Analyzing the context of large-scale educational assessments using multilevel latent variable modeling

Inaugural-Dissertation

in der Fakultät Humanwissenschaften

der Otto-Friedrich-Universität Bamberg

vorgelegt von

Theresa Rohm

aus Bad Salzungen

Bamberg, den 07.10.2022

Tag der mündlichen Prüfung: 28.07.2022

Dekan: Prof. Dr. Claus-Christian Carbon

Erstgutachter: Prof. Dr. Claus H. Carstensen

Zweitgutachter: Dr. Timo Gnambs

URN: urn:nbn:de:bvb:473-irb-559049
DOI: https://doi.org/10.20378/irb-55904

Content

Abstract

Large-scale assessments in education often involve the measurement of latent competence, with the aim to perform comparisons between groups in subsequent analyses. The assessment of latent competence is thereby often performed within the institutionalized context of schools, or with the help of interviewers visiting respondents' homes. In addition, sample selection for educational assessments might focus on groups as sampling units. For example, German secondary school types are used as primary sampling units in a multi-stage sampling procedure in educational studies. Subsequently, several schools per school type are selected for student competence assessment and in consequence, students are nested within clusters of schools. Hence, hierarchical data structures in large-scale educational assessments are caused by contexts of assessments, the sampling design or the combination of both. Regarding the resulting measures of latent competence, construct-irrelevant variance could be introduced by the context of educational assessments in the form of variance between clusters. This can potentially reduce the quality of measurement and impair a fair and valid interpretation of test scores. Further uncertainty in latent competence assessment arises by using as set of item indicators to measure the latent trait of interest. Besides the overall latent trait, also single items can vary by assessment contexts, indicated by random item effects that denote item variation across clusters. In addition, item differences can occur between groups (i.e., fixed item effects) and, if severe, indicate measurement non-invariance in the form of differential item functioning between groups. Item bias is indicated when item differences between groups favor a particular group. If item bias is found between groups that were used as sampling units (e.g., school type), item bias is associated with the context of competence assessment. Hence, group differences of items are then found on the cluster-level of the hierarchical data structure.

For the detection of such sources of construct-irrelevant variance stemming from the assessment context, as well as item bias and item variances associated with the assessment context, multilevel latent variable models are presented and applied to competence and cognitive ability assessments. Competence measurements in the domains mathematic and reading are examined, as well as a measure of the cognitive ability perceptual speed, all assessed within the National Educational Panel Study. In the first study, interviewer and area clusters were investigated in an adult mathematic competence assessment, as hierarchical structures that might introduce construct-irrelevant variance. The examination of cross-classified multilevel item response theory models showed substantial interviewer variance in mathematic competence, while area effects were small. Subsequent analyses revealed some interviewers with undue influence, that were in addition associated with the respondents' number of missing values in the assessed test and participation rates at the subsequent competence assessment. The second study investigated consequences of item bias, in the form of cluster-level group differences in items by school type, for students reading competence development from fifth to ninth grade of German secondary school. Measurement non-invariance occurred especially between the highest and lowest school type of German secondary schools at all measurement occasions. Nevertheless, the school type comparisons of reading competence development were not sensitive to found measurement non-invariance between school types and a parallel development of reading competence between German secondary school types was presented. In the third study, in addition to cluster-level group differences in item estimates by school type (i.e., fixed item effects), also random item effects across school clusters per school type were investigated for three items measuring perceptual speed for German secondary school students at ninth grade. Fixed- and random-group differential item functioning was investigated in comparison of students from several types of regular schools and students with special educational needs. Random-group differential item functioning was found for two out of the three items, indicating that estimated item difficulties differed across school clusters. Such differences across assessment contexts (i.e., school clusters) might stem from problems of standardized test assessment. Finally, the results of the three studies are discussed with regard to test standards for educational and psychological testing. The results are furthermore compared to empirical evaluations of context effects in other educational large-scale assessments.

Key words: assessment context, item response theory, multilevel latent variable modeling, large-scale educational assessments, National Educational Panel Study

1. Introduction

Large-scale assessments in educational research are used for comparisons within or across educational systems and compare at regional, national or international levels with the purpose to evaluate effectiveness of educational systems (Tierney, 2016). A central aim of large-scale educational assessments is in this regard to compare test results by groups, and furthermore, explain found group differences by relevant variables. It is thereby of interest, how competence assessments have to be designed, executed and evaluated, in order to assess the theoretically defined constructs at justifiable efforts for the respondents, while having high measurement precision (i.e., reliability), validity and fair group comparisons.

While group differences signify differences between respondents who share a specific characteristic (e.g., gender, school type, migration background), also clusters of respondents can exist and can be associated with differences regarding constructs of competence assessments. A cluster refers to respondents that share a similar or the same context of an assessment (e.g., respondents interviewed by the same interviewer, students visiting the same school or class). Respondents comprising a cluster are often more similar regarding characteristics of themselves and their environment. For example, students of the same class visit the same school type, share a learning environment at school, can be more similar with regard to their socioeconomic background and live in geographical proximity. In the evaluation of large-scale assessments of educational research, little attention has been given to the context of competence assessment, even though educational assessments of school students regularly assess tests within clusters of schools or classes. Apart from institutionalized contexts of testing, adult respondents are for example supervised by interviewers during test assessment within their private homes. Due to these various assessment contexts, when comparing individuals across different contexts, evidence should be presented that administered measures are measurement invariant across clusters, groups and groups comprised of various clusters (e.g., school types comprising several schools where the assessment was conducted). Measurement invariance means that a psychological instrument assesses the unobserved construct equally across different groups of individuals or over time and that therefore the scale equally mirrors the latent construct in all instances (Finch, 2014). Measurement non-invariance in the form of differential item functioning (DIF; Camilli, 2006; Zumbo, 2007) across clusters and groups (i.e., item by group interactions), as well as different item variance across clusters within groups (i.e., cluster bias) can affect valid comparisons and in the worst-case lead to wrong conclusions (Borsboom, 2006). The validity of inferences about changes in proficiency levels by longitudinal large-scale assessments might be impaired by effects related to the overall context of the assessment and several investigated groups can be affected to a different extent by context effects.

Therefore, the here presented models and methods follow the aim to detect sources of construct-irrelevant variance stemming from the assessment context, as well as detecting item bias between cluster-level groups (e.g., school type groups comprising several school clusters) and item variances associated with the assessment context for comparisons between groups. As item variance refers to random effects, such as items varying across clusters, item bias is present when item differences between groups mainly favor one group in particular. Such sources of variance and bias might limit comparisons between assessed groups, as well as group comparisons on latent competence development. Please note, that measurement invariance and absence of DIF are necessary but not sufficient conditions for test fairness (Borsboom, 2006; Meredith, 1993). Test fairness not solely concerns item functioning in statistical terms, but all aspects of educational testing, including test construction, execution and administration with equal treatment of all test takers, as well regarding test evaluation.

In case of longitudinal assessments, comparability over time within and between groups is an additional challenge. Furthermore, it is of question when measurement invariance matters

(Borsboom, 2006) and therefore, beyond the investigation of the existence of cluster-level or group effects in educational LSAs, research strategies are presented to assess if found effects impair longitudinal competence assessment. Besides the necessity to identify items that work to the advantage or disadvantage of any particular group the test is assessed to, it is necessary to evaluate the impact of found effects on group comparisons for single-occasion competence measurements as well as competence development.

The next section portrays context effects within standards for educational and psychological testing. In the subsequent section, empirical assessments of various context effects in (international) LSAs are presented. Thereafter, statistical models and methods for the investigation of context effects in latent competence and ability assessment are introduced and furthermore, the three manuscripts of this dissertation that applied these methods and models for the examination of context effects are summarized. Finally, the findings of the manuscripts are discussed in light of test standards and comparable empirical assessments of context effects in large-scale educational assessments.

2. Context effects in the realm of test standards

The Standards for Educational and Psychological Testing (Standards; AERA, APA, NCME, 2014), were first published in 1966, last revised in 2014 and are edited jointly by the American Educational Research association (AERA), the American Psychological Association (APA) and the National Council on Measurement in Education (NCME). The foundation of the Standards mainly concerns validity, reliability and fairness of educational and psychological testing and provide criteria for test development and evaluation. As criterion for the evaluation of large-scale assessments, validity is defined as the degree to which a test measures what it was intended to measure, reliability concerns the consistency and dependability of assessment results and fairness means absence of bias toward any particular group of examinees (Klein and Hamilton, 1999). Fairness in educational assessments concerns all parts of the testing procedure, from construction, administration and scoring of a test to evaluation of test results and their use for performance comparisons (Dorans and Cook, 2016). It signifies equity between groups under study given the whole test procedure, evaluation and use of results, but no equity with regard to test results.

Construct validity, as one of the main types of validity evidence (e.g., besides content and criterion validity), defines how appropriate inferences are, that are made based on observations and measurements. It therefore determines if a test measures the intended construct and therein construct validity and the fairness principle cohere: if construct validity is given across all groups of interest, measurement bias is absent and fairness of the test instrument toward any particular group of examinees can be deduced. The Standards furthermore contain guidelines for assessing the validity of test score interpretation and discuss measurement bias as a primary threat to fairness (Zimmermann and Grudnitski, 2017). As group comparisons are the goal of LSAs in education, consistency in the form of equal treatment during the test procedure is necessary (Tierney, 2016). For example, in group-mean comparisons of latent constructs, item bias can severely affect mean differences. The size of bias is decisive and criteria for the evaluation of bias need to be relative in size to the group-effect of interest. It is furthermore recommended to study robustness of effects under investigation of various levels of measurement bias (Borsboom, 2006) and mean comparisons of latent variables can even be validly made with partial or approximate measurement invariance (van de Schoot et al., 2013). However, investigations of measurement invariance are a prerequisite for the assumption of (partial or approximate) measurement invariance.

Fairness issues first appeared in year 1974 in the Standards (Nisbet and Shaw, 2019) and define principles, but no statistical methods or analysis strategies that adhere to these principles. Hence, there is a demand for the development of statistical methods and research strategies to assure that the fairness principles of the Standards are met. With regard to the here presented work on context effects, fairness is among others defined in the Standards by fairness in treatment during the testing procedure and fairness as a lack of measurement bias. For the first principle, standardized test procedures and administration conditions as well as scoring procedures are essential to have comparable contexts in which test takers can demonstrate the abilities to be measured. The principle of absence of measurement bias can be assessed through DIF and DTF (differential test functioning) detection and evaluation. Test items favoring one side, group or person are considered biased or measurement non-invariant. Bias concerns systematic error in test scores and it is the case when factors irrelevant to the intended construct of an assessment influence the results. Such evidence of bias can be found through the statistical analysis of item differences. In this regard, measurement invariance is given when in each group the same attribute relates in the same way to the same set of observations. More precisely, in each group under comparison the mathematical function relating latent variables to observations must be the same (Borsboom, 2006). Aside from these theoretical definitions the assessment of this “sameness” is not straightforward, as measurement invariance and item bias encompass very different statistical models from item response theory (IRT) and factor-analytic approaches. Such methods for DIF detection have different statistical assumptions (e.g., relation between observed and latent variables, shape of population distribution on the latent variable) and use different modeling strategies and selection criteria for biased items (Borsboom, 2006). In applications to educational research, choices of DIF detection methods might make a difference to the findings on measurement invariance.

A further threat to the fair and valid interpretation of test scores, as examined by the Standards (2014), is when tests or test procedures produce construct-irrelevant variance in scores. Inappropriate score interpretations between groups might be the consequence, as scores might be lowered or raised. Sources of construct-irrelevant components of scores can be introduced in various ways, for example through inappropriate sampling of test content, aspects of the test context such as lack of clarity in test instructions or problems with standardized test instructions, item complexities unrelated to the measured construct and further characteristics of test items, scoring criteria favoring one group over another, interactions of the test examiner with the test taker, or complex demands in test processing that are unrelated to the measured construct (e.g., demands of language proficiency). Empirical research investigating test context effects is rather limited even though the test context encompasses various aspects of tests and the test environment, which might influence the test takers performance and thereby increase construct-irrelevant variance in test scores.

When latent constructs are compared between groups and over time, reliability and validity are associated, as the level of reliability indicates the consistency of estimated scores across replications of the testing procedure (precision). Thereby, reliability can depend on the population and the overall variability of the construct being measured. In the evaluation of group performances, variation stemming from the sampling of individuals can be a major error source, but it is anticipated to diminish with growing sample sizes. However, even with high precision in individual test scores, error associated with the sampling of individuals can be a source of error in inferences about the assessed construct. The estimation of group means can involve error sources that are different from those at the individual-level (Standards, 2014). Particular sources of error on the group-level can for example stem from clusters actively sampled of populations (e.g., school clusters) or arising from the assessment procedure (e.g., interviewer clusters). This can lead to varying reliabilities in comparison of populations. Therefore, reliability of a measured construct depends on assessment procedures and the population being assessed. Standard error estimation and conditional standard error estimation of measurement can thereby be informative for conclusions drawn from assessments, as they reflect variation in the measured construct from error sources, similar to reliability and generalizability coefficients. As not all sources of error can be identified or captured, the

Standards recommend to evaluate those sources of error that are likely to have the highest impact. Overall, reliability is high, when variability stemming from errors is small in comparison to observed variation in scale scores. When tests have differential reliability across groups, associations with the observed score might be different across groups, as regression coefficients will be constricted in groups where reliability is lower (Borsboom, 2006). Hence, besides overall standard error of measurement, conditional standard errors of measurement should be reported when there is an expectation that standard errors are not constant across subgroups. Furthermore, the extent and impact of such differences should be investigated and reported when feasible (Standards, 2014).

3. Context effects in large-scale educational assessments

In research using data from large-scale educational assessments, sources of construct-irrelevant variance or measurement invariance are not routinely investigated and some researchers assume measures to be invariant across groups without providing evidence for this assumption (Borsboom, 2006). Subtle violations of measurement invariance can lead to erroneous conclusions, especially when point hypotheses (e.g., population means of a test score are equal) are researched (Borsboom, 2006). However, some issues with regard to context effects of testing have been investigated for (international) large-scale assessments and results were sometimes discussed in light of test fairness across the compared groups. For international LSAs special attention is thereby on consequences for country comparisons, while studies focusing on national comparisons investigate consequences for comparisons of subpopulations.

In Germany, several LSAs assess latent competences for national and international comparisons. LSAs conducted within respondents' homes in Germany are for example the *Programme for the International Assessment of Adult Competencies* (PIAAC), and several cohorts of the *National Educational Panel Study* (NEPS). Among the most prominent LSAs conducted in school settings are the *Programme for International Student Assessment* (PISA), *Progress in International Reading Literacy Study* (PIRLS), *Trends in International Mathematics and Science Study* (TIMMS), the *National Educational Panel Study* (NEPS) and *IQB Trends in Student achievement* comparing German federal states. All face methodological challenges, such as missing values in competence or questionnaire data, measurement invariance over time and across groups (i.e., DIF), stability of longitudinal development (e.g., correct linking methods), response times and speed sensitivity of administered tests, the evaluation of test context effects (e.g., effects of test design and mode of administration), or rater effects in paper based assessments and constructed-response items in computer based assessments (Shin, von Davier and Yamamoto, 2019).

