the british psychological society
promoting excellence in psychology

British Journal of Mathematical and Statistical Psychology (2021), 74, 157–175

www.wileyonlinelibrary.com

A semiparametric approach for item response function estimation to detect item misfit

Carmen Köhler[1*], Alexander Robitzsch[2,3], Katharina Fährmann[1], Matthias von Davier[4] and Johannes Hartig[1]

[1]DIPF – Leibniz Institute for Research and Information in Education, Frankfurt, Germany
[2]IPN – Leibniz Institute for Science and Mathematics Education, Kiel, Germany
[3]Centre for International Student Assessment (ZIB), Munich, Germany
[4]National Board of Medical Examiners (NBME), Philadelphia, Pennsylvania, USA

When scaling data using item response theory, valid statements based on the measurement model are only permissible if the model fits the data. Most item fit statistics used to assess the fit between observed item responses and the item responses predicted by the measurement model show significant weaknesses, such as the dependence of fit statistics on sample size and number of items. In order to assess the size of misfit and to thus use the fit statistic as an effect size, dependencies on properties of the data set are undesirable. The present study describes a new approach and empirically tests it for consistency. We developed an estimator of the distance between the predicted item response functions (IRFs) and the true IRFs by semiparametric adaptation of IRFs. For the semiparametric adaptation, the approach of extended basis functions due to Ramsay and Silverman (2005) is used. The IRF is defined as the sum of a linear term and a more flexible term constructed via basis function expansions. The group lasso method is applied as a regularization of the flexible term, and determines whether all parameters of the basis functions are fixed at zero or freely estimated. Thus, the method serves as a selection criterion for items that should be adjusted semiparametrically. The distance between the predicted and semiparametrically adjusted IRF of misfitting items can then be determined by describing the fitting items by the parametric form of the IRF and the misfitting items by the semiparametric approach. In a simulation study, we demonstrated that the proposed method delivers satisfactory results in large samples (i.e., $N \geq 1{,}000$).

1. Introduction

In the two most recent waves of the Programme for International Student Assessment (PISA; OECD, 2017) and the Programme for the International Assessment of Adult Competencies (PIAAC; Yamamoto, Khorramdel, & von Davier, 2013), the root mean squared deviation (RMSD) has been used to identify misfitting items. In both studies, parametric item response theory (IRT) models are used to scale the data: the

*Correspondence should be addressed to Carmen Köhler, DIPF – Leibniz Institute for Research and Information in Education, Rostocker Str. 6, 60323 Frankfurt, Germany (email: carmen.koehler@dipf.de).

DOI:10.1111/bmsp.12224

two-parameter logistic (2PL) model (Birnbaum, 1968) for dichotomous items and the generalized partial credit model (Muraki, 1992) for polytomous items. The RMSD assesses whether the parametric model applied, which assumes a specific form of the item response function (IRF), is appropriate by comparing observed item responses with model-predicted item responses. This assessment of item fit is a necessary step in the process of test development to ensure adequate test inferences (American Educational Research Association, American Psychological Association, & National Council for Measurement in Education, 2014; Hambleton & Han, 2005).

For the cognitive outcomes in the PISA and PIAAC main studies, the reported cut-off values for the RMSD are 0.12 and 0.15, respectively (OECD, 2017; Yamamoto et al., 2013). For the questionnaire data, the cut-off value was 0.30 (OECD, 2017). In the PISA field trial for the cognitive outcomes, RMSD values exceeding 0.20 were considered as non-negligible deviations (OECD, 2015). In these studies, the RMSD describes the deviation between the observed IRF within a country and the predicted IRF from calibrations which use either the international or group-specific item parameter. These examples and the lack of reasoning for defining these specific values as critical cut-off values show that to date, there is no consensual RMSD value for identifying misfitting items.

Köhler, Robitzsch, and Hartig (2020) examined the performance of the RMSD in detecting deviations between the predicted 2PL model IRF and the observed IRF. Note that there are different approaches for estimating the observed IRF, which we discuss in more detail in the next subsection. Using the approach of individual posterior probabilities, the authors found that the empirical RMSD using individual posteriors depends on sample size and test length; it is larger for small sample sizes and long tests. The fact that the empirical RMSD is not an unbiased estimator of the population RMSD but strongly depends on characteristics of the data set is an undesirable property when the RMSD fit statistic is used to determine the size of misfit. The reason for this lack of consistency is that the estimated person posterior distributions are based on the fitted parametric model. The parametric model, however, might not hold for all items and thus might lead to incorrect individual posteriors. Since the computation of the RMSD statistic relies on these posterior probabilities, the estimated RMSD statistic becomes biased.

The main aim of this study is to adapt the RMSD estimation to obtain unbiased RMSD values. Instead of basing the posterior probability on the parametric model, a semiparametric approach is applied. We therefore draw on Rossi, Wang, and Ramsay (2002), who used the basis function expansion approach (Ramsay & Silverman, 2005) on logit functions. This approach is semiparametric in so far as the IRF is still parametric, but the unknown functional form of the IRF is approximated by a finite number of basis functions. The idea of using this semiparametric model is that it automatically identifies items that need a semiparametric adjustment of the IRF. Once these items are identified, the final model can include a mix of parametric and semiparametric IRF estimation, where the fitting items are forced to conform to a specific IRF. This allows for a more precise estimation of the RMSD for the misfitting items.

1.1. The RMSD and the RISE statistic

The idea underlying all approaches to quantifying the fit of an IRF is to calculate the distance between the nonparametric IRF and the parametric IRF across the examinee trait level θ. The root of the integrated square between these two functions of item i ($i = 1,\ldots, I$) quantifies the distance between them and is defined as the population RMSD item fit statistic:

$$\mathrm{RMSD} = \sqrt{\int \left[P_i(\theta) - P_i^*(\theta;\boldsymbol{\gamma}_i)\right]^2 w(\theta)\mathrm{d}\theta}, \tag{1}$$

where $P_i(\theta)$ represents the true IRF, $P_i^*(\theta;\boldsymbol{\gamma}_i)$ is the parametrically fitted IRF depending on a vector $\boldsymbol{\gamma}_i$ of item parameters, and w denotes the density function of the latent trait θ, which is usually chosen as the density of the standard normal distribution. Note that the definition of RMSD has also appeared in the literature as the root integrated squared error (RISE; Douglas & Cohen, 2001; Sueiro & Abad, 2011).

