Person Misfit, Test Anxiety and Test-Taking Motivation in a Large-scale Mathematics Proficiency Test

Christian Spoden[1], Jens Fleischer[2], & Andreas Frey[3]

[1]German Institute for Adult Education - Leibniz Centre for Lifelong Learning, Bonn, Germany

[2]University of Duisburg-Essen, Essen, Germany

[3]Goethe University Frankfurt (Main), Germany, and Centre for Educational Measurement (CEMO) at the University of Oslo, Norway

Corresponding author:

Christian Spoden

German Institute for Adult Education - Leibniz Centre for Lifelong Learning, Bonn, Germany

Heinemannstr. 12-14

53175 Bonn

Telefon: +49 (0)228 32 94-141

Email: spoden@die-bonn.de

Person Misfit, Test Anxiety, and Test-Taking Motivation in a Large-Scale Mathematics Proficiency Test for Self-Evaluation

In many countries the international large-scale assessments like the Programme for International Student Assessment (PISA; OECD, 2019) or the Trends in International Mathematics and Science Study (TIMSS; Mullis, Martin, & Loveless, 2016) have become the pulsebeat of educational monitoring and the starting point for various educational reforms. In Germany, for instance, an unexpectedly low performance of the German students in the PISA 2000 assessment effectuated the development of educational standards (e.g., Neumann, Kauertz, & Fischer, 2010; XXX), which are now continuously assessed and monitored based on student samples as a measure of quality assurance in education (National Assessment of Proficiencies; e.g., Stanat et al., 2019). Furthermore, state-wide assessments of curricular competencies (Leutner, Fleischer, Spoden, & Wirth, 2007; Spoden & Leutner, 2012), which each student in Germany usually takes twice at relevant dates in their schooldays (3rd and 8th grades), measure which standard-based competencies all students really have achieved. These assessments are designed as a form of self-evaluation (Spoden & Leutner, 2012) within schools. They provide feedback for individual teachers about the aggregated proficiency levels of students in their courses in order to then initiate processes of school improvement against the background of the national educational standards. Along with the National Report on Education reporting on additional indicators for educational progress (Authoring Group Educational Reporting, 2016), these different assessments constitute a comprehensive structure of educational monitoring and quality assurance in the German school system, providing, in general, both reliable and valid aggregated information on proficiencies of student cohorts or learning groups. These elements of the monitoring system are not designed for individual test score interpretation and feedback to the individual student (Spoden & Leutner, 2012), even though frequent requests by teachers acknowledge their wish to use the testing instrument for the purpose of grading the individual. The tests do not achieve a sufficient level of measurement precision at the individual level (XXX). However, this is not the only drawback for individual test score interpretation from such a study as will be outlined in the following.

Various studies in educational research have demonstrated that affective and motivational variables in a test situation, such as test anxiety or test-taking motivation, have impact on the individual results of standardized proficiency tests (e.g., Cassady, 2010; Eklöf, 2008; Knekta, 2016; Seipp, 1991; Sundre & Kitsantas, 2004; Wise & DeMars, 2005). These results indicate that there are students for whom the performance measured in the test is below their best possible performance (e.g., Asseburg & Frey, 2013). This distorts Cronbach's (1970) conceptual idea of measuring maximum performance as an indicator of the underlying latent ability, which underlies most achievement measures, at least implicitly. Test developers have a range of statistical methods at their disposal to assess the validity of test score interpretations, especially within the framework of the item response theory (IRT; e.g., van der Linden, 2016), which is underlying the test scoring in large-scale studies as assessments of educational evaluation. The IRT is a modern alternative to the more traditional and still more frequently used approach of test design under the classical test theory (CTT; e.g., DeMars, 2018). One of the strengths of IRT is that the model assumptions (local stochastical independence, monotonicity, unidimensionality; see van der Linden, 2016) can be checked and tested in the data to prove the validity of the conclusions drawn about the scores. For example, item fit measures (e.g., Chalmers & Ng, 2017) provide information on the fit of each single test item to identify poor performing and potentially inaccurately designed items. The results from IRT item fit analysis oftentimes bring up results, which are very much in line with results from the well-known point-biserial item-/total-score correlation prevalently used traditional test design. An IRT-based approach previously discussed as a method to examine the validity of individual test score interpretations is the person fit analysis (Schmitt, Chan, Sacco, McFarland, & Jennings, 1999). Person fit measures (e.g., Meijer & Sijtsma, 2001; Tendeiro, Meijer, & Niessen, 2016) have been developed under the IRT framework to assess the model fit of a student's individual response pattern to the estimated item parameters. The development of person fit measures is also based on the assumption that substantial discrepancy between expected and observed item responses of an individual student may indicate undesired effects occuring during test administration that affected the students' test performance (XXX). Low person fit gives some indication to treat the estimated ability levels with caution and undertake further actions to validate or modify the scoring, or even decide for re-testing of the individual (Smith, 1985). In the context of educational monitoring and school evaluation this gives an additional reason to be reluctant about individual test score reporting and possibly even adjust aggregated results (e.g., by trimming or weighting the response patterns before aggregation; see Brown & Villareal, 2007; Rudner, Bracey, & Skaggs, 1996). This is especially relevant when the significance of affective or motivational characteristics that may affect the performance in the test situation is considered. In this article data from a state-wide assessments of curricular competencies used as an instrument of school evaluation are investigated with the goal of analyzing whether substantial person misfit exists and variance in person fit can be statistically related to test anxiety in mathematics and to test-taking motivation. The goal is to give further empirical evidence that the described issue is in fact present in large-scale data used for evaluation purposes in education and may provoke incorrect conclusions possibly drawn from these tests.