Focusing on the evaluation of context effects as one challenge, that may have undesired effects on test performance, surrounding conditions of the assessment setting (e.g., noises, disturbances during test assessment, presence of third parties during assessment), motivational factors and test endurance associated with context groups (e.g., school types, test administrators) can be an issue. Furthermore, the context can be associated with the design of the test instrument. For example, there can be differences in booklet design across groups, effects of item difficulty (e.g., because of the sequencing of items or sections), effects of testing time, or the ordering of response options (Nagy, Lüdtke and Köller, 2016). To conclude, the context can affect the assessed instrument as a whole, several sections, or single items of an assessed test.

The assessment setting of large-scale educational assessments can be broadly divided into studies that are conducted in respondents' homes and studies conducted in institutional settings such as schools. In both cases, test administrators or interviewers are regularly present and there are differences in the performed tasks in comparison of these settings. For studies assessed within respondents' homes, interviewers conduct the interview based on a questionnaire, while the assessment of cognitive tests is typically self-administered and the interviewer should stay rather passive throughout. As a study conducted in respondents' homes, the PIAAC Survey of

Adult Skills is conducted in over 40 countries (with each N = 5,000 participants) and is assessed as a computer-delivered survey for adult respondents (from 16 to 65 years of age), including multi-stage adaptive testing (Kirsch and Lennon, 2017). Comparable scales are created for the assessment of cognitive skills (i.e., literacy, numeracy and problem-solving), based on IRT methods. Even though PIAAC is self-administered by respondents using the computer, the quality of the assessment relies on interviewers working independently in the field. Hibben, Pennell and Scott (2018) described how PIAAC interviewers could deviate from survey protocols, such as the task to select household members for surveying, location of the assessment, and how much time a respondent may take to complete the assessment (e.g., experienced interviewers may encourage respondents to rush through the assessment or terminate it), as well as the presence of a third party that respondents could draw upon during the assessment to help with answers.

For the first PIAAC data collection cycle (between 2011-2018), the testing conditions (e.g. assessed in respondents' homes or some other agreed-upon location) and the quality of interviewers were described as factors that potentially influence the data quality and the validity of country comparisons (Keslair, 2018). The testing conditions (i.e., frequency of distractions and the type of space in which tests were assessed) were found to vary between countries but were not expected to impact quality differences among countries. In contrast, interviewer assignment size and interviewer effects were assumed to exert influences. However, the report (Keslair, 2018) only assumed that cognitive measures were affected by these differences, but no empirical evidence (e.g., statistical analyses considering cognitive measures) was presented. Furthermore, personal information on interviewers was not collected in the first PIAAC cycle, sampling areas and interviewer assigned areas overlapped and information why an interview was not completed (e.g., due to respondent refusal, unavailability, person could not be reached) was not reported with the PIAAC data.

In addition, a qualitative study based on audio interview recordings found for the PIAAC fieldwork in Germany, conducted in 2011 to 2012, that interviewer behavior systematically deviated from standardized interviewing techniques (Ackermann-Piek and Massing, 2014). Unfortunately, the researchers only examined the behavior for the background questionnaire and did not consider the data collected in the skills assessment. Furthermore, estimating interviewer variance (varying effects across interviewers) is not feasible, as the sampling design was not interpenetrated (i.e., interviewers are administered to several areas and areas are visited by various interviewers) and in consequence, area clusters and interviewer clusters mostly overlapped (Hibben et al., 2018; Keslair, 2018). Hence, the sampling variance is confounded with interviewer variance.

Research on studies conducted in school settings also reported several context effects. Evaluating PISA data from 2006 and 2009, Oliveri and von Davier (2017) found violations of the assumption of complete item parameter invariance by language of test administration, item type, item function and item format. They therefore proposed for an internationally comparable scale the use of models that acknowledge partial invariance and demanded for further research on partial invariance models in score scale calibration approaches. Based on the PISA Assessment 2012, item position effects were found to vary between countries and schools, with position effects being stronger in groups with lower average achievement (Debeer, Buchholz, Hartig and Jannsen, 2014; Hartig and Buchholz, 2012) and item position effects being associated with the students' school type (Qian, 2014). Hence, the occurrence of test context effects can be related to the observed achievement scores, making it difficult to measure the 'true' proficiency for each group under study. Furthermore, longitudinal comparisons of student competence by the German PISA 2012 assessment with a longitudinal extension to year 2013, was found to be affected by item position effects (i.e., items becoming more difficult when later positioned in a test) stemming from a booklet design, that varied between assessment occasions and school type groups (Nagy et al., 2016). Ignoring this test context effect in the IRT estimation led to an underestimation of achievement gains. This was most pronounced in the domains reading and science, as they were most affected by item position effects. Furthermore, these effects were different in comparison of school types (i.e., classified within two categories: upper secondary track "Gymnasium" vs. middle secondary school track "Realschule" or comprehensive schools "Gesamtschulen"). As a result, these test context effects had consequences for assessing group differences in reading and science competence, because when ignoring them, a larger achievement difference in favor for the upper secondary school track resulted. The assessment of mathematic competence was on the contrary robust to found position effects. A further underestimation of achievement gains occurred because of selective sample dropout in the science domain, where science competence development for students of the middle secondary or comprehensive school track would be underestimated when it was not corrected for this dropout effect (Nagy, Köller, Lüdtke and Heine, 2017).

For the PISA 2015 study, the assessment mode changed from paper-based assessment to computer-based assessment and Robitzsch et al. (2017) concluded from a re-analysis of the PISA data that declines in mathematics and science competence among German students could be associated with an unfamiliarity with computer use in the classroom setting, as computer-administered items were more difficult when compared to paper-based assessed items. In addition, a further investigation on the change in the PISA assessment mode between 2012 and 2015 found that trend changes for the assessment of mathematic competence in the Netherlands was to some degree related to DIF between modes (Feskens, Fox and Zwitser, 2019). However, it was concluded that that the methodology adopted by PISA was sufficient to account for these mode effects.

Analyzing results from PIRLS of year 2006 on reading comprehension in fourth grade of school, Oliveri and von Davier (2014) found that for a large proportion of groups (i.e., 33 out of 40 participating countries) 75 percent of items were well-fitting and hence, common between all countries. Therefore, a quasi-international scale was created with a minority of items having country-unique parameters. Assuming partial measurement invariance and using a subset of country-based parameters led to improved model fit of the data. However, it was found that country comparisons were not altered by using the model considering some country-unique parameters in comparison to a model with item parameters constrained to be the same for all countries. On a further note, they found that groups with higher levels of item misfit were at the upper and lower spectrum of achievement and Oliveri and von Davier (2014) assumed that it is due to less information at both ends of the achievement spectrum.

To conclude from these (international) LSAs, construct-irrelevant variance can be introduced by the assessment context (i.e., presence of test administrators or interviewers). Investigations of such sources of variance and its consequences for the assessed competence can be unfeasible due to the study design (i.e., non-interpenetrated samples). Furthermore, differences in item parameters stemming from the test context can be associated with estimates of group comparison and mean estimates of change. In addition, context effects can be unequal across groups under study and in consequence unequally impact estimates of mean trend change.

4. Analyzing context effects with multilevel latent variable modeling

For the investigation of context effects on assessments of latent competence, statistical models and methods, as well as research strategies are needed. A central goal of the here presented models and methods is the detection of sources of construct-irrelevant variance stemming from the assessment context, as well as item bias and item variances associated with the assessment context. Therefore, multilevel structural equation modeling (MSEM; Kamata and Vaughn, 2010) is beneficial when cluster effects and cluster-level group effects are of interest. The MSEM is the more general framework to multilevel IRT models (Fox, 2010; Fox and Glas 2001; Kamata and Vaughn, 2010; Lu, Thomas, and Zumbo, 2005; Muthén and Asparouhov,

2012) and a translation of item thresholds and loadings to difficulty and discrimination parameters of the IRT normal ogive model was presented by Lord, Novick and Birnbaum (1968). Further information on the equivalence of the presented factor-analytic model to the normal ogive model can be found in McDonald (1967) and Finch (2005).

To investigate context effects in competence measurement, a MSEM framework is proposed, where the measurement part is comparable to a two-parameter IRT model:

$$y_{ijc}^{*} = \nu_{jc} + \Lambda_j\, \theta_{ic} + \varepsilon_{ijc}. \tag{1}$$

Here, y_{ijc}^{*} represents a J x 1 vector of unobserved latent response scores for respondent $i \in 1 \ldots I$, nested within cluster $c \in 1 \ldots C$ with j $\in 1 \ldots J$ items in the test. Thereby, ν_{jc} is the intercept of indicator variable j which can be allowed to vary across clusters c, θ_{ic} is a K x 1 vector of ability scores for K latent factors; in case of a unidimensional model, $K = 1$. Furthermore, Λ_j is a J x K matrix of discrimination parameters (factor loadings) and the ε_{ijc} are the zero mean normally distributed residuals.

Using a unidimensional IRT model with $K = 1$, both Λ_j and ε_{ijc} are J x 1 vectors. In case of observed dichotomous responses, y_{ji} is defined as

$$\begin{aligned} &y_{ji} = 1, \text{ if } y_{ji}^{*} \geq \tau_j \text{ and} \\ &y_{ji} = 0, \text{ if } y_{ji}^{*} < \tau_j, \end{aligned} \tag{2}$$

with τ_j denoting the item difficulty parameter (threshold) for item j $\in 1 \ldots J$. Likewise, for polytomously scored items, the scoring categories range from 0 to M with

$$\begin{aligned} &y_{ji} = M, \text{ if } y_{ji}^{*} \geq \tau_{jM,} \\ &y_{ji} = M\text{-}1, \text{ if } \tau_{j(M-1)} \leq y_{ji}^{*} \leq \tau_{jM,,} \\ &\ldots \\ &y_{ji} = 1, \text{ if } \tau_{j1} \leq y_{ji}^{*} \leq \tau_{j2,} \text{ and} \\ &y_{ji} = 0, \text{ if } y_{ji}^{*} < \tau_{j1}. \end{aligned} \tag{3}$$

The structural model part of this two-level SEM allows to regress latent factors θ_{ic} on the cluster-level covariates x_c (e.g., school type, interviewer or area characteristics):

$$\theta_{ic} = \alpha_c + \Gamma_c x_c + u_{ic}. \tag{4}$$

Intercepts are given by α_c and when no latent variable is specified as a predictor, the intercepts simply become factor means. The slopes for the observed cluster-level covariates are denoted by Γ_c. The residuals u_{ic} are assumed to be normally distributed with zero mean and *K x K* covariance matrix. In cases of two (or more) sources of clustering, which are not hierarchically nested within each other, a cross-classified MSEM is presented for the measurement of two separate cluster-levels (i.e., interviewer and area clusters exist as two separate levels besides each other). The structural model part allows for varying intercepts and regression coefficients across separate clusters with $c \in 1 \ldots C$ denoting one cluster-level (i.e., interviewers) and $g \in 1 \ldots G$ indicating another, parallel cluster-level (i.e., area clusters). This can be expressed as:

$$\theta_{icg} = \alpha_{cg} + \Gamma_{cg} x_{icg} + u_{icg}$$

$$\text{with } c = 1, 2, \ldots, C \text{ and } g = 1, 2, \ldots, G, \tag{5}$$

where the latent factor θ_{icg} for respondent *i* nested in clusters *c* and *g* is regressed on individual-level covariates x_{icg}. The intercept α_{cg} and the slopes Γ_{cg} are allowed to vary across each cluster-level as a function of between-cluster variables w_c and w_g:

$$\alpha_{cg} = \alpha_{00} + A_c w_c + A_g w_g + \varepsilon_c + \varepsilon_g, \tag{6}$$

$$\Gamma_{cg} = \Gamma_{00} + \Gamma_c w_c + \Gamma_g w_g + \xi_c + \xi_g. \tag{7}$$

The residual u_{icg} is assumed to be normally distributed with zero mean and variance $Var(u_{icg}) = \sigma^2_u$, whereas the residuals for the cluster-levels, ε_c, ξ_c, ε_g and ξ_g, are each multivariate normally distributed with zero means and variance-covariance structures $\Sigma = \begin{pmatrix} \sigma^2_\varepsilon & \sigma_{\varepsilon\xi} \\ \sigma_{\varepsilon\xi} & \sigma^2_\xi \end{pmatrix}$. When denoted as an unconditional cross-classified model without predictor variables, (5), (6) and (7) reduce to the mixed-effects formulation:

$$\theta_{icg} = \alpha_{00} + \varepsilon_c + \varepsilon_g + u_{icg}. \tag{8}$$

Here, the achievement of respondent i equals the sum of the grand-mean achievement of all respondents α_{00}, the random effect ε_c introduced by cluster c, the random effect ε_g of the cluster g, and a random respondent effect u_{icg}. These random effects introduced by clusters are indicators of the amount of construct-irrelevant variance in latent factors, stemming from the assessment context.

Using the MSEM formula (1), cluster specific item parameters can be allowed:

$$\nu_{jc} = \omega_j + \Lambda_j \theta_c + \varepsilon_{jc}. \tag{9}$$

where ω_j is the cross-cluster grand intercept of indicator variable j, which equals the grand mean of the latent variable when the between-level latent variable is zero. The term Λ_j refers to the cluster-level factor loading of indicator variable j with θ_c being the score of cluster c on the cluster-level latent variable. The cluster-level error term ε_{jc} is the random intercept denoting item variance across clusters. Item bias across cluster-level groups can be detected when cluster-level covariates are added that interact with the intercept of indicator variable j:

$$\nu_{jc} = \omega_j + \Lambda_j \theta_c + \beta_j x_c + \Upsilon_j \theta_c x_c + \varepsilon_{jc}. \tag{10}$$

A significant regression coefficient β_j would indicate non-invariant intercepts (i.e., latent means) of the latent trait between groups and a significant regression coefficient Υ_j indicates if there is a difference in response probabilities across groups (i.e., non-invariant factor loadings). Setting this extension by a grouping variable into formula (1), a MSEM considering different intercepts and item properties across the between-level groups results:

$$y_{ijc}^{*} = \omega_j + \Lambda_j \theta_c + \beta_j x_c + \Upsilon_j \theta_c x_c + \Lambda_j \theta_{ic} + \varepsilon_{jc} + \varepsilon_{ijc}. \tag{11}$$

Furthermore, using this modeling framework, item variances associated with the assessment context can be examined by investigating random loading and threshold effects across clusters (Hartig, Köhler and Naumann, 2020; Verhagen, Levy, Millsap and Fox, 2015). Through inclusion of an additional random item effect $\varepsilon_{jc.}$ such item variance can be examined and random item variance is present when $\sigma^2_{\varepsilon jc}$ is larger than 0 per item. However, slight deviations from zero might not impair group comparisons. In addition, random item effects can be different in comparison of cluster-level groups, which is termed cluster bias (Jak, Oort, and Dolan, 2014).

With regard to estimation methods, Bayesian approaches have shown to be beneficial in the estimation of MSEM when categorical variable models are estimated and when there are several measurement occasions, as is the case using longitudinal data (Muthén and Asparouhov, 2012; Steele and Goldstein, 2006). Furthermore, Bayesian analysis is helpful to obtain correct variance estimates in MSEM and to obtain admissible estimates where a frequentist estimation method is likely to fail (Kaplan and Deapoli, 2013; Deapoli and Clifton, 2015; Verhagen et al., 2015).

