Fährmann *et al.*
Large-scale Assessments in Education (2022) 10:7
https://doi.org/10.1186/s40536-022-00124-w

Large-scale Assessments in Education

METHODOLOGY Open Access

Practical significance of item misfit and its manifestations in constructs assessed in large-scale studies

Katharina Fährmann[1], Carmen Köhler[1*], Johannes Hartig[1] and Jörg-Henrik Heine[2]

*Correspondence:
carmen.koehler@dipf.de

[1] DIPF Leibniz Institute for Research and Information in Education, Rostocker Str. 6, 60323 Frankfurt, Germany
[2] Centre for International Student Assessment (ZIB), Munich, Germany

Abstract

When scaling psychological tests with methods of item response theory it is necessary to investigate to what extent the responses correspond to the model predictions. In addition to the statistical evaluation of item misfit, the question arises as to its practical significance. Although item removal is undesirable for several reasons, its practical consequences are rarely investigated and focus mostly on main survey data with pre-selected items. In this paper, we identify criteria to evaluate practical significance and discuss them with respect to various types of assessments and their particular purposes. We then demonstrate the practical consequences of item misfit using two data examples from the German PISA 2018 field trial study: one with cognitive data and one with non-cognitive/metacognitive data. For the former, we scale the data under the GPCM with and without the inclusion of misfitting items, and investigate how this influences the trait distribution and the allocation to reading competency levels. For non-cognitive/metacognitive data, we explore the effect of excluding misfitting items on estimated gender differences. Our results indicate minor practical consequences for person allocation and no changes in the estimated gender-difference effects.

Keywords: Item fit, Practical significance of item fit, Item response theory, PISA 2018 field trial data, Large scale assessment, Educational measurement

Introduction

When using item response theory (IRT) models for scaling and data evaluation, according to Wainer and Thissen (1987), one of the prerequisites for valid statements is the fit of a measurement model to the data. Hambleton and Han (2005) suggest an assessment and evaluation of this model fit in five steps. Local item fit is involved in two of those steps, namely checking item fit to test for (1) statistical significance and (2) practical significance of item misfit.

In both research and application, usually only step (1) is applied (Zhao & Hambleton, 2017). In the stage of scale construction, misfitting items are frequently excluded from the item pool (Sinharay & Haberman, 2014). This seems to be premature insofar as item fit statistics as an evaluation criterion for misfit are problematic. Different fit statistics with various cut-off values exist, yet, no clear guidelines

for these cut-off values exist (de Ayala, 2009). This is also demonstrated by the inconsistent use of both the fit statistics and their cut-off values, for example, in large scale assessments (LSAs; e.g., ACARA, 2013; Allen et al., 1999; OECD, 2018a 2018b). While the properties of different item fit statistics have been investigated intensively over the past decades, the focus has recently expanded to the practical significance of item misfit (e.g., Hambleton & Han, 2005; Zhao & Hambleton, 2017). Sinharay and Haberman (2014) define the practical significance of item misfit as "an assessment of the extent to which the decisions made from the test scores are robust against the misfit of the IRT models" (p. 23). Hence, item misfit is of practical significance if it leads to practical consequences for the measure of interest. It should be noted that there is not *the one* practical consequence of item misfit. The question is whether the amount of statistical item misfits leads to practical consequences for the intended application (Zhao & Hambleton, 2017). Therefore, no single quantity or threshold for evaluating practical significance exists. Instead, practical significance of item misfit can be examined in terms of various aspects depending on the purpose of the test. This also means that differences between the data and the model assumptions might have no practical consequences for the intended test outcome at all (Liang et al., 2014; van der Linden & Hambleton, 1997).

On the one hand, no agreement on the best methods to conduct practical significance analyses exists (Swaminathan et al., 2006; Zhao & Hambleton, 2017). On the other hand, findings from the few empirical studies in this research field have shown that item misfit is frequently of no practical significance (Crișan et al., 2017; Köhler & Hartig, 2017; Liang et al., 2014; Sinharay & Haberman, 2014; Sinharay et al., 2011; Tendeiro & Meijer, 2015; van Rijn et al., 2016; Zhao, 2016). The first aim of this paper is to derive criteria to assess practical consequences of item misfit regarding individual and population level comparisons. In addition, we explore how to deal with discovered practical consequences of item misfit. Furthermore, most of the existing studies examine main survey data with pre-selected items when investigating practical consequences of item misfit. This research gap is frequently mentioned in the limitations section of those studies (e.g., Sinharay et al., 2011; Sinharay & Haberman, 2014). To our knowledge, only Sinharay and colleagues (2011) used pretest data for some of their analyses. Field trial studies in which items are tested for the first time likely show a higher proportion of misfitting items than finalized versions of the assessment. Therefore, field trial studies can provide more insight into the practical consequences of item misfit. As a second aim, we apply the derived criteria to PISA 2018 field trial data of the German PISA sub sample to address this research gap for both cognitive and noncognitive/metacognitive data.

The manuscript is organized as follows: Firstly, arguments for testing the practical significance of item misfit are discussed. Secondly, possible criteria for evaluating practical significance of item misfit regarding individual and population level comparisons in LSAs are presented. Thirdly, the criteria for the population level are applied to two empirical examples, one for cognitive and one for non-cognitive/metacognitive measurements of the latent trait. In closing, the results and consequences of the study are laid out.

Arguments for testing practical significance of item misfit

There are various arguments for testing practical consequences of item misfit. A first general argument is that in empirical settings, perfect fit between the actual data and the prediction of the model is impossible (Molenaar, 1997). It is self-evident that no model predicts the data perfectly or the model would otherwise be over identified. These circumstances raise the question of how much item misfit is justifiable, which is closely related to the question of the practical significance of item misfit. Box and Draper (1987) argue that while all models are wrong, this does not mean that some of them are not still useful. For example, if the sample size is sufficiently large, no IRT model will fit the data perfectly according to the model fit criteria (Sinharay et al., 2011), and a large proportion of items is frequently flagged as misfitting (Hambleton & Swaminathan, 1985; Liang et al., 2014). For these reasons, several researchers suggest evaluating the practical significance of model and item misfit (e.g., Sinharay & Haberman, 2014; Zhao & Hambleton, 2017).

Additional arguments are that item removal is not desirable for economic reasons because item development costs time and money (Köhler & Hartig, 2017). Furthermore, items are developed to sufficiently represent the construct to be measured, thus a limited number of items is available for each subdomain, and item removal could lead to the case that the construct is no longer represented sufficiently (Crişan et al., 2017; Köhler & Hartig, 2017). Thus, from a more practical perspective, the content validity of the measurement may become distorted. For a test taker, disadvantages are possible if, for example, statistically misfitting items that have been answered correctly are subsequently removed (Crişan et al., 2017). Zhao and Hambleton (2017) point out that it is sometimes necessary to keep a misfitting item in a test for balancing content coverage or for particular IRT analyses such as IRT-based equating, differential item functioning (DIF), and computer-adaptive testing (CAT).

