RMSD}}_i = \sqrt{\sum_t \left[\hat{P}_i(\theta_t) - \hat{P}_i^*(\theta_t;\boldsymbol{\gamma}_i)\right]^2 w_t}. \tag{6}$$

Note that the definition of the population RMSD mirrors the definition of the root integrated squared error (RISE; see Douglas & Cohen, 2001). Computation of the RMSD statistic relies on posterior probabilities resulting from marginal maximum likelihood estimation, while the RISE statistic compares fully nonparametrically estimated IRFs with parametrically estimated IRFs.

Research Motivation and Research Questions

Statistical inference for the RMSD statistic has thus far not been investigated in detail. The RMSD values are difficult to interpret since no generally accepted cutoff value for misfit exists. The distribution of the empirical RMSD under exact model fit is unknown, and it is not yet investigated whether this distribution is affected by sample size, number of items, or the presence of misfitting items in the data. Furthermore, analogue to the goodness-of-fit statistic standardized root mean squared residual that is used to evaluate the fit of structural equation models (see Maydeu-Olivares, 2017), the RMSD statistic suffers from finite sample bias. We give a brief heuristic explanation as to why the expected sample RMSD differs from the population RMSD statistic: The sample RMSD is based on squared terms (see Equation 6). It holds that

$$\hat{P}_i(\theta_t) - \hat{P}_i^*(\theta_t) = \underbrace{[\hat{P}_i(\theta_t) - P_i(\theta_t)]}_{B_1} + \underbrace{[P_i(\theta_t) - P_i^*(\theta_t)]}_{B_2} + \underbrace{[P_i^*(\theta_t) - \hat{P}_i^*(\theta_t)]}_{B_3}, \tag{7}$$

where only B_2 contains a term that also appears in the population RMSD. In small samples, there will be essential sampling variability of the fitted parametric IRFs in the third term B_3 and in the first term B_1. Squaring these terms will result in additional sampling variability of the sample RMSD statistic. Thus, the expected value of the estimated statistic will exceed the corresponding population statistic.

The presents study investigates three research questions:

Research Question 1: How is the empirical RMSD affected by various characteristics of the data?

Research Question 2: Can the finite sample bias be corrected using nonparametric bootstrapping?

Research Question 3: Can the RMSD serve as a reliable test statistic to correctly identify item misfit by applying parametric bootstrapping methods?

Research Question 3 also entails a comparison to other, more common fit statistics.

The first simulation study (Study 1) revolves around Research Question 1 and investigates whether the empirical RMSD is sensitive to varying data conditions such as sample size, item size, and number of misfitting items in the data. In the second simulation study (Study 2), we use bootstrapping procedures to answer Research Questions 2 and 3. In general, bootstrapping methods can provide consistent estimators of the distribution of a statistic (Efron, 1979). They can further be used to obtain asymptotic refinements in order to reduce finite sample bias (Habing, 2001). Regarding Research Question 2, the nonparametric bootstrapping method is used to construct a bias-corrected RMSD ($\text{RMSD}_{\text{np.bs}}$). To answer Research Question 3, we apply the parametric bootstrap to construct the sampling distribution of the RMSD under the null hypothesis, thus obtaining critical values for evaluating item fit ($\text{RMSD}_{\text{p.bs}}$). In Study 2, we also compare the performances of the infit and outfit statistics, $S - X^2$, and RMSD using critical values obtained from the parametric bootstrap regarding their Type I error rates and power. Note that, thus far, the RISE approach is not implemented in any available software or R package, which is why it was not included in this study.

Study 1

Method

Simulation design. Study 1 was designed to investigate the characteristics of the empirical RMSD under different data conditions when the population RMSD is known. We varied the sample size (500; 5,000; 100,000), the number of items (50; 200; 500), and the number of misfitting items in the data set (1; 10; 20), resulting in 27 conditions (3 sample sizes $\times$ 3 item sizes $\times$ 3 number of misfits). The number of replications r varied with sample size so that for $N = 500$, $r = 500$; for $N = 5{,}000$, $r = 350$; and for $N = 100{,}000$, $r = 20$. Furthermore, the study was conducted separately for two types of misfit: items with a guessing parameter and items with a nonmonotone IRF. Data generation and analyses were conducted in the open-source software R (R Core Team, 2018).

Data generation. The fitting items were generated under the 2PL model (Birnbaum, 1968). The person ability parameters θ_p were drawn from a standard normal distribution $\theta \sim N(0, 1)$. The slope parameters a_i were randomly drawn from a log-normal distribution with $a_i \sim LN(0, 0.5)$; the difficulty parameters b_i were randomly drawn from a standard normal distribution $b_i \sim N(0, 1)$. To avoid extreme item parameters that might result in simulated data in which either no persons or all persons answered the item correctly, we excluded outliers and redrew a_i and b_i until the parameters lay within two standard deviations from the mean of the distributions from which they were drawn.

Two types of misfitting items were generated and investigated separately. The first type of misfitting item was generated under the three-parameter logistic (3PL) model (Birnbaum, 1968), which is given by

$$P(X_{pi} = 1|\theta_p) = c_i + (1 - c_i)\frac{\exp(a_i(\theta_p - b_i))}{1 + \exp(a_i(\theta_p - b_i))}, \tag{8}$$

where X_{pi} corresponds to the observed item responses, and c_i is the guessing parameter and the lower asymptote of the IRF. This type of IRF likely occurs in tests with multiple-choice items, where all response options seem equally plausible to examinees with low θ values. If all examinees at the lower end of the θ continuum simply took a guess at the correct response, the probability of success on that item would not fall under the guessing asymptote.

The second type of misfitting item had a nonmonotonic IRF (see Wainer & Thissen, 1987) and can be described as

$$P(X_{pi} = 1|\theta_p) = \frac{c_i}{1 + \exp[a_i(\theta_p - (b_i + d_i))]} + \frac{1}{1 + \exp[-a_i(\theta_p - b_i)]}, \tag{9}$$

where d_i is a positive number creating a dip in the IRF. This means that the probability of a correct response decreases at some point on the theta continuum. A likely scenario for this occurrence is an item where the distractors work especially well for examinees at medium theta levels, thus reducing the chance of success for this ability group. Other possibilities concern misconceptions of certain content matter that are only prevalent in respective ability groups.