Nevertheless, a problem of model identification is inherent when using the thus far presented MSEM for DIF detection between fixed groups, for example through the inclusion of groups as cluster-level covariates and regressing them on item parameters: as one item needs to be constraint, DIF actually measures the difference between the constraint and the respective jth item (Penfield and Camilli, 2007). Further details on this indeterminacy of latent variables and solutions for measurement invariance testing using an approach of relative DIF are discussed by Bechger and Maris (2015) and Schulze and Pohl (2020). Methods to select item sets free of DIF for anchor selection in multiple group comparison were presented by Kopf, Zeileis and Strobl (2015) or Huelmann, Debelak and Strobl (2019).

When interested in item bias across cluster-level variables, applying the method of Alignment Optimization (AM), as proposed by Asparouhov and Muthén (2014), is another way to investigate if measurement invariance is violated through DIF between the cluster-level covariates. A comparison of AM to other methods of measurement invariance testing with many groups was presented by Kim et al. (2017). This maximum likelihood-based method follows the logic of multiple group factor analysis and searches for an optimal set of measurement parameters. Items with invariant factor means and variances across multiple groups are detected while nested data structures can be fitted, because standard errors are computed using the Huber-White sandwich estimator (Huber, 1967; White, 1980).

Measurement invariance is investigated by starting with a configurational model assuming no invariance, with zero mean and factor variance set to one in all groups. Subsequently, factor means and variances are allowed to vary across the groups, with the attempt to find as much invariance as possible. Finally, an alignment fitting function identifies the model by performing pairwise comparisons for all groups of both intercept and loading parameters, thereby finding the optimal measurement invariant model by minimizing the amount of measurement non-invariance:

$$F = \sum_{p} \sum_{g1<g2} w_{g1,g2}\ f\left(a_{pg1} - a_{pg2}\right)$$

$$+ \textstyle\sum_p \sum_{g1<g2} w_{g1,g2}\ f\left(\delta_{pg1} - \delta_{pg2}\right). \qquad (12)$$

With *p* being the number of observed indicators, *g1* and *g2* represent the group for every pair of groups in the data, a_{pg1} and a_{pg2} represent the factor loadings of the compared groups, while δ_{pg1} and δ_{pg2} denote the intercepts of the respective groups. For every measurement parameter in function F, the difference of every pair of groups is scaled through the component loss function *f*, defined as:

$$f(x) = \sqrt{\sqrt{x^2 + 0.01}}\,. \qquad (13)$$

This also called total loss, simplicity or component loss function (Jennrich, 2006) accumulates the total measurement non-invariance over the items and optimizes at a few large non-invariant loading and intercept parameters and many approximately invariant parameters. Hence, the problem of constraining one item used as a reference is circumvented. A detailed presentation and a further IRT application using the AM approach is presented in Muthén and Asparouhov (2014).

5. Manuscripts of this dissertation

This dissertation comprises three manuscripts that all concern context effects in educational LSAs of latent competence, but focus on various consequences of the presence of such effects for the assessment. For each manuscript, the title, a brief summary of the research question and main results are presented. The here depicted topics and results are discussed in the subsequent section in light of test standards, comparable research on context effects of LSAs, as well as methods and analysis strategies using multilevel latent variable modeling.

5.1. Manuscript 1: Construct-irrelevant variance from the assessment context

Rohm, T., Carstensen, C. H., Fischer, L. & Gnambs, T. (2021). Disentangling Interviewer and Area Effects in Large-Scale Educational Assessments using Cross-Classified Multilevel Item Response Models. *Journal of Survey Statistics and Methodology*, 9(4), 722-744. https://doi.org/10.1093/jssam/smaa015

Summary

As outlined, construct variance can be associated with the context of the assessment. Construct-irrelevant variances stemming from the context factors can be a threat to the fair and valid interpretation of test scores. For the assessment of mathematic competence to the adult cohort of the German National Educational Panel Study (NEPS), interviewer and area clusters were identified as two possible sources of variance stemming from the assessment context. Hence, an analysis of construct variance stemming from two separate cluster-levels was conducted. Both levels were disentangled using cross-classified multilevel IRT models and Bayesian methods were applied. The measurement model of mathematic competence was thereby specified as a two-parameter logistic IRT model (Kamata and Vaughn, 2010), which in comparison to the Rasch Model (Rasch, 1960) allows for unequal item discrimination among respondents with different abilities. In educational and psychological contexts, such multilevel IRT models were presented by Fox and Glas (2001), Fox (2003), and Skrondal and Rabe-Hesketh (2004). In the structural part of the model, latent mathematical ability was regressed on individual-level covariates of respondent characteristics (i.e., age, gender, ethnicity, educational level, employment status, cultural capital, political area size per respondent) and interviewer characteristics (i.e., age, gender, educational attainment, working experience as an interviewer).

Besides regression coefficients and the overall amount and share of variance due to both cluster-levels, the Bayesian residual estimates of each cluster unit were of interest. These residual estimates demonstrate heterogeneity in mathematical competence per specific cluster. Furthermore, they reflect the similarity between the measurements of respondent mathematical competence within the clusters. To identify exceptional clusters, random effects were drawn for each cluster from the posterior distribution of mathematic competence estimation. Furthermore, the posterior standard deviation was used to obtain confidence intervals for the random intercepts of cluster specific competence values.

Results of the estimated models were threefold: (1) Observed variance in mathematical competence was much higher between interviewers (6.6 percent) than between areas (0.8 percent). Hence, substantial interviewer variance in the assessment of adult mathematic competence is present, while area effects are small. (2) While all respondent characteristics exerted significant effects on mathematic competence, none of the observed interviewer characteristics was found to be significantly related to respondent mathematical competence. (3) Using interviewer residuals sampled from the posterior distribution and investigating their posterior probability interval (PPI) revealed some interviewers with undue influence on the measurement of respondent mathematic competence. Four out of the 200 interviewers had a PPI above zero and twelve interviewers had a PPI below zero, indicating that their estimated competence intercept deviates significantly from the survey population mean. Such deviations were not found for area clusters. To assess the impact of the deviating interviewers, the respondents' numbers of missing values in the administered mathematic competence test and the participation rates at the subsequent competence assessment (about five to six years later) were compared. Respondents interviewed by interviewers with significantly higher residual estimates had, in comparison to the respondents interviewed by non-outlying interviewers, significantly lower missing values in the competence test, but also lower participation rates at the subsequent measurement occasion (five to six years later). For respondents who were interviewed by interviewers with significantly lower residual estimates, no significant differences were found.

To sum up, while area effects were negligible, a noteworthy amount of construct-irrelevant variance stemming from the assessment context was present and a considerable number of interviewers with effects on the assessment were found. However, a comparison between individual assessment scores of a model considering the random interviewer effects and a model ignoring the nested structure, showed that interviewer effects do not distort the individual assessments of mathematical competence. Nevertheless, variance in mathematic competence is higher due to interviewer presence. Our work demonstrates the effects of clusters of the assessment context on the measurement of a latent construct (i.e., mathematic competence), using multilevel IRT modeling. Construct-irrelevant variance due to the context of the assessment was observed and specific clusters were found to exert an undue influence on competence measurement.

5.2. Manuscript 2: Context-level group differences in competence development

Rohm, T., Carstensen, C. H., Fischer, L. & Gnambs, T. (2021). The achievement gap in reading competence: the effect of measurement non-invariance across school types. *Large-scale Assessments in Education*, 9(23). https://doi.org/10.1186/s40536-021-00116-2

Summary

Context-level group differences in the measurement of latent constructs (i.e., item bias) might affect the outcome of measurements. The groups in this study were school types of students and they are associated with the assessment context insofar, as the school type groups comprise the sampled school clusters (i.e., specific schools that students visit) of the assessment. Differences of item estimates can be present when items of the measured construct function differently across groups. In effect, the results of comparisons between the here investigated school type groups might be associated with the item differences. When some items are non-invariant between groups, the underlying relationship among item indicators and the measured construct might be unequal between the groups under study. As the absence of measurement bias is one precondition for the principle of test fairness, non-invariances between groups can limit inferences that are made based on assessments containing item bias and inference about group differences with regard to the measured construct can be invalid.

In the conducted study, it was assumed that differences between groups in measurement constructs might impact the longitudinal measurement of competence development. The groups were cluster-level covariates (i.e., school types), as clusters could be attributed to a specific group (i.e., schools belonging to a specific school type). Students' reading competence development from fifth to ninth school grade was compared between school types and it was examined if differential item functioning between school types affects estimates of reading competence development from fifth to ninth school grade. For the identification of differential item functioning and for the longitudinal comparison of reading competence, the clustered data structure was taken into account, as students were nested within schools. Reading competence was thereby assessed at three measurement occasions: at fifth, seventh and ninth school grade, with N = 7,276 participating students. Differential item functioning between school types was assessed using the alignment optimization method (Asparouhov and Muthén 2014; Kim et al. 2017; Muthén and Asparouhov 2014) and multilevel structural equation models (Kaplan, Kim and Kim, 2009; Marsh et al. 2009; Rabe-Hesketh, Skrondal and Zheng, 2007) were used to assess weather measurement non-invariance between school types is associated with reading competence development. Therefore, item information of items functioning differently between school types was treated as unique to the respective non-aligning school type versus the remaining school types.

Results of differential item functioning indicated some measurement non-invariance between school types for each measurement occasion and most non-invariance occurred between the lowest and highest school type (i.e., lower secondary school vs. high school). One-third of the items assessed in fifth grade and about one-fourth of the items in seventh and ninth grade exhibited measurement invariance. Furthermore, most item non-invariances were present for items positioned in the last part of the competence tests. Nevertheless, taking these measurement non-invariances into account by estimating school type specific item discrimination and difficulty parameters, did not alter the parallel pattern of competence development between the school types. Hence, school type comparisons of reading competence development were not sensitive to found measurement non-invariance between school types. However, measurement non-invariance between school types was more frequent in fifth grade where one test form was assessed to all students and less frequent in seventh and ninth grade when two booklets (i.e., easy and difficult test version) were administered.

As the study focused on the measurement of reading competence development between school types, it was found that differences in reading competence development from fifth to ninth grade follow a parallel pattern. Thereby, three distinct school type groups can be identified (i.e., high school vs. middle secondary school/comprehensive school/schools offering all school tracks vs. lower secondary school). Besides the assessment of differential item functioning across school types, it was tested if the found pattern of competence development between the different school types is robust to the ignorance of the hierarchical structure (i.e., students nested in schools). Misleading conclusions about the development of reading competence between German secondary school types resulted, when the hierarchical data structure was not considered, either through multilevel modeling or cluster-robust standard error estimation.

The conducted study adds to the research on reading competence development among German secondary school types, where either a parallel development (e.g., Retelsdorf and Möller 2008; Schneider and Stefanek 2004) or a widening gap (e.g., Pfost and Artelt, 2013) for the same schooling years was found. Previous studies used cluster-robust maximum likelihood estimation methods, but did not report analyses of measurement invariance between school types. We therefore emphasize the necessity to present evidence and discuss findings of measurement invariance for such group-mean comparisons in studies on competence development. Findings should thereby be discussed in the light of comparability of the measured construct across groups and over measurement occasions, as constructs should have the same meaning for all groups and at all assessments.

5.3. Manuscript 3: Random item effects in ability assessment

Gnambs, T., Scharl, A., & Rohm, T. (in press). Comparing perceptual speed between educational contexts: The case of students with special educational needs. *Psychological Test Adaptation and Development.* Accepted for publication. https://doi.org/10.1027/2698-1866/a000013

Summary

Besides cluster-effects and group-level differences in large-scale educational assessments, random item effects across clusters can be another source for differences in the measurement of latent constructs across contexts that impair principles of fair and valid assessment. Item differences can thereby be divided into fixed-group and random-group DIF, where the former indicates item bias across groups and the latter indicates random item effects across clusters. Our study compared three items measuring perceptual speed between students of regular schools (i.e., basic secondary schools, intermediate schools, upper secondary schools) and students with special educational needs for German secondary school students at ninth grade

(N = 3,312). Using the Rasch (1960) Poisson counts model with a zero-inflation process, parameterized as a general linear mixed-effects model (Fox, 2010), differential item functioning between school types was investigated for zero-inflation and item difficulty parameters. Furthermore, random school-level item effects were examined for item difficulty parameters. Model comparisons revealed that the most complex model, considering both fixed- and random-group DIF effects, had the best fit.

Only modest fixed-group DIF between special and regular school students of basic and intermediate secondary schools was present. Students from a specific school track were not systematically disadvantaged, as the direction of DIF was unequal between the two items exhibiting DIF. However, random-group DIF was found for two out of the three items. Especially for the first item, the estimated item difficulties differed across school clusters. These differences across assessment contexts (i.e., school clusters) might indicate problems of standardized test assessment. Found random variances might be due to differences in test instruction and test execution, as assessments were supervised by trained test administrators. However, it remains unclear if these random effects are associated with the specific school or with the quality of test administration. Nevertheless, the findings highlight the importance of standardized test assessments to obtain comparable test information on groups of interest for educational and psychological research.

6. Discussion

As declared in the Standards for Educational and Psychological Testing (Standards, 2014), test fairness, reliability and validity can be impaired by features of the test context, such as measurement non-invariance between cluster-level groups, error sources distorting standard error estimation and construct-irrelevant variance. However, studies of educational and psychological LSAs have rarely investigated effects stemming from the assessment context on measurement invariance in (longitudinal) group comparisons.

For a valid assessment of competence and valid inferences for group comparisons, equal treatment during the test procedure is one premise (Tierney, 2016). In light of competence assessment settings, equal treatment might be impaired when interviewers or test administrators do not adhere to standardized protocols of test assessment. The first manuscript presented evidence for consequences of unequal treatment of respondents through interviewers in mathematic competence assessment for the first assessment cycle of the NEPS adult cohort. Construct-irrelevant variance in mathematic competence was introduced to a considerable extent by interviewers, but not by areas in which the respondents were nested. It also showed that unequal treatment of respondents during the test assessment is not necessarily associated with bias, as none of the interviewer variables was significantly related to the respondent's mathematic competence. Similarly, the principle of fairness presupposes fairness in treatment during the testing procedure and the found associations between some interviewers with significantly higher residual estimates with characteristics of their respondents (i.e., significantly decreased number of missing values; significantly lower participation rates at the subsequent measurement occasion five to six years later) challenge the fairness principle. It is of question if these results occurred due to problems with standardized test assessments and a lack of interviewer training, or due to unwarranted interactions of the test examiner with the test taker. From this study, that was conducted in respondents' homes, it can be concluded that the test context (i.e., presence of interviewers) affected the assessed test as a whole for some respondents. Deviations from standardized test assessment procedures through a limited number of interviewers can lead to considerable construct-irrelevant variance, even though the tests were self-administered and interviewers were advised to stay passive during test assessment.

A comparable test setting is employed by the PIAAC Survey of Adult Skills and likewise, the quality of assessed competence depends on independently working interviewers. However, for the assessment of competence in PIAAC, effects stemming from interviewers were only assumed and potential sources of error introduced by interviewers were described for example by Hibben, Pennell and Scott (2018) and Keslair (2018), or solely investigated for their impact on the questionnaire data (Ackermann-Piek and Massing, 2014). Hence, some proposed impairments for the PIAAC study (i.e., interviewer work experience, interviewer assignment size and interviewer effects) were investigated with the here presented study (Rohm, Carstensen, Fischer and Gnambs, 2021) and found to exert influence on adult competence assessment (i.e., interviewer effects). However, as no statistical evidence was presented for interviewer effects in PIAAC assessments, results cannot be comparatively discussed. Furthermore, for the first cycle of PIAAC competence assessment, the sampling areas and interviewer assigned areas overlapped and personal information on interviewers was not collected. One can therefore conclude that the detection of potential context effects stemming from the deployment of interviewers in the field work process is not a mere statistical task. The detection of context effects already starts with the planning and design implementation of competence assessment studies. Elaborate statistical models and methods are of no use when potential error sources of the assessment context cannot be distinguished (i.e., when sampling and interviewer variance cannot be divided as the sampling design is not interpenetrated). To conclude, strategies for fair group comparisons start before data collection and evaluation, already with the design of the study and sampling of the respondents.