Methods

Criteria for evaluating practical significance of item misfit

Practical significance of item misfit involves the consequences of excluding single items on the basis of item fit criteria (i.e., item selection). In order to assess whether item selection has practical consequences with regard to the distribution of the measured trait, criteria for evaluating these consequences are considered. In scenarios with aggregated or individual interpretations of results, different foci of the investigation are set. In scenarios where test takers receive personal feedback, results at the individual level are of interest (e.g., college entrance tests or language proficiency tests), whereas in scenarios with aggregated interpretations of the results, results at the population level are relevant (e.g., tests for competence comparisons in LSAs). We first consider general criteria that are applicable to both application scenarios. We subsequently discuss criteria that only apply to the respective scenarios.

General premises for evaluating practical significance of item selection

Practical significance of item selection can be represented by meaningful differences in the measures of interest when misfitting items are excluded from the data set. To

quantify a meaningful difference, we propose that a greater emphasis should be put on the effect size of this difference. Traditionally, an observed significant difference between two parameters or parameter vectors can be evaluated by the effect size (Cohen, 1988), without—in contrast to statistical significance—being dependent on the sample size. This allows to state whether statistically significant differences are large enough and practically significant as well (Kirk, 1996; Peeters, 2016). Although an effect size, for instance between the different means of two measured trait distributions with or without item selection, cannot be understood as a synonym for practical significance of item selection, it can nevertheless serve as a basis for its examination. As Kirk (1996) argues, however, practical significance is simultaneously less definite than the statistical significance because depending on the application, a different approach for evaluating practical significance may be the most appropriate. A popular approach to interpret the effect size is Cohen's d (1988, 1992). Although there is a general framework for the interpretation of Cohen's d, these rules of thumb are discussed controversially in the literature. It remains to be clarified whether the classical categorization of the effect size according to Cohen (1988) is also applicable to the practical significance of item selection. Thompson (2007) argues that there are situations in which even small effect sizes can have large impacts, so there are limitations with regard to the context. Accordingly, he suggests comparing the calculated effect sizes with effect sizes from similar sources instead of using rules of thumb. Such a procedure helps to classify effect sizes in the context of item misfit.

Criteria depending on application scenario

In the following, criteria for evaluating the practical significance of item selection on the measured trait for tests focusing on the individual level and tests focusing on the population level are put up for discussion. Figure 1 gives an overview of the criteria discussed in this chapter. In scenarios with aggregated interpretations of results, conclusions about the distribution of person abilities and the assignment to proficiency levels are relevant. In addition, relationships with other variables are investigated. In scenarios with individual interpretations of results, the emphasis lies on rankings, score estimates or classifications of the individual. Note that the classifications in Fig. 1 are meant to provide guidance and should not be taken rigidly—ultimately, the assessment intentions should dictate the criterion of interest.

In the following, we make a distinction between cognitive and noncognitive/metacognitive scales. For both types of scales, general criteria for assessing the practical consequences of item selection are presented first. Subsequently, additional criteria for both types of scales are discussed separately.

General criteria

In order to evaluate the practical consequences of item selection, a first indicator is the extent to which the trait estimates correlate with each other when items are included versus when they are excluded from the data set. Without practical significance of item selection, the rank correlation should be close to 1, while with increasing practical consequences this correlation decreases. The use of correlations to

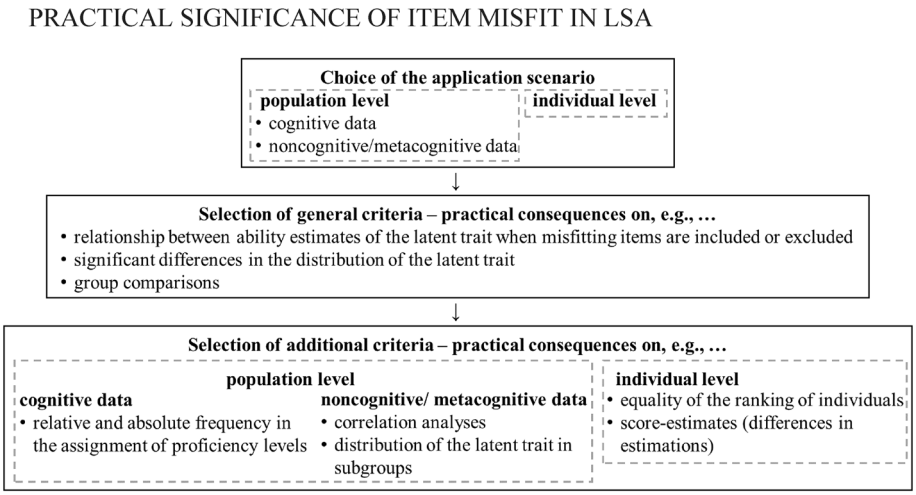

Fig. 1 Criteria that can be used to evaluate the practical consequences of item misfit

investigate the practical significance of item selection can be found in various studies. For example, Tendeiro and Meijer (2015) analyzed the practical consequences of item selection on the performance ranking by (1) determining the overlap between the highest ranked test takers according to the true ability, (2) estimating the correlations between top-ranking true ability and the corresponding approximations based on estimated ability values and their scores. In the study from Crişan et al. (2017), the practical consequences of selection were examined with Spearman rank correlations. Sinharay et al. (2011) investigated various educational studies and focused on score estimates and correlation coefficients to determine the practical consequences of item selection. Zhao (2016) analyzed medical data: The practical significance of item selection on score estimates and the classification of the severity of the diagnosis at the group level was investigated using correlation analysis.

It should further be examined whether meaningful differences in the mean values or variances of the distribution occur. Liang et al. (2014) compared the differences in test score distributions and concluded that the values were close to each other, indicating that item selection had no practical consequences on differences in test score distributions. With respect to the location of the distribution, group mean values can be calculated and compared (with and without the misfitting items). The differences in the mean values can be evaluated in terms of statistical significance and effect size, with the effect size clearly playing the more important role. Methods for mean value comparisons for dependent samples like the *t*-test or the Wilcoxon signed-rank test with subsequent Cohen's *d* estimations are suitable. Such a procedure is used in the study by Zhao (2016). This approach can also be found in the study by van Rijn et al. (2016), who evaluated practical consequences based on comparisons of competences between specific subgroups.

Additional criteria for testing on population level

Cognitive data

In order to evaluate the practical consequences of item selection with respect to the ability distribution when assessing cognitive data, differences can be reflected in the assignment of persons to the proficiency levels. This is closely related to comparing the location and variance of the distribution of the measured trait. The proficiency level a person is assigned to can depend on whether or not a misfitting item is used for estimating ability scores. The McNemar–Bowker test for pairwise symmetry examines whether individuals change between two category classes of proficiency levels symmetrically or asymmetrically. Asymmetrical changes mean that a significantly differing number of individuals change between two proficiency levels or category classes. Accordingly, it can be investigated whether these shifts of individuals between different proficiency levels are significant, and which effect size (Cohen's *g*) they show.[1]

Noncognitive/metacognitive data

Since background questionnaire (BQ) variables are often used as predictors in multiple regressions or in correlational analyses, it is useful to calculate the correlation between the measured trait and another variable when the statistically misfitting items are included and when they are excluded. Practical significance of item selection is evident if the correlations significantly differ (van Rijn et al., 2016). A method for quantifying potential bias of relationship estimates (e.g., correlation coefficients) due to item misfit in low-stakes achievement tests was developed by Köhler and Hartig (2017), who provide upper and lower boundaries for the change in the regression coefficient when misfitting items remain in the analysis.