As is evident from the equations, both IRFs are unknown and need to be estimated based on the empirical data. Given a finite grid of θ values with quadrature nodes θ_t for $t = 1,\ldots,T$, the integral in the RMSD can be replaced by a discrete summation

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

where the weights w_t are normalized values of the normal density function w at θ_t. The estimated parametric IRF $\hat{P}_i^*$ is based on a unidimensional item response model fitted by marginal maximum likelihood (MML) estimation. The estimator of $P_i^*(\theta;\boldsymbol{\gamma}_i)$ is denoted by $\hat{P}_i^*(\theta,\hat{\boldsymbol{\gamma}}_i)$. Hence, a sample-based RMSD statistic can be defined as

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

Three different approaches to estimating the true IRF P_i by an estimator $\hat{P}_i$ can be distinguished in the computation of the RMSD statistic. First, the IRF P_i can be fully nonparametrically estimated. The general advantage of nonparametric approaches is their flexibility, which allows a better description of the true IRF underlying the data. Several approaches to nonparametric IRF estimation exist: kernel smoothing (Douglas & Cohen, 2001; Lee, Wollack, & Douglas, 2009; Ramsay, 1991), isotonic and smoothed isotonic regression (Barlow, Bartholomew, Bremmer, & Brunk, 1972; Lee, 2007), and approaches based on posterior probabilities (Stone, 2000; Sueiro & Abad, 2011). The estimation is similar to joint maximum likelihood estimation because estimated abilities $\hat{\theta}_p$ for persons $p = 1,\ldots,N$, are involved in the estimation, which are often computed in a first step. The most commonly used nonparametric method is kernel smoothing. Ramsay's (1991) kernel estimate for a dichotomous IRF is as follows:

$$\hat{P}_i(\theta_t) = \frac{\sum_{p=1}^N K\left(\left(\theta_t - \hat{\theta}_p\right)/\kappa\right) x_{pi}}{\sum_{p=1}^N K\left(\left(\theta_t - \hat{\theta}_p\right)/\kappa\right)}, \tag{4}$$

where $\mathbf{x}_p = \left(x_{p1},\cdots,x_{pI}\right)$ is the vector of item responses for person $p = 1,\ldots,N$, and $K = K(x)$ is the kernel function, which can be uniform, Gaussian or quadratic (Ramsay, 1991; Lee et al., 2009). The bandwidth parameter κ determines how rapidly the weights approach zero and thus determines the degree of smoothing (Ramsay, 1991). Larger values of κ decrease the variance of the estimated function at each point (Sueiro & Abad, 2011).

Second, the fitted parametric model involving all estimated item parameters $\hat{\boldsymbol{\gamma}}_i$ ($i = 1,\ldots,I$) can be used to empirically reconstruct the true IRF P_i based on individual posterior probabilities (Köhler et al., 2020; Sueiro & Abad, 2011; von Davier, 2016), which

are part of the model output from the MML estimation. Individual posterior probabilities $h_p^*(\theta_t)$ for the RMSD statistic are calculated as

$$h_p^*(\theta_t) = P(\theta_t|\mathbf{x}_p) \propto \phi(\theta_t) \prod_{i=1}^{I} \left(\hat{P}_i^*(\theta_t;\hat{\boldsymbol{\gamma}}_i)^{x_{pi}} \left(1 - \hat{P}_i^*(\theta_t;\hat{\boldsymbol{\gamma}}_i)\right)^{1-x_{pi}} \right), \tag{5}$$

where ϕ denotes the density of the standard normal distribution. The true IRF at a particular node θ_t is subsequently estimated by

$$\hat{P}_i(\theta_t) = \frac{\sum_{p=1}^{N} h_p^*(\theta_t|\mathbf{x}_p) x_{pi}}{\sum_{p=1}^{N} h_p^*(\theta_t|\mathbf{x}_p)}. \tag{6}$$

The similarity between the computation in equations (4) and (6) is evident, since the individual kernel contribution $K\left((\theta_t - \hat{\theta}_p)/\kappa\right)$ in equation (4) can be interpreted as an approximation of the individual posterior distribution $h_p^*(\theta_t|\mathbf{x}_p)$ in equation (6).

In this paper, we employ a third approach, in which we implement a semiparametric estimation using MML. Estimated IRFs $\hat{P}_i$ from the semiparametric item response model are compared with the parametrically estimated functions $\hat{P}_i^*(\theta_t;\hat{\boldsymbol{\gamma}}_i)$ in the computation of the RMSD statistic. We focus on the performance of this third approach because studies have shown that the estimated RMSD statistic based on the first and second estimation approaches depends on sample size and test length (Köhler et al., 2020; Lee et al., 2009; Sueiro & Abad, 2011). Significance tests using parametric bootstrapping have been obtained for both approaches (Köhler et al., 2020; Lee et al., 2009). Although this method has shown promising results in terms of Type I error rates and power to detect misfit, it does not provide an effect size for misfit. We use the term *effect size* in the sense of a consistent quantitative measure that informs about the magnitude of the deviation between the predicted IRF and the true IRF. Hence, the effect size quantifies the degree of item misfit. We aimed to fill this gap by investigating the third approach, which we outline in the following subsection.

1.2. Research question and aim of the study

To obtain more consistent estimates of the distance between the two IRFs, we aim to combine the parametric and a semiparametric approach. Instead of approximating the IRF of all items semiparametrically, only those items that deviate from the parametric model are fitted with semiparametric IRF. To select items for which the parametric IRF is not flexible enough, we use the statistical technique of regularization (Hastie, Tibshirani, & Wainwright, 2015). Regularization is a common technique in many areas, for example, for selecting variables in regression models (Oelker & Tutz, 2017), exploring the dimensional structure of latent variables (Sun, Chen, Liu, Ying, & Xin, 2016), assessing differential item functioning (Tutz & Schauberger, 2015), or estimating graphical networks of latent variables (Epskamp, Rhemtulla, & Borsboom, 2017).

Rossi et al. (2002) use a basis function approach for semiparametric estimation of a unidimensional item response model (see also Falk & Cai, 2016; Liang & Browne, 2015). The basic idea is that the logit transform of the item response probability is composed of a linear part that corresponds to the parametric 2PL model and a more flexible part

constructed using linear combinations of basis functions, which corresponds to the misfit. Misfitting items are detected if the part using basis function expansions contributes considerably to the IRF estimation. In this study, we investigate the optimal number of basis functions and the optimal penalty to detect which items deviate from the parametric model and hence should be approximated semiparametrically. These items are ideally the items producing misfit. After having identified these items, all others can be fitted parametrically.

Our goals are: (1) to determine whether the basis function expansion approach works well with the group lasso to identify misfitting items; (2) to investigate how to select the optimal penalty value; (3) to establish whether an application of the method results in more stable RMSD estimates of the items that actually show misfit, which might hence be used as a form of effect size. We therefore simulate data sets with varying conditions to provide the ideal number of basis functions and the selection criterion for choosing the ideal penalty value. We subsequently examine whether the basis function expansion approach produces consistent RMSD values, and compare its performance with the kernel smoothing approach and the regular RMSD approach using individual posterior probabilities.