Besides fairness in the treatment during the test procedure, the absence of measurement bias is another requirement for the principle of fairness in test assessment. With regard to the context of the assessment, item differences can be attributable to groups comprised of cluster units. For example, in student competence assessments, item differences can occur across school types and the school types are comprised of school clusters that were the primary unit of sampling. Hence, methods for the statistical analysis of item bias between school types shall at best take these dependencies of clustered sampling into account. For the analysis of item bias in cluster-level variables, the AM method was presented in chapter 4 and applied for DIF detection in the second manuscript of this dissertation (Rohm, Carstensen, Fischer and Gnambs, 2021). When compared to other DIF detection methods, for example using a SEM or IRT approach as presented in chapter 4, the AM method has the advantage that none of the items needs to be constrained to be used as reference, as the simplicity function optimizes at non-invariant loading and intercept parameters after accumulating the total measurement non-invariance over the items. However, one has to be cautious when applying the AM procedure, as this method produces satisfying results only if the majority of items is measurement invariant across the groups (Asparouhov and Muthén, 2014). As proposed before, the context can affect the assessed instrument as a whole, several sections, or single items of an assessed test. Applying the AM to three measurement occasions of reading competence in NEPS from fifth to ninth grade, most measurement non-invariance occurred for items positioned in the last part of the competence tests. Hence, measurement non-invariance associated with school type groups affected only several sections of the competence tests and not the instrument as a whole. Furthermore, most measurement non-invariance occurred between the highest and lowest school type group and one can therefore add that the test context can be associated differently across sampled groups.

One of the main questions of this dissertation is if found measurement non-invariance related to context groups or clusters matters. Hence, after the identification of items that work on the advantage or disadvantage of school types using the AM, the MSEM framework was used to examine if some item bias between school types affects results of reading competence development in school type comparison. This strategy to investigate consequences of measurement non-invariance showed that patterns of competence development among German secondary school types were not altered when school type specific item estimates were included for items exhibiting DIF. Nevertheless, as students were sampled within their school context (i.e., students nested in schools), it was found to be important to statistically correct for these dependencies, either through multilevel modeling or cluster-robust standard error estimation. When this error source introduced by sampling of the individuals is not considered, misleading conclusions about reading competence development between different school types resulted. Therefore, ignoring nested data structures in group comparisons would reduce reliability of the results. The error source of students nested within school contexts is quite known and regularly accounted for in school type comparisons of reading competence, mostly through cluster-robust estimation methods (e.g., Pfost and Artelt, 2013; Retelsdorf and Möller, 2008; Schneider and Stefanek, 2004). In comparison, interviewer clusters introducing errors is not as frequently researched or accounted for, which might be due to the fact that school clusters are actively sampled from the population while interviewer clusters arise from the test assessment procedure.

Comparing the findings of the second manuscript to similar research on item bias associated with the assessment context in reading competence measurement, varying item effects across school types have been previously reported in evaluation of PISA 2012 (Debeer et al., 2014; Hartig and Buchholz, 2012; Qian, 2014) and the German longitudinal PISA extension from 2012 to 2013 (Nagy et al., 2016). When statistical models did not account for these test context effects (i.e., item position effects across secondary school types), larger achievement differences favoring the upper secondary school track resulted (Nagy et al., 2016). The study on NEPS reading competence assessment in secondary school, presented through the second manuscript of this dissertation, did find item bias across school types. In comparison to the PISA 2012 and 2012 to 2013 studies, it was hereby found that accounting for theses item effects with statistical modeling did not alter the differences in reading competence development across school types. Nevertheless, please note that the studies using PISA data focused on item position effects that varied across school types while the second manuscript of this dissertation investigated item bias across school types and item positions did not vary in the reading competence assessments of the NEPS.

In addition, results of the second manuscript can be compared to results on item effects in PIRLS (Oliveri and von Davier, 2014) insofar, that in both cases the number of measurement invariant items across assessment contexts (i.e., school type comparisons in NEPS vs. country comparisons in PIRLS) was quite high and the consideration of partial measurement invariance improved model fit but did not alter group comparisons. In both studies the number of measurement invariant reading competence items amounted to 75 percent (i.e., PIRLS assessment at fourth grade; NEPS assessment at seventh and ninth grade) with an exception of NEPS assessment in fifth grade, where about two-third of items were measurement invariant across the groups (i.e., school types). Another similarity between the studies is given therein, that higher levels of misfit were found for groups at the upper and lower spectrum of achievement, where Oliveri and von Davier (2014) assumed this might be due to less information at both ends of the achievement spectrum. Likewise, for all three reading competence assessments in NEPS from fifth to ninth grade, most measurement non-invariance was found between school types with lowest and highest average achievement, hinting to this assumption.

Besides measurement non-invariance arising from fixed groups (i.e., school types), random-group DIF can be present, which is measurement non-invariance across clusters (i.e., school clusters). When random-group DIF is observed, item properties such as item difficulty vary across cluster units, which might indicate problems of standardized test assessment and limit principles of fairness and validity in competence assessments. As stated in chapter 2, the choice of selection criteria for item bias and also the choice of DIF detection methods could make a difference to findings on measurement invariance. For example, the third manuscript of this dissertation (Gnambs, Scharl and Rohm, in press), demonstrated the presence of random school-level item effects for item difficulty parameters, which is random-group DIF. Variance in item difficulty across school clusters was found for two out of the three items and especially pronounced for the first assessed item. When indicating fixed-group DIF through school type by item interactions, the first item would regularly be set as the constrained reference item. With the presence of such pronounced random item variance across school clusters, this is not a suitable choice and the item with random-group measurement invariance was set as reference. This choice of reference group can make a difference to the detection of DIF between the fixed groups (i.e., school types).

Altogether, the analyses of the third manuscript on random-group DIF also demonstrate how construct-irrelevant variance stemming from the assessment context is associated with single items of the assessed test. Such random deviations of item parameters have rarely been investigated as violations of measurement invariance. The development of Bayesian methods has enabled the estimation of item specific variance components and the inclusion of covariances of random effects on the group-level (Hartig et al., 2020; Verhagen et al., 2015). Nevertheless, item variances across cluster units and their associations with estimates of group comparisons have rarely been investigated or discussed. Hence, investigations on the relevance of random item effects across clusters are desiderata for future research.

7. Conclusion

In this dissertation, effects on competence and ability measurement in psychological and educational testing stemming from the context of the assessment have been analyzed using multilevel latent variable modeling methods. The three presented studies focusing on (1) construct-irrelevant variance from the assessment context, (2) context-level group differences in item estimates and their relevance for the measurement of competence development and (3) random item effects across clusters, have been set in relation to standards for educational and psychological testing, was well as to results of similar LSA studies. Regarding the content of the three studies, contexts exerting effects on assessments can be manifold and for example stem from interviewer or test administrator presence, where clusters are a feature of the assessment procedure. Furthermore, clusters can arise from the sampling procedure, as is the case when schools or areas are the primary sampling unit. However, clustered data does not necessarily present context effects on assessments and there are statistical methods to correct for dependencies in the data stemming from clustered sampling (e.g., cluster-robust standard error estimation methods or multilevel models).

As demonstrated through three studies, the presence of context effects can have consequences for competence and ability assessments and unequal treatment of respondents across contexts can for example lead to construct-irrelevant variance, item bias between context groups or random item variance. Because of manifold contexts in which competence is assessed in LSAs and the various effects it might have on results, researchers have to decide which assessment contexts to investigate and how to examine possible effects on the assessment. As stated in the standards for educational and psychological testing (2014), not all error sources can be identified or captured and it is therefore recommended to evaluate those which are likely to have the highest impact. Statistical methods, research strategies and some likely error sources associated with contexts of competence and ability assessment in LSAs have been presented in this dissertation. Thereby, the here presented work on context effects in LSAs is interdisciplinary, as the work is associated with different research disciplines. While the investigation of interviewer and area effects is in the area of survey statistics, the work on reading competence development between secondary school types is in the field of educational research and the comparison of the perceptual speed assessment between school types is among pedagogical and psychological test assessment.

What can be generalized over all these research disciplines with regard to context effects is, that critical evaluations of assessment practices and reflections on study designs are important to recognize features of the assessment that might lead to increased variance of the construct, increased item variance, or item bias (Tierney, 2016). Hence, thoughtful preparation and critical reflection of assessments are desired. Finally, following the recommendation from Borsboom (2006), educational research needs to transition from the question if a test is biased to 'does the amount of bias in that test matter?'. As the first question is empirical, the second question demands from researchers to critically review the purpose of assessments and reflections on sources of biasing effects.

8. References

Asparouhov, T., & Muthén, B. (2014). Multiple-Group Factor Analysis Alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(4), 495–508. doi:10.1080/10705511.2014.919210

Ackermann-Piek, D., & Massing, N. (2014). Interviewer behavior and interviewer characteristics in PIAAC Germany. *Methods, Data, Analyses*, 8(2), 199-222. doi:10.12758/mda.2014.008

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

Bechger, T. M., & Maris, G. (2015). A statistical test for differential item pair functioning. *Psychometrika*, 80, 317–340. doi:10.1007/s11336-014-9408-y

Borsboom, D. (2006). When Does Measurement Invariance Matter? *Medical Care*, 44(11), 176-181.

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (pp. 221–256). Westport, CT: Praeger.

Depaoli, S., & Clifton, J. (2015). A Bayesian Approach to Multilevel Structural Equation Modeling with Continuous and Dichotomous Outcomes. *Structural Equation Modeling: A Multidisciplinary Journal*, 22, 1-25. doi:10.1080/10705511.2014.937849

Debeer, D., Buchholz, J., Hartig, J., & Janssen, R. (2014). Student, school, and country differences in sustained test-taking effort in the 2009 PISA reading assessment. *Journal of Educational and Behavioral Statistics*, 39, 502-523.

Dorans, N.J., & Cook, L.L. (2016). *Fairness in Educational Assessment and Measurement* (1st ed.). Routledge. https://doi.org/10.4324/9781315774527

Feskens, R., Fox, J. P., & Zwitser, R. (2019). Differential Item Functioning in PISA Due to Mode Effects. In Veldkamp B., & Sluijter C. (Eds.), *Theoretical and Practical Advances in Computer-based Educational Measurement. Methodology of Educational Measurement and Assessment.* Springer, Cham. doi:10.1007/978-3-030-18480-3_12

Finch, H. (2014). Measurement Invariance. In Michalos A.C. (Ed.). *Encyclopedia of Quality of Life and Well-Being Research*. Springer, Dordrecht. doi:10.1007/978-94-007-0753-5_1759

Finch, H. (2005). The MIMIC Model as a Method for Detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT Likelihood Ratio. *Applied Psychological Measurement*, 29(4), 278–295. doi:10.1177/0146621605275728

Fox, J.-P. (2010). *Bayesian Item Response Modeling: Theory and Applications*. New York: Springer.

Fox, J.-P. (2003). Stochastic EM for Estimating the Parameters of a Multilevel IRT Model. *British Journal of Mathematical and Statistical Psychology*, 56, 65-81.

Fox, J.-P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using gibbs sampling. *Psychometrika*, 66, 271-288.

Hartig, J., & Buchholz, J. (2012). A multilevel item response model for item position effects and individual persistence. *Psychological Test and Assessment Modeling*, 54, 418-431.

Hartig, J., Köhler, C., & Naumann, A. (2020). Using a multilevel random item Rasch model to examine item difficulty variance between random groups. *Psychological Test and Assessment Modeling*, 62, 11-27.

Hibben, C. K., Pennell, B.-E., & Scott, L. (2018). Interviewer effects in multicultural, multinational and multiregional surveys. *Quality Assurance in Education*, 26(2), 278-289. doi:10.1108/QAE-06-2017-0032

Huber, P.J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability: Statistics*, 1(1), 221-233, Berkeley: University of California Press.

Huelmann, T., Debelak, R., & Strobl, C. (2019). A Comparison of Aggregation Rules for Selecting Anchor Items in Multigroup DIF Analysis. *Journal of Educational Measurement*, 57(2), 185-215. doi:10.1111/jedm.12246

Jak, S., Oort, F. J., & Dolan, C. V. (2014). Measurement bias in multilevel data. *Structural Equation Modeling*, 21(1), 31–39. doi:10.1080/10705511.2014.856694

Jennrich, R.I. (2006). Rotation to Simple Loadings Using Component Loss Functions: The Oblique Case. *Psychometrika*, 71, 173–191. doi:10.1007/s11336-003-1136-B

Kamata, A., & Vaughn, B.K. (2010). Multilevel IRT Modeling. In J.J. Hox & J.K. Roberts (Eds.), *Handbook of Advanced Multilevel Analysis* (pp. 41-57). Abingdon: Routledge.

Kaplan, D., & Depaoli, S. (2013). Bayesian statistical methods. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods (Vol. 1): Foundations* (pp. 407–437). Oxford University Press.

Kaplan, D., Kim, J.-S., & Kim, S.-Y. (2009). Multilevel latent variable modeling: Current research and recent developments. In R. E. Millsap & A. Maydeu-Olivares (Eds.), *The Sage handbook of quantitative methods in psychology* (p. 592–612). Sage Publications Ltd. doi:10.4135/9780857020994.n24

Keslair, F., (2018). Interviewers, test-taking conditions and the quality of the PIAAC assessment. *OECD Education Working Papers*, No. 191. Paris: OECD Publishing. doi:10.1787/5babb087-en

Kim, E., Cao, C., Wang, Y. & Nguyen, D. (2017). Measurement Invariance Testing with Many Groups: A Comparison of Five Approaches. *Structural Equation Modeling: A Multidisciplinary Journal*, 24(4), 524-544. doi:10.1080/10705511.2017.1304822

Kirsch, I., & Lennon, M.L. (2017). PIAAC: a new design for a new era. *Large-scale Assessments in Education*, 5. doi:10.1186/s40536-017-0046-6

Klein, S. P., & Hamilton, L. S. (1999). *Large-Scale Testing: Current Practices and New Directions*. Santa Monica, CA: RAND Corporation, 1999.

Kopf, J., Zeileis, A.,& Strobl, C. (2015). A Framework for Anchor Methods and an Iterative Forward Approach for DIF Detection. *Applied Psychological Measurement*, 39(2), 83-103. doi: 10.1177/0146621614544195

Lord, F.M., Novick, M.R., & Birnbaum, A. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Lu, I. R. R., Thomas, D. R., & Zumbo, B. D. (2005). Embedding IRT in Structural Equation Models: A Comparison With Regression Based on IRT Scores. *Structural Equation Modeling: A Multidisciplinary Journal*, 12(2), 263–277. doi:10.1207/s15328007sem1202_5

Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov. T., Muthén, B., & Nagengast, B. (2009). Doubly-latent models of school contextual effects: Integrating multilevel and structural equation approaches to control measurement and sampling error. *Multivariate Behavioral Research*, 44, 764-802.

McDonald, R. P. (1967). *Nonlinear factor analysis*. Psychometric Monographs, No. 15.