In order to choose the values of $a_i, b_i, c_i,$ and d_i parameters for simulating the misfitting items, we conducted an additional preliminary simulation study. Since the size of misfit depends on the combination of the item parameters, we needed to identify which parameter constellations result in small, medium, and large item misfit. The item parameters we investigated in the preliminary simulation were: $a_i[0.2,\ 5]$ at equally spaced intervals of .05; $b_i[-3,\ 3]$ at equally spaced intervals of .05; and $c_i[0.1,\ 0.5]$ at equally spaced intervals of .1. For generating nonmonotone IRFs, $d_i[1,\ 3]$ at equally spaced intervals of .2 was used. These values were chosen in accordance with previous studies (Orlando & Thissen, 2003; Sueiro & Abad, 2011). For each of the possible parameter combinations, we compared the IRF of the generating model with the IRF approximated by the 2PL model using the true θ values. We then calculated the difference between the two curves—that is, the discrete RMSD (see Equation 3). Generating misfit under the 3PL model resulted in an RMSD range from approximately 0 to .137 and a mean of .038. When generating the nonmonotone items, the RMSD_{pop} range was from approximately 0 to .266, with a mean of .047. We therefore decided on the following definitions of the size of misfit: $\text{RMSD}_i < .02$ negligible misfit, $.02 \leq \text{RMSD}_i < .05$ small misfit, $.05 \leq \text{RMSD}_i < .08$ medium misfit, and $\text{RMSD}_i \geq .08$ large misfit.

To generate the misfitting items in Study 1, we randomly drew item parameter combinations that produced a medium misfit of about .05 in each replication. Thus, the parameter combinations differed across replications and across items.

Computation of empirical RMSD. After generating the data sets, we used the 2PL model implemented in the TAM package (Robitzsch, Kiefer, & Wu, 2017) to analyze the data and to estimate the empirical RMSD. We investigated how the RMSD performs under the 2PL model, since the major studies PISA and PIAAC, which both investigate item fit using the RMSD, scale the data under the 2PL model. The model estimation used 31 quadrature nodes from -5 to 5 for conducting the numerical integration, six M-steps for item parameter estimation, and a convergence criterion of .001 maximum change in the deviance value. In a subsequent step, the RMSD values for all fitting and all misfitting items were averaged, respectively, across all replications.

Results

Figure 1A and 1B show the empirical mean RMSD values for the fitting items in the 3PL condition and the nonmonotone condition, respectively. One of the main interesting findings is that the empirical RMSD of the fitting items only closely approximates the population RMSD of 0 in the condition with 1 misfitting item and a large sample size. This means that the RMSD of fitting items is overestimated for all $N < 100{,}000$. Besides this, the three main results worth noting are, firstly, that the empirical RMSD depends on sample size. As the sample size increases, the RMSD decreases, which is due to the reduction of the finite sample bias. Secondly, the empirical RMSD also depends on the total number of items in the data set. More items in a data set lead to higher RMSD values, which mirrors the findings of Sueiro and Abad (2011). This effect decreases as sample size increases. An important factor in the estimation of the RMSD is the ratio between the number of items and sample size. The least favorable ratio of number of persons to number of items was 500:500, which produced the highest RMSD values. Finally, the empirical RMSD hardly depends on the number of misfitting items in the data set.[2]

Figure 2 illustrates that the empirical RMSD of the misfitting items only closely approximates the population RMSD of 0.05 in the conditions with many items ($I = 500$) and a medium sample size ($N = 5{,}000$). In all other conditions, the empirical RMSD either over- or underestimates the population RMSD. For the misfitting items, the first and second main results mirror the main results we found for the fitting items, namely, that the empirical RMSD depends on sample size and that the empirical RMSD depends on the total number of items in the data set. Regarding the third main result, the empirical RMSD of the misfitting items depends to some extent on the number of misfitting items in the data set.

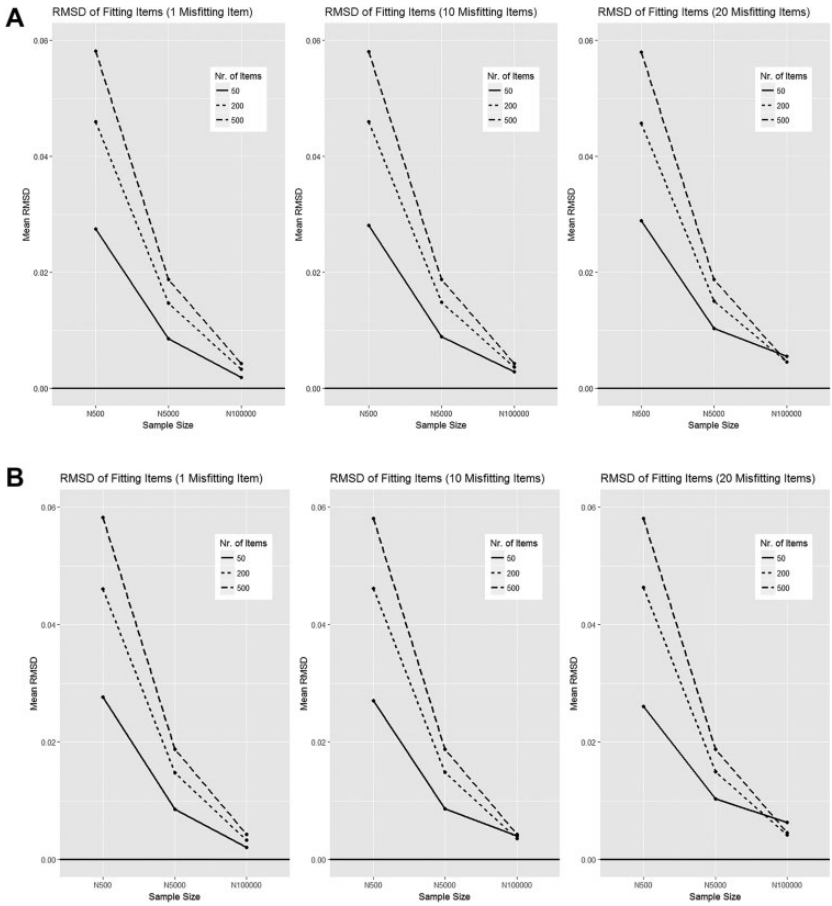

FIGURE 1. *Mean root mean squared deviation (RMSD) for fitting items when (A) generating misfit under the three-parameter logistic and (B) generating nonmonotone items. The horizontal line marks the population RMSD of the fitting items.*

When there was only one misfitting item in the data set, the RMSD performed especially badly in detecting this item when it was generated under a 3PL model in the condition with $I = 50$ and $N = 500$ and when the item had a nonmonotone IRF in the conditions with $I = 50$ and $N = 500$, with $I = 50$ and $N = 100{,}000$, and with $I = 200$ and $N = 100{,}000$.