If significant differences occur in the correlation analyses, it is of interest to examine the distributions of the latent trait in the individual subgroups in more detail, and whether individual groups are affected differently by item selection. For example, the removal of one item could lead to a shift in the distribution of the measured trait for males but not for females. To evaluate the practical consequences of item selection with respect to the distribution of the trait in the subgroups, it is reasonable to test whether meaningful differences in the mean values or variances per subgroup are observable.

Additional criteria for testing on individual level

Item selection can lead to shifts in the rankings of individuals, which can have consequences for the test takers. The Jaccard Index (Jaccard, 1912) can be used to make statements about the similarity of two scored sets, what are in this case the rankings. If significant differences between rankings become evident, it indicates a practical significance of item selection. Crișan et al. (2017) use this procedure to supplement the calculation of rank correlations. A further supplement to the rank correlation and comparison of scored sets are graphical evaluations, in which latent trait estimates or scored estimates are plotted against each other in a scatterplot with and without the inclusion of the misfitting items. This gives more insight into which parts of the trait continuum are

[1] Cohen's *g* (Cohen, 1988) is designed for applications where the expected proportion is 50%.

influenced the most. Such graphical evaluations can be found in the studies conducted by Sinharay et al. (2011). In addition, the McNemar-Bowker test for pairwise symmetry can be applied. A significant change of individuals between the category classes indicates on the one hand that changes in the ranking are observable and on the other hand that these changes can be identified for the respective category classes. The effect size Cohen's *g* can be calculated to classify those differences.

Score estimates often serve as a basis for ranking or classifying individuals. Item selection can result in a significant practical change in the score estimates. To compare score estimates, the difference that matters (dtm) method can be applied. Dorans and Feigenbaum (1994) used dtm to evaluate the practical consequences of item selection on score estimates. A difference between an equated score and a criterion score greater than 0.5 of the reported scale unit is referred to as a significant difference in scores (Dorans & Feigenbaum, 1994). This procedure is also used by Sinharay et al. (2011), Sinharay and Haberman (2014) and Zhao and Hambleton (2017).

Dealing with practical significance of item selection

If a practically significant effect of item selection is found, the question of how to deal with it arises. The occurrence of a practically significant change in the results due to item selection indicates that the constructs slightly differ in their meaning when the statistically misfitting items are included versus when they are excluded. If the items measured the exact same construct, correlations with other variables would not be affected. Therefore, it has to be clarified whether the model is more suitable to represent the intended construct with or without the statistically misfitting items. In our view, this decision depends on the construct of interest. To make this decision, both the (1) model fit and the (2) item content need to be considered.

The item fit analyses should be considered in light of the overall model evaluation, since the item fit analyses are only one step in testing the (1) model fit. Other investigations are to test for unidimensionality, DIF, and local independence.[2] Especially the latter two analyses provide additional insight on whether an affected item is beneficial to the model and should be kept in the data set, or whether the suspicion that the item should be excluded from the test is strengthened.

In addition, the items should be examined more closely on the (2) content level.[3] On the one hand, it has to be checked if the item was constructed as intended during item development. This means that it should be ruled out that ambiguities in the formulation of the item or translation errors have occurred. On the other hand, it is important to check the extent to which the item differs from the other items in terms of content and whether the item in question actually covers the desired construct. If it represents the desired construct, it is also to be clarified whether the item has an added value for measuring the construct. If it does, this argues in favor of keeping the item.

In conclusion, please note that the causes of item misfit and its practical consequences through item selection on the measured trait can be diverse. Other model fit indicators and considerations at the content level can aid in deciding whether to exclude an item.

[2] A comprehensive overview of these test steps can be found in Swaminathan et al. (2006).

[3] An overview of evaluation steps can be found in Hartig et al. (2020).

Data example 1 for testing on population level: cognitive

The aim of this empirical example is to demonstrate the application of the discussed criteria to investigate practical consequences that item selection has on the assignment of persons to proficiency levels. It is based on the domain *reading literacy* of the cognitive German PISA field trial data 2018. [4] In this subsection, we describe the methodological procedure for evaluating the practical significance of item selection for cognitive data. We first give an account on the sample and the assessment, which is followed by a description on item calibration and item fit calculation. We then describe how individuals were assigned to proficiency levels on the basis of their abilities and how we compared whether there were practically significant differences between the inclusion and exclusion of the misfitting items. All analyses were conducted with the open sources software R, version 4.0 (R Core Team, 2022). For the IRT analyses, we used the R package TAM, version 3.5–19 (Robitzsch et al., 2020). The syntax for all our analyses is available on OSF (https://osf.io/r23dt/).[5]

Sample and assessment

A sample size of $N = 1890$ was tested with 402 items, of which 88 were trend items and 314 were new items. It should be noted that 65 of the new items were reading fluency items, which were constructed as very easy items in order to improve measurement precision at the lower end of the proficiency levels (OECD, 2018b). Out of all 402 items, 33 were polytomous and 369 were dichotomous. The survey was computer-based and employed a multi-matrix design, meaning that not all items were processed by all participants. The number of responses per item lay between 155 and 736. The number of responses per person lay between seven and 111.

Data scaling

In accordance with the procedure for scaling the data in the PISA 2018 study (OECD, 2020), the trend items and new items were treated differently. Since 2015, a new IRT approach has been used for scaling the PISA data, combining the Rasch model (Rasch, 1960/1980) and the partial credit model (PCM; Masters, 1982) with the two-parameter logistic model (2PL; Birnbaum, 1968) and the generalized partial credit model (GPCM; Muraki, 1992). The trend items from the 2000–2012 PISA cycles were recalibrated under the Rasch model and the PCM and the new items were scaled under the 2PL model/GPCM (OECD, 2015). In our study, the item parameters from the recalibration of the trend items were available; therefore, we did not carry out this step ourselves.

The 314 new items were calibrated in two steps, similar to the procedure in PISA. In the initial step, the reading fluency items were excluded for scaling the data, because they were expected to be very easy. We fixed the item parameters for the trend items and calibrated the 249 new items (without the 65 new reading fluency items) under the 2PL

[4] Although the two examples presented were based on the German PISA field trial data 2018, the fact that only the data from Germany were used means that an exact reproduction of the PISA results was not intended. Note, usually, the field trial data are not conceptualized to draw conclusions about the assignment to proficiency levels (OECD, 2018b, 2020).

[5] A public sharing of the data in the context of our manuscript is, unfortunately, not possible. The sharing of PISA data as a) public use files or b) after an official request to receive the data (from the Research Data Center [FDZ] at IQB) is only intended for data from the main study.

model/GPCM. In the second stage, the reading fluency items were included in the data set. For scaling the reading fluency items, the item parameters for both the trend items and the 249 new items were fixed and the reading fluency items were calibrated under the 2PL model/GPCM.