The text is organized as follows. First, the rationale for person fit analysis and its usefulness for assessing the validity of individual test scores are described. Then, previous empirical results are presented that illustrate the state of research on correlations of person fit and individual test behavior. This is followed by an empirical study investigating person fit and its correlates in a large-scale assessment. The text ends with a discussion of the results and practical recommendations concerning individual and group-specific test score interpretations in large-scale studies aiming to provide information for school evaluation.

The Rationale for Person Fit Analysis

Person fit measures provide information on the validity of test score interpretations as a measure of the model fit of the student's individual response pattern (Meijer & Sijtsma, 2001; see also Tendeiro, Meijer, & Niessen, 2016 for a software implementation). This is easily illustrated by a short example (quite similar to the example previously given in Meijer, 1996). Assume a fictitious test with ten items and, for reasons of simplicity, percentages of correct responses (item difficulty) equally distributed in decreasing order from .95 to .05 taken by four students. Table 1 shows the fictitious response patterns. The first two response patterns show expected responses where the probability of a correct response, coded as *one* (incorrect responses coded as *zero*), decreases with increasing item difficulty (Pattern 1: 1111110000; Pattern 2: 1111011000). These response patterns obtain high levels of person fit given the underlying IRT model. The third and fourth response patterns, displayed in the lower part of the table (Patterns 3 and 4), on the other hand, each indicate aberrant response patterns and, most probably, person misfit. The third response pattern of a student with medium proficiency (four correct item responses) includes incorrect responses to three easier items of difficulty .95, .85 and .75 at the beginning of the test (Pattern 3: 0001111000); this might be an indication of nervousness at the beginning of the test. In the last response pattern (Pattern 4: 1010111110) of a person with higher proficiency (seven correct item responses), incorrect responses were given to isolated easier items of difficulty .85 and .65, possibly because of inattention or lack of motivation and effort (Haladyna, 2004; Meijer, 1996). Person fit measures express these differences in the fit or misfit of a response pattern to the item difficulty in a single statistical coefficient, which can be compared across individuals, even individuals with different proficiency. Person fit analysis can thus be utilized to identify students who require adjustment of the response pattern or re-testing before estimating their proficiency levels. Smith (1985) summarized four typical options to handle measurement disturbances in the individual response patterns, depending on the specific testing situation. These include reporting several subtest abilities, modifying the response string (e.g., eliminating unreached items) and re-estimating the student's ability, not reporting any score and retest the individual, and deciding that the error introduced into the estimation by the measurement disturbance has only marginal impact on the students's total proficiency estimate. In educational contexts results from person fit analysis have previously also been applied to identify individuals who experienced different learning opportunities (Harnisch & Linn, 1981) and to check for the fairness of adjustments in testing situations (e.g., for students with disabilities; Engelhard, 2008).

--- please insert Table 1 here ---

Correlates of Person (Mis-)Fit

Several studies have also provided empirical evidence that person misfit is indeed related to validity issues (e.g., Meijer 1997; Schmitt, Chan, Sacco, McFarland, & Jennings, 1999). Individual characteristics most often discussed as potential origins of aberrant responding (and therefore as possible correlates of person fit) include test anxiety and test-taking motivation (see also Haladyna, 2004, p. 237-242). The term test anxiety describes emotional, physiological, and behavioral responses to concerns of the students about a possible failure in an evaluative situation like a test or exam (XXX). It is assumed that test-anxious students experience stress-induced interfering thoughts about this failure during test administration, are cognitively affected (by decreasing the working memory performance; Angelidis et al., 2019), and this leads to difficulties in correctly solving even items of lower difficulty (similar to Response Pattern 4 in Table 1). As a consequence, the test performance is deteriorated (e.g. Cassady, 2010; Seipp, 1991) and aberrant response patterns may occur. Students with a high level of test-taking motivation work on test items very carefully, and this behavior can be assumed to be accompanied by response patterns similar to the Guttman pattern (XXX). The Guttman pattern implies a deterministic response behavior: up to a difficulty level corresponding to their ability level, all item responses are correct but all responses to items of higher difficulty are incorrect (similar to Pattern 1 in Table 1). On the other hand, students with a low level of test-taking motivation are expected to give up on the test prematurely, to make careless mistakes, or to randomly guess (XXX). All of these types of behavior result in incorrect item responses even to easy items and lead to an underestimation of the true ability level (e.g. Wise, Pastor, & Kong, 2009; see also Meijer, 1996, for similar theoretical assumptions on item response behavior).

Findings on the effects of test anxiety and motivation on person fit in standardized testing are somewhat inconsistent. Analyzing a proficiency test (Metropolitan Achievement Test) with the subscales reading, mathematics, and science, Schmitt and Crocker (1984) found that low-performing test-anxious students displayed lower fit and high-performing test-anxious students displayed higher person fit than nonanxious students. Petridou and Williams (2007) and also Dodeen and Darabi (2009) investigated both test anxiety and test-taking motivation as possible correlates of person fit in mathematics tests. Petridou and Williams (2007) considered compositional effects in students aged 5 to 11. In their study, the person fit measures were aggregated at the level of school classes. In a regression analysis at the individual level, significant correlations between test anxiety or motivation and person fit were demonstrated in a newly developed mathematics test (odds ratios between 0.85 and 1.10; the person fit measure was split into a binary variable). In the subsequent multilevel analysis at the individual and school-class level, taking the class composition with respect to both characteristics into account, no significant effect was found. Dodeen and Darabi (2009) examined the correlations between test anxiety and the test-taking motivation of 10th-grade students with person fit in a mathematics test. They found weak correlations of $r = .20$ with test anxiety and a moderate negative correlation of $r = -.41$ with test-taking motivation. Given the polarity of the variables, these correlations indicate aberrant response patterns in anxious and less motivated students.