Overall, the results demonstrate that the empirical RMSD is not an unbiased estimator of the population RMSD but largely depends on characteristics of the data set. The two main influences are sample size and the number of items in the data.

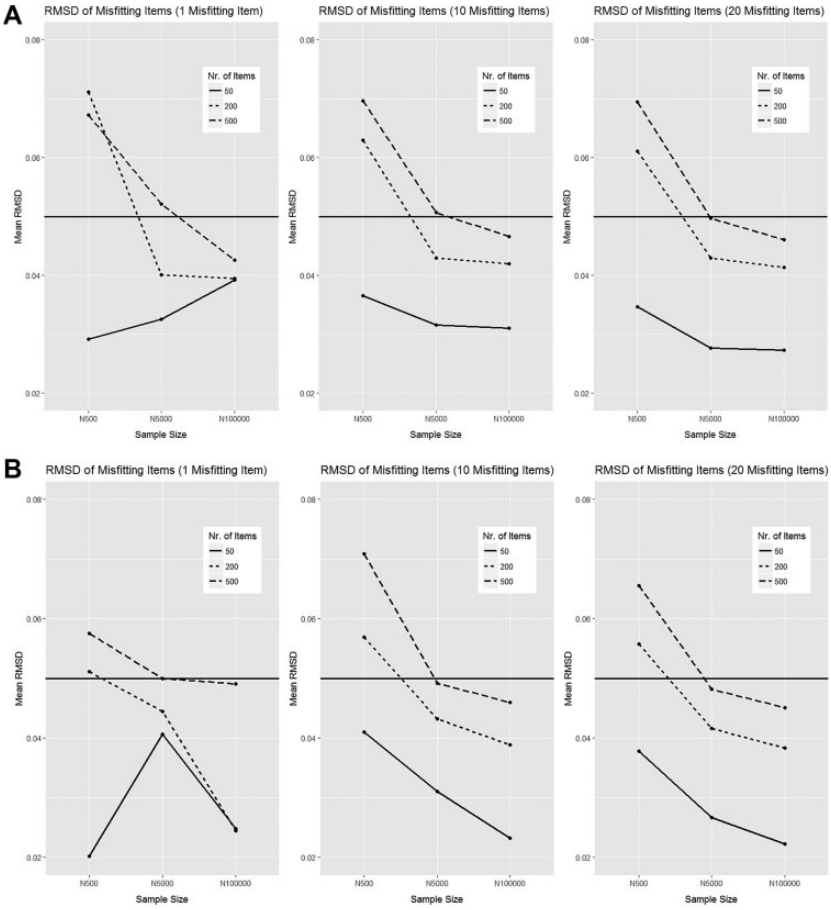

FIGURE 2. *Mean root mean squared deviation (RMSD) for misfitting items when (A) generating misfit under the three-parameter logistic and (B) generating nonmonotone items. The horizontal line marks the population RMSD of the misfitting items.*

Constructing Sampling Distribution for RMSD

Nonparametric Bootstrap

The nonparametric bootstrap can be used to construct the sampling distribution of a statistic in the underlying population of the given sample (Efron, 1979). The idea is that the empirical data represent a random sample drawn from the population distribution and that this unknown population distribution can be estimated via the empirical data. Therefore, the empirical data are treated as a

population from which samples are drawn. The general procedure for the nonparametric bootstrap involves three steps (see, e.g., Habing, 2001):

(1) Fit a parametric item response model to the observed data and estimate the sample statistic of interest, T,
(2) use Monte Carlo simulation to generate R data sets by drawing samples of size N with replacement from the empirical data set, and
(3) calculate the statistic of interest, T^*, for each of the samples in Step 2.

The bias in T equals the expected difference between T and Θ, where Θ represents the population parameter. Using the bootstrap, we can approximate this expectation so that $B^* \equiv E^*(T^* - T)$, where E^*T^* is the estimated expected value of T^*—that is, the average of the calculated T^* for each of the bootstrap samples. The bias-corrected estimator of T is given by $T - B^*$, with standard error $E^*[(T^* - T)^2]^{1/2}$. In this way, the bootstrap improves first-order asymptotic approximations and serves as a tool to reduce an estimator's finite sample bias (Horowitz, 2001). Su, Scheu, and Wang (2007), for example, proposed applying this method to construct confidence intervals around the unstandardized infit and outfit statistics. Raykov (2005) applied a bias-correction nonparametric bootstrap estimator to measures of global misfit in structural equation modeling.

Using the nonparametric bootstrap, we

(1) calculated $\widehat{\text{RMSD}}$ from the parametric scaling model for each item i,
(2) drew random answering patterns with replacement of size n from the empirical data set, resulting in $R = 200$ data sets, and
(3) calculated RMSD_b^* for each R.

The $\text{RMSD}_{\text{np.bs}}$ was estimated as

$$\text{RMSD}_{np.bs} = \widehat{\text{RMSD}} - (E^*\text{RMSD}^* - \widehat{\text{RMSD}}) = 2\,\widehat{\text{RMSD}} - E^*\text{RMSD}^*. \quad (10)$$

The term within parentheses in Equation 10 represents the bias B^* of the uncorrected RMSD estimator. Subtracting the bias from the uncorrected estimator leads to the bias-corrected estimator $\text{RMSD}_{\text{np.bs}}$.