Item fit analyses

For the item fit analyses, we used the RMSD as an item fit statistic. For the PISA field trial analyses, a cut-off value of RMSD ≥ 0.20 is reported for cognitive data (OECD, 2018b). In the PISA main study, the cut-off is 0.12 (OECD, 2020). In both the PISA field trial and the main analyses, the RMSD value describes the deviation between the observed item response function (IRF) within a country and the predicted IRF from calibrations that use either the international or a group specific item parameter. This means that the RMSD in PISA is mainly used as a detection method for DIF (OECD, 2020; Tijmstra et al., 2020). In our research, however, we explore the field trial data from only Germany—without making group comparisons—and use the RMSD to evaluate item quality (i.e., item fit) in this single subsample. Using the RMSD for this purpose affects its size and hence the cut-off value. Thus far, no generally agreed upon cut-off value for the RMSD exists when it is used to detect item misfit. Köhler et al. (2020) investigated the RMSD as an item fit statistic and used simulation studies in combination with resampling techniques across varying data-set conditions. They showed that the RMSD depends on characteristics of the data set. In each condition of their study, the mean RMSD values for both fitting and misfitting items were much lower than 0.12. Based on their results, we used a conservative cut-off of 0.06 and a cut-off value of 0.08 to exclude solely items with a large misfit.[6]

Assignment to proficiency levels

We used multiple imputed values (plausible values; PVs) as proficiency estimates, which is the method also employed in PISA (see, e.g., OECD, 2020). In the PISA study, PISA scores were estimated in order to assign students to proficiency levels. These scores are intended to enable comparability both across cycles and domains (OECD, 2020). For this purpose, the approach of a conditioning model was used, where PVs were calculated by incorporating a background model based on scales from the BQ. Taking the results from the participating countries into account, the PVs were transformed into the PISA scores using a linear transformation. The scores were reported on a scale with a mean of 500 score points and a standard deviation of 100 score points.[7] In our study, we used the reported PISA scale as a basis to calculate scores for the individual persons. While we also estimated PVs, we did not apply a background model because we aimed to avoid distortions by a misspecified background model. Rutkowski (2014) support this approach since the inclusion of a background model, especially in secondary analyses, can be a source of error. For the transformation to the PISA scores, we calculated both

[6] Note that we investigate the practical significance of varying degrees of severity of misfit without recommending the applicability of a specific RMSD cut-off value.

[7] In the first PISA study in 2000, the OECD mean was set at 500 and the standard deviation at 100. However, in later assessments, the OECD mean is no longer exactly 500, but has changed, for example, due to a different response pattern of the participants or the increase in the number of OECD countries (Reiss, Weis, Klieme, & Köller, 2019).

the mean item difficulty and the variance to determine the two factors (slope and intercept) required for the linear transformation. Since the item difficulty and ability are on a joint logit scale, the calculated factors were used to transform each person's mean PV to a PISA score.

The PISA main study 2018 reports six proficiency levels, of which the first is divided into three sublevels (1a–c). The cut-scores representing the minimum score of a proficiency level are 189 for Level 1c, 262 for Level 1b, 335 for level 1a, 407 for Level 2, 480 for Level 3, 553 for Level 4, 626 for Level 5, and 697 for Level 6. We applied these cut-scores and calculated the percentage of people at each proficiency level.

Rescaling and reanalyzing

In the second stage, the statistically misfitting items were removed from the data set based on the two different cut-off values (1) RMSD > 0.06 and (2) RMSD > 0.08. For the comparability of the results, it was necessary to use a linking procedure. We applied the fixed anchor calibration approach, in which the item parameters of the fitting trend items were fixed for the rescaling (this was 13 trend items for RMSD > 0.06 and 28 for RMSD > 0.08, which we describe in more detail in the results section). The respective recalibration of the two selected data sets was conducted analogous to the procedure of the first scaling using the 2PL/GPCM. The item parameters for the remaining new items (without reading fluency items) were estimated. Subsequently, the item parameters for both the trend items and the new items were fixed and the reading fluency items were calibrated. The assignment of respondents to the proficiency levels was carried out analogously to the procedure described above. The factors calculated in the first scaling were used for the linear transformation in order to ensure comparability in the assignment of persons to the proficiency levels.

Evaluation of the practical significance of item selection

The relevant criteria for evaluating the practical significance of item selection in this first example were the correlation between the trait estimates, differences in the distribution of the latent trait, and the frequencies of person assignment to proficiency levels. To examine the extent of the relationship between the trait estimates with and without item removal, we calculated the (1) Spearman rank correlation between the trait estimates. We used the rank correlation instead of the Pearson correlation because we did not assume strict linearity for two reasons: (a) Item removal does not affect all ability estimates equally, and (b) impacts are unequally strong for each item. In addition to the correlation, we examined whether differences in (2) the mean values and (3) the variance of the ability distribution occurred. Therefore, we conducted a *t*-test for dependent samples and an *F*-test of equality of variances both with a significance level of 5 %.[8] To test for substantial differences between means, we calculated Cohen's *d*. With regard to (4) frequencies of the assignment of individuals to proficiency levels, we examined whether differences occurred when statistically misfitting items were included or excluded. As mentioned in the defined criteria, the McNemar-Bowker test for pairwise symmetry was

[8] When applying inferential statistics to compare the measures of interest (i.e., means), we use test statistics for repeated measures, since the underlying data set and hence the people are the same.

applied with a significance level of 5 %. We calculated the effect size Cohen's *g* (1988) to investigate relevant changes of allocations to proficiency levels.

For a better understanding of how the assignment of individuals to proficiency levels changes as a result of item selection, the results were examined at the individual level. For this purpose, weighted likelihood estimates (WLEs) are used instead of PVs, since PVs are explicitly not suitable for reporting individual ability scores because the values are selected at random from a distribution (Lüdtke & Robitzsch, 2017; Wu, 2005). This estimation procedure can introduce bias in the estimated abilities of individuals; its use in scenarios with individual interpretations of the trait is not recommended. The scaling and assignment of persons to the proficiency levels based on WLEs was carried out as described above. The percentage of persons who were grouped at the same proficiency level before and after item selection was calculated as well as the proportion of people at each proficiency level who shifted to a higher or lower proficiency level.

Data example 2 for testing on population level: non-cognitive/metacognitive

The aim of this empirical example is to investigate two scales of the BQ and determine whether item selection has practical consequences for the distribution of the measured trait or for estimated correlations between the measured trait and another variable.

In this subsection, we describe the methodological procedure for evaluating the practical significance of item selection for scales with a small number of items. Firstly, the sample and measure as well as the process of scaling the data including all items, the item fit analyses, and the correlation analyses are presented. Afterwards, the procedure for rescaling without the statistically misfitting items and the reanalyses of the data are outlined. Lastly, the evaluation of the practical significance of item selection is presented. As for Example 1, the analyses were conducted in R (R Core Team, 2022).

Sample and measure

We chose two exemplary scales of the German PISA field trial data 2018, namely the scale "General fear of failure" (ST183) and the scale "Discriminating school climate" (ST223). Both scales were reported for the first time in 2018, showed a noticeable statistical item misfit, and are thus well suited to investigate the practical significance of item selection. The scale ST183 was assessed with five Likert-type items with four categories. 772 students gave responses to the ST183 items, of which 374 were female (48.45%). The scale ST223 was based on 10 Likert-type items with four categories. 804 students gave responses to the ST223 items, of which 393 were female (48.88%).