2. Determining the number of basis functions and penalty

2.1. Semiparametric IRF estimation using B-spline functions

Using the basis function expansion approach (Ramsay & Silverman, 2005), the logit function can be described as

$$\text{logit}\, P(X_i = 1|\theta) = g_i(\theta) = \underbrace{a_i\theta + d_i}_{g_{i,\text{fit}}} + \underbrace{\sum_{k=1}^{K} \beta_{ik}\phi_k(\theta)}_{g_{i,\text{misfit}}}, \tag{7}$$

where the ϕ_k are K quadratic B-spline functions (see also Rossi et al., 2002). The 2PL model corresponds to fitting $g_{i,\text{fit}}$, whereas model misfit is represented in the basis function expansions $g_{i,\text{misfit}}$. The number K of basis functions equals the number of interior knots plus the order of the function. More basis functions lead to more inter-knot intervals and hence to more flexibility. This also means a greater computational burden, however, and requires a sufficient amount of observations within each inter-knot interval. Fewer basis functions lead to smoother fitted curves that might miss some irregularities in the logit function.

Instead of having the number of knots determining the smoothness of the fitted curve, smoothing methods and roughness penalty approaches have been developed (see, for example, Eilers & Marx, 1996; O'Sullivan, 1986; Meier, van de Geer, & Bühlmann, 2008; Ramsay, 1991; Tibshirani, 1996). Regularization techniques can be used to select only those items which deviate from the 2PL model (i.e., items which need a specification of the semiparametric term). In penalized maximum likelihood estimation (see Sun, Chen, Liu, Ying, & Tao, 2016), item parameter estimates are obtained by maximizing the difference of the log-likelihood function l and a regularization term

$$l_{\text{pen}}(\mathbf{a}, \mathbf{d}, \boldsymbol{\beta}) = l(\mathbf{a}, \mathbf{d}, \boldsymbol{\beta}) - N\lambda \sum_{i=1}^{I} \text{Pen}(\boldsymbol{\beta}_i), \tag{8}$$

where $\mathbf{a}$ and $\mathbf{d}$ are the vectors containing the item slopes and intercepts of the linear part of the IRFs, respectively. The vector $\boldsymbol{\beta}_i$ contains all basis coefficients for item i. The scalar λ

is a regularization parameter that controls the roughness of IRFs. In our study, we use the group lasso penalty (Hastie et al., 2015):

$$\text{Pen}(\boldsymbol{\beta}_i) = \sqrt{\sum_{k=1}^{K} \beta_{ik}^2}. \tag{9}$$

By multiplying this group lasso penalty with a regularization parameter λ, the penalization contribution increases as λ increases, forcing the function towards the 2PL model. Small values of λ reduce the penalization, thus allowing more roughness in the fitted IRF. The group lasso method was originally proposed to select groups of relevant variables in regression models (Yuan & Lin, 2006). In our context, the size of the penalty on an item determines for each item individually whether all B-spline coefficients are set to zero or all coefficients are estimated freely (see Tutz & Gertheiss, 2016). The method is thus in and of itself informative about which items deviate from a parametric IRF and can hence be used for model selection. Items for which all B-spline coefficients are set to zero are perfectly fitted by the 2PL model and receive an estimated RMSD of zero.

Similarly to Sun et al. (2016), the iterative EM algorithm (Dempster, Laird, & Rubin, 1977) is applied. Each iteration in the EM algorithm consists of two steps. In the E-step, the expected log-likelihood is computed with respect to the posterior distribution of the latent variable θ; in the M-step, the expected log-likelihood computed from the E-step is maximized. In the E-step, the integration with respect to the continuous latent variable θ is approximated by an evaluation of the log-likelihood function at discrete values θ_t ($t = 1,\ldots,T$). Because all item parameters are specific to each item, the M-step consists of independent maximizations with respect to item parameters for each item. For each item-specific set of parameters, the maximization boils down to a problem of regularized estimation with a group lasso for a weighted logistic regression. We employ a coordinate descent approach for maximization (Breheny & Huang, 2015; Yang & Zou, 2015) that cycles between the estimation of item intercept, item slope, and the set of parameters for the basis function of the semiparametric part of the IRF.

2.2. Selection of the number of B-splines and the regularization parameter

The size of the penalty on an item depends on the number of B-spline functions K and on the regularization parameter λ. In our application, both parameters are ideally chosen such that the parameters of the basis functions of the fitting items are all forced to zero but are freely estimated for the misfitting items. When working with penalties, the number of knots is trivial, since the amount of smoothness is controlled by the penalization. Aguilera and Aguilera-Morillo (2013) investigated the performance of B-spline approaches in functional data analysis and showed that using 5, 15 or 25 knots resulted in no substantial differences for penalized B-splines.

To determine the value of the regularization parameter, different selection criteria have been proposed. One method is to use information criteria such as the Bayesian information criterion (BIC; Schwartz, 1978), the Akaike information criterion (AIC; Akaike, 1973), and modifications or generalizations thereof (see, for example, Eilers & Marx, 1996; Lloyd-Jones, Nguyen, & McLachlan, 2018; Sun et al., 2016; Zhang, Li, & Tsai, 2010). The degrees of freedom needed in AIC and BIC calculation are given by the number of item parameters that are estimated to be different from zero.

Another common method is (generalized) cross-validation (see, e.g., Aguilera & Aguilera-Morillo, 2013; Oelker & Tutz, 2017; Zhang, Li, & Tsai, 2010). This method takes into account the predictive performance of a model (Simonoff, 1996). Instead of just focusing on minimizing the mean squared error when fitting the model to the observed data, λ can be chosen such that the model performs well in describing new data. To implement the so-called *k*-fold cross-validation, the available data are separated into *k* sets of equal sizes. The model is fitted *k* times, where the *j*th data subset is excluded from the whole data set in the *j*th fit. The predicted values are computed for the cases from the *j*th data subset to calculate a cross-validation error. In the case of IRT models, cross-validation can be computed by the predicted log-likelihood. The sum of the cross-validation errors determines a measure of overall fit of the model for a given regularization parameter λ. An optimal parameter value of λ minimizes the cross-validation error.

3. Simulation study

3.1. Design

The primary aim of the present study was to examine the performance of the approach to fitting a regularized semiparametric IRF under varying conditions, namely sample size (N = 500, 1,000 and 5,000), the number of items in the data set (I = 20, 60 and 100), and the proportion of misfitting items in the data set (0%, 10% and 30%). The values were chosen to cover a wide range of data properties found in realistic settings. Given these three manipulated factors, the number of overall conditions was $3 \times 3 \times 3 = 27$. One thousand replications were conducted in each condition.