Parametric Bootstrap

The parametric bootstrap provides the sampling distribution of a statistic under the null hypothesis and thus allows testing of whether the empirically derived value exceeds, for example, 95% of the values generated under the null hypothesis. Steps 1 and 3 were identical to the nonparametric bootstrap. Step 2 of the parametric bootstrap was as follows:

(2) Given the estimated parameters of the IRT model in Step 1, use Monte Carlo simulation to generate a large number of R data sets.

The distribution of the obtained T^* is the bootstrap distribution. Since the R data sets were constructed under the assumption of a fitting model, the distribution of T^* is the distribution of the statistic under the null hypothesis. This distribution can be used to calculate critical values by taking, for example, the 2.5th and 97.5th percentile. Stone (2000) applied parametric bootstrapping to obtain a null chi-squared distribution for testing item fit. Using a similar approach, Habing (2001) constructed significance tests for a local dependence assessment based on residual covariances of item pairs. Studies investigating different variations of the RISE statistic also employed parametric bootstrapping methods to perform hypothesis tests (Douglas & Cohen, 2001; Lee, Wollack, & Douglas, 2009; Sueiro & Abad, 2011; Wells & Bolt, 2008).

In our study, the specific procedure for obtaining the sampling distribution of the RMSD under the null hypothesis was as follows:

(1) We calculated $\widehat{\text{RMSD}}$ under the 2PL model.
(2) Using the same parametric model, the estimated item difficulty, and item discrimination parameters, $\hat{\beta}_i$ and $\hat{\alpha}_i$, and randomly drawn person parameters θ of size N from a normal distribution $\theta \sim N(0, \hat{\sigma}^2)$, we simulated $R = 200$ data sets.
(3) We calculated RMSD_b^* for each simulated data set R. Collectively, the RMSD_b^* values thus obtained approximate the sampling distribution of the $\widehat{\text{RMSD}}$ under the null hypothesis.

As the RMSD can only take positive values, we took the 95th percentile of this distribution to obtain the critical value for a one-sided test with a significance level of $\alpha = .05$.

Study 2

The second simulation study investigated the bias correction using the nonparametric bootstrap ($\text{RMSD}_{\text{np.bs}}$) and compared the performance of Orlando and Thissen's (2000) $S - X^2$, the infit and outfit proposed by Wu (1997), and the RMSD (von Davier, 2005), using parametric bootstrapping as a test statistic by conducting critical values around the RMSD ($\text{RMSD}_{\text{p.bs}}$). Note that the infit and outfit are typically applied in the context of the Rasch model (Rasch, 1960), since in the Rasch model persons' sum scores are sufficient statistics for the trait level (Ames & Penfield, 2015; Swaminathan, Hambleton, & Rogers, 2007; Wu & Adams, 2013). However, they can and have been applied to 2PL models as well, and our aim was to investigate their performance under the 2PL model.

Method

Simulation design. The test length was fixed to 50 items. The following factors were manipulated: sample size (500; 1,000; 5,000), number of misfitting items (0; 5; 15), and size of misfit (small; medium; large). As the condition with 0 misfit does not cross with size of misfit, the number of conditions was 21 (3 sample sizes $\times$ 2 conditions with different numbers of misfits $\times$ 3 sizes of misfit $+$ 3 sample sizes $\times$ 1 condition with no misfit). One thousand replications were conducted in each condition. As for Study 1, the simulation was conducted twice, using two types of misfit (guessing parameter and nonmonotone IRF).

Data generation. The parameters for the fitting items were analogous to those in Study 1. As in Study 1, the misfitting items were generated under the 3PL model (Birnbaum, 1968) and the model producing nonmonotonic IRFs (see Equation 9). To vary the size of misfit in Study 2, we randomly drew 15 item parameter combinations of $a_i, b_i, c_i,$ and d_i that, according to our preliminary study, resulted in small, medium, and large RMSD values, respectively (see Tables A1 and A2 of the Appendix in the online version of the article). In the condition with only 5 misfitting items, only the first five item parameter combinations were used to generate the item responses of the misfitting items. Note that compared to Study 1, we kept the item parameter combinations fixed across the 1,000 replications. In this way, we were able to separate variance that might be due to different item parameter combinations from variance across the replications.

Computation of item fit. Four statistics were estimated and evaluated: the correction of the RMSD using nonparametric bootstrapping ($\text{RMSD}_{\text{np.bs}}$), the weighted (infit) and unweighted (outfit) mean squares (MNSQ) fit statistics as defined by Wu (1997), Orlando and Thissen's $S - X^2$ (2000), and the RMSD (von Davier, 2005) using the parametric bootstrap to obtain critical values ($\text{RMSD}_{\text{p.bs}}$). All indices were estimated in the open-source software R (R Core Team, 2018). Infit and outfit, $S - X^2$, and the RMSD are implemented in the TAM package (Robitzsch et al., 2017); critical values using the parametric bootstrap and bias corrections of the RMSD were implemented in R by the authors. As in Study 1, we used a 2PL model with 31 quadrature nodes from -5 to 5 to conduct the numerical integration, six M-steps for item parameter estimation, and a convergence criterion of .001 maximum change in the deviance value.

Type I error rates. To examine Type I error rates of infit and outfit, $S - X^2$, and RMSD, the proportion of fitting items that were identified as misfitting at a significance level of $\alpha = .05$ was calculated in each condition. For the weighted and unweighted MNSQ, we used the typically employed cutoff criterion of 1.15 to determine whether an item showed misfit (see, e.g., OECD, 2012, 2015; Pohl & Carstensen, 2012). We also calculated the Type I error rates of the transformed infit/outfit t values. For the $S - X^2$ statistic, the empirical p values of the χ^2 test

TABLE 1.
Mean RMSD and $RMSD_{np.bs}$ of the Fitting Items in the Conditions With Only Fitting Items

N	RMSD	$RMSD_{np.bs}$
500	.038	.023
1,000	.027	.016
5,000	.012	.007

Note. RMSD = root mean squared deviation; np.bs = nonparametric bootstrap.

were used. The critical values obtained from the parametric bootstrapping procedures were employed to evaluate the statistical significance of the RMSD, denoted as $RMSD_{p.bs}$.