Data scaling and item fit analyses

Following the procedure used in the PISA study for scaling the BQ data (OECD, 2020), the two scales were each calibrated under the 2PL model/GPCM. For the item fit analyses, we estimated the RMSD. In the PISA main study (2015, 2020), an RMSD cut-off value of RMSD > 0.30 was used for the BQ data. As described above, this cut-off value is not directly transferable, which is why we used a cut-off value of RMSD > 0.04 (see Köhler et al., 2020).

Correlation analyses

According to the previously defined criteria, we calculated the WLE for each student. Since BQ data are frequently used as predictors in multiple regressions or for correlational analyses, we examine whether estimated correlations change when misfitting items remain in the scale. We used the gender variable as a demonstration because gender is one of the most prominant independent variables in educational research. We calculated the correlation between the measured trait and gender once when misfitting items were included and once when misfitting items were excluded. In addition, the mean and variance of the distribution of the measured trait were calculated within each gender group in order to investigate whether the item selection had a group-specific impact. Since this is a demonstration of the application of the criteria, only this one variable was used to evaluate the impact of item misfit. In real applications, one would perform this analysis for all relevant variables.

Rescaling and reanalyzing

For rescaling, we removed the statistically misfitting items from the scales based on the cut-off value RMSD > 0.04. For comparing the results with and without the misfitting items, the fixed anchor calibration approach was used for the linking, meaning that item parameters of the fitting items were fixed for the recalibration. For the scaling process, the 2PL model/GPCM was applied. Subsequently, reanalysis of the correlation between the measured trait and gender was conducted. Also, the means and the variances of the measured trait within each gender group were compared.

Evaluation of practical significance of item selection

Applying the defined criteria for evaluating the practical significance of item selection for non-cognitive/metacognitive data, a *t*-test for dependent samples was used to investigate the differences in means. To obtain an idea of whether these differences were meaningful, we calculated the effect size Cohen's *d*. We applied an *F*-test to test for equality of variances of the measured trait. To investigate the practical significance of item selection with respect to the similarity between the WLE estimates, we compared the WLE estimates of the calibrations with and without statistically misfitting items by calculating the rank correlation. Furthermore, the correlation between gender and the measured trait was investigated. To determine if differences were statistically significant, Fishers' *Z*-transformation of the correlation coefficients was conducted. In addition, we investigated whether the distributions of the measured trait in the subgroups were affected by item removal.

Results

Data example 1 for testing on population level: cognitive

Firstly, we present the general results on the RMSD as well as the number of items with a statistical item misfit, and the results on item difficulties. Afterwards, the results on the evaluation of the practical significance of item selection are shown.

Table 1 Comparison of the item difficulties before and after item selection

	Trend and new items				Trend items only			
	$\sum_{Items}$	Score Range	*M*	*SD*	$\sum_{Items}$	Score Range	*M*	*SD*
No selection	402	191.52–798.25	500.00	100.00	88	402.40–696.78	537.52	57.02
Cut-off = 0.06	266	201.01–797.88	483.98	114.20	13	450.93–623.98	527.42	41.13
Cut-off = 0.08	323	207.26–800.73	492.52	105.23	28	402.40–623.98	515.45	44.54

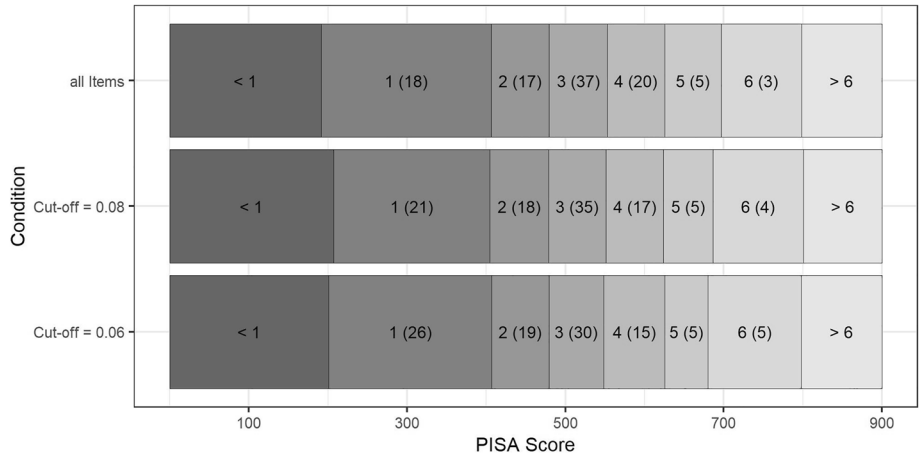

Fig. 2 Proficiency levels and the percentage of items (in parentheses) before and after item selection using the two cut-off criteria RMSD > 0.08 and RMSD > 0.06

General results

In the item fit analyses, the RMSD values of the items ranged between 0.010 and 0.329. Using a cut-off value of RMSD > 0.06, 136 items showed an item misfit, so that after the item selection, 266 items remained in the data set, 13 of which were trend items. Using a cut-off value of RMSD > 0.08, 79 items showed a statistical item misfit, so that after item selection 323 items remained in the data set, 28 of which were trend items.

For the trend items, Table 1 shows that using the different cut-off values lead to a change in the score range, mean and standard deviation. Whereas the mean decreases after item selection, the standard deviation increases for trend and new items but decreases for trend items only. This is probably because trend items lie more in the middle of the ability spectrum than the new items, especially the easy reading fluency items. The impact of item selection on the coverage of item difficulties at the specific proficiency levels is shown in Fig. 2. A reduction in the number of items was particularly noticeable at the intermediate proficiency levels. However, despite the reduced number of items, the item difficulty ranges per proficiency level were covered in a comparable manner. This means that the representation of item difficulties at the edges of the proficiency level was not noticeably affected by item selection.

Practical significance of item selection

Using RMSD > 0.06 as the cut-off value, the correlation between the ability estimates with and without item selection was $r = 0.92$, while using RMSD > 0.08, the correlation

was $r = 0.93$, indicating a highly monotonous relationship between the ability estimates. Including all items, the mean value for all PVs was 0.760 with a variance of 0.949. Using RMSD > 0.06 as a cut-off value, the mean was 0.766 and the variance 0.939. A cut-off value of RMSD > 0.08 resulted in a mean of 0.754 and a variance of 0.945 for the PVs. For both cut-off values, the results of the *t*-test for dependent samples showed no significant difference for the mean PVs at a significance level of 5 % before and after item selection (for RMSD > 0.06: $t(1889) = -0.79$, $p = 0.43$; for RMSD > 0.08: $t(1889) = 0.76$, $p = 0.45$). The effect sizes were negligible (for RMSD > 0.06: $d < 0.001$; for RMSD > 0.08: $d < 0.001$). For both cut-off values, the results of the *F*-test with $\alpha = 0.05$ indicated no significant differences in the variances of the ability estimates (for RMSD > 0.06: $F(1889, 1889) = 1.02$, $p = 0.67$; for RMSD > 0.08: $F(1889, 1889) = 1.01$, $p = 0.87$). Thus, there was no practical significance of item selection with respect to the location and variance of the ability distribution.