3.2. Data generation

We generated the fitting items under the two-parameter logistic (2PL) model (Birnbaum, 1968), randomly drawing the person ability parameters θ_p from a standard normal distribution $\theta \sim N(0,1)$. The item parameters were drawn once and were then fixed for every condition and every repetition: the difficulty parameters b_i were randomly drawn from a standard normal distribution $b_i \sim N(0,1)$, and the slope parameters a_i were randomly drawn from a log-normal distribution with $a_i \sim LN(0,0.5)$. To avoid extreme item parameters that would result in data where 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 around the mean of the distribution they were drawn from. The descriptive statistics of the item parameters in the three conditions with only fitting items are listed in Table 1.

Table 1. Descriptive statistics of the fitting items in the conditions with only fitting items

I	$\overline{b}$ (SD_b)	Min_b	Max_b	$\overline{a}$ (SD_a)	Min_a	Max_a
20	−0.30 (0.69)	−1.21	1.08	1.24 (0.52)	0.45	2.06
60	−0.37 (0.80)	−1.63	1.65	1.29 (0.50)	0.33	2.18
100	−0.23 (0.79)	−1.63	1.65	1.22 (0.49)	0.30	2.18

Note. $\overline{b}$= mean of difficulty parameters, $\overline{a}$ = mean of discrimination parameters, SD = standard deviation, Min = minimum, Max = maximum

To induce some variation in the type of misfit, misfitting items were generated in two different ways. Note that in each condition with misfitting items, the proportion from both methods was the same. One type of misfitting item was generated under the three-parameter logistic (3PL) model (Birnbaum, 1968), which is given by

$$P\left(X_{pi}=1|\theta_p\right)=c_i+(1-c_i)\frac{\exp\left(a_i\left(\theta_p-b_i\right)\right)}{1+\exp\left(a_i\left(\theta_p-b_i\right)\right)}, \tag{10}$$

where X_{pi} corresponds to the observed item responses, and c_i is the guessing parameter. The other type of misfitting item was generated to describe a non-monotone IRF. The IRF in this case is given by

$$P\left(X_{pi}=1|\theta_p\right)=\frac{c_i}{1+\exp[a_i(\theta_p-(b_i+d_i))]}+\frac{1}{1+\exp[-a_i(\theta_p-b_i)]}, \tag{11}$$

where d_i is a positive number that creates a dip in the IRF (Orlando & Thissen, 2003; Wainer & Thissen, 1987). This means that at the point $\theta_p=d_i$, the probability of a correct response decreases, and persons with a higher ability have a lower success probability than people with a lower ability. Based on Köhler et al. (2020), we chose the values for a_i, b_i and c_i to generate misfit under the 3PL model, and a_i, b_i, c_i and d_i for the non-monotone IRF so that they would produce a population RMSD of 0.05, meaning that all misfitting items had a similar size of misfit, regardless of the type of misfit. A population RMSD of 0.05 can be considered small to medium (Köhler et al., 2020). We only simulated this rather small size of misfit since if the proposed method works well for small sizes of misfit, it will perform even better for larger sizes of misfit. The exact values for generating the misfitting items are given in Tables S1 and S2 in the Appendix S1.

3.3. Calculation of RMSD using three different approaches

For each of the generated data sets, we calculated (i) the RMSD with kernel smoothing as described above, using the *KernSmoothIRT* package in R (Mazza, Punzo, & McGuire, 2014), (ii) the RMSD based on individual posterior probabilities as implemented in the *TAM* package in R (Robitzsch, Kiefer, & Wu, 2020), and (iii) the RMSD with the semiparametric approach. Note that (i) and (ii) are only relevant for answering research question (3) regarding the use of the statistics as an effect size. In the following, we focus on the semiparametric approach. Research question (3) is addressed in Section 4.4.

For the semiparametric approach, we included an additional function in the TAM package (i.e., `TAM::tam.np()`) and implemented the penalized basis function expansions with group lasso for regularization of the flexible term as described above. We used basis functions of order 3, that is, piecewise quadratic. In several preliminary runs, we varied the number of B-splines. In line with findings from Aguilera and Aguilera-Morillo (2013), this had no effect on our estimates, so we fixed the number of B-splines to 4. The knots were located at −3.0, 0, 3.0. To investigate the performance of the AIC, BIC, and cross-validation (CV) to select the optimal regularization parameter λ, each data set was analysed with λ varying between 0.01 and 0.30 at equally spaced intervals of 0.01. We thus ran 30 models for each repetition in each condition, obtaining different RMSD values due to the different regularization parameter.

3.4. Computing Type I error rates and power for semiparametric RMSD

To evaluate whether the proposed semiparametric approach performs well in identifying the misfitting items using any of the selection criteria, we calculated detection rates separately for each selection criterion. Although we are not conducting formal hypotheses tests, we labelled the ratio of fitting items that were incorrectly detected as misfitting items the *Type I error rate* and the items that were correctly detected as misfitting the *power rate*, since these terms are widely used in similar studies, and easily understood. To determine Type I error, we selected the model considered optimal according to the selection criterion in each simulation condition and calculated the proportion of fitting items that were identified as misfitting. Since the group lasso method constrains all items that were identified as fitting close to zero, all RMSD values above 0.01 were items the method identified as misfitting. Note that this RMSD value is not to be interpreted as a typical cut-off criterion that marks the point at which the item contains too much misfit, but as a selection criterion that identifies whether or not the nonlinear part of the response function was involved in the IRF approximation. Whether or not the nonlinear part of the model has coefficients that are sufficiently close to zero is determined by the group lasso and the regularization parameter. We only use the RMSD to detect for which items and with what size of the regularization parameter the group lasso considered the nonlinear part as negligible.

Overall, the goal of the simulation was not just to determine whether the proposed methods works or not, but to test whether one of the selection criteria works well in selecting the smoothing parameter under which the method performs well.

4. Results

We calculated Type I error rates and power for the semiparametric approach in all conditions for each selection criterion. Note that in some instances, the minimum value of a selection criterion was the same for several sizes of λ. If this was the case, we selected the solution for the largest λ when estimating Type I error rate and power. This simply means that our results reflect the lowest possible Type I error and the lowest possible power—given the particular minimum—since a larger λ generally lead to fewer items being identified as misfitting.