Research Aim

The aforementioned studies relied on data sets generated in different testing contexts, provided still inconsistent results with regard to correlates of person fit, and are thus not significant in terms of whether or not consequences need to be drived for large-scale assessments as part of the German educational monitoring system. Especially state-wide assessments of curricular competencies involve feedback to teachers on their learning groups of potentially smaller size (sometimes less than 15 students) and are periodically requested to provide even individual feedback about the students' proficiency level. In this study it is aimed to provide evidence for the potential thread to validity of the test score interpretations on individual level in these and similarly designed evaluation studies, which can be attributed to relevant individual measurement disturbances arising from test anxiety and test-taking motivation. Two hypotheses were investigated:

(1) A relevant number (> 5 %) of students shows misfitting response patterns.

(2) Person fit is negatively correlated with test anxiety and positively correlated with test motivation.

Lower person fit by students with higher levels of test anxiety and lower levels of test motivation would indicate that these students give responses that cannot be validly mapped to a proficiency level. This corrupts to some degree the test score interpretations that were suggested by the educational administrators to the teachers. Especially in educational assessments with a reference to curricular guidelines or national educational standards, aberrant response patterns are noticeable as the test content involves a (learning) progression from items requiring more basic skills to items requiring the same basic skills plus much more advanced skills. This induces a specific order of item difficulties. Person fit measures are sensitive to violations against the property of invariant item ordering, especially the person fit statistic H_j^T applied in this study (Sijtsma, 1986; Sijtsma & Meijer, 1992). The statistic H_j^T was used as a person fit measure to ensure the best possible detection of person misfit when testing the two hypotheses. The power of this measure to detect misfit has been proven by extensive Monte Carlo simulations and outperformed any of the statistics applied in the aforementioned studies (Karabatsos, 2003; Dimitrov & Smith, 2006).

Method

Sample

The data used was obtained from a large-scale assessment of mathematics proficiency in the eighth grade; this data was collected for one federal state of Germany (North Rhine-Westphalia). Data on affective and motivational variables were available from a subsample of school classes that completed supplementary questionnaires (see below) on a voluntary basis, because the assessment is usually not accompanied by extensive background questionnaires for reasons of data protection. The schools were invited by a member of the research team to voluntarily participate in the study. Two to three classes from four school tracks (*Gymnasium*, *Gesamtschule*, *Realschule*, *Hauptschule*[1]) were included in this sample, resulting in a total of 10 classes. The classes comprised a total of 228 students (53.1% female). After the test was administered, all students were asked by a research assistant about their feelings and their motivation during the test.

Instruments

[1] The Gymnasium, the Realschule and the Hauptschule are the highest, the middle and the lowest secondary school track in the state of North Rhine-Westphalia. The Gesamtschule is an integrative school concept, where students of different proficiency either attend the same courses (up to some school grade) or are assigned to different courses within the same school, depending on their profiency level.

The assessment framework of the assessment was based on the curricula of the federal state of North Rhine-Westphalia, which relies on the national educational standards in Germany. As outlined above, the primary aim of the assessment was to provide schools with valid and reliable information on the proficiency of their students in order to then initiate processes of school improvement (Leutner, Fleischer, Spoden, & Wirth, 2007; Spoden & Leutner, 2012). To a large extent, the way in which the schools decided to handle the results was up to them but they did need to report their conclusions about the assessment results and justify their upcoming activities for school improvement to the school supervisory board.

Items from the test instrument were developed to map content areas of secondary education in mathematics and were repeatedly checked for any shortcomings by teams composed of content matter experts from mathematics education and professional test developers. The proficiency test in mathematics comprised a total of 33 items, which were subdivided into two booklets of two different levels of mean item difficulty. The booklet with an item sample of lower mean difficulty was administered to the Realschule, the Hauptschule and the more basic level courses in the Gesamtschule; the booklet with an item sample of higher difficulty was administered to the Gymnasium and the more advanced level courses in the Gesamtschule. Thus, the booklets were assigned depending on the school tracks and according to the expected mean proficiency levels of students in these tracks. Booklet and school track were consequently confounded. However, a common test metric was established by an overlap of 15 items in the two booklets and a common (Rasch) scaling. Given that a sufficient test time was assured and not-reached items were therefore avoided, omitted items were coded as incorrect, indicating nonperformance on this item. Compared to the original scaling procedure, three items originally scored as partial credit were dichotomized for the present study, so that all items were available in dichotomous response format to facilitate the interpretation of the fit measure.

Test anxiety and test-taking motivation were assessed by means of an additional survey administered to the participating schools. In response to the requests expressed by these schools, established short scales previously applied in the German national test of the PISA 2006 assessment (Ramm et al., 2006) were used to keep the students' stress in the test situation as low as possible. Test anxiety in mathematics (TAM) was assesed using five items measuring test anxiety from the (academic) anxiety scale originating from the Project for the Analysis of Learning and Achievement in Mathematics (Pekrun et al., 2005). Test-taking motivation (TMO) was measured by three items originating from the (posttest) scale on invested effort of the on-line motivation questionnaire (Boekaerts & Otten, 1993).

Satisfactory reliabilities were found for all scales. The large-scale proficiency test in mathematics (sample statistics: M = -0.29; SD = 1.14) obtained a reliability of rel_{EAP} = .86. The TAM scale (M = 0.03, SD = 0.89) showed an EAP reliability of rel_{EAP} = .76 with item-score correlations between .70 and .86. The reliability of the TMO scale (M = 0.00, SD = 0.84) was rel_{EAP} = .68 with item-score correlations between .78 and .86. The correlations between TAM or TMO and MATH were as expected: TAM was negatively correlated to MATH (r = -.25, p = .02), TMO was positively correlated to MATH (r = .13, p = .09), and TAM and TMO were negatively correlated (r = -.18, p = .05).