Table 2 displays the relative and absolute frequencies in the assignment of persons to the proficiency levels for the respective item selection criterion. Since with and without item selection no individuals were located at competence levels 1a and 1b, the proficiency levels 1a–c were merged into proficiency level 1. The results of the McNemar–Bowker test indicated that the null hypothesis was retained for both RMSD cut-off values. No significant differences were observed between the assignment to proficiency levels and the inclusion/exclusion of misfitting items, with only one exception: With the cut-off RMSD > 0.06, asymmetrical changes between proficiency levels 3 and 4 were observed insofar that significantly more people changed to level 4 than to level 3, with $p = 0.033$. The effect size Cohen's *g* lay below 0.1 for all comparisons, thus indicating small effects (Cohen, 1988).

The median PISA scores for the corresponding proficiency levels are provided in Table 2.[9] Table 2 shows that the ability distributions changed predominantly at the extreme proficiency levels. These shifts might result from a lower item representation per proficiency level or from measurement inaccuracies in the boundary areas of the proficiency levels. As described in the above mentioned criteria, it should be examined how the construct has changed due to item selection. Table 2 further illustrates how many individuals were assigned to a different proficiency level after item selection. The assignment to proficiency levels was relatively inconsistent: At least 17% of the persons per proficiency level shifted to another proficiency level after item selection.

Overall, the results show no practical significance of item selection on the ability distribution and on the frequencies of assigning individuals to the proficiency levels. Note that results at the population level are of predominant interest in scenarios with aggregated interpretations of results. Results at the individual level are not reported back to the student in PISA. Nevertheless, we found that single individuals frequently changed the proficiency levels after item removal.

Data example 2 for testing on population level: non-cognitive/metacognitive

Firstly, we present general results regarding the RMSD values and the WLE reliability. Secondly, results of the evaluation of the practical significance are shown.

[9] We used the median instead of the mean because of the skewed distribution of the abilities per proficiency level.

Table 2 Assignment to the proficiency levels for no item selection and selection with the RMSD cut-off 0.06 and RMSD cut-off 0.08

PL	Cut	All Items		Cut-off = 0.06					Cut-off = 0.08				
		N_{Level} (%)	*Md*	N_{Level} (%)	*Md*	N_{remain} (%)	N_{down} (%)	N_{up} (%)	N_{Level} (%)	*Md*	N_{remain} (%)	N_{down} (%)	N_{up} (%)
1	189	17 (0.95)	393.47	3 (0.16)	398.00	3 (17.65)	–	14 (82.35)	9 (0.48)	392.90	8 (47.06)	–	9 (52.94)
2	407	146 (7.83)	457.90	106 (5.63)	458.14	81 (55.48)	–	65 (44.52)	124 (6.77)	459.06	107 (73.23)	1 (0.68)	38 (26.03)
3	480	539 (28.62)	524.24	526 (27.93)	527.87	410 (76.07)	11 (2.04)	118 (21.89)	498 (26.35)	525.96	429 (79.59)	8 (1.48)	101 (18.91)
4	553	776 (41.16)	587.78	844 (44.82)	588.02	642 (82.73)	50 (6.44)	84 (10.82)	791 (42.01)	589.93	647 (83.38)	30 (3.87)	99 (12.76)
5	626	371 (19.63)	647.84	370 (19.65)	646.86	275 (74.12)	83 (22.37)	13 (3.50)	412 (21.80)	650.90	308 (83.02)	42 (11.32)	21 (5.66)
6	697	34 (1.80)	714.58	34 (1.81)	716.90	19 (55.88)	11 (44.12)	–	49 (2.59)	718.52	27 (79.41)	7 (20.59)	–

The results in the columns N_{Level} and *Md* are based on PVs, the columns N_{remain}, N_{down}, and N_{up} on WLE. N_{Level} = number of persons per proficiency level; N_{remain} = number of persons who remain on the same proficiency level before and after item selection; N_{down} = number of persons who change to a lower proficiency level after item selection; N_{up} = number of persons who change to a higher proficiency level after item selection; PL = proficiency level

Table 3 Descriptive results for the two scales *General Fear of Failure and Discriminating School Climate* before and after item selection

	General fear of failure			Discriminating school climate		
	Before item selection (5 items)	After item selection (3 items)	Δ	Before item selection (10 items)	After item selection (3 items)	Δ
θ mean	0.01	0.02	− 0.01	0.33	− 0.07	0.40***
θ sd	1.24	0.99	0.25***	1.17	1.07	0.10***
correlation of gender and WLE (SE)	− 0.19 (0.03)	− 0.21 (0.03)	0.02	− 0.06 (0.04)	− 0.04 (0.03)	− 0.02
θ mean (female)	0.22	0.24	− 0.02	0.10	− 0.03	0.07***
θ sd (female)	1.11	0.99	0.12	1.13	1.03	0.10
θ mean (male)	− 0.19	− 0.19	0.00	− 0.03	− 0.11	0.08**
θ sd (male)	1.08	0.95	0.13	1.21	1.11	0.10

A *t*-test was used to calculate significant differences between the means, an *F*-test of equality of variances was used for significant differences between the variances, and a Fishers' *Z*-transformation of the correlation coefficients was used for differences between the correlations.

θ = latent trait based on WLE estimation

** $p < .05$

*** $p < .01$

General results

Scaling all items, the RMSD values of the items of scale ST183 ranged from 0.014 to 0.063. Two items showed a higher RMSD value than 0.04 and were removed from the scale. For scale ST223, the items showed RMSD values in a range from 0.018 to 0.068. Seven of the ten items had a higher RMSD value than 0.04, with two items showing an RMSD value greater than 0.06 and three items greater than 0.05. These values indicate that the scale exhibited a high extent of item misfit. Since we assume that an increasing extent of item misfit results in the most severe practical consequences, we removed all seven items from the scale. As Table 3 displays, the WLE reliabilities decreased for both scales after item removal. Especially for scale ST223, the value decreased substantially, which is due to the large number of items we removed.

Practical significance of the item selection

Table 3 illustrates the results before and after item selection. The *t*-test for dependent samples indicated no significant difference between the means of ST183 at a significance level of 5% ($t(771) = 0.93$, $p = 0.351$). The effect size was negligible ($d < 0.01$). The result of the *F*-test indicated a significant difference in the variances of the distributions of the measured trait at a significance level of 5 % ($F(771, 771) = 0.79$, $p < 0.001$). However, the distributions of the respective WLE estimates overlap very strongly, with a correlation of $r = 0.99$, indicating no practically significant difference between the WLE estimates. A correlation analysis between the WLE estimates (with and witihout misfitting items) and gender was calculated. We detected no significant difference in the correlations (see Table 3), meaning that item selection had no practical consequences. No significant differences in the means and variances of the measured trait for females and males with regard to the categories were observed, either (see Table 3). Overall, hardly any practical significance of item selection was found for scale ST183.

A significant difference in the mean values of the distribution of the measured trait with and without misfitting items was detected for scale ST223 ($t(803) = -4.37$, $p < 0.001$). The effect size Cohen's d was 0.154, meaning that item selection had small practical consequences. The result of the F-test indicated a significant difference in the variances of the measured trait ($F(803, 802) = 0.85$, $p = 0.017$). The rank correlation between the estimates with and without item selection was high $r = 0.83$, but indicated differences between the WLE estimates. With respect to the relationship between the measured trait and gender before and after item selection, Table 3 shows that no meaningful differences between the correlations existed. Regarding the specific categories of the variable gender, the results of the t-tests for dependent samples for females and for males indicated significant differences in the mean values when comparing the inclusion and exclusion of misfitting items ($t(392) = 3.86$, $p < 0.01$ for females and $t(410) = 2.29$, $p = 0.022$ for males). For the difference in mean values, Cohen's d was 0.19 for females and 0.11 for males, meaning that item selection had small practical consequences. No significant differences in the variance of the measured trait were observed in the distributions of females and males ($F(392, 392) = 1.21$, $p = 0.055$ for females; $F(409, 410) = 1.15$, $p = 0.147$ for males).