4.1. Type I error rate

Table 2 displays the results for the Type I error rate and the power for each selection criterion. All selection criteria consistently showed a low to acceptable Type I error rate, which means that the majority of the fitting items were correctly declared as fitting. The AIC and BIC showed almost identical Type I error rates in the conditions with only fitting items. When using these criteria to select λ, the Type I error rate was low except for small sample sizes with a medium or a large number of items. In the conditions with $N = 5{,}000$, both statistics resulted in the correct identification of fitting items in almost all cases. The highest Type I error rate occurred in the condition with $N = 500$ and 100 items (.087), which can still be considered acceptable, however. The results for CV generally showed higher Type I error rates than the AIC and BIC. They were high for medium and large numbers of items when the sample size was $N = 1{,}000$ or $N = 5{,}000$. The best performance was found for a sample size of $N = 5{,}000$ and $I = 100$ (.027). Acceptable

Table 2. Type I error rate and power for each selection criterion

N	*I*	Number of misfitting items	Type I error AIC	Power AIC	Type I error BIC	Power BIC	Type I error CV	Power CV
500	20	0	.019	NA	.018	NA	.073	NA
500	20	2	.018	.289	.017	.286	.129	.470
500	20	6	.026	.288	.022	.276	.249	.562
500	60	0	.055	NA	.055	NA	.075	NA
500	60	6	.059	.531	.058	.529	.141	.630
500	60	18	.091	.521	.085	.509	.504	.848
500	100	0	.087	NA	.087	NA	.092	NA
500	100	10	.118	.648	.117	.646	.206	.730
500	100	30	.160	.653	.155	.647	.580	.905
1,000	20	0	.002	NA	.001	NA	.067	NA
1,000	20	2	.004	.210	.001	.178	.135	.626
1,000	20	6	.009	.261	.001	.184	.294	.716
1,000	60	0	.008	NA	.008	NA	.034	NA
1,000	60	6	.012	.506	.008	.469	.183	.808
1,000	60	18	.068	.634	.013	.447	.601	.975
1,000	100	0	.014	NA	.014	NA	.030	NA
1,000	100	10	.028	.644	.019	.592	.259	.911
1,000	100	30	.117	.760	.035	.597	.663	.989
5,000	20	0	.001	NA	0	NA	.085	NA
5,000	20	2	.011	.783	0	.204	.212	.962
5,000	20	6	.087	.789	0	.286	.459	.961
5,000	60	0	.000	NA	0	NA	.037	NA
5,000	60	6	.041	.951	.001	.748	.217	.989
5,000	60	18	.218	.994	.053	.950	.390	.994
5,000	100	0	.000	NA	0	NA	.027	NA
5,000	100	10	.049	.976	.005	.916	.239	.992
5,000	100	30	.247	.996	.106	.988	.402	.996

Note. For Type I error rate, $< .05$, $< .10$. For power, $> .80$, $< .70$. NA = not available (conditions with no misfitting items).

values were observed for the remaining conditions, where the highest Type I error rate existed for $N = 500$ and $I = 100$ (.092).

As is evident from Table 2, the Type I error rates for the AIC and BIC were predominantly low or acceptable for conditions in which both fitting and misfitting items were present in the data set. For the AIC, the smallest error occurred in the case of $N = 1{,}000$ and $I = 20$ (.004), meaning that in this condition, less than 1% of the fitting items were incorrectly declared as misfitting. Larger amounts of misfitting items in the data set resulted in inflated Type I error rates. The highest Type I error rates were observed for $N = 5{,}000$ with 60 and 100 items of which 30% were misfitting (.218 to .247). In these conditions, around a quarter of the fitting items were incorrectly identified as misfitting. The BIC performed best in conditions with $N = 5{,}000$ and $I = 20$, where no incorrect identifications were found. The Type I error rate increased as the number of items and the number of misfitting items in the data set increased. Overall, it was too high in three conditions. For $N = 500$ and $I = 100$ (.155), the values were the highest and slightly inflated. Overall, the highest Type I error rates were observed for CV. Using this selection criterion resulted in inflated Type I error rates across all conditions, with exceptionally large values for large amounts of misfit in the data. The lowest Type I error rate occurred in the condition $N = 500$ and $I = 20$ with 10% misfitting items (.129), which was already too high. In the remaining conditions, the Type I error rate increased up to .663 for $N = 1{,}000$ and $I = 100$ with 30% misfitting items, meaning that around two-thirds of the fitting items were incorrectly declared as misfitting.

4.2. Power

Table 2 also summarizes the power of the semiparametric approach to detect misfitting items in each of the conditions that contained misfitting items. Note that since the power of the two types of misfit was of a similar magnitude, we aggregated over both types of misfit and only present the overall results.

For all three selection criteria, the power generally increased with increasing sample size and an increasing amount of misfit. As is evident from Table 2, the AIC had high power in conditions with large sample sizes and a large or medium number of items. The power was still acceptable for a large sample size and a small number of items and for $N = 1{,}000$ and $I = 100$ with 30% misfitting items (.760). In other conditions, the power to detect misfitting items was too low, especially for small numbers of items. Here, a correct identification occurred in less than 30% of cases. The BIC performed best in conditions with large sample sizes and a large number of items: More than 90% of the items with misfit were correctly detected in these conditions. For $N = 5{,}000$ and $I = 60$ with 10% misfit, the power was still acceptable (.748), while in all other conditions, the power was smaller than .70. Similarly to the AIC, a small number of items lead to the lowest power. Out of the three selection criteria, CV showed the best performance in detecting misfitting items. For the majority of conditions, the power was higher than 90% or displayed at least acceptable values. Only in four conditions was the power below 70%.

4.3. Overall results

Although the CV showed the best performance for the power, the Type I error rates were too high when using this selection criterion. While the AIC and BIC had lower power in most conditions, their Type I error rate performance was superior, meaning that the majority of fitting items were correctly identified as fitting items. We recommend using

Table 3. Mean RMSD values (across replications) of correctly identified misfitting items based on the AIC model selection criterion using the semiparametric approach

			Type of misfit	
N	*I*	Number of misfitting items	3PL	Non-monotone IRF
500	20	2	0.084	0.082
500	20	6	0.080	0.082
500	60	6	0.066	0.060
500	60	18	0.053	0.062
500	100	10	0.059	0.057
500	100	30	0.052	0.056
1,000	20	2	0.075	0.072
1,000	20	6	0.071	0.071
1,000	60	6	0.060	0.057
1,000	60	18	0.047	0.055
1,000	100	10	0.055	0.053
1,000	100	30	0.048	0.052
5,000	20	2	0.054	0.037
5,000	20	6	0.053	0.041
5,000	60	6	0.049	0.037
5,000	60	18	0.046	0.045
5,000	100	10	0.049	0.042
5,000	100	30	0.047	0.047

Note. For the calculation of the mean RMSD estimates across the replications using the semiparametric approach only the RMSD values of items that were correctly detected as misfitting were included.

the AIC because it showed acceptable power rates in more conditions than the BIC. Note, however, that the Type I error rate was slightly higher when using the AIC compared to the BIC. Note also that the AIC was only appropriate if the sample size was large enough ($N > 1{,}000$ and $I > 100$, or $N > 5{,}000$), and at the same time, the amount of misfit did not greatly exceed 10%. Since we used results from the model with the highest λ when several models were identified as showing the lowest minimum, it is possible that using the model with the lowest λ instead would have resulted in slightly higher Type I error and power rates. This might have improved the results in some conditions with very low Type I error rates and low power rates (e.g., for $N = 1{,}000$ and $I = 100$ with 10% misfitting items).