Data Analyses

In the first step of the data analysis, the items of the mathematics scale from both test booklets were scaled according to the (dichotomous) Rasch model. The scalability criterion was a weighted item mean squared error between 0.8 and 1.2. In the second step of the data analysis, the person fit measure H_j^T (Sijtsma, 1986) was computed for both booklets of the complete sample ($N \approx$ 53,000, Booklet A, and $N \approx$ 135,000, Booklet B) to identify person misfit. Considering a student j with a given response pattern and a second student from the sample, l = 1,…, n, the index H_j^T is given by

$$H_j^T = \frac{\sum_{l \neq j}(\beta_{jl} - \beta_j \beta_l)}{\sum_{l \neq j} \min\{\beta_j(1-\beta_l),(1-\beta_j)\beta_l\}}, \qquad [1]$$

where β_j and β_l are the individual proportions of correct responses given by student j and by student l, respectively, and β_{jl} is the proportion of correct responses shared by student j and student l. The term $\min\{\beta_j(1-\beta_l\ ),(1-\beta_j)\beta_l\}$ refers to the lower of two products of correct response patterns of students j and l. The maximum of the measure is H_j^T = 1, which indicates a perfect matching in the item responses of student j and those of all of the remaining students. A measure of H_j^T = 0 indicates, on average, no correlation between student j's responses and those of other students (Sijtsma & Meijer, 1992). Referring to a rule of thumb by Karabatsos (2003), measures of H_j^T<.22 indicate violations against the Rasch model. The person fit measure H_j^T indicates sensitivity to misfit due to violations against the property of invariant item ordering (Sijtsma & Meijer, 1992), which is crucial for the Rasch model. Karabatsos (2003) found H_j^T to outperform 35 person fit measures in terms of detecting misfit (violations against the Rasch model) in the response pattern. Dimitrov and Smith (2006) replicated the advantage of H_j^T in the misfit detection compared to typical and adjusted Rasch person fit statistics. Karabatsos (2003) argued that H_j^T outperformed other powerful person fit statistics because it quantifies the conformity between an individuals's response pattern with the response patterns of all the remaining respondents while other powerful statistics compare the pattern against the typical response pattern summarized over correct response proportions in the sample of respondents. Thus, H_j^T is more sensitive to all individual item response patterns. Additionally, H_j^T outperformed many other (parametric) person fit statistics that relied on estimated model parameters as these parametric statistics use the same data set twice to estimate the model parameters and construct the estimated predictions for correct responses and to measure the model fit to the same predictions of the fit of the data to the estimated item parameters. Nonparametric person fit statistics like H_j^T do not rely on these dependence between data and parameter estimates which seems to be an essential aspect to detect inconsistent patterns (Karabatsos, 2003).

With regard to the TAM and TMO scales, the responses to all items were coded as missing unless at least three completed item responses were available. On the longer TAM scale, missing values on not more than two items (i.e., less than 50% of the items from this scale) were imputed using the two-way method (Sijtsma & van der Ark, 2003). The TMO scale was not imputed due to the low overall number of items. This involved remaining non-responses on each of the variables in a low percentage (TAM: 20 individuals, 9 %; TMO: 18 individuals, 8 %). Subsequently, different polytomous IRT models were estimated for the item responses from these two scales and were then compared on the basis of information criteria. A restricted graded-response model (Samejima, 2016) with a constraint discrimination parameter displayed the best model fit for both the TAM and TMO scales.

Misfitting response patterns were determined by computing the percentage of response patterns with H_j^T<.22 (see Karabatsos, 2003). The statistical effects of TAM and TMO on H_j^T were analyzed using regression models with robust standard errors, adjusting for the nested data structure (students nested in school classes). Three analysis models were compared by means of information criteria: Due to the separate computation of H_j^T for each booklet, an effect-coded booklet indicator (in the following: BOOK) was included in Model 1 as a control variable. In addition, the expected a posteriori (EAP; Bock & Mislevy, 1982) estimate of mathematical proficiency was dichotomized by median split and included in the model as a predictor (in the following: MATH_Dummy) in order to be able to statistically consider the interaction effects of TAM and TMO with different levels of test performance (low test performance, high test performance). The variable was dichotomized as it allows for a simpler interpretation of interactions and to investigate effects in more detail for different performance groups. Also note that a previous study reported negative correlations between H_j^T and the raw score (De Leeuw & Hox, 1994). For the predictor TAM, Model 1 is given as follows (the individual index j was omitted here for reasons of clarity):

$$\text{Model 1: } H^T = B_0 + B_1 \text{ BOOK} + B_2 \text{ MATH\_Dummy}. \qquad [2]$$

In Model 2, ability level estimates for TAM were added as predictors:

$$\text{Model 2: } H^T = B_0 + B_1 \text{ BOOK} + B_2 \text{ MATH\_Dummy} + B_3 \text{ TAM}. \qquad [3]$$

In Model 3, interaction effects between TAM and the mathematical ability estimate were also included in the model:

$$\begin{aligned}\text{Model 3: } H^T = B_0 + B_1 \text{ BOOK} + B_2 \text{ MATH\_Dummy} + B_3 \text{ TAM} + \qquad [4]\\ + B_4 \text{ MATH\_Dummy TAM}.\end{aligned}$$

For the predictor TMO, the same models were estimated, but with TAM replaced by TMO (Model 4-6). Each of these regression models was estimated by means of the lm.cluster function from the R package “miceadds” (Robitzsch, Grund, & Henke, 2017).