Power. The power to detect item misfit was estimated by calculating the proportion of correctly detected misfitting items across the replications in each of the conditions containing misfitting items. The same cutoff criterion and significance levels as for estimating Type I error rates were used.

Results

Bias reduction using nonparametric bootstrap. As is evident from Tables 1 and 2, the nonparametric bootstrap correction $RMSD_{np.bs}$ was smaller than the RMSD in each condition, indicating that the reduction of the finite sample bias was successful. However, the $RMSD_{np.bs}$ values of the fitting items still significantly exceeded 0, thus overestimating the population RMSD. As in Study 1, the empirical RMSD depended on sample size; also, the empirical RMSD of the misfitting items was either under- or overestimated in most conditions. The bias reduction using nonparametric bootstrapping is thus not efficient in recovering the population RMSD in order to use it as a form of an effect size. The nonparametric bootstrap procedure should also not be used to determine exact inference—that is, to evaluate whether or not an item shows misfit.

Type-I error rates in fit condition. Regarding the different performances of the fit statistics for exact inference, Table 3 shows that Type I error rates for infit/outfit were deflated, meaning that hardly any items were (incorrectly) identified as misfitting. The $S - X^2$ statistic and the RMSD using the critical values from the parametric bootstrapping method ($RMSD_{p.bs}$) showed acceptable Type I error rates. Note that, overall, the results were independent of sample size.

Type-I error rates in misfit condition. In the conditions where misfitting items were included in the data set, results regarding Type I error rates were similar to the results in the fit condition (see Table 4; results for the conditions with items

TABLE 2.
Mean RMSD and $RMSD_{np.bs}$ of the Fitting and Misfitting Items, Respectively, in the Conditions With Misfitting Items

			Fitting Items		Misfitting Items	
% Misfit	Size	*N*	RMSD	$RMSD_{np.bs}$	RMSD	$RMSD_{np.bs}$
10	Small	500	.027	.016	.057	.053
		1,000	.019	.012	.055	.052
		5,000	.009	.006	.053	.052
	Medium	500	.028	.016	.039	.032
		1,000	.019	.012	.035	.032
		5,000	.009	.005	.032	.031
	Large	500	.028	.016	.027	.018
		1,000	.020	.012	.021	.015
		5,000	.009	.005	.014	.012
30	Small	500	.028	.017	.052	.048
		1,000	.020	.012	.050	.048
		5,000	.011	.008	.048	.048
	Medium	500	.028	.016	.036	.030
		1,000	.020	.012	.033	.029
		5,000	.009	.006	.030	.029
	Large	500	.028	.016	.028	.018
		1,000	.019	.011	.021	.015
		5,000	.009	.005	.014	.012

Note. RMSD = root mean squared deviation; np.bs= nonparametric bootstrap.

TABLE 3.
Type I Error Rates in Fitting Item Condition

N	Infit	Infit_t	Outfit	Outfit_t	$S-X^2$	$RMSD_{p.bs}$
500	0	0	.019	.001	.073	.061
1,000	0	0	.010	.001	.079	.065
5,000	0	0	.001	.001	.076	.063

Note. t = *t* value of the infit/outfit statistics, respectively; RMSD = root mean squared deviation; p.bs = parametric bootstrap.

generated using a nonmonotone function are displayed in Table A3 of the Online Appendix). Infit/outfit showed deflated Type I error rates in all conditions. The $S-X^2$ statistic displayed acceptable Type I error rates except in the conditions with 30% misfit and medium/large sizes of misfit. The $RMSD_{p.bs}$ method performed similarly to the $S-X^2$ statistic and also showed acceptable results. It increased as sample size and the number of misfitting items

TABLE 4.
Type I Error Rates in Misfitting Item Condition (Misfit Generated Using the 3PL Model)

% Misfit	Size	N	Infit	Infit_t	Outfit	Outfit_t	$S - X^2$	$RMSD_{p.bs}$
10	Small	500	.000	.000	.023	.028	.073	.085
		1,000	.000	.000	.012	.033	.079	.087
		5,000	.000	.000	.002	.043	.078	.088
	Medium	500	.000	.000	.021	.027	.075	.068
		1,000	.000	.000	.011	.031	.079	.070
		5,000	.000	.000	.001	.042	.084	.077
	Large	500	.000	.000	.022	.028	.079	.063
		1,000	.000	.000	.010	.033	.087	.070
		5,000	.000	.000	.001	.049	.092	.091
30	Small	500	.000	.000	.020	.027	.074	.095
		1,000	.000	.000	.010	.032	.076	.101
		5,000	.000	.000	.001	.043	.072	.109
	Medium	500	.000	.000	.020	.027	.087	.088
		1,000	.000	.000	.010	.033	.103	.097
		5,000	.000	.000	.001	.047	.182	.144
	Large	500	.000	.000	.026	.038	.127	.076
		1,000	.000	.000	.014	.045	.160	.096
		5,000	.000	.000	.002	.086	.385	.236

Note. t = *t* value of the infit/outfit statistics, respectively; RMSD = root mean squared deviation; p.bs = parametric bootstrap.

increased. It was thus slightly too high in the conditions with 30% misfit and large sample sizes.

Power. As is evident from Table 5, the infit and its respective *t* value hardly detected item misfit (results for the conditions with items generated using a nonmonotone function are displayed in Table A4 of the Online Appendix). The outfit detection rates lay between 20% and 37% (23% and 54% regarding outfit *t* values) for items generated under the 3PL model and between 3% and 33% (9% and 53% regarding outfit *t* values) for items generated under a nonmonotone function. These results were still unsatisfactory. The $S - X^2$ statistic had high power rates in conditions with large sample sizes and large sizes of misfit. The power to detect misfitting items in data sets with sample sizes of $N = 500$ and $N = 1{,}000$ was low to moderate. The $RMSD_{p.bs}$ method performed well in detecting misfitting items in all conditions with large sizes of misfit and in all conditions with a large sample size. The power of the $RMSD_{p.bs}$ to detect misfit was lowest in the condition with 15 misfitting items, small sizes of misfit, and $N = 500$. For the most part, an increase in the number of misfitting items had a slightly negative effect on their power to detect misfit.