Meredith, W. (1993) Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58, 525–543. https://doi.org/10.1007/BF02294825

Muthén, B. and Asparouhov, T. (2012). Bayesian SEM: A more flexible representation of substantive theory. *Psychological Methods*, 17, 313-335.

Muthén, B., & Asparouhov, T. (2014). IRT studies of many groups: The alignment method. *Frontiers in Psychology*, *5*, 978. doi:10.3389/fpsyg.2014.00978

Nagy, G., Lüdtke, O., & Köller, O. (2016). Modeling test context effects in longitudinal achievement data: Examining position effects in the longitudinal German PISA 2012 assessment. *Psychological Test and Assessment Modeling*, 58(4). doi: 10.25656/01:12804

Nagy, G., Lüdtke, O., Köller, O., & Heine, J-H. (2017). IRT-Skalierung der Tests im PISA-Längsschnitt 2012/2013: Auswirkungen von Testkontexteffekten auf die Zuwachsschätzung. *Zeitschrift für Erziehungswissenschaft*, 20(2), 229-258. doi: 10.1007/s11618-017-0749-z

Nisbet, I., & Shaw, S. D. (2019). Fair assessment viewed through the lenses of measurement theory. *Assessment in Education: Principles, Policy & Practice*, 26(5), 612-629. doi:10.1080/0969594X.2019.1586643

Oliveri, M. E., & von Davier, M. (2014). Toward increasing fairness in score scale calibrations employed in international large-scale assessments. *International Journal of Testing*, 14(1), 1–21. doi: 10.1080/15305058.2013.825265

Oliveri, M. E., & von Davier, M. (2017). Analyzing the invariance of item parameters used to estimate trends in international large-scale assessments. In H. Jiao, & R. W. Lissitz (Eds.), *Test fairness in the new generation of large-scale assessment* (pp. 121–146). New York: Information Age Publishing.

Penfield, R.D., & Camilli, G. (2007). Differential item functioning and item bias. In S. Sinharay & C. R. Rao (Eds.), *Handbook of Statistics, Volume 26: Psychometrics* (pp. 125-167). New York, NY: Elsevier.

Pfost, M., & Artelt, C. (2013). Reading literacy development in secondary school and the effect of differential institutional learning environments. In M. Pfost, C. Artelt & S. Weinert

(Eds.), *The Development of Reading Literacy from Early Childhood to Adolescence. Empirical Findings from the Bamberg BiKS Longitudinal Studies* (pp. 229-278). Bamberg: University of Bamberg Press.

Qian, J. (2014). An Investigation of Position Effects in Large-Scale Writing Assessments. *Applied Psychological Measurement*, 38(7), 518-534. doi:10.1177/0146621614534312

Rabe-Hesketh, S., Skrondal, A.,& Zheng, X. (2007). Multilevel Structural Equation Modeling. In S.-Y. Lee (Ed.), *Handbook of Latent Variable and Related Models* (pp.209-227). North-Holland: Elsevier.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press.

Retelsdorf, J., & Möller, J. (2008). Entwicklungen von Lesekompetenz und Lesemotivation: Schereneffekte in der Sekundarstufe? *Zeitschrift Für Entwicklungspsychologie und Pädagogische Psychologie*, 40(4), 179–188. doi:10.1026/0049-8637.40.4.179

Robitzsch, A., Lüdtke, O., Köller, O., Kröhne, U., Goldhammer, F., & Heine, J.-H. (2017). Herausforderungen bei der Schätzung vonTrends in Schulleistungsstudien. *Diagnostica*, 63(2), 148–165. doi:10.1026/0012-1924/a000177

Schneider, W., & Stefanek, J. (2004). Entwicklungsveränderungen allgemeiner kognitiver Fähigkeiten und schulbezogener Fertigkeiten im Kindes- und Jugendalter. *Zeitschrift Für Entwicklungspsychologie und Pädagogische Psychologie*, 36(3), 147–159. doi:10.1026/0049-8637.36.3.147

Schulze, D., & Pohl, S. (2020). Finding clusters of measurement invariant items for continuous covariates. *Structural Equation Modeling: A Multidisciplinary Journal*, 28(2), 219-228. doi: 10.1080/10705511.2020.1771186

Shin H.J., von Davier M., & Yamamoto K. (2019) Investigating Rater Effects in International Large-Scale Assessments. In Veldkamp B., & Sluijter C. (Eds.). *Theoretical and Practical Advances in Computer-based Educational Measurement. Methodology of Educational Measurement and Assessment*. Springer, Cham. doi:10.1007/978-3-030-18480-3_13

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models*. Boca Raton, FL: Chapman & Hall/CRC.

Steele, F., & Goldstein, H. (2006). A Multilevel Factor Model for Mixed Binary and Ordinal Indicators of Women's Status. *Sociological Methods & Research*, 35(1), 137–153. doi:10.1177/0049124106289112

Tierney, R. D. (2016). Fairness in educational assessment. In M. A. Peters (Ed.), *Encyclopedia of Educational Philosophy and Theory*. Singapore: Springer Science+Business Media. doi:10.1007/978-981-287-532-7_400-1

van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthén, B. (2013). Facing off with Scylla and Charybdis: A comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in Psychology*, 4, 1–15. doi:10.3389/fpsyg.2013.00770

Verhagen, A. J., Levy, R., Millsap, R. & Fox, J.-P. (2015). Evaluating evidence for invariant items: A Bayes factor applied to testing measurement invariance in IRT models. *Journal of Mathematical Psychology*, 72. doi:10.1016/j.jmp.2015.06.005.

White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*, 48(4), 817. doi:10.2307/1912934

Zumbo, B. D. (2007). Three generations of differential item functioning (DIF) analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, 4, 223–233.

Zimmermann, L., & Grudnitski, P.C., (2017). Considerations in Making next generation assessments accessible and fair. In Jiao, H. & Lissitz, W. (Eds.). *Test Fairness in the New Generation of Large Scale Assessments*. The MARCES Book Series.

Authors' contributions to manuscripts

Rohm, T., Carstensen, C. H., Fischer, L. & Gnambs, T. (2021). Disentangling Interviewer and Area Effects in Large-Scale Educational Assessments using Cross-Classified Multilevel Item Response Models. *Journal of Survey Statistics and Methodology*, 9(4), 722-744. https://doi.org/10.1093/jssam/smaa015

The development of the research idea, literature review, data analysis, result interpretation and writing of significant parts of the manuscript was done by Theresa Rohm. Claus H. Carstensen, Luise Fischer and Timo Gnambs substantially revised the manuscript and provided substantial input regarding the statistical analyses. All authors read and approved the final manuscript.

Rohm, T., Carstensen, C. H., Fischer, L. & Gnambs, T. (2021). The achievement gap in reading competence: the effect of measurement non-invariance across school types. *Large-scale Assessments in Education*, 9(23). https://doi.org/10.1186/s40536-021-00116-2

Theresa Rohm analyzed and interpreted the data used in this study. Theresa Rohm conducted the literature review and drafted significant parts of the manuscript. Claus H. Carstensen, Luise Fischer and Timo Gnambs substantially revised the manuscript and provided substantial input regarding the statistical analyses. All authors read and approved the final manuscript.

Gnambs, T., Scharl, A., & Rohm, T. (in press). Comparing perceptual speed between educational contexts: The case of students with special educational needs. *Psychological Test Adaptation and Development.* Accepted for publication. https://doi.org/10.1027/2698-1866/a000013

Timo Gnambs developed the research idea and conducted significant parts of data analysis and result interpretation. Timo Gnambs drafted significant parts of the manuscript. Anna Scharl and Theresa Rohm provided substantial input to the statistical analysis and result interpretation. All authors contributed substantially to literature review and writing of the manuscript. All authors read and approved the final manuscript.

Appendix

Copyright for manuscripts

Rohm, T., Carstensen, C. H., Fischer, L. & Gnambs, T. (2021). Disentangling Interviewer and Area Effects in Large-Scale Educational Assessments using Cross-Classified Multilevel Item Response Models. *Journal of Survey Statistics and Methodology*, 9(4), 722-744. https://doi.org/10.1093/jssam/smaa015

The main author obtained the license to use the material in the dissertation by Oxford University Press on October 07, 2022. License number: 5403581114780

Rohm, T., Carstensen, C. H., Fischer, L. & Gnambs, T. (2021). The achievement gap in reading competence: the effect of measurement non-invariance across school types. *Large-scale Assessments in Education*, 9(23). https://doi.org/10.1186/s40536-021-00116-2

The article is licensed under a Creative Commons Attribution 4.0 International License. No changes were made to the original material.

Gnambs, T., Scharl, A., & Rohm, T. (in press). Comparing perceptual speed between educational contexts: The case of students with special educational needs. *Psychological Test Adaptation and Development.* Accepted for publication. https://doi.org/10.1027/2698-1866/a000013

The article is licensed under a Creative Commons Attribution 4.0 International License. No changes were made to the original material.

Journal of Survey Statistics and Methodology (2021) **9**, 722–744

DISENTANGLING INTERVIEWER AND AREA EFFECTS IN LARGE-SCALE EDUCATIONAL ASSESSMENTS USING CROSS-CLASSIFIED MULTILEVEL ITEM RESPONSE MODELS

THERESA ROHM*
CLAUS H CARSTENSEN
LUISE FISCHER
TIMO GNAMBS

In large-scale educational assessments, interviewers should ensure standardized settings for all participants. However, in practice many interviewers do not strictly adhere to standardized field protocols. Therefore, systematic interviewer effects for the measurement of mathematical competence were examined in a representative sample of $N = 5{,}139$ German adults. To account for interviewers working in specific geographical regions, interviewer and area effects were disentangled using cross-classified multilevel item response models. These analyses showed

Theresa Rohm is Research Assistant at the Leibniz Institute for Educational Trajectories, Wilhelmsplatz 3, 96047 Bamberg, Germany, and Research Assistant at the University of Bamberg, Germany. Claus H. Carstensen is Professor for Psychological Methods of Educational Research atthe University of Bamberg,Wilhelmsplatz 3, 96047 Bamberg, Germany. Luise Fischer is Research Assistant at the Leibniz Institute for Educational Trajectories, Wilhelmsplatz 3, 96047 Bamberg, Germany, and Research Assistant at the University of Bamberg, Germany. Timo Gnambs is head of the Educational Measurement Unit at the Leibniz Institute for Educational Trajectories, Wilhelmsplatz 3, 96047 Bamberg, Germany, and professor for psychology at Johannes Kepler University Linz, Altenberger Straße 69, 4040 Linz, Austria.

This article uses data from the National Educational Panel Study (NEPS): Starting Cohort 6 – Adults, doi: 10.5157/NEPS: SC6:3.0.1. From 2008 to 2013, NEPS data were collected as part of the Framework Programme for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research. As of 2014, the NEPS survey is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network. The research project was partially funded by the Deutsche Forschungsgemeinschaft (DFG; http://www.dfg.de) within the Priority Programme 1646 entitled “A Bayesian model framework for analyzing data from longitudinal large-scale assessments” under Grant No. CA 289/8-2 (awarded to Claus H. Carstensen). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

*Address correspondence to Theresa Rohm, Leibniz Institute for Educational Trajectories, Wilhelmsplatz 3, 96047 Bamberg, Germany; E-mail: theresa.rohm@lifbi.de.

doi: 10.1093/jssam/smaa015 Advance access publication 4 September 2020

that interviewer behavior distorted competence measurements, whereas regional effects were negligible. On a more general note, it is demonstrated how to identify conspicuous interviewer behavior with Bayesian multilevel models.

KEYWORDS: Administration effects; Competence measurement; Interviewer effects; Large-scale assessment; Multilevel item response theory.

1. INTRODUCTION

Interviewer behavior is an essential factor in large-scale educational assessments to guarantee valid measurements of, for example, cognitive abilities, motivations, or attitudes (Moss, Girard, and Haniford 2006). By adhering to standardized field protocols, interviewers need to accomplish a variety of tasks such as creating comparable settings that avoid unnecessary disruptions or providing similar assistance to all participants who does not give an undue advantage to some respondents. Typically, not all interviewers are equally capable; depending on their abilities or motivations, some interviewers might be more likely to succeed in creating standardized assessment conditions than others (Schaeffer, Dykema, and Maynard 2010; West and Olson 2010; West, Kreuter, and Jaenichen 2013; Turner, Sturgis, Martin, and Skinner 2014). If specific interviewer behavior affects the responses of some participants, responses from different respondents being assessed by the same interviewer are likely correlated and, thus, exhibit a systematic interviewer-specific variance. This variance might even depend on specific interviewer characteristics (e.g., age or experience) or interactions between interviewer and respondent characteristics. Consequently, a test taker's responses not only reflect the construct of interest (i.e., attitudes, cognitive abilities) but also context effects introduced by nonstandardized assessment conditions. Ignoring these dependencies in the analysis of respondent data risks underestimating standard errors and, in turn, overestimating the statistical significance of effects (Durrant, Groves, Staetsky, and Steele 2010; Finch and Bolin 2017).

Because interviewers often work in a specific geographical region, interviewer effects can be confounded with regional characteristics (O'Muircheartaigh and Campanelli 1998). To deal with the possible relatedness of respondents being assessed by the same interviewer and of respondents living in the same sampling area, the present study adopts cross-classified multilevel models to disentangle both sources of variance. In this way, dependencies introduced by interviewer behavior and geographical areas are distinguished by estimating separate random effect structures (Maas and Hox 2004). We demonstrate cross-classified multilevel modeling in a German large-scale assessment of mathematical competences and evaluate the impact of interviewers on competence measurement.

2. INTERVIEWER AND AREA EFFECTS IN LARGE-SCALE ASSESSMENTS

Domain-specific competences such as mathematical or reading competence represent central factors for successful performance in many educational and professional situations (Hartig, Klieme, and Leutner 2008). They explain educational trajectories, occupational choices, and even differences in wages (Heckman, Stixrud, and Urzua 2006) and, thus, determine the social and economic success of individuals. Moreover, from the perspective of cross-country comparisons, enhanced cognitive skills improve economic well-being of nations (Hanushek and Woessmann 2008). Therefore, various large-scale assessments such as the *Programme for International Student Assessment* (PISA), the *Trends in International Mathematics and Science Study* (TIMMS), the *Progress in International Reading Literacy Study* (PIRLS), or the *Programme for the International Assessment of Adult Competencies* (PIAAC) have been initiated to identify determinants of skill inequality and provide policy makers recommendations for political action. The study of competences requires standardized measurements that allow for the estimation of reliable competence scores (Pohl and Carstensen 2013). Importantly, respective test scores should only reflect individual differences in the measured competence and not situational influences or context effects from, for example, different assessment modes (e.g., computer versus paper; cf. Wang, Jiao, Young, Brooks, and Olson 2007), distractive environments (e.g., disturbance by other test takers or media devices; cf. Shelton, Elliott, Eaves, and Exner 2009), or different forms of assistance (e.g., lengthy versus limited test instructions). In this regard, interviewers are assigned an essential role. They are responsible for the implementation of standardized assessment settings for all participants and, thus, should give each test taker equal opportunities to achieve good test scores.