Results

Results Concerning Person Misfit Measured by H_j^T

The distribution of the person fit measure H_j^T in the sample investigated (*M* = .38; *SD* = 0.13) is presented as a histogram in Figure 1. It should be noted that five response vectors could not be interpreted in terms of person fit due to perfect scores (2.2 %). This, however, is a valid result in terms of the fit measure and is not a missing value caused by a nonresponse. Thus, it was decided not to impute these measures but to reduce the sample size to 223 students with clearly interpretable response patterns. Figure 1 illustrates that the majority of these response patterns (195 response patterns; 87.4 %) obtained satisfying model fit given the cutoff of .22 proposed by Karabatsos (2003), but a subsample of students (28 response patterns; 12.6 %) showed person misfit below this cutoff that was in need of explanation.

--- please insert Figure 1 here (monochrome color scheme) ---

Results Concerning the Effects of TAM on H_j^T

The results of the regression models estimating the effects of TAM on H^T are given in Table 2. It is obvious from the results of Model 1 that not only students with different booklets differed in terms of person fit but also that different levels of math ability (MATH_Dummy; coded *zero* for below median mathematics ability and coded *one* for above median mathematics ability) had a negative and significant effect on H^T. The results of Model 2 clarified that, in addition to the effects identified with regard to Model 1, TAM was not related to H^T in general. Model 3 indicated not only significant differences between both booklets and, again, a negative effect of MATH_Dummy on H^T, but also a significant interaction effect for TAM and the MATH_Dummy variable. The TAM main effect, now indicating effects of TAM for students with below median mathematics ability, is nonsignificant. While the BIC favored Model 1 in a statistical comparison of the three models, Model 3 outperformed Model 1 and Model 2 in terms of model fit by the AIC (Model 1: AIC = -276.41, BIC = -262.78; Model 2: AIC = -277.07, BIC = -260.03; Model 3: AIC = -282.22, BIC = -261.78); Model 3 also displayed the lowest percentage of residual variance (Model 1: R^2 = 6.5 %; Model 2: R^2 = 7.6 %; Model 3: R^2 = 10.5 %). To further illustrate this finding and give a more comprehensible measure for the relationship of H^T and TAM, the correlations in the group of students with a below median EAP was $r = .10$ ($p = .52$), while the correlation in the group of students with an above median EAP was $r = .28$ ($p = .01$).

--- please insert Table 2 here ---

Results Concerning the Effects of TMO on H_j^T

The results of the regression models 4-6 estimating the effects of TMO on H^T are given in Table 3. Note that Model 4 is equal to Model 1 in Table 2. Model 5 and Model 6 involved the known differences between booklets and different mathematics ability levels also visible in Table 2. Neither Model 5 nor Model 6 involved any evidence for a main effect of TMO on

H^T or an interaction effect of TMO with MATH_Dummy (Model 1: AIC = -276.41, BIC = -262.78; Model 2: AIC = -274.43, BIC = -257.39; Model 3: AIC = -273.21, BIC = -252.77); results on the residual variance likewise did not support TMO as a relevant predictor: R^2 (Model 1) = 6.5 %; R^2 (Model 2) = 6.5 %; R^2 (Model 3) = 6.8 %.

--- please insert Table 3 here ---

Discussion

Standardized large-scale assessments are an important building block in the educational monitoring system in many countries (e.g., ACAR, 2019; NCES, 2019), including Germany. While sample-based assessments like the National Assessment of Proficiencies (Stanat et al., 2019) address the educational administration in particular, state-wide assessments of curricular competencies taken by all students in Germany provide information for all teachers on the proficiency levels of their students against the national educational standards. Considering the limited measurement precision for individual score reporting, feedback in state-wide assessments of curricular competencies is given about the proficiency levels of larger learning groups like classes and courses. This study, based on reanalyses of data from a mathematics proficiency test and a supplementary survey has pointed to another issue related to individual score reporting: the lack of fit in the response behavior of some students, even with psychometrically sound tests. The study also revealed that test anxiety had a negative effect on the powerful person fit measure H_j^T in a group of high ability students. This finding supports previous results from different test contexts and is an indication that test-anxious students not only have lower test scores but also respond to test items differently (in terms of response patterns) compared to students with lower test anxiety. In contrast to earlier studies, however, no effect of test-taking motivation on person misfit was found. This finding might be due to the low-stakes character of the assessment for the students. Potential future studies aiming to better understand this effect might either involve an experimental variation of the test stakes (e.g., by offering incentives) or investigate matched groups from high- and low-stakes tests.

The results from this study indicate that test score interpretations at the individual level are not equally valid for all students, but depend to some degree on the level of test anxiety. This is an important message to both teachers, who possibly wish to utilize these tests for individual grading, and the educational administration who are monitoring the results and preparing the feedback to the teachers. While there are options to optimize the tests in terms of measurement precision (e.g., by means of increasing the test time and the number of administered test items or, in a more advanced manner, by means of computerized adaptive testing; Frey, in press; van der Linden & Glas, 2000), there is not much that can be done about the misfit at the individual level in this particular context. In fact, the design of large-scale assessments for school eveluation does not facilitate any of the suggested actions (see above) to handle misfit, as modifying the response patterns for some students or non-scoring responses corrupts the principle of test fairness and retesting might usually not be possible due to the high administrative work. This brings up a substantial challenge for individual score reporting, which cannot be ignored. This substantial problem is somewhat mitigated at the aggregated level when results are reported as the percentage of students from a school class measured at certain proficiency levels (based on their test scores) or as aggregated statistics computed across all students. For example, Rudner, Bracey, and Skaggs (1996) reported that person fit analysis did not change general statistical information computed from data from the National Assessment of Educational Progress (NAEP) State Trial Assessment substantially (although the authors acknowledge that different conclusions may be drawn about the individual). However, in some geographical regions and some school tracks, the classes are quite small, so that biased test scores due to person misfit might also have an impact on score reporting at this level. Due to these biased results, which follow from effects of specific emotional or motivational student characteristics irrelevant to the actual test content, teachers might come to incorrect conclusions concerning the mean proficiency level of the students in their courses. To prevent this, Brown and Villareal (2007) proposed using credibility functions to weight the response pattern of students based on the level of person fit when reporting proficiency at an aggregated level. Response patterns with low fit are provided with a lower weight. Based on the results of this and previous studies, we encourage test administrators, depending on the – potentially various – goals of the school evaluation assessment, to apply person fit weights. Especially for measures of school evaluation, we believe that the usage of these weights to increase the validity of the intended test score interpretations is beneficial.