4.4. RMSD values from semiparametric approach as an effect size

To investigate whether the estimated RMSD obtained from the semiparametric approach can be used as an effect size, we compared the average RMSD estimate in each condition with the population RMSD of 0.05, since we used item parameters to generate this size of misfit in the data. Because the AIC performed best in selecting the regularization parameter, we only examined the RMSD values that resulted under this method. We calculated the mean RMSD estimates across the replications in each condition separately for fitting and (the type of) misfitting items. We chose to report the mean RMSD estimates of the misfitting items based on only those items that were correctly identified as misfitting since values for incorrectly identified items are constrained to near zero. Averaging across correctly and incorrectly identified items would distort the information of the average size

Table 4. Mean RMSD values (across replications) using individual posterior probabilities (IPP) and kernel smoothing (KS)

			Fitting items		Misfitting items by type of misfit			
					3PL		Non-monotone IRF	
N	*I*	Number of misfitting items	IPP	KS	IPP	KS	IPP	KS
500	20	0	0.017	0.083	NA	NA	NA	NA
500	20	2	0.016	0.068	0.029	0.086	0.028	0.208
500	20	6	0.016	0.065	0.024	0.114	0.027	0.179
500	60	0	0.029	0.096	NA	NA	NA	NA
500	60	6	0.029	0.090	0.044	0.100	0.046	0.138
500	60	18	0.029	0.075	0.038	0.105	0.045	0.127
500	100	0	0.034	0.082	NA	NA	NA	NA
500	100	10	0.034	0.078	0.051	0.100	0.053	0.125
500	100	30	0.033	0.070	0.045	0.114	0.050	0.129
1,000	20	0	0.012	0.079	NA	NA	NA	NA
1,000	20	2	0.011	0.062	0.027	0.078	0.021	0.203
1,000	20	6	0.011	0.057	0.021	0.108	0.022	0.170
1,000	60	0	0.020	0.091	NA	NA	NA	NA
1,000	60	6	0.020	0.085	0.039	0.092	0.038	0.126
1,000	60	18	0.020	0.070	0.034	0.102	0.039	0.120
1,000	100	0	0.024	0.076	NA	NA	NA	NA
1,000	100	10	0.024	0.072	0.045	0.093	0.046	0.116
1,000	100	30	0.024	0.063	0.040	0.109	0.043	0.122
5,000	20	0	0.005	0.075	NA	NA	NA	NA
5,000	20	2	0.005	0.056	0.026	0.071	0.014	0.198
5,000	20	6	0.005	0.048	0.020	0.102	0.018	0.163
5,000	60	0	0.009	0.086	NA	NA	NA	NA
5,000	60	6	0.009	0.079	0.035	0.082	0.030	0.113
5,000	60	18	0.010	0.063	0.031	0.096	0.033	0.109
5,000	100	0	0.011	0.069	NA	NA	NA	NA
5,000	100	10	0.011	0.065	0.040	0.086	0.039	0.106
5,000	100	30	0.012	0.055	0.035	0.102	0.037	0.112

Note. NA = not available (conditions with no misfitting items).

of the RMSD value in those cases where the method works. Stated more precisely, the power rates displayed in Table 2 tell us in which conditions the semiparametric method works, whereas Table 3 tells us how well it works given that the identification was correct.

As is evident from Table 3, the estimated RMSD using the semiparametric approach was not wholly independent of the properties of the data set. The RMSD decreased as the number of items in the data set increased, as sample size increased, and as the proportion of misfitting items in the data set increased. Overall, however, the RMSD did not deviate drastically from the population RMSD except in the conditions with $N = 500$ and $I = 20$, and $N = 1{,}000$ and $I = 20$. Both types of generated misfit led to similar mean RMSD estimates, with mostly slightly lower RMSD values for items with a non-monotone IRF than mean RMSD estimates for items with misfit generated under the 3PL model.

Overall, results showed that in conditions with a large sample size, conditions with a large number of items, and conditions with a medium number of items with 30% misfit, items estimated under the semiparametric approach showed an RMSD similar to the population RMSD. In these cases, the RMSD estimates can be used as an effect size.

In order to evaluate the performance of the semiparametric approach in comparison to other approaches, we additionally estimated the RMSD using individual posterior probabilities (IPP) and kernel smoothing (KS). As is evident from Table 4, the IPP approach tends to underestimate the population RMSD of the misfitting items; it varies between 0.014 and 0.053, especially depending on the number if items in the data set. Due to the underestimation, the separation between fitting and misfitting items becomes impossible. The KS approach overestimates the population RMSD of the misfitting items; it varies between 0.071 and 0.203. Note that results for the misfitting items from Tables 3 and 4 are not directly comparable, since the RMSD estimates in Table 3 are based solely on items that were correctly identified as misfitting. In contrast, estimates in Table 4 are based on all misfitting items.

5. Concluding remarks

The application of the group lasso method for detecting item misfit was proposed. For large sample sizes, the method accurately provides information not only about whether the item significantly deviates from the model proposed IRF but also allows the size of item misfit to be classified. In contrast to previously proposed item fit statistics, the method is independent of sample size in so far as, once the misfitting items are identified, the computation of the RMSD for those items is based on a semiparametric approach.

Although this study investigated the method in conjunction with the RMSD, several other approaches to classifying the size of misfit after having determined the misfitting items are possible. Instead of using the RMSD as defined in equation (1), misfit could be directly quantified in the logit metric that might enhance detection of misfit at the extreme ends of the trait continuum (see, for example, Haberman, Sinharay, & Chon, 2013). Also, the weights w_t could be chosen differently. Instead of following a normal density function where extreme scores are assigned near-zero weights, weights could be based on the ability density or even a uniform function. Other parametric, semiparametric, or nonparametric approaches can potentially be used as well (see, for example, Douglas & Cohen, 2001; Lee, 2007; Lee et al., 2009; Ramsay, 1991). To evaluate the size of the difference between the observed IRF and the model IRF one could use value area measures (Raju, 1988), graphical displays (see, for example, Haberman et al., 2013; Swaminathan, Hambleton, & Rogers, 2006), and possibly even approaches that are typically used to evaluate the size of differential item functioning (for an overview, see, for example, DeMars, 2011).