TABLE 5.
Power to Detect Misfitting Items (Misfit Generated Using the 3PL Model)

% Misfit	Size	*N*	Infit	Infit_t	Outfit	Outfit_t	$S - X^2$	$RMSD_{p.bs}$
10	Small	500	.000	.000	.367	.351	0.073	0.468
		1,000	.000	.000	.363	.362	0.098	0.513
		5,000	.000	.135	.349	.389	0.230	0.770
	Medium	500	.000	.000	.266	.262	0.245	0.563
		1,000	.000	.000	.262	.305	0.368	0.812
		5,000	.000	.000	.279	.442	0.934	1.000
	Large	500	.000	.000	.220	.232	0.466	0.868
		1,000	.000	.000	.214	.280	0.709	0.992
		5,000	.000	.000	.201	.467	1.000	1.000
30	Small	500	.000	.000	.293	.248	0.075	0.381
		1,000	.000	.000	.271	.277	0.088	0.427
		5,000	.000	.020	.224	.354	0.166	0.733
	Medium	500	.000	.000	.334	.340	0.134	0.532
		1,000	.000	.000	.329	.384	0.204	0.728
		5,000	.000	.000	.325	.465	0.666	0.999
	Large	500	.000	.000	.255	.324	0.241	0.803
		1,000	.000	.000	.249	.411	0.425	0.981
		5,000	.000	.000	.255	.540	0.998	1.000

Note. t = *t* value of the infit/outfit statistics, respectively; RMSD = root mean squared deviation; p.bs = parametric bootstrap; 3PL = three-parameter logistic.

Empirical Example

We applied the item fit statistics to real PISA 2009 data to demonstrate the differences between the fit statistics in detecting misfit and to test applicability of the parametric bootstrap for the RMSD ($RMSD_{p.bs}$). Please note that the $S - X^2$ statistic is only defined for complete response data; since the PISA data contain missing item responses due to a multiple matrix design, we were unable to apply this statistic to the data. We used 88 items from reading literacy tests in Albania and 93 items from the same test in the United States. We scaled the countries separately under the 2PL model. Sample sizes were $N = 3{,}820$ for Albania and $N = 5{,}233$ for the United States.

Results showed that neither the infit nor its critical *t* value was exceeded for any of the items. The outfit indicated misfit for 1 item in Albania and for 3 items in the United States, all of which also had critical *t* values. The RMSD values in the sample from Albania had a mean of 0.017 ($SD = 0.009$; range: 0.004–0.048); in the sample from the United States, the mean was 0.019 ($SD = 0.011$; range: 0.003–0.058). According to our classification into small, medium, and large bias, 23 items showed small misfit, and there was no medium or large misfit in the Albanian sample; 27 items showed small misfit and 3 items showed medium

misfit in the U.S. sample. In terms of $RMSD_{p.bs}$, a larger number of the items were considered misfitting (38 in Albania and 55 in the United States). This means that up to half of the items (in the United States, more than half) did not conform to the 2PL model. An inspection of the expected and observed IRFs showed that the $RMSD_{p.bs}$ penalized even the slightest deviation from the expected 2PL function. Most expected IRFs only showed slight deviations, with occasional dips but mostly monotonic curves. In the two samples, the fit statistics identified different items as misfitting.

Overall, the empirical example shows that the purely statistical evaluation of misfit using the parametric bootstrap can result in rather conservative item fit decisions. Practitioners should decide on how much misfit they are willing to accept and could use the 1st percentile of the parametric bootstrap distribution, hence lowering the critical value of $\alpha = .05$ to $\alpha = .01$.

General Discussion

Detecting model aberrant items is an important step in the process of test evaluation. Many of the common item fit statistics have been criticized for their inadequate Type I error rates and their weak power to detect misfit. A more recent fit statistic, the RMSD, is currently used for assessing PISA and PIAAC data; thus far, it has hardly been investigated. In this article, we explored how the empirical RMSD is influenced by various characteristics of the data set (Study 1). Furthermore, nonparametric and parametric bootstrap procedures were applied to the RMSD to correct for finite sample bias and to obtain accurate critical values. In a second simulation study (Study 2), these approaches were compared to more common approaches—the infit and outfit proposed by Wu (1997) and Orlando and Thissen's (2000) $S - X^2$—according to their Type I error rates and power.

Results from Study 1 illustrate that the empirical RMSD is not an unbiased estimator of the population RMSD, and bias depends on characteristics of the data set. Study 2 showed that of the approaches considered for the RMSD, the parametric bootstrap yielded the most desirable results, with only slightly inflated Type I error rates and moderate to high power rates. Infit and outfit MNSQ and t values produced deflated Type I error rates and low power rates. The $S - X^2$ statistic showed slightly inflated Type I error rates, especially for large sample sizes, and acceptable power to detect item misfit for large sample sizes.

The results from Study 1 show that the empirical RMSD in a sample deviates from the population RMSD. The reason for this deviation lies in the estimation of the posterior distribution. When the true IRF contains a guessing parameter but the fitted model is a 2PL model, the true IRF will differ from the fitted parametric IRF. Estimating a person's posterior distribution based on the fitted parametric model will result in an estimate that differs from the person's individual posterior distribution based on the truly nonparametric IRF. As a result, the empirical

RMSD only approximates the population RMSD in conditions with a large number of fitting items in the data set and a vast sample size ($N = 100{,}000$ in our example). A solution that avoids estimating the RMSD based on parametrically fitted IRFs is to calculate the posterior distribution based on nonparametric IRFs (Rossi, Wang, & Ramsay, 2002) as postulated in the RISE approach (see Sueiro & Abad, 2011). The reason the RMSD depends on the number of items in the data set—a result that was also found by Sueiro and Abad (2011)—needs to be investigated in more detail.