References

Angelidis, A., Solis, E., Lautenbach, F., van der Does, W., & Putman, P. (2019). I'm going to fail! Acute cognitive performance anxiety increases threat-interference and impairs WM performance. *PloS one*, *14*(2), e0210824. doi: 10.1371/journal.pone.0210824

Asseburg, R., & Frey, A. (2013). Too hard, too easy, or just right? The relationship between effort or boredom and ability-difficulty fit. *Psychological Test and Assessment Modeling*, *55*, 92–104.

Australian Curriculum, Assessment and Reporting Authority (ACARA) (2019). *Measurement Framework for Schooling in Australia 2019*. Sydney: ACARA. Retrieved from: https://www.acara.edu.au/docs/default-source/default-document-library/measurement-framework-for-schooling-in-australia-2019773213404c94637ead88ff00003e0139.pdf?sfvrsn=0

Authoring Group Educational Reporting (2016). *Education in Germany. An indicator-based report including an analysis of education and migration.* Bielefeld: Bertelsmann. Retrieved from: https://www.bildungsbericht.de/en/archive/the-national-report-on-education-2016/NationalReportonEducation2016.pdf

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*, 431–444. doi: 10.1177/014662168200600405

Boekaerts, M., & Otten, R. (1993). Handlungskontrolle und Lernanstrengung im Schulunterricht [Action control and learning-related effort in the classroom]. *Zeitschrift für Pädagogische Psychologie*, *7*, 109–116.

Brown, R. S., & Villareal, J. C. (2007). Correcting for person misfit in aggregated score reporting. *International Journal of Testing*, *7*, 1–25. doi: 10.1080/15305050709336855

Cassady, J.C. (2010). Test anxiety: Contemporary theories and implications for learning. In J.C. Cassady (Ed.), *Anxiety in schools: The causes, consequences, and solutions for academic anxieties* (pp. 7-26). New York, NY: Peter Lang.

Chalmers, R. P. & Ng, V. (2017). Plausible-value imputation statistics for detecting item misfit. *Applied Psychological Measurement*, *41*, 372-387. doi: 10.1177/0146621617692079

Cronbach, L. J. (1970). *Essentials of psychological testing*. New York: Harper & Row.

De Leeuw, E. D., & Hox, J. J. (1994). Are inconsistent respondents consistently inconsistent? A study of several nonparametric person fit indices. In J. J. Hox & W. Jansen (Eds.), *Measurement problems in social and behavioral research* (pp. 67–87). Amsterdam: SCO-KI.

DeMars, C. E. (2018). Classical test theory and item response theory. In P. Irwing, T. Booth & D. J. Hughes (Eds.), *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale, and test development* (pp. 49-73). Hoboken, NJ: John Wiley & Sons Ltd. doi: 10.1002/9781118489772.ch2

Dimitrov, D. M., & Smith, R. M. (2006). Adjusted Rasch person-fit statistics. *Journal of Applied Measurement*, *7*(2), 170-183.

Dodeen, H., & Darabi, M. (2009). Person-fit: Relationship with four personality tests in mathematics. *Research Papers in Education*, *24*, 115–126.

Eid, M., & Zickar, M. J. (2007). Detecting response styles and faking in personality and organizational assessments by mixed Rasch models. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models. Extensions and applications* (pp. 255–270). New York: Springer.

Eklöf, H. (2008). Test-taking motivation on low-stakes tests: A Swedish TIMSS 2003 example. In M. von Davier & D. Hastedt (Eds.), *Issues and methodologies in large-scale assessments* (Vol. 1) (pp. 9–21). Hamburg: IEA-ETS Research Institute.

Engelhard, G. (2008). Using item response theory and model—data fit to conceptualize differential item and person functioning for students with disabilities. *Educational and Psychological Measurement*, *69*, 585-602. doi: 10.1177/0013164408323240

Frey, A. (in press). Computerisiertes adaptives Testen [computerized adaptive testing]. In H. Moosbrugger & A. Kelava (Eds.), *Testtheorie und Fragebogenkonstruktion* (3. ed.). Berlin/Heidelberg, Germany: Springer.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd edition). Mahwah, NJ: LEA.

Harnisch, D. L., & Linn, R. L. (1981). Analysis of item response practice. Questionable test data and dissimilar curriculum practice. *Journal of Educational Measurement*, *18*, 133-146. doi: 10.1111/j.1745-3984.1981.tb00848.x

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person fit statistics. *Applied Measurement in Education, 16*, 277–298. doi: 10.1207/S15324818AME1604_2

Knekta, E. M. (2017). Are all students equally motivated to do their best on all tests? Differences in reported test-taking motivation within and between tests with different stakes. *Scandinavian Journal of Educational Research*, *61*, 95-111. http://dx.doi.org/10.1080/00313831.2015.1119723

Leutner, D., Fleischer, J., Spoden, C. & Wirth, J. (2007). Landesweite Lernstandserhebungen zwischen Bildungsmonitoring und Individualdiagnostik [XXX]. *Zeitschrift für Erziehungswissenschaft*, *Special Issue 8*, 149-167. doi: 10.1007/978-3-531-90865-6_9

Meijer, R. R. (1996). Person fit research: An introduction. *Applied Measurement in Education*, *9*, 3-8.