The proposed method is an implementable and computationally affordable procedure. A user-friendly form of implementation could be to automatically fit the regularized semiparametric IRT model using different values for the regularization parameter. A function that automatically compares the AIC of each of the models would tell us about the minimum and hence the appropriate regularization parameter. The user then only needs to run the model with that specific parameter and estimate the RMSD values. Although the power to detect item misfit was somewhat unsatisfactory for sample sizes below 5,000, note that we used a small to medium size of misfit for generating the misfitting items. This means that the power to detect more severely misfitting items with the proposed method is much higher for moderate and possibly even small sample sizes.

There are a few limitations to this study that might be addressed in future research. Firstly, we only examined two types of misfitting items, namely items generated under the 3PL and items with a non-monotone IRF. Although we found no severe differences between the performance of the method regarding the type of misfit, it would be worth investigating whether the method performs equally well for other types of misfit. Secondly, we point out that using the AIC as a selection criterion for the regularization parameter produced satisfactory results for medium to large sample sizes (i.e., $N \geq 1{,}000$). Other group lasso selection criteria such as Nishii's (1984) generalized information criterion or the C_p-type criterion suggested by Yuan and Lin (2006) might prove to be more effective (for a summary of lasso methods and lasso penalties, see also Petersen & Witten, 2019). Thirdly, all remaining items were generated under perfect fit. Future studies could answer the research question of whether the method still performs well – in terms of small Type I errors – when remaining items in the data set show slight misfit or even other types of misspecifications such as local dependence. In this respect, it might be worth exploring other group lasso methods, for example, the adaptive group lasso as suggested by Wei and Huang (2010), which in addition to its consistency showed promising results for small sample sizes, which was not the case for our proposed method. Furthermore, it would be worth investigating alternative forms of regularization for maximum penalized-likelihood estimation. The application of the group lasso method implies a binary decision whether an item fits a parametric model or shows misfit. A less segregating form of regularization such as smoothing splines or P-splines (Aguilera & Aguilera-Morillo, 2013; Eilers & Marx, 1996; Rossi et al., 2002), or the use of a ridge penalty of basis function parameters that would allow some extent of misfit in each item might further improve the statistical properties of the RMSD estimate. However, this would increase the number of parameters that need to be estimated, and there would be no clear decision criterion as to whether an item shows misfit. As for large-scale assessments like PISA, a pending line of investigation revolves around the application of the proposed approach to empirical data. It would be interesting to test how well the method performs not only in detecting item misfit but also in evaluating differential item functioning for which the RMSD statistic is also employed (e.g., in large-scale studies such as PISA or PIAAC). Lastly, a comparison to fit statistics based on sum scores (e.g., $S - X^2$; Orlando & Thissen, 2000), pseudo-observed scores (e.g., X^{2*}; Stone, 2000), or other model-based item fit statistics (e.g., Lagrange multiplier test; Glas & Suárez Falcón, 2003) might be warranted.

We would like to point out that the use of any fit statistic should be functional rather than stringent. The test of model and item fit should be considered a multifaceted process, not one that relies on a single evaluation of a fit statistic (Hambleton & Han, 2005). It encompasses a re-evaluation of the identified items and careful consideration of whether the indicated misfit calls for item revision or outright item removal. The few studies that have evaluated the practical consequences of item removal for relevant outcomes of high- and low-stakes large-scale assessments have found limited effects (Köhler & Hartig, 2017; Liang, Wells & Hambleton, 2014; Sinharay & Haberman, 2014; van Rijn, Sinharay, Haberman, & Johnson, 2016).

Acknowlegements

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).

Conflicts of interest

All authors declare no conflict of interest.

Author contributions

Carmen Köhler, Ph.D. (Formal analysis; Funding acquisition; Methodology; Project administration; Supervision; Writing – original draft; Writing – review & editing); **Alexander Robitzsch** (Conceptualization; Methodology; Resources; Software; Writing – original draft; Writing – review & editing); **Katharina Fährmann** (Formal analysis; Writing – original draft) **Matthias von Davier** (Conceptualization); **Johannes Hartig** (Conceptualization; Funding acquisition).

Data availability statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Aguilera, A. M., & Aguilera-Morillo, M. C. (2013). Comparative study of different B-spline approaches for functional data. *Mathematical and Computer Modelling*, *58*, 1568–1579. https://doi.org/10.1016/j.mcm.2013.04.007

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csáki (Eds.), *Proceedings of the Second International Symposium on Information Theory* (pp. 267–281). Budapest: Akadémiai Kiadó.

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.

Barlow, R. E., Bartholomew, D. J., Bremmer, J. M., & Brunk, H. D. (1972). *Statistical inference under order restrictions*. New York, NY: John Wiley & Sons. https://doi.org/10.1111/j.1467-9574.1973.tb00228.x

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.

Breheny, P., & Huang, J. (2015). Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and Computing*, *25*(2), 173–187. https://doi.org/10.1007/s11222-013-9424-2

DeMars, C. E. (2011). An analytic comparison of effect sizes for differential item functioning. *Applied Measurement in Education*, *24*, 189–209. https://doi.org/10.1080/08957347.2011.580255

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, *39*, 1–38. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Douglas, J., & Cohen, A. (2001). Nonparametric item response function estimation for assessing parametric model fit. *Applied Psychological Measurement*, *25*(3), 234–243. https://doi.org/10.1177/01466210122032046

Eilers, P. H. C., & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, *11*(2), 89–121.

Epskamp, S., Rhemtulla, M. T., & Borsboom, D. (2017). Generalized network psychometrics: Combining network and latent variable models. *Psychometrika*, *82*, 904–927. https://doi.org/10.1007/s11336-017-9557-x

Falk, C. F., & Cai, L. (2016). Maximum marginal likelihood estimation of a monotonic polynomial generalized partial credit model with applications to multiple group analysis. *Psychometrika*, *81*, 434–460. https://doi.org/10.1007/s11336-014-9428-7

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. https://doi.org/10.1177/0146621602250530

Haberman, S. J., & Sinharay, S. (2013). Generalized residuals for general models for contingency tables with application to item response theory. *Journal of American Statistical Association*, *108*, 1435–1444. https://doi.org/10.1080/01621459.2013.835660

Haberman, S. J., Sinharay, S., & Chon, K. H. (2013). Assessing item fit for unidimensional item response theory models using residuals from estimated item response functions. *Psychometrika*, *78*, 417–440. https://doi.org/10.1007/s11336-012-9305-1

Hambleton, R., & 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.