The results from Study 2 align with and enhance prior research. The infit and outfit statistics have been shown to depend on sample size (Wu, 1997; Wu & Adams, 2013). The current findings demonstrate that the statistics also depend on the size of misfit. Since infit and outfit were primarily developed for the Rasch model, their poor performance in accurately detecting item misfit might not be surprising. It is also obvious from the results that the cutoff criterion of 1.15, which is frequently used to determine whether an item shows misfit, is too liberal and should be adapted. Whether the formulas provided by Wu and Adams (2013), which take sample size into account when calculating the infit and outfit MNSQ, also hold for the 2PL model should be investigated in future studies. Previous studies regarding the $S - X^2$ statistic also showed acceptable Type I error rates and low to moderate power to detect misfitting items (Orlando & Thissen, 2000; Wells & Bolt, 2008). The good performance using the parametric bootstrap approach for the RMSD is in line with studies that also applied this approach (Douglas & Cohen, 2001; Habing, 2001; Lee et al., 2009; Stone, 2000; Sueiro & Abad, 2011; Wells & Bolt, 2008). The nonparametric bootstrap-corrected RMSD displayed high power rates but also highly inflated Type I error rates. The inflated Type I error rates resulted from the fact that the nonparametric bootstrap correction fails to completely eliminate the finite sample bias. The expected RMSD values still significantly exceeded 0, thus producing too many false rejections of the null hypothesis.

Overall, the authors propose using the parametric bootstrap to define critical values for the RMSD. This procedure is easily implemented for simple study designs and needs more elaborate Monte Carlo simulations for generating the bootstrap samples as the complexity of the study design increases.

One limitation of the research presented is that the design of the second simulation is rather simple and fails to map closely the more complex study designs that are typically employed in large-scale assessments. PISA, for example, uses a multimatrix sampling design in which examinees are presented with only a subset of items (von Davier, Sinharay, Beaton, & Oranje, 2006). This results in large numbers of nonadministered items, which, in turn, might have an effect on the fit statistics.

Another possible enhancement of the research involves investigating more types of item misfit. In this study, misfit was generated using the 3PL model and a model for nonmonotone IRF. Other possible reasons for misfit are IRFs with

plateaus and IRFs with upper asymptotes (see, e.g., Douglas & Cohen, 2001; Lee et al., 2009; Sueiro & Abad, 2011). However, the results between the two types of misfit we investigated were almost identical, which points to a generalizability of our findings. It would also be interesting to investigate the performance of the fit statistics with several types of item misfit within the same data set. So far, most studies have examined the misfitting item types in separate simulation conditions. However, in real data settings, a mix of fitting and different types of misfitting items is more likely. The Type I error and power rates might change in the presence of different types of misfitting item.

The performance of the RMSD using parametric bootstrap methods should be further evaluated with respect to polytomous items. Also, application of the item fit indices to more complex models such as multidimensional models constitutes a relevant research area. Furthermore, normal trait distribution was assumed in our simulation. This assumption might not hold in each empirical setting and could have an effect on the fit statistics (see, e.g., Liang et al., 2014). How the fit indices perform for other types of trait distribution needs future research.

Lastly, we would like to stress that the testing of model fit involves several steps, the testing of statistical item fit being only one of them (Hambleton & Han, 2005). Fit evaluation is a multifaceted process that includes a thorough investigation of why an item has been identified as misfitting. Test developers should study these items carefully and consider whether the reasons for misfit necessitate their removal. Recent studies also examine practical consequences of item misfit when making decisions on item removal, thus taking the purpose of the test and the implications from the assessment into account (Köhler & Hartig, 2017; Liang et al., 2014; Sinharay & Haberman, 2014; van Rijn, Sinharay, Haberman, & Johnson, 2016).

Declaration of Conflicting Interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as part of the research project "Statistical and Practical Significance of Item Misfit in Educational Testing" (Grant No. KO 5637/1-1).

Notes

1. Wu (1997) proposed an alternative calculation of the infit and outfit item fit statistics within the multidimensional random coefficients multinomial logit model by Adams, Wilson, and Wang (1997). Instead of defining the fit statistics based on individual person ability estimates (e.g., the weighted

likelihood estimates; Warm, [1989]), the calculation of infit and outfit is based on individual posterior ability distributions (Wu, 1997). This approach is implemented in the software ConQuest (Wu, Adams, Wilson, & Haldane, 2007) and in the R package TAM (Robitzsch et al., 2017). Note that other software packages (e.g., the R package MIRT; Chalmers, 2012) use individual person ability estimates. The different approaches to calculating infit and outfit often lead to substantially different estimates of the fit statistics for small to moderate numbers of items.

2. Note that we conducted additional analyses to investigate the influence of the number of misfitting items in the data set. Instead of keeping the number of misfitting items in the data set fixed at 1, 10, and 20, we fixed the percentage of misfitting items at 10% and 20% of the total number of items in the data set. However, this hardly influenced the results, which is a further indication that the number of misfitting items only plays a minor role.

References

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*, 1–23. doi:10.1177/0146621697211001

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.

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

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*, 1–29. Retrieved from http://www.jstatsoft.org/v48/i06/

Chon, K. H., Lee, W., & Ansley, T. N. (2013). An empirical investigation of methods for assessing item fit for mixed format tests. *Applied Measurement in Education*, *26*, 1–15.

DeMars, C. E. (2005). Type I error rates for PARSCALE's fit index. *Educational and Psychological Measurement*, *65*, 42–50. doi:10.1177/0013164404264849

Douglas, J., & Cohen, A. (2001). Nonparametric item response function estimation for assessing parametric model fit. *Applied Psychological Measurement*, *25*, 234–243.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, *7*, 1–16.

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

Glas, C. A. W., & Suárez Falcón, J. C. (2003). A comparison of item-fit statistics for the three parameter logistic model. *Applied Psychological Measurement*, *27*, 87–106.

Habing, B. (2001). Nonparametric regression and the parametric bootstrap for local dependence assessment. *Applied Psychological Measurement*, *25*, 221–233.

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.

Horowitz, J. L. (2001). The bootstrap. In J. J. Heckman & E. Leamer (Eds.), *Handbook of econometrics* (pp. 3159–3228). Amsterdam, the Netherlands: Elsevier Science, B.V.