Meijer, R. R. (1997). Person fit and criterion-related validity: An extension of the Schmitt, Cortina and Whitney study. *Applied Psychological Measurement*, *21*, 99–113.

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

Mullis, I. V. S., Martin, M. O., & Loveless, T. (2016). *20 Years of TIMSS: International Trends in Mathematics and Science Achievement, Curriculum, and* Instruction. Chestnut Hill, MA: IEA.

National Center for Educational Statistic (NCES) (2019). An overview of NAEP? Retrieved from: https://nces.ed.gov/nationsreportcard/subject/about/pdf/NAEP_Overview_Brochure_2018.pdf

Neumann, K., Kauertz, A. & Fischer, H. E. (2010). From PISA to educational standards. The impact of large-scale assessments on science education in Germany. *International Journal of Science and Mathematics Education*, *8*, 545-563.

OECD (2019). *PISA 2018 results (Volume I): What students know and can do*. Paris, France: OECD Publishing. doi: 10.1787/5f07c754-en

Pekrun, R., Jullien, S., Lichtenfeld, S., Frenzel, A. C., Götz, T., v. Hofe, R. & Blum, W. (2005). *Skalenhandbuch PALMA: 4. Messzeitpunkt (8. Klassenstufe)* [Technical Report PALMA: 4th Measurement Occasion]. Universität München: Institut für Pädagogische Psychologie.

Petridou, A., & Williams, J. (2007). Accounting for aberrant test responses using multilevel models. *Journal of Educational Measurement*, *44*, 227–247. doi: 10.1080/0969594X.2010.516606

Ramm, G., Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., Neubrand, M., Pekrun, R., Rolff, H.-G., Rost, J., & Schielefe, U. (Hrsg.) (2006). *PISA 2003. Dokumentation der Erhebungsinstrumente* [PISA 2003 – Technical report]. Münster: Waxmann.

Robitzsch, A., Grund, S., & Henke, T. (2017). miceadds: Some additional multiple imputation functions, especially for mice. R package version 2.5-9.

Rudner, Bracey, & Skaggs (1996).

Samejima, F. (2016). Graded response models. In W. J. van der Linden (Hrsg.), *Handbook of item response theory. Volume one: Model*s (pp. 95–107). Boca Raton: Chapman & Hall/CRC.

Schmitt, A. P., & Crocker, L. (1984, April). *The relationship between test anxiety and person fit measures*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Schmitt, N., Chan, D., Sacco, J. M., McFarland, L.A., & Jennings, D. (1999). Correlates of person fit and effect of person fit on test validity. *Applied Psychological Measurement*, *23,* 41–53. doi: 10.1177/01466219922031176

Seipp, B. (1991). Anxiety and academic performance: A meta-analysis of findings. *Anxiety Research*, *4*, 27–41. doi: 10.1080/08917779108248762

Sijtsma, K. (1986). A coefficient of deviant response patterns. *Kwantitative Methoden, 7,* 131–145.

Sijtsma, K., & Meijer, R. R. (1992). A method for investigating the intersection of item response functions in Mokken's nonparametric IRT model. *Applied Psychological Measurement*, *16*, 149–157. doi: 10.1177/014662169201600204

Sijtsma, K., & van der Ark, L. A. (2003). Investigation and treatment of missing item scores in test and questionnaire data. *Multivariate Behavioral Research*, *38*, 505–528. doi: 10.1207/s15327906mbr3804_4

Spoden, C. & Leutner, D. (2012). Vergleichsarbeiten als Instrument schulischer Selbstevaluation [State-wide assessments of curricular competencies as measures of self-evaluation within schools]. In A. Helmke, T. Helmke, D. Leutner, G. Pham, T. Riecke-Baulecke & C. Spoden (Hrsg.), *Schulmanagement Handbuch* (144. ed.) (p. 65-84). München, Germany: Oldenbourg.

Sundre, D. L., & Kitsantas, A. L. (2004). An exploration of the psychology of the examinee: Can examinee self-regulation and test-taking motivation predict consequential and non-consequential test performance? *Contemporary Educational Psychology*, *29*, 6–26. doi: 10.1016/S0361-476X(02)00063-2

Stanat, P., Schipolowski, S., Mahler, N., Weirich, S., Henschel, S. (Eds.) (2019). *IQB Trends in Student Achievement 2018. The Second National Assessment of Mathematics and Science Proficiencies at the End of Ninth Grade.* Münster, Germany: Waxmann.

Tendeiro, J. N., Meijer, R. R., and Niessen, A. S. M. (2016). PerFit: An R Package for Person-FitAnalysis in IRT. *Journal of Statistical Software*, *74*(5), 1–27.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*, 1–17. doi: 10.1207/s15326977ea1001_1

Wise, S. L., Pastor, D. A., & Kong, X. J. (2009). Correlates of rapid-guessing behavior in low-stakes testing: implications for test development and measurement practice. *Applied Measurement in Education, 22*, 185–205. doi: 10.1080/08957340902754650

van der Linden, W. J. (Eds.). (2016). *Handbook of item response theory. Volume one: Model*s. Boca Raton: Chapman & Hall/CRC.

van der Linden, W. J., & Glas, C.A.W. (Eds.). (2000). *Computerized adaptive testing: Theory and practice.* St. Paul, MN: Assessment Systems Corporation.

Table 1

Example of expected response patterns (Patterns 1 - 2) and unexpected response patterns (Patterns 3 - 4) in a fictitious proficiency test

	Item	Item	Item	Item	Item	Item	Item	Item	Item	Item
Response Patterns	1	2	3	4	5	6	7	8	9	10
Pattern 1	1	1	1	1	1	1	0	0	0	0
Pattern 2	1	1	1	1	0	1	1	0	0	0
Pattern 3	0	0	0	1	1	1	1	0	0	0
Pattern 4	1	0	1	0	1	1	1	1	1	0
Proportion of correct responses	.95	.85	.75	.65	.55	.45	.35	.25	.15	.05

Note. Code “0” is an incorrect response, code “1” is a correct response.