2.1 Interviewer Effects

Interviewers can affect the quality of the obtained data through the contact with possible respondents (nonresponse error) and the actual process of interviewing (interviewer bias). Nonresponse error is produced because interviewers influence the propensity of the respondents to participate in the survey (Schaeffer et al. 2010; West and Olson 2010; West et al. 2013; Vassallo, Durrant, Smith, and Goldstein 2015). In contrast, interviewer bias is introduced during the administration of the questionnaire or test. Various directly observable interviewer characteristics such as the interviewers' age, gender, or ethnicity as well as unobservable characteristics (e.g., experiences, stereotypes about the respondent, attitudes toward the surveyed topic, expectations about item difficulty) can exert nonnegligible effects on survey responses (Groves 1989; Hox 1994; O'Muircheartaigh and Campanelli 1998; Rosenthal 1967, 2002; Tourangeau and Yan 2007; Brunton-Smith, Sturgis, and Williams 2012). For example, a well-known systematic influence on survey results are interpersonal expectancy effects (Rosenthal 1994). Interviews are a social process: not only do respondents provide information to the interviewer, but interviewers also provide information to the respondents. If interviewers hold certain beliefs about the topic of a survey, they might unintentionally communicate subtle hints (e.g., by body language, tone of voice) to which respondents might react. In survey research, interviewer effects have sometimes been found to be small, often explaining less than 10 percent of variance in nationally representative household surveys (e.g., Groves 1989; Brunton-Smith, Sturgis, and Leckie 2016). Rarely, cross-country studies show intra-interviewer variance approaching 20 percent (e.g., Beullens and Loosveldt 2014, 2016). However, even small effects can have an undue impact on the quality of the obtained data, particularly when each interviewer surveys many respondents (Kish 1965; Collins 1980; Hox, de Leeuw, and Kreft 1991; Schaeffer et al. 2010).

Differences in interviewer behavior can also systematically bias the assessment of competences (Rosenthal 1994, 2002). Although great effort is invested into standardizing large-scale assessments, for example, with the help of administration manuals and mandatory interviewer trainings, empirical investigations on the effectiveness of these efforts are rather limited. One exception is an analysis of interviewer effects within institutional contexts (classroom setting) of student educational assessments (Lüdtke, Robitzsch, Trautwein, Kreuter, and Ihme 2007). These authors found negligible interviewer effects in the 2002 PISA assessment of mathematical competence, explaining less than 1 percent of variance. Furthermore, neither interviewer characteristics (e.g., gender, experience) nor interactions between respondents' and interviewers' gender yielded an effect on the observed achievement scores. So far, little is known about interviewer effects on competence measurement in noninstitutional individual settings. Because of the less standardized assessment situation in the respondents' private homes, differences in interviewer behavior might have stronger effects on competence measurements.

2.2 Area Effects

Another source of imprecision in the estimation of respondents' proficiency is variance introduced through the sampling of respondents through regional clusters. In complex sampling designs, respondents are selected from a population using multistage cluster sampling. Thereby, the responses of survey participants belonging to the same area cluster can be correlated. The homogenizing effect of sampling points is also termed “spatial homogeneity” (Schnell and Kreuter 2005). It results from similar sociodemographic characteristics of respondents who live in the same area (Schnell and Kreuter 2002; Gabler and Lahiri 2009), as well as socioeconomic and cultural characteristics, accessibility or factors of urbanicity (Haunberger 2010). For example, within a regional cluster income, age and ethnicity of the respondents are likely to be more similar than across different clusters (Lee, Forthofe, and Lorimor 1989); consequently, measured attitudes, proficiencies, and behaviors related to these characteristics are likely to be correlated with the regional clustering.

In face-to-face surveys, interviewers are often assigned to respondents based on spatial proximity. However, when each interviewer works in a specific region, effects of interviewers and areas can be confounded. This confounding could be minimized with the use of an interpenetrated design (Mahalanobis 1946; Hox 1994), where interviewers are assigned at random to respondents, living in different areas. Consequently, explanatory variables on the interviewer and area level can be assumed to be no longer correlated. However, this is often rather impractical for national surveys because this design is associated with high travel expenses for interviewers. In contrast, partially or limited interpenetrated designs allow to empirically disentangle interviewer and area effects, although interviewer and area clusters do overlap to some extent. As a requirement for this limited interpenetrated design, some interviewers work in more than one area and areas are visited by more than one interviewer. For example, a recent simulation study (Vassallo, Durrant, and Smith 2017) on cross-classified multilevel logistic models predicting survey nonresponse suggests that three areas per interviewer can be sufficient interviewer dispersion across areas, resulting in good precision of survey estimates. Particularly, the random variance structure can be severely biased when interviewers work in only one area, whereas intercept estimates seem to be less affected by restrictive interviewer allocation schemes.

Therefore, empirical analyses of interviewer effects need to account for possible clustering of interviewers within specific areas (Durrant et al. 2010; Brunton-Smith et al. 2012; Turner et al. 2014). In survey research, joint estimations of interviewer and area effects typically found that interviewers made a higher contribution to the homogenizing effect in survey estimates as compared to sampling point clusters (Hansen, Hurwitz, and Bershad 1961; O’Muircheartaigh and Campanelli 1999; Schnell and Kreuter 2002). These studies were similar in trying to randomly allocate respondent addresses to interviewers within geographical pools or districts. Main differences were the purpose of the study (accuracy of US census data versus refusal and noncontact in the British Household Panel Study versus a design-effects study), the strategy of random allocation of interviewers to areas (e.g., the amount of interviewers allocated to respondents within and across areas), and the statistical method used to separate the interviewer and area effects (F-test versus multilevel cross-classified models versus three-level models). However, in large-scale educational assessments of adult competencies, confounded interviewer and area effects have rarely been investigated.

3. IDENTIFICATION OF INTERVIEWER AND AREA EFFECTS

Multilevel modeling is useful to separate construct variance from context effects. If substantial interviewer or area effects occur, individual observations are not completely independent. These dependencies can be acknowledged in the modeled error structure by specifying different random effects (Maas and Hox 2004). Multilevel cross-classified models allow for more than one effect of nesting to occur at the same level (Raudenbush 1993; Rasbash and Goldstein 1994; Goldstein 2011). Hence, they can alleviate the problem of confounded effects that occur from interviewer nesting and spatial clustering (Hox and de Leeuw 1994; O'Muircheartaigh and Campanelli 1998; Durrant et al. 2010). Especially when the implementation of a completely interpenetrated design (respondents are randomly assigned to interviewers, independent of any regional allocation) is not feasible, multilevel modeling approaches are beneficial to obtain unbiased estimates from partially interpenetrated designs.

To investigate cross-classified interviewer and area effects in competence measurement, we adopt a multilevel structural equation modeling (SEM) framework where the measurement part is specified as a two-parameter item response theory (IRT) model as $y_i{}^* = \Lambda \cdot \theta_i + \varepsilon_i$ (Kamata and Vaughn 2011). Here, $y_i{}^*$ represents the vector of J unobserved latent response variables for respondent $i \in 1 \ldots I$ that gives rise to the observed dichotomous responses y_i such that for item j $y_{ij} = 1$ if $y_{ij}{}^* \geq \tau_j$ and $y_{ij} = 0$ if $y_{ij}{}^* < \tau_j$. The latent variable θ_i is a vector of K factor scores representing the measured ability; in the case of a unidimensional model, $K = 1$. Finally, Λ is an $I \times K$ matrix of discrimination parameters and ε are the I zero mean normally distributed residuals.

The structural model part allows for varying intercepts and regression coefficients across C interviewers and G area clusters (for further details see Kamata and Vaughn 2011; Kaplan 2014). This can be expressed as

$$\theta_{icg} = \alpha_{cg} + \Gamma_{cg} x_{icg} + u_{icg} \text{ with } c = 1,\ 2, \ldots, C \text{ and } g = 1,\ 2, \ldots, G, \quad (1)$$

where the latent factor θ_{icg} for respondent i nested in interviewer c and area g is regressed on individual-level covariates x_{icg}. The intercept α_{cg} and the slopes Γ_{cg} are allowed to vary across interviewers and areas as a function of between-interviewer variables w_c and between-area variables w_g:

$$\alpha_{cg} = \alpha_{00} + \mathrm{A}_c w_c + \mathrm{A}_g w_g + \varepsilon_c + \varepsilon_g, \quad (2)$$

$$\Gamma_{cg} = \Gamma_{00} + \Gamma_c w_c + \Gamma_g w_g + \xi_c + \xi_g. \quad (3)$$

The residual u_{icg} is assumed to be normally distributed with zero mean and variance $\mathrm{Var}(u_{icg}) = \sigma^2{}_u$, whereas the residuals for the interviewer, ε_c and ξ_c, and area cluster, ε_g and ξ_g, are each multivariate normally distributed with zero means and variance–covariance structures $\Sigma = \begin{pmatrix} \sigma^2_\varepsilon & \sigma_{\varepsilon\xi} \\ \sigma_{\varepsilon\xi} & \sigma^2_\xi \end{pmatrix}$. The three residual structures $\sigma^2{}_u$, Σ_c, and Σ_g are usually assumed to be independent. As the interviewer-to-area distribution is not random due to the design of the analyzed study, Σ_c and Σ_g might be correlated. Consequently, even though theoretically assumed, the model cannot reveal interviewer-by-region interaction effects. When the model is presented as an unconditional cross-classified model without predictor variables, (1), (2), and (3) reduce to the mixed-effects formulation in (4):

$$\theta_{icg} = \alpha_{00} + \varepsilon_c + \varepsilon_g + u_{icg}. \tag{4}$$

Here, the achievement of respondent i equals the sum of the grand-mean achievement of all respondents α_{00}, the random effect ε_c introduced by interviewer c, the random effect ε_g of the region g, and a random respondent effect u_{icg}.

Cross-classified multilevel models can be estimated in a Bayesian framework with a Markov Chain Monte Carlo (MCMC) algorithm. Parameter estimates are obtained from posterior distributions using a Gibbs-sampler that are generated by repeated sampling from conditional distributions based upon observed data given prior information about the parameters. Thus, the uncertainty about parameter estimates is reflected in the posterior distribution. This allows for the calculation of point estimates (posterior mean) and posterior credibility intervals, which do not rely on a normal approximation of the posterior distribution (Van den Noortgate, De Boeck, and Meulders 2003). Nevertheless, using noninformative priors leads to results that are asymptotically equivalent to respective maximum likelihood estimates (Muthén and Asparouhov 2016). The Bayesian method using MCMC estimation has several advantages: For one, complex multilevel models can be fitted to the data that might not be estimable using likelihood-based frequentist methods (Finch and Bolin 2017). Furthermore, the method is helpful for noncontinuous (binary) item responses with missing values and unbalanced designs. MCMC-based Bayesian models for binary responses have been examined by Fox and Glas (2001) or Goldstein and Browne (2005); respective MCMC-based Bayesian approaches for continuous and ordinal responses are described in Lee and Song (2004). The flexibility of MCMC-based Bayesian methods for model fitting is especially beneficial for the structural model part of multilevel models, as it does not rely on asymptotic theory, presents posterior distributions for random effects, and results in more accurate parameter estimates (Muthén and Asparouhov 2012; for an application of a Bayesian multilevel SEM, see Kaplan 2014).

Cross-classified multilevel latent variable models are not often presented in large-scale educational assessments. There are applications, for example, in the context of school effectiveness research (Fox 2010) and for the measurement of attainment targets of Dutch reading comprehension for students at the end of primary school (Van den Noortgate et al. 2003). In addition, multilevel cross-classified testlet models were explored to analyze the dependency of items from clustering factors, such as testlet and content areas, as well as person factors (Jiao, Kamata, and Xie 2016). Furthermore, there are applications to longitudinal data, where, for example, students' performance scores are clustered within students and within teachers (Luo and Kwok 2012). In addition, cross-classified structural equation models were examined for a longitudinal measurement of teacher-ratings of US students' aggressive–disruptive behavior (Asparouhov and Muthén 2016).

In the present study, multilevel models with cross-classifications of interviewer and area clusters are presented to account for possibly confounded effects on the measurement of adult mathematic competence. These analyses have two aims: first, we want to identify to what degree competence measurements in large-scale assessments are distorted by interviewer and area effects. For this purpose, interviewer and area residual variance ($\sigma^2{}_{\varepsilon}$ and $\sigma^2{}_{\xi}$) are set in relation to the overall residual variance ($\sigma^2{}_{u}$) of the latent factor (θ_{icg}). Second, we demonstrate with a hands-on example how to identify interviewers that unduly influence the test results, based on the random effect variance ($\sigma^2{}_{\varepsilon}$) introduced by interviewer *c*.

4. PRESENT STUDY

Our methodological goal is to identify interviewer effects on mathematic achievement through multilevel cross-classified analysis using Bayesian MCMC methods. This is very similar to the aim of Lüdtke et al. (2007). Nevertheless, our study differs in classification factors (test administrator and school versus interviewer and area), study population (school students versus adults), setting (group testing versus face-to-face settings), as well as the modeling of the latent construct. While Lüdtke et al. (2007) used manifest mathematics scores at the respondent level that were scaled in advance of the analysis, we incorporate the measurement model directly into our cross-classified multilevel model. As the mathematics construct cannot be assessed directly, it is measured by a set of items reflecting the hypothetical construct. Thereby, the variable of interest cannot be measured perfectly and, in effect, measurement error is present. Using a cross-classified multilevel latent variable model, we account for measurement error of the latent variable. The measurement part of our model is specified as a two-parameter logistic IRT model, which is a model that is frequently presented in educational and psychological measurement on multilevel IRT modeling (e.g., Fox and Glas 2001; Fox 2003;

Skrondal and Rabe-Hesketh 2004). In comparison to the Rasch Model (Rasch 1980), the assumption of equal item discriminations is relaxed to allow that items discriminate unequally among respondents with different abilities.

5. METHODS

5.1 Sample and procedure

The participants were part of the *National Educational Panel Study* (NEPS; Blossfeld, Roßbach, and von Maurice 2011) that included a representative sample of German adults (see Hammon, Zinn, Aßmann, and Würbach 2016, for details on the stratified multistage sampling procedure). Primary sampling units of a two-stage sampling procedure served as area clusters (strata) in the analyses. Respondents were randomly drawn from local registers of residents within each area cluster and a private research institute supervised the distribution of addresses to interviewers. Although respective information was not explicitly provided, we assume that respondents were allocated to interviewers based on proximity of the living addresses.

Originally, 5,245 respondents participated. However, about 2 percent of the sample was excluded due to an excessive number of missing values on the competence test ($n = 24$), background information on the respondents or interviewers ($n = 81$), or failure to match respondent records to an interviewer ($n = 1$). Thus, the analyses are based on a sample of $N = 5,139$ respondents (50.9 percent women) aged between 25 and 72 years (Mdn = 52.33, SD = 10.96). Nearly half of the sample attained matriculation standard or holds a graduate degree (47.1 percent). Overall, the respondents lived in ninety-two area clusters (strata) with an average of Mdn = 37.5 (Min = 1, Max = 360) persons per regional cluster. The respondents were interviewed by 200 different interviewers (40.0 percent women) who each tested Mdn = 21 (Min = 1, Max= 123) persons (three interviewers interviewed only one respondent). Furthermore, the interviewers visited Mdn = 2 regions (Min = 1, Max = 8); each region was visited by Mdn = 3 interviewers (Min = 1, Max = 30). The distribution of regions per interviewer (see online Supplementary Material) shows that most of the interviewers ($n = 114$, 57 percent) worked in at least two different regions. Nevertheless, a considerable number of interviewers ($n = 86$, 43 percent) worked in only one region. The respondents were tested individually in their private homes by a professional survey institute. The interviewers had the complex task of administering the competence test within a computer-assisted personal interview. Thus, they had to switch from being responsive to the respondent during the completion of the computerized questionnaire to the application of strict rules of standardization during the subsequent paper-based competence test (Fellenberg, Sibberns, Jesske, and Hess 2016). Important tasks for the interviewers during the personal interview were to motivate the respondent and to present the items and response options, whereas they had to standardize the competence assessment by minimizing disturbances in the respondents' home environment. Further details on the data collection process and the survey execution are provided on the project website (http://www.neps-data.de).