Köhler, C., & Hartig, J. (2017). Practical significance of item misfit in low-stakes educational assessment. *Applied Psychological Measurement*, *41*, 388–400. doi:10.1177/0146621617692978

Lee, Y.-S., Wollack, J., & Douglas, J. (2009). On the use of nonparametric ICC estimation techniques for checking parametric model fit. *Educational and Psychological Measurement*, *69*, 181–197.

Liang, T., Wells, C. S., & Hambleton, R. K. (2014). An assessment of the nonparametric approach for evaluating the fit of item response models. *Journal of Educational Measurement*, *51*, 1–17.

Maydeu-Olivares, A. (2017). Assessing the size of model misfit in structural equation models. *Psychometrika*, *82*, 533–558. doi:10.1007/s11336-016-9552-7

Maydeu-Olivares, A., & Joe, H. (2005). Limited- and full-information estimation and goodness-of-fit testing in 2^n contingency tables: A united framework. *Journal of the American Statistical Association*, *100*, 1009–1020. doi:10.1198/016214504000002069

McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement*, *9*, 49–57.

Organization for Economic Cooperation and Development. (2012). *PISA 2009 technical report*. Paris, France: Author. doi:10.1787/9789264167872-en

Organization for Economic Cooperation and Development. (2015). *PISA 2015 field trial analysis report: Outcomes of the cognitive assessment. Meeting of the Technical Advisory Group Paris (TAG(1506)1 Field Trial Cognitive Outcomes)*. Paris, France: Author.

Organization for Economic Cooperation and Development. (2016). *PISA 2015 results (volume I): Excellence and equity in education*. Paris, France: Author. doi:10.1787/9789264266490-en

Organization for Economic Cooperation and Development. (2017). *PISA 2015 technical report*. Paris, France: Author.

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

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement*, *27*, 289–298.

Pohl, S., & Carstensen, C. H. (2012). *NEPS technical report—Scaling the data of the competence tests (NEPS Working Paper No. 14)*. Bamberg, Germany: Otto-Friedrich-Universität, Nationales Bildungspanel.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Nielsen & Lydiche. (Expanded edition, 1980)

Raykov, T. (2005). Bias-corrected estimation of noncentrality parameters of covariance structure models. *Structural Equation Modeling*, *12*, 120–129.

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0. Retrieved from http://www.R-project.org/

Reiser, M. (2008). Goodness-of-fit testing using components based on marginal frequencies of multinomial data. *British Journal of Mathematical and Statistical Psychology*, *61*, 331–360. doi:10.1348/000711007X204215

Robitzsch, A., Kiefer, T., & Wu, M. (2017). *TAM: Test analysis modules* [R package Version 2.4-9]. Retrieved from http://cran.r-project.org/web/packages/TAM/index

Rossi, N., Wang, X., & Ramsay, J. O. (2002). Nonparametric item response function estimates with the EM algorithm. *Journal of Educational and Behavioral Statistics*, *27*, 291–317.

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

Stone, C. A. (2000). Monte Carlo based null distribution for an alternative goodness-of-fit test statistic in IRT models. *Journal of Educational Measurement*, *37*, 58–75.

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*, *4*, 331–352.

Su, J.-H., Scheu, C.-F., & Wang, W.-C. (2007). Computing confidence intervals of item fit statistics in the family of Rasch models using the bootstrap method. *Journal of Applied Measurement*, *8*, 190–203.

Sueiro, M. J., & Abad, F. J. (2011). Assessing goodness of fit in item response theory with nonparametric models: A comparison of posterior probabilities and kernel-smoothing approaches. *Educational and Psychological Measurement*, *71*, 834–848.

Swaminathan, H., Hambleton, R. K., & Rogers, H. J. (2007). 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, the Netherlands: 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*, 10. doi:10.1186/s40536-016-0025-3

von Davier, M. (2005). *mdltm: Software for the general diagnostic model and for estimating mixtures of multidimensional discrete latent traits models* [Computer software]. Princeton, NJ: Educational Testing Service.

von Davier, M., Sinharay, S., Oranje, A., & Beaton, A. (2006). Statistical procedures used in the National Assessment of Educational Progress (NAEP): Recent developments and future directions. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics (Vol. 26): Psychometrics* (pp. 1039–1055). Amsterdam, the Netherlands: Elsevier.

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

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427–450. doi:10.1007/BF02294627

Wells, C. S., & Bolt, D. M. (2008). Investigation of a nonparametric procedure for assessing goodness-of-fit in item response theory. *Applied Measurement in Education*, *21*, 22–40.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*, 1–25.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis. Rasch measurement*. Chicago, IL: 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 Master's dissertation). University of Melbourne, Australia.

Wu, M., & Adams, R. J. (2013). Properties of Rasch residual fit statistics. *Journal of Applied Measurement*, *14*, 339–355.

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

Yamamoto, K., Khorramdel, L., & von Davier, M. (2016). Scaling PIAAC cognitive data. In *Technical report of the survey of adult skills (PIAAC)* (2nd edition, Chapter 17). Paris, France: OECD Publishing.

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

Authors

CARMEN KÖHLER is a postdoctoral research associate at the DIPF | Leibniz Institute for Research and Information in Education, Rostocker Str. 6, 60323 Frankfurt, Germany; email: carmen.koehler@dipf.de. Her primary research interests include item response theory, evaluation of model fit, dealing with missing data, and methods of education research.

ALEXANDER ROBITZSCH is a research scientist at the Leibniz Institute for Science and Mathematics Education (IPN), Olshausenstraße 62, 24118 Kiel, Germany; email: robitzsch@ipn.uni-kiel.de. His primary research interests include national and international large-scale assessments, item response modeling, missing data, multiple imputation, and multilevel analysis.

JOHANNES HARTIG is head of the unit Educational Measurement at DIPF | Leibniz Institute for Research and Information in Education and also professor for psychology at Goethe University in Frankfurt am Main, Rostockerstr. 6, 60232 Frankfurt am Main, Germany; email: hartig@dipf.de. His primary research interests include psychometric models for educational achievement tests, context and position effects in achievement tests and questionnaires.

Manuscript received August 18, 2017
First revision received August 15, 2018
Second revision received April 16, 2019
Accepted September 18, 2019