Table 2

Summary of regression models estimating the statistical effects of test anxiety in mathematics on the person fit measure H_j^T (N = 223)

	Model 1				Model 2				Model 3			
Variable	*B*	*SE**	*t*	*p*	*B*	*SE**	*t*	*p*	*B*	*SE**	*t*	*p*
Intercept	0.39	0.02	23.43	0.00	0.39	0.02	23.23	0.00	0.39	0.02	23.80	0.00
BOOK	0.04	0.01	2.90	0.00	0.04	0.01	3.48	0.00	0.04	0.01	3.49	0.00
MATH_Dummy	-0.06	0.03	-1.81	0.07	-0.07	0.04	-1.88	0.06	-0.07	0.03	-2.02	0.04
TAM					-0.02	0.02	-0.98	0.33	0.01	0.02	0.38	0.71
MATH*TAM									-0.05	0.02	-2.23	0.03

Notes. R^2 (Model 1) = 0.065, R^2 (Model 2) = 0.076, R^2 (Model 3) = 0.105. BOOK = booklet indicator, MATH_Dummy = mathematical proficiency. TAM = test anxiety in mathematics. TMO = test-taking motivation. *SE** = robust standard errors for nested data. Model 1 cannot be compared with Model 2 and 3 by means of the *F* statistic due to different underlying data matrices; Model 2 versus Model 3: *F*(1, 218) = 7.106, *p* < 0.01).

Table 3

Summary of regression models estimating the statistical effects of test motivation on the person fit measure H_j^T (N = 223)

	Model 4				Model 5				Model 6			
Variable	*B*	*SE**	*t*	*p*	*B*	*SE**	*t*	*p*	*B*	*SE**	*t*	*p*
Intercept	0.39	0.02	23.43	0.00	0.39	0.02	23.22	0.00	0.39	0.02	23.25	0.00
BOOK	0.04	0.01	2.90	0.00	0.04	0.01	2.90	0.00	0.03	0.01	2.90	0.00
MATH_Dummy	-0.06	0.03	-1.81	0.07	-0.06	0.03	-1.84	0.07	-0.06	0.03	-1.77	0.08
TMO					0.00	0.01	-0.10	0.92	-0.01	0.02	-0.71	0.48
MATH*TMO									0.02	0.03	0.61	0.54

Notes. R^2 (Model 1) = 0.065, R^2 (Model 2) = 0.065, R^2 (Model 3) = 0.068. BOOK = booklet indicator, MATH_Dummy = mathematical proficiency. TAM = test anxiety in mathematics. TMO = test-taking motivation. *SE** = robust standard errors for nested data. Model 1 cannot be compared with Model 2 and 3 by means of the *F* statistic due to different underlying data matrices; Model 2 versus Model 3: *F*(1, 218) = 7.106, *p* < 0.01).

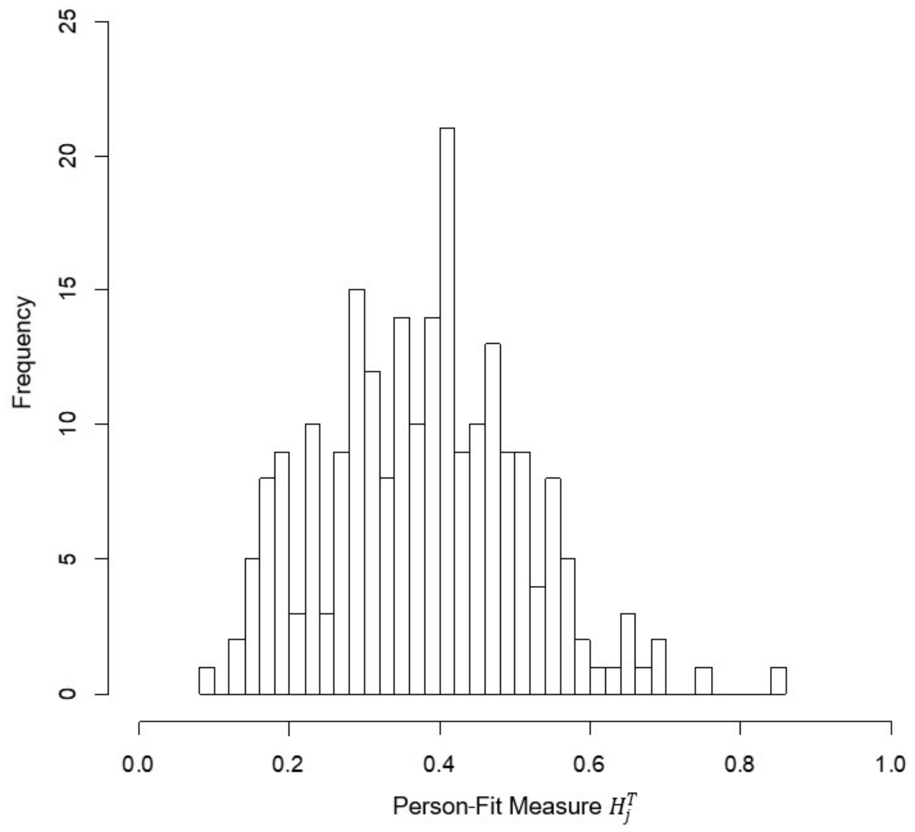

Figure 1. Histogram of the person fit measure H_j^T (N = 223).

[Monochrome color scheme should be used for this figure.]