5.2 Instruments

Mathematical competence was measured with a paper-based achievement test including twenty-one items that were specifically constructed for administration in the NEPS. All items were accompanied by multiple choice or short-constructed response formats that were dichotomously scored. The construction rationale and development of the test are described by Neumann, Duchhardt, Grüßing, Heinze, Knopp, et al. (2013). Following the NEPS framework for mathematical competence, each item belonged to one of the four content areas: (1) quantity, (2) space and shape, (3) change and relationships, and (4) data and chance. Thereby, the content areas of the NEPS do not follow the canonical categorization of mathematical disciplines (e.g., geometry, algebra, analysis, probability theory) but refer to four content areas encompassing everyday problems. Mathematical competence in adulthood is characterized by a strong focus on the literacy aspect (e.g., apply mathematical concepts to a variety of contexts) as compared to younger cohorts (e.g., students at school). Hence, the measured concept is assumed to have high variance among the adult population, with some items covering mathematical issues that are necessary for everyday life and other items being very specific for typical contexts (e.g., relevant for specific careers/occupations). In addition, but not related to the content areas, six cognitive components were required to solve the tasks: (1) mathematical communication, (2) mathematical argumentation, (3) modeling, (4) using representational forms, (5) mathematical problem solving, and (6) technical abilities and skills. These cognitive processes condition the mathematical ability of adults as they need to be activated when solving the respective item. Both dimensional concepts, the four content areas and the six cognitive components, are closely linked to the PISA framework (OECD 2004; for details see Neumann et al. 2013). Despite the different components specified in the construction rationale, these are not assumed to represent distinct dimensions. Rather, the test is dominated by a single mathematical factor. In-depth psychometric analyses corroborated a unidimensional structure and measurement invariance across several respondent characteristics (see Jordan and Duchhardt 2013). In addition, hierarchical IRT models with random discrimination and threshold effects did not indicate substantial item variances across interviewer or area clusters.

We acknowledged several *respondent characteristics* that might be associated with mathematical competence: age (in years), gender (coded 0 for men and 1 for women), ethnicity (coded 0 for no migration background and 1 otherwise), educational level (with four categories: lower secondary degree or less, secondary education, matriculation standard, graduate degree), employment status (coded 0 as employed and 1 otherwise), and cultural capital (as reflected by the number of books in the household). In addition, the political area size per respondent (measured as number of inhabitants of the respondents' municipality with seven categories ranging from "below 2,000 inhabitants" to "500,000 and more inhabitants") was acknowledged. Considering these individual characteristics in our analyses should increase the comparability between regional clusters because geographical areas might differ on key sociodemographic characteristics. Otherwise, differences between interviewers could reflect differences between areas.

Finally, several *interviewer characteristics* were available. Besides gender (coded 0 for men and 1 for women), the interviewers' age and educational attainment were each measured with three categories using either "less than 50 years," "50–65 years," and "older than 65 years" or "up to lower secondary degree," "secondary education," and "matriculation standard." Work experience as an interviewer, recorded as the general interviewing experience of being employed at the private institute that supervised the assignment of interviewers to the sampled respondents of the NEPS, was indicated on four categories including "up to 2 years," "2–3 years," "4–5 years," and "more than 5 years." Descriptive statistics for these variables are summarized in the online Supplementary Material.

5.3 Statistical Analyses

As previous analyses supported a unidimensional scale (Jordan and Duchhardt 2013), a unidimensional two-parametric IRT model (Kamata and Vaughn 2011) was fitted to the mathematical test. Continuous predictors of the latent ability (i.e., respondents' age, number of books in the household, and political area size) were grand-mean centered. Because interviewer effects were expected to be statistically confounded with effects at the area level, cross-classified multilevel models with MCMC and noninformative priors were estimated in *Mplus* 8 (Muthén and Muthén 1998–2017). All variance parameters were estimated using inverse-gamma priors IG $(-1, 0)$, while loading and threshold parameters were estimated with normal distribution priors of zero mean and variance of 5, N (0, 5). The prior for the parameters of all first- and second-level covariates was the normal distribution with zero mean and infinity variance, N $(0, \infty)$. A discussion and additional model estimation results on the sensitivity of variance components to the choice of prior are found in the online Supplementary Material.

All parameter estimates and standard errors are the means and standard deviations of two parallel MCMC chains using a burn-in of half of the minimum

5,000 iterations. Thinning of the chains was applied to reduce autocorrelations (use of every 20th iteration). A convergence criterion of 0.05 was set for each model, indicating that parameter convergence is achieved when the *Potential Scale Reduction* (PSR) values fall below 1.05. Trace plots were used for each parameter to evaluate successful convergence of the estimates. Likewise, autocorrelation function plots were investigated to determine whether the estimated models delivered reliable estimates. For the evaluation of model parameters, the mean of the posterior distribution and the Bayesian 95 percent credibility interval were used. Posterior predictive checks that compared the predictive distribution to the observed data involved the PSR criterion (Gelman and Rubin 1992) for which values below 1.1 indicate convergence (Gelman, Carlin, Stern, and Rubin 2004) and the Kolmogorov–Smirnov test. The latter evaluates the hypothesis that both MCMC chains have an equal distribution, using 100 draws from each of the two chains per parameter.

Finally, the Bayesian residual estimates are used to visualize heterogeneity stemming from interviewer and area clusters, as well as the dependence between units nested within the clusters. To identify exceptional interviewer and area clusters, the posterior standard deviations are used as standard errors for making inferences about the random interviewer and area effects of interest. Random effects are drawn for each cluster based on the posterior distribution of θ_{icg} given the observed data for the cluster. The random effects distribution can thereby be viewed as mirroring the variation in θ_{icg} in the survey population (Skrondal and Rabe-Hesketh 2009). In addition, the posterior standard deviation is used to form confidence intervals for the estimated random intercepts of interviewer- and area-specific measured competence values.

5.4 Data Availability and Analyses Syntax

The complete data set analyzed in this study is available at http://www.neps-data.de. Moreover, the analyses syntax used to generate the reported results are provided in an online repository at https://doi.org/10.17605/OSF.IO/FKA9X.

6. RESULTS

Because respondents were nested in interviewers and geographical areas, mathematical competence was modeled in a cross-classified multilevel IRT framework as outlined above. We estimated a series of increasingly complex models to evaluate potential interviewer effects (see table 1). The trace and autocorrelation plots for all models indicated sufficient convergence of the parameter estimation. Moreover, after 1,000 iterations, the PSR criterion fell below 1.1 for all parameters and the Kolmogorov–Smirnov statistics were not significant (all *p*s $>$ .01). Thus, the models showed appropriate posterior predictive quality for the parameters on the within and between level.

Table 1. Results of Cross-Classified Multilevel IRT Models Estimating Adult Mathematical Achievement(−0.202, −0.147)

	Model 1			Model 2			Model 3		
	M	SD	95% PPI	*M*	SD	95% PPI	*M*	SD	95% PPI
Fixed effects									
Age				−0.175	0.014	(−0.202, −0.147)	−0.175	0.014	
Gender (ref. male)				−0.325	0.012	(−0.348, −0.301)	−0.324	0.012	(−0.348, −0.300)
Migration background (ref. no)				−0.064	0.013	(−0.089, −0.040)	−0.064	0.013	(−0.090, −0.039)
Educational attainment (ref. secondary education)									
No degree or lower sec. degree				−0.149	0.015	(−0.178, −0.120)	−0.149	0.015	(−0.179, −0.120)
Matriculation standard				0.176	0.014	(0.146, 0.204)	0.175	0.014	(0.146, 0.203)
Graduate degree				0.347	0.015	(0.318, 0.376)	0.347	0.015	(0.318, 0.376)
Employment status (ref. employed)				−0.044	0.014	(−0.070, −0.017)	−0.044	0.014	(−0.071, −0.017)
Cultural capital				0.165	0.015	(0.136, 0.194)	0.165	0.015	(0.136, 0.194)
Political area size				−0.041	0.019	(−0.078, −0.005)	−0.041	0.019	(−0.079, −0.004)
Interviewer-level covariates									
Gender (ref. male)							−0.139	0.088	(−0.308, 0.040)
Age (ref. up to 49 years)									
50–65 years							−0.091	0.102	(−0.293, 0.109)
Older than 65 years							−0.007	0.107	(−0.197, 0.217)
Educational attainment (ref. lower sec. degree)									
Secondary education							0.150	0.128	(−0.095, 0.401)
Matriculation standard							0.011	0.129	(−0.239, 0.260)
Work experience as interviewer (ref. up to 2 years)									
2–3 years							0.087	0.129	(−0.169, 0.345)
4–5 years							0.248	0.126	(−0.016, 0.488)
More than 5 years							0.045	0.126	(−0.202, 0.289)
Variance components of random effects									
Respondents	0.425	0.033	(0.363, 0.492)	0.236	0.020	(0.197, 0.275)	0.244	0.019	(0.209, 0.284)
Interviewers	0.030	0.006	(0.020, 0.045)	0.032	0.006	(0.022, 0.046)	0.033	0.006	(0.023, 0.046)
Areas	0.004	0.003	(0.000, 0.013)	0.001	0.001	(0.000, 0.005)	0.001	0.001	(0.000, 0.005)

NOTE.— Standardized results are presented for fixed effects. *M*, posterior mean; SD, posterior standard deviation; PPI, posterior probability interval (2.5th and 97.5th percentiles of the posterior distribution).

6.1 Interviewer and Area Effects

In the first step, we estimated the amount of variance in competence measurement that is attributable to the different interviewers and areas without considering any predictors (i.e., a null model; see Model 1 in table 1). The impact of clustering on the outcome variable was investigated using intraclass correlation coefficients (ICCs) that indicate the proportion of variance attributable to a higher-order cluster (i.e., interviewers, areas) in the total variance. Larger ICCs indicate larger dependencies for interviewer or area clusters and, thus, a greater need for multilevel analyses (Hox 2010; Finch and Bolin 2017). The variance in mathematical achievement between interviewers was much higher than the variance between areas: about 6.6 percent of the observed variance in mathematical competence was attributable to interviewers, whereas only 0.8 percent was attributable to the nesting of respondents in geographical areas.

Moreover, the design effect highlights the accuracy of the results in comparison to random sampling; at the same time, it denotes how much larger the sample size must be to obtain the same precision in survey estimates (Schnell and Kreuter 2005). For example, a design effect of 2 is assumed to reduce the effective sample size by half (Schaeffer et al. 2010). In the present study, the design effects for the interviewer and area clusters were 2.60 and 1.44, respectively. Thus, there was substantial interviewer variance, but negligible area effects.

In the second step (see Model 2 in table 1), the respondent characteristics and the size of the political area the respondents live in were added as fixed effects. This revealed significantly ($p < .05$) worse achievement for women, respondents with migration background, lower education, or a lower socioeconomic status, and those without employment. Moreover, test takers living in smaller areas (as measured by political area size) achieved slightly better mathematical competence as compared to people living in strongly populated areas. Although the inclusion of these variables reduced the respondent-specific random variance by nearly a half, the interviewer variance remained unaffected.

6.2 Identification of Influential Interviewers

Even though the variance in mathematical competences traceable to interviewer presence was rather high, none of the investigated interviewer characteristics (e.g., gender, age, education, and work experience) was found to be significantly related to the latent competence of the respondents (see Model 3 in table 1). Furthermore, the interaction of interviewer and respondent gender did not affect mathematical achievement. Thus, sociodemographic differences were unable to identify interviewers with aberrant test administration behaviors. Therefore, we used the interviewer residual terms (second-level errors) that were sampled from the posterior distribution of our estimated multilevel model (Model 1) to identify exceptional interviewers. Because these residuals

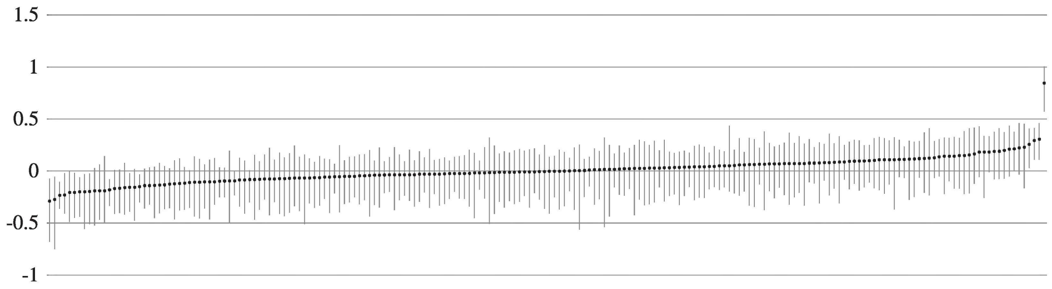


Figure 1. Residuals of Interviewers with Corresponding Posterior Probability Interval (2.5th and 97.5th Percentiles of the Posterior Distribution).

were sample estimates and, therefore, incorporated a level of uncertainty (e.g., they depend on the number of interviewed respondents and on the amount of within- and between-interviewer variation), we ranked the interviewers according to the interviewer residual effects with their 95 percent probability interval (see figure 1). Residuals whose posterior probability intervals do not overlap with the general mean indicate interviewers with undue influence on the competence measurement of the respondents. Out of 200 interviewers, 4 had an interval above and 12 had an interval below zero. Hence, their estimated competence intercept deviates from the survey population mean.

To confirm that the results did not depend on the number of regions an interviewer worked in, we refitted Models 1–3 to data collected by the 57 percent of interviewers who worked in at least two different regions. These results did not indicate substantial differences from the findings reported above, as the estimated fixed and random effects remained nearly identical (see table S12 in the online Supplementary Material). In addition, interviewer residual effects whose 95 percent posterior probability interval did not overlap with the general mean in the original estimation also had significant deviation in their residual effect in these sensitivity analyses.

In multilevel IRT analyses, shrinkage to the general mean can be expected for interviewer residuals if a large number of respondents were assigned to an interviewer. Then, the posterior mean resembles the intercept of a separate regression for this interviewer. Hence, the identification of exceptional interviewers also depends on the group size (i.e., the number of respondents per interviewer), which is also termed sensitivity of interviewer residuals to group size (Pickery and Loosveldt 2004). However, in our case, the interviewer residual was not correlated to the number of completed test administrations, $r = .07$, $p = .30$. Thus, the amount of uncertainty on the interviewer level did not depend on the size of the clusters. In addition, as 3 out of the 200 interviewers only interviewed one respondent, we refitted all models excluding these three cases. The exclusion did not alter the results presented above. Finally, we tested the sensitivity of individual assessment scores to the presence of random interviewer effects, by using posterior means of estimated mathematical competence. Comparing these estimated values of individual assessments between

(1) the model with random interviewer effects and (2) the model ignoring the nested structure (hence, the 2PL model) resulted in a high correlation ($r = .97$, $p \leq .001$), an average mean deviation in individual competence scores of .00, and a root-mean-squared error of .22. In conclusion, interviewer differences do not cause distortions in the individual assessments of mathematical competence, although variance in the outcome is higher due to interviewer presence.