Discussion

The two main objectives were to define criteria to evaluate the practical significance of item selection and to apply these criteria to the PISA 2018 field trial data of the German subsample. The use of field trial data to investigate practical consequences of item selection is crucial, since they represent the basis for decisions regarding the items in the main surveys, and a higher extent and amount of item misfit can be expected here. Therefore, these data were well suited to demonstrate the application of the established criteria. Although our results regarding the absence of practical significance cannot be generalized to other studies, they add to the scarce existing research. Another contribution of our study are the criteria we established, which can be applied to other assessments. The methods we propose contribute to the point made by Hambleton and others (e.g., Hambleton & Han, 2005, Sinharay & Haberman, 2014; Zhao & Hambleton, 2017), who stress the necessity of evaluating IRT models in terms of practical significance.

Our finding that the statistical item misfit was seldom associated with practical consequences for the distribution of the measured trait concurs with previous studies (Crişan et al., 2017; Köhler & Hartig, 2017; Liang et al., 2014; Sinharay et al., 2011; Sinharay & Haberman, 2014; Tendeiro & Meijer, 2015; van Rijn et al., 2016; Zhao, 2016). Since these studies predominantly investigated main survey data, we wanted to test practical implications when using field test data—expecting more severe consequences, which was not the case.

When evaluating the practical significance of item selection, various methodological approaches are applied in the literature. This is certainly also due to the fact that, depending on the application scenario, different research foci are set, each of which is also accompanied by different test outcomes that are of interest. The latter is reflected in the different methods for evaluating the practical consequences of item selection. Using the same criteria for evaluating the practical consequences of item selection for the same application scenarios and the same test outcomes can also lead to greater comparability

of the results obtained with regard to the practical significance of item selection. A first step towards this can be our established criteria, which provide options for evaluating the practical consequences of item selection. These criteria do not claim to be exhaustive and can be extended for other intended test uses. The chosen criteria certainly need to comply with the intended use of the test.

Whenever practical consequences of item selection are evident, the question arises how to deal with statistically misfitting items. Crișan et al. (2017) point out that there are three ways in which misfitting items can be dealt with. One is that the item misfit can be ignored, which may affect the accuracy of the model parameter estimates. Second, the items with misfit can be removed, but this may lead to insufficient construct coverage. A third option is to use a better-fitting model, which could cause estimation problems, among others. From our point of view, the choice of one of these strategies is closely related to the question of whether the model is more appropriate with or without the misfitting items. From our point of view, it is not possible to give a general answer to the question whether a scale is fundamentally better by keeping or removing misfitting items by analyzing the statistical and practical significance of item fit, because the occurrence of practical consequences indicates that two slightly differing constructs are being measured. Whether the construct with or without misfitting items is the construct to be mapped has to be examined substantially.

If, in contrast, there are no practical consequences of item misfit, the question of how to deal with the misfitting items arises. In general, a user can ask two questions: (1) How many items are in the data set and how many of them show a misfit? (2) Is the misfitting item needed to represent the content of the construct? For example, if there is only one misfitting item in a data set, or if there is only a very small percentage of misfitting items in a medium or large data set and they contribute to construct coverage, the item could be kept in the data set. The item then contributes to higher test information without biasing the results. In addition, a single misfitting item is unlikely to bias test outcomes. If many items show a misfit, it should be examined, especially for the items with the largest item misfit, whether they are needed for construct coverage and whether they are inconspicuous regarding other item checks (e.g., DIF). When making the decision of item removal based on the testing of practical consequences, please note that those consequences of are calculated for specific test outcomes. An item selection could have consequences on test outcomes not investigated.

As with any study, this work is subject to limitations. First, it is debatable whether our operationalization of practical significance of item selection meets the core of practical significance. To evaluate whether an item selection is practically significant, the criteria we established again involve statistical analyses. This approach could be seen as shifting the problem of evaluating statistical item misfit to other inferential analyses. Note, however, that only changes in the outcome of interest allow determining practical consequences for a particular test assessment. Furthermore, we emphasize that particular attention should be payed to effect sizes and not to inferential analyses.

Another limitation of the study is that the results presented here refer to an exemplary empirical field trial data set. This naturally limits the generalizability of the

results. In future studies, it should be tested whether the established criteria are also applicable in practice beyond our data examples, and which adjustments or additions are necessary.

Note also that a major aim of international LSAs is to make country comparisons. In this study, we only used PISA 2018 field trial data of the German subsample and only one test cycle, meaning that a comparison across countries or test cycles was not possible. Future research could shed light on whether country rankings are affected differently by the presence of misfitting items. Furthermore, it would be interesting to examine in which way the development of country-specific proficiency levels over time are affected by practical consequences of item selection.

We solely focused on practical significance of item misfit in LSAs, and discussed evaluation criteria for this specific study design. In future research, it could be examined whether and which adjustments of the criteria are necessary for other study designs such as CAT. In addition, it would be of interest to examine to what extent factors such as interactions with other model violations such as multidimensionality affect the practical significance of the item selection. Another important consideration to make is that misfitting items have potential nefarious effects on the performance of otherwise well-fitting items. Studies have shown that increased proportions of aberrant responses affect parameter estimates of fitting items as well (Silva Diaz et al., 2022; Su et al., 2007), which can affect reliability of the scale as a whole. In this regard, it is important to note that we only used item misfit as a selection criterion. Further assessment of test quality at the scale level was not conducted. Another option could be to crosscheck to what extent our results differ when we randomly select items. If similar results are obtained, this would suggest that the practical consequences result from the selection of the items themselves rather than the exclusion of misfitting items.

Conclusion

In this work, the evaluation of practical significance of item selection based on item misfit was investigated, which is an important step in the evaluation of model fit. First, criteria were put up for discussion that can be used to investigate the practical consequences of item selection for various application scenarios. Second, the application of these criteria was demonstrated with two data examples from the PISA 2018 field trial study of the German subsample for both cognitive and non-/metacongnitive data. Our results indicate that item selection based on item misfit is not necessarily associated with practical consequences for the latent trait. Nevertheless, the evaluation of the practical consequences should be included in the item fit analyses in order to make a well-founded decision whether a statistically misfitting item should be removed from the test or not. Overall, the criteria we established can be a first step for comparative evaluations of the practical consequences of item misfit.

Acknowledgements

We would like to thank Jorge N. Tendeiro and the anonymous second reviewer for their helpful suggestions.