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity. The lasso and generalizations*. Boca Raton, FL: CRC Press. https://doi.org/10.1201/b18401

Köhler, C., & Hartig, J. (2017). Practical significance of item misfit in low-stakes educational assessment. *Applied Psychological Measurement*, *41*(5), 388–400. https://doi.org/10.1177/0146621617692978

Köhler, C., Robitzsch, A., & Hartig, J. (2020). A bias-corrected RMSD item fit statistic: An evaluation and comparison to alternatives. *Journal of Educational and Behavioral Statistics*, *45*, 251–273. https://doi.org/10.3102/1076998619890566

Lee, Y.-S. (2007). A comparison of methods for nonparametric estimation of item characteristic curves for binary items. *Applied Psychological Measurement*, *31*, 121–134. https://doi.org/10.1177/0146621606290248

Lee, Y., 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. https://doi.org/10.1177/0013164408322026

Liang, L., & Browne, M. W. (2015). A quasi-parametric method for fitting flexible item response functions. *Journal of Educational and Behavioral Statistics*, *40*(1), 5–34. https://doi.org/10.3102/1076998614556816

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), 1–17. https://doi.org/10.1111/jedm.12031

Lloyd-Jones, L. R., Nguyen, H. D., & McLachlan, G. J. (2018). A globally convergent algorithm for lasso-penalized mixture of linear regression models. *Computational Statistics and Data Analysis*, *119*, 19–38. https://doi.org/10.1016/j.csda.2017.09.003

Mazza, A., Punzo, A., & McGuire, B. (2014). KernSmoothIRT: An R package for kernel smoothing in item response theory. *Journal of Statistical Software*, *58*(6), 1–34. http://www.jstatsoft.org/v58/i06

Meier, L., van de Geer, S., & Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of Royal Statistical Society, Series B*, *70*, 53–71. https://doi.org/10.1111/j.1467-9868.2007.00627.x

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–176. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x

Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Annals of Statistics*, *12*, 758–765. https://doi.org/10.1214/aos/1176346522

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems (with discussion). *Statistical Science*, *1*, 502–527. https://doi.org/10.1214/ss/1177013525

OECD. (2015). *PISA 2015 field trial analysis report: Outcomes of the cognitive assessment (JT03371930)*. Paris: OECD.

OECD. (2017). *PISA 2015 technical report*. Paris: OECD.

Oelker, M.-R., & Tutz, G. (2017). A uniform framework for the combination of penalties in generalized structured models. *Advances in Data Analysis and Classification*, *11*(1), 97–120. https://doi.org/10.1007/s11634-015-0205-y

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

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*(4), 289–298. https://doi.org/10.1177/0146621603027004004

Petersen, A., & Witten, D. (2019). Data-adaptive additive modeling. *Statistics in Medicine*, *38*(4), 583–600. https://doi.org/10.1002/sim.7859

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*, 495–502. https://doi.org/10.1007/BF02294403

Ramsay, J. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, *56*, 611–630. https://doi.org/10.1007/BF02294494

Ramsay, J. O., & Silverman, B. W. (2005). *Functional data analysis* (2nd ed.). New York, NY: Springer. https://doi.org/10.1002/0471667196.ess3138

Robitzsch, A., Kiefer, T., & Wu, M. (2020). *TAM: Test Analysis Modules. R package version 3.6-8*. Retrieved from https://CRAN.R-project.org/package=TAM

Rossi, N., Wang, X., & Ramsay, J. O. (2002). Nonparametric item response function estimates with the EM algorithm. *Journal of Educational and Behavioral Statistics*, *27*(3), 291–317. https://doi.org/10.3102/10769986027003291

Schwartz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464. https://doi.org/10.1214/aos/1176344136

Simonoff, J. S. (1996). *Smoothing methods in statistics*. New York, NY: Springer. https://doi.org/10.1007/978-1-4612-4026-6

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. https://doi.org/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*(1), 58–75. https://doi.org/10.1111/j.1745-3984.2000.tb01076.x

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. *Educational and Psychological Measurement*, *71*(5), 834–848. https://doi.org/10.1177/0013164410393238

Sun, J., Chen, Y., Liu, J., Ying, Z., & Tao, X. (2016). Latent variable selection for multidimensional item response theory models via L1 regularization. *Psychometrika*, *81*(4), 921–939. https://doi.org/10.1007/s11336-016-9529-6

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.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, *58*, 267–288. https://doi.org/10.1111/j.1467-9868.2011.00771.x

Tutz, G., & Gertheiss, J. (2016). Regularized regression for categorical data. *Statistical Modelling*, *16*(3), 161–200. https://doi.org/10.1177/1471082X16642560

Tutz, G., & Schauberger, G. (2015). A penalty approach to differential item functioning in Rasch models. *Psychometrika*, *80*(1), 21–43. https://doi.org/10.1007/s11336-013-9377-6

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), 1–23. https://doi.org/10.1186/s40536-016-0025-3

von Davier, M. (2016). *Software for multidimensional discrete latent trait models – mdltm*. Draft manual.

Wainer, H., & Thissen, H. (1987). Estimating ability with the wrong model. *Journal of Educational Statistics*, *12*, 339–368. https://doi.org/10.3102/10769986012004339

Wei, F., & Huang, J. (2010). Consistent group selection in high-dimensional linear regression. *Bernoulli*, *16*, 1369–1384. https://doi.org/10.3150/10-BEJ252

Yamamoto, K., Khorramdel, L., & von Davier, M. (2013). Scaling PIAAC cognitive data. In OECD (Ed.), *Technical report of the survey of adult skills (PIAAC)* (2nd ed., pp. 1–33). Paris: OECD Publishing.

Yang, Y., & Zou, H. (2015). A fast unified algorithm for solving group-lasso penalized learning problems. *Statistics and Computing*, *25*, 1129–1141. https://doi.org/10.1007/s11222-014-9498-5

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, *68*, 49–67. https://doi.org/10.1111/j.1467-9868.2005.00532.x

Zhang, Y., Li, R., & Tsai, C. (2010). Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association*, *105*, 312–323. https://doi.org/10.1198/jasa.2009.tm08013

Received 6 March 2019; revised version received 17 September 2020

Supporting Information

The following supporting information may be found in the online edition of the article:

Table S1. Generating item parameters for misfitting items generated under the 3PL model

Table S2. Generating item parameters for misfitting items generated under a nonmonotone IRF model