Author contributions

KF, CK, and JH designed the conceptual framework. JHH provided the study materials of the PISA field trial 2018 of the German subsample and advised on the data preparation. KF planned the the study design. KF conducted the analyses. KF wrote the original draft. CK contributed to the writing of the manuscript. JH, and JHH reviewed and edited the paper. CK supervised the study. Project funding was acquired by CK and JH. All authors have agreed both to be personally

accountable for the author's own contributions and to ensure that questions related to the accuracy or integrity of any part of the work, even ones in which the author was not personally involved, are appropriately investigated, resolved, and the resolution documented in the literature. All authors read and approved the final manuscript.

Funding
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).

Availability of data and materials
The data that support the findings of this study are available from the German PISA national center but restrictions apply to the availability of these data, which were used under license for the current study, and are not publicly available.

Declarations

Ethics approval and consent to participate
In this manuscript, secondary analyses were conducted using the PISA 2018 field trial data of the German PISA sub sample. This dataset was used under license from the German PISA national center for the current study, and are not publicly available. Therefore, neither consent to participate or consent for publication nor ethics approval were required for the reported analyses.

Competing interests
The authors declare that they have no competing interests.

Received: 26 April 2021 Accepted: 22 June 2022
Published online: 12 July 2022

References

ACARA. (2013). *National assessment program - science literacy technical report 2012*. Australian Curriculum, assessment and reporting authority.
Allen, N. L., Carlson, J. E., & Zelenak, C. A. (1999). *The NAEP 1996 technical report*. National center for educational statistics.
Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Addison-Wesley.
Box, G. E., & Draper, N. R. (1987). *Empirical model-building and response surfaces*. Wiley.
Cohen, J. (1988). Statistical power analysis for the behavioral sciences. *Routledge*. https://doi.org/10.4324/9780203771587
Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159.
Crişan, D. R., Tendeiro, J. N., & Meijer, R. R. (2017). Investigating the practical consequences of model misfit in unidimensional IRT models. *Applied Psychological Measurement, 41*(6), 439–455. https://doi.org/10.1177/0146621617695522
De Ayala, R. J. (Ed.). (2009). *Methodology in the social sciences. The theory and practice of item response theory*. New York: Guilford Press.
Dorans, N. J., & Feigenbaum, M. D. (1994). *Equating issues engendered by changes to the SAT and PSAT/NMSQT. ETS Research Memorandum*. Princeton, NJ: Educational Testing Service.
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*. Degnon Associates.
Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: principals and applications*. Springer.
Hartig, J., Frey, A., & Jude, N. (2020). Validity of test value interpretations. In H. Moosbrugger & A. Kelava (Eds.), *Testtheorie und fragebogenkonstruktion*. Springer.
Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New Phytologist, 11*(2), 37–50. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x
Kirk, R. E. (1996). Practical significance: a concept whose time has come. *Educational and Psychological Measurement, 56*(5), 746–759. https://doi.org/10.1177/0013164496056005002
Köhler, C., & Hartig, J. (2017). Practical significance of item misfit in educational assessments. *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*(3), 251–273. https://doi.org/10.3102/1076998619890566
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
Lüdtke, O., & Robitzsch, A. (2017). An introduction to the plausible values technique for psychological research. *Diagnostica, 63*(3), 193–205. https://doi.org/10.1026/0012-1924/a000175
Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174. https://doi.org/10.1007/BF02296272
Molenaar, I. W. (1997). Lenient or strict application of IRT with an eye on practical consequences. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 38–49). Waxmann Verlag.
Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159–176.
OECD. (2015). *PISA 2015 field trial analysis report: Outcomes of the cognitive assessment (Meeting of the technical advisory group)*. Paris: OECD Publishing.
OECD. (2018a). *PISA 2015: PISA results in focus*. OECD Publishing.

OECD. (2018b). *PISA 2018 Field trial analysis report for the cognitive assessment*. OECD Publishing.
OECD. (2020). *PISA 2018 technical report*. OECD Publishing.
Peeters, M. J. (2016). Practical significance: Moving beyond statistical significance. *Currents in Pharmacy Teaching and Learning, 8*(1), 83–89. https://doi.org/10.1016/j.cptl.2015.09.001
R Core Team (2022). R: A language and environment for statistical computing [Computer software] R foundation for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/
Rasch, G. (1960/1980). Probabilistic models for some intelligence and attainment tests. (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.
Reiss, K., Weis, M., Klieme, E., & Köller, O. (Eds.). (2019). *PISA 2018: Grundbildung im internationalen Vergleich [PISA 2018: Basic education in international comparison]*. Waxmann.
Robitzsch, A., Kiefer, T., & Wu, M. (2020). TAM: Test analysis moduls. *R package version, 3*, 5–19. Computer software.
Rutkowski, L. (2014). Sensitivity of achievement estimation to conditioning model misclassification. *Applied Measurement in Education, 27*(2), 115–132. https://doi.org/10.1080/08957347.2014.880440
Silva Diaz, J. A., Köhler, C., & Hartig, J. (2022). Performance of Infit and outfit confidence intervals calculated via parametric bootstrapping. *Applied Measurement in Education*. https://doi.org/10.1080/08957347.2022.2067540
Sinharay, S., Haberman, S. J., & Jia, H. (2011). *Fit of item response theory models: A survey of data from several operational tests (Research Report* (Vol. No. RR-11-29)). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2011.tb02265.x.
Sinharay, S., & Haberman, S. J. (2014). How often is the misfit of item response theory models practically significant? *Educational Measurement: Issues and Practice, 33*(1), 23–35. https://doi.org/10.1111/emip.12024
Su, Y. H., Sheu, C. F., & Wang, W. C. (2007). Computing cis of item fit statistics in the family of rasch models using the bootstrap method. *Journal of Applied Measurement*, *8*(2), 190–203. https://www.ncbi.nlm.nih.gov/pubmed/17440261
Swaminathan, H., Hambleton, R. K., & Rodgers, H. J. (2006). Assessing the fit of item response theory models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Psychometrics* (pp. 683–718). Elsevier.
Tendeiro, J. N., & Meijer, R. R. (2015). How serious is IRT misfit for practical decision-making? *LSAC Research Report Series, 15*(4), 1–22.
Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools, 44*(5), 423–432. https://doi.org/10.1002/pits.20234
Tijmstra, J., Bolsinova, M., Liaw, Y.-L., Rutkowski, L., & Rutkowski, D. (2020). Sensitivity of the RMSD for detecting item-level misfit in low-performing countries. *Journal of Educational Measurement, 57*(4), 566–583. https://doi.org/10.1111/jedm.12263
Van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. Springer. https://doi.org/10.1007/978-1-4757-2691-6
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.
Wainer, H., & Thissen, D. (1987). Estimating ability with the wrong model. *Journal of Educational Statistics, 12*(4), 339–368. https://doi.org/10.3102/10769986012004339
Wu, M. L. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation, 31*(2–3), 114–128. https://doi.org/10.1016/j.stueduc.2005.05.005
Zhao, Y. (2016). Impact of IRT item misfit on score estimates and severity classifications: an examination of PROMIS depression and pain interference item banks. *Quality of Life Research, 26*(3), 555–564. https://doi.org/10.1007/s11136-016-1467-3
Zhao, Y., & Hambleton, R. K. (2017). Practical consequences of item response theory model misfit in the context of test equating with mixed-format test data. *Frontiers in Psychology, 8*, 1–11. https://doi.org/10.3389/fpsyg.2017.00484

Publisher's Note