6.3 Impact of Influential Interviewers

The impact of the identified interviewers were examined by evaluating (1) the number of missing values in the administered mathematics test and (2) participation rates in a subsequent competence assessment about five to six years later. First, respondents tested by interviewers with significantly higher residual estimates ($M = 2.18$, SD = 2.23, $N = 223$) had significantly ($p < .05$) missing values on the competence test as compared to respondents tested by non-outlying interviewers [$M = 3.15$, SD = 3.67, $N = 4{,}406$; $t(286.72) = 6.15$, $p < .001$, $d = 0.27$]. In contrast, for respondents tested by interviewers with significantly lower residual estimates ($M = 3.42$, SD = 3.68, $N = 510$), no significant difference in the number of missing values was found [$t(631.99) = -1.57$, $p = .118$, $d = 0.07$]. Second, we compared the average participation rates for the subsequent competence assessment. For the respondent group tested by interviewers with non-outlying residual estimates, 37.20 percent did not participate at the next assessment as compared to 47.98 percent for the interviewers with significantly higher residual estimates and 40.78 percent for the interviewers with significantly lower residual estimates. These differences in response rates were significant at [$z(1) = 11.21$, $p < .001$], for the interviewers with significantly higher residual estimates, but not significant for the interviewers with significantly lower residual estimates [$z(1) = 3.00$, $p = .083$].

7. DISCUSSION

Interviewers play a decisive role in social surveys and educational large-scale assessments. Particularly, in household studies that visit respondents in their private homes, interviewers have a great responsibility and need to create standardized settings while administering questionnaires and achievement tests under comparable conditions. If specific interviewer behavior affects the responses of participants, the validity of the measured constructs might be called into question. Therefore, survey managers need to evaluate the interview process and identify interviewers with an undue impact on respondent behavior.

The present study examined interviewer effects on mathematical achievement in a German large-scale assessment. Our Bayesian estimation of higher-order random effects in adult mathematical achievement identified a considerable number of interviewers that exhibited pronounced effects on the competence measurement, while area effects were negligible. These interviewer effects can yield important consequences. For one, statistical analyses that ignore the multilevel structure, especially the clustering of respondents in different interviewers, might result in underestimation of standard errors and, consequently, in the overestimation of statistical significance of found effects (Durrant et al. 2010; Finch and Bolin 2017). As an alternative to multilevel modeling a Huber/White correction to obtain robust standard errors in statistical analysis is appropriate (Huber 1967; White 1982) if estimates of second-level standard errors are biased. More information on that procedure can be found in Goldstein (2011) and Raudenbush and Bryk (2002).

7.1 Implications for Large-Scale Assessments

What are the implications of the presented results? First, practitioners engaged in large-scale assessments need to minimize interviewer effects on competence measurements. Even though large efforts are already invested into interviewer training and standardization of test situations, our results stress the need for further improvements, with the goal of achieving comparable settings for all test takers. Interviewer abilities are decisive in obtaining answers from different respondents that can be aggregated and compared across respondents to derive generalizable conclusions about population effects. Nevertheless, considering the sensitivity of individual assessment scores in the presented study, the presence of interviewer variance does not lead to bias in individual assessments.

To reduce interviewer variance on population effects, educational measurement could be improved by switching to an institutionalized setting that tests all respondents in highly standardized test centers. Further studies are needed that compare adult competence measurements in both individual and institutional settings. This might give invaluable insight into interviewer effects introduced by different modes of administration. In comparison, large-scale educational data administered to students in a classroom setting found less than 1 percent of interviewer variance (Lüdtke et al. 2007).

Second, we presented a versatile methodological approach to empirically quantify interviewer effects on competence measurement. Bayesian analyses of cross-classified multilevel IRT models allowed us to disentangle interviewer from regional effects. Moreover, by investigating posterior draws of interviewer-level random effect structures, inferences about effects from specific interviewers on competence testing can be made. Our study found that respondents interviewed by interviewers with significantly higher residual estimates had, in comparison to the respondents interviewed by non-outlying interviewers, significantly lower missing values in the competence test, but also lower participation rates at the subsequent measurement occasion. For respondents who were interviewed by interviewers with significantly lower residual estimates, no significant differences were found. So far, the precise reasons why outlying interviewers exerted these effects are unclear. It might be the case that they (unintentionally) interfered with the competence assessment (e.g., gave unrequested assistance) that bothered respondents and refrained them from further participation. Survey managers can use this approach as a tool for intervention, by having regular updates of the posterior distributions during data collection. As the posterior distributions point to interviewers with a significant effect on the survey measures, these interviewers can be additionally trained. Furthermore, to minimize the relatedness between interviewer and area clusters, we recommend a sampling design where each interviewer works in more than one area and each area is visited by more than one interviewer.

7.2 Limitations and Directions for Future Research

As a limitation of our study, a considerable number of interviewers worked in only one sample area and unobservable confounding of interviewer and area effects exists. Hence, the dependencies cannot be fully distinguished by the measurement of separated random effect structures. Consequently, our results might be slightly distorted as compared to results obtained from a design, where interviewers are randomly distributed across areas. Nevertheless, this design limitation is common to national surveys. Moreover, a recent simulation study (Vassallo et al. 2017) found that three regions per interviewer are sufficient dispersion to obtain accurate estimates.

Interviewers were assigned to respondents based on spatial proximity of the living addresses, limiting the validity of our results. The multilevel cross-classified model assumes that the residual structures ($\sigma^2{}_u$, Σ_c, and Σ_g) are independent, but given the design of the study the interviewer-to-area distribution is not random. Hence, the assumption of independent residual structures (Σ_c and Σ_g) is violated by the design of interviewer-to-respondent allocation. Even though we assume a limited interpenetrated design as being sufficient to disentangle interviewer and area clusters as sources of variance, unobservable confounding remains. In a fully interpenetrated design, where interviewers are assigned randomly to respondents, differences in interviewer means would allow a causal interpretation. With such a design, differences in interviewer means would reflect true differences in interviewer behavior. Unfortunately, a random allocation of interviewers across areas implies high costs for nationwide studies.

Although the presented results highlighted the influence of interviewer behavior on competence measurements, more research is needed to identify potential predictors of non-ignorable interviewer effects. In our analyses, the heterogeneity of survey estimates across interviewers was not related to observed interviewer characteristics. Therefore, future research should examine additional background information on the interviewers and the test administration process to understand the origin of interviewer effects. This might help alleviate respective effects by adapting the study design or improving the recruitment and training of interviewers. Moreover, our approach of identifying influential interviewers could be refined. Finding some interviewers with larger random effects might be expected because of the assumption of multivariate normally distributed intercepts on the second level of our multilevel model (Finch and Bolin 2017). So far, it is unknown whether significantly outlying interviewers also adversely affect the validity of the competence estimates in large-scale educational assessments. Future research needs to develop measures that give further insights into the amount of deviation per interviewer; especially measures of severity for found outliers are needed.

8. CONCLUSION

The presented analyses reemphasize the conclusion of Schaeffer et al. (2010): interviewers are important in complex samples, helpful when critical response rates are expected, and especially useful in complex measurements. Therefore, we recommend intensified training and close monitoring for all tasks performed by the interviewers in the field, starting from respondent recruitment and persuasion for survey participation up to the standardized test administration. For this purpose, our Bayesian multilevel approach can be implemented to identify conspicuous interviewers during the ongoing data collection process.

Supplementary materials

Supplementary materials are available online at academic.oup.com/jssam. The online supplementary material contains summary statistics of selected variables by hierarchical level, a summary of area to interviewer distribution, a summary of interviewer to area distribution, a discussion on the sensitivity of variance components to prior choice, tables of random item effects for interviewer and area clusters, estimation results for the sample of interviewers having worked in at least two different regions (57% of the interviewers) as well as a figure on residuals of area clusters with corresponding posterior probability interval.

REFERENCES

Asparouhov, T., and B. Muthén (2016), "General Random Effect Latent Variable Modeling: Random Subjects, Items, Contexts, and Parameters," in *Advances in Multilevel Modeling for*

Educational Research: Addressing Practical Issues Found in Real-World Applications, eds. J. Harring, L. Stapleton and S. Beretvas, Chapter 6, pp. 163–129, Charlotte, NC: Information Age Publishing.

Beullens, K., and G. Loosveldt (2014), "Interviewer Effects on Latent Constructs in Survey Research," *Journal of Survey Statistics and Methodology*, 2, 433–458.

———. (2016), "Interviewer Effects in the European Social Survey," *Survey Research Methods*, 10, 103–118.

Blossfeld, H.-P., H.-G. Roßbach, and J. von Maurice (2011), "Education as a Lifelong Process – The German National Educational Panel Study (NEPS)," *Zeitschrift Für Erziehungswissenschaft*, 14, 19–34.

Brunton-Smith, I., P. Sturgis, and J. Williams (2012), "Is Success in Obtaining Contact and Cooperation Correlated with the Magnitude of Interviewer Variance?," *Public Opinion Quarterly*, 76, 265–286.

Brunton-Smith, I., P. Sturgis, and G. Leckie (2016), "Detecting and Understanding Interviewer Effects on Survey Data by Using a Cross-Classified Mixed Effects Location-Scale Model," *Journal of the Royal Statistical Society. Series A*, 180, 551–568.

Collins, M. (1980), "Interviewer Variability: A Review of the Problem," *Journal of the Market Research Society*, 22, 77–95.

Durrant, G. B., R. M. Groves, L. Staetsky, and F. Steele (2010), "Effects of Interviewer Attitudes and Behaviors on Refusal in Household Surveys," *Public Opinion Quarterly*, 74, 1–36.

Fellenberg, F., H. Sibberns, B. Jesske, and D. Hess (2016), "Quality Assurance in the Context of Data Collection", in *Methodological Issues of Longitudinal Surveys: The Example of the National Educational Panel Study*, eds. H.-P. Blossfeld, J. von Maurice, M. Bayer and J. Skopek, vol. 1, Chapter 5, pp. 579–593, Wiesbaden: Springer.

Finch, W. H., and J. E. Bolin (2017), *Multilevel Modeling Using Mplus*, Boca Raton: Champan & Hall/CRC.

Fox, J.-P. (2003), "Stochastic EM for Estimating the Parameters of a Multilevel IRT Model," *British Journal of Mathematical and Statistical Psychology*, 56, 65–81.

———. (2010), *Bayesian Item Response Modeling: Theory and Applications*, New York: Springer.

Fox, J.-P., and C. A. W. Glas (2001), "Bayesian Estimation of a Multilevel IRT Model Using Gibbs Sampling," *Psychometrika*, 66, 271–288.

Gabler, S., and P. Lahiri (2009), "On the Definition and Interpretation of Interviewer Variability for a Complex Sampling Design," *Survey Methodology*, 35, 85–99.

Gelman, A., and D. B. Rubin (1992), "Inference from Iterative Simulation Using Multiple Sequences," *Statistical Science*, 7, 457–472.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2004), *Bayesian Data Analysis* (2nd ed.), Boca Raton: Chapman & Hall.

Goldstein, H. (2011), *Multilevel Statistical Models* (4th ed.), Chichester: Wiley.

Goldstein, H., and W. Browne (2005), "Multilevel Factor Analysis Models for Continuous and Discrete Data," in *Contemporary Psychometrics*, eds. A. Maydeu-Olivares and J.J. McArdle, Chapter 14, pp. 453–475, New Jersey: Lawrence Erlbaum Assoc. Publishers.

Groves, R. M. (1989), *Survey Errors and Survey Costs*, New York: Wiley.

Hammon, A., S. Zinn, C. Aßmann, and A. Würbach (2016), "Samples, Weights, and Nonresponse: The Adult Cohort of the National Educational Panel Study (Wave 2 to 6)," NEPS Survey Paper No. 7, Bamberg, Germany: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Hansen, M. H., W. N. Hurwitz, and M. A. Bershad (1961), "Measurement Errors in Census and Surveys," *Bulletin of the International Statistical Institute*, 38, 351–374.

Hanushek, E., and L. Woessmann (2008), "The Role of Cognitive Skills in Economic Development," *Journal of Economic Literature*, 46, 607–668.

Hartig, J., E. Klieme, and D. Leutner (2008), *Assessment of Competencies in Educational Contexts*, Göttingen: Hogrefe Publishing.

Haunberger, S. (2010), "The Effects of Interviewer, Respondent and Area Characteristics on Cooperation in Panel Surveys: A Multilevel Approach," *Quality & Quantity*, 44, 957–969.

Heckman, J., J. Stixrud, and S. Urzua (2006), “The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior,” *Journal of Labor Economics*, 24, 411–482.

Hox, J. J. (1994), “Hierarchical Regression Models for Interviewer and Respondent Effects,” *Sociological Methods & Research*, 22, 300–318.

———. (2010), *Multilevel Analyses: Techniques and Applications* (2nd ed.), New Jersey: Lawrence Erlbaum Assoc. Publishers.

Hox, J. J., and E. D. de Leeuw (1994), “A Comparison of Nonresponse in Mail, Telephone, and Face-to-Face Surveys. Applying Multilevel Modelling to Meta-Analysis,” *Quality & Quantity*, 28, 329–344.

Hox, J. J., E. D. de Leeuw, and I. I. G. Kreft (1991), “The Effect of Interviewer and Respondent Characteristics on the Quality of Survey Data: A Multilevel Model”, in *Measurement Errors in Surveys*, eds. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz and S. Sudman, pp. 439–462, New York: Wiley.

Huber, P. J. (1967), “The Behavior of Maximum Likelihood Estimates under Nonstandard Conditions”, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pp. 221–233, Berkeley: University of California Press.

Jiao, H., A. Kamata, and C. Xie (2016), “A Multilevel Cross-Classified Testlet Model for Complex Item and Person Clustering in Item Response Modeling,” in *Advances in Multilevel Modeling for Educational Research: Addressing Practical Issues Found in Real-World Applications*, eds. J. Harring, L. Stapleton, and S. Beretvas, Chapter 5, pp. 139–162, Charlotte, NC: Information Age Publishing.

Jordan, A.-K., and C. Duchhardt (2013), “NEPS Technical Report for Mathematics—Scaling Results of Starting Cohort 6–Adults,” NEPS Working Paper No. 32, Bamberg: University of Bamberg, National Educational Panel Study.

Kamata, A., and B. K. Vaughn (2011), “Multilevel IRT Modeling,” in *Handbook of Advanced Multilevel Analysis*, eds. J. J. Hox and J. K. Roberts, Chapter 3, pp. 41–57, New York: Routledge.

Kaplan, D. (2014), *Bayesian Statistics for the Social Sciences*, New York: Guilford Press.

Kish, L. (1965), *Survey Sampling*, New York: Wiley.

Lee, E. S., R. N. Forthofe, and R. J. Lorimor (1989), *Analyzing Complex Survey Data*, Newbury Park: Sage.

Lee, S. Y., and X. Y. Song (2004), “Evaluation of the Bayesian and Maximum Likelihood Approaches in Analyzing Structural Equation Models with Small Sample Sizes,” *Multivariate Behavioral Research*, 39, 653–686.

Lüdtke, O., A. Robitzsch, U. Trautwein, F. Kreuter, and J.-M. Ihme (2007), “Are There Test Administrator Effects in Large-Scale Educational Assessments? Using Cross-Classified Multilevel Analysis to Probe for Effects on Mathematic Achievement and Sample Attrition,” *Methodology*, 3, 149–159.

Luo, W., and O. Kwok (2012), “The Consequences of Ignoring Individuals’ Mobility in Multilevel Growth Models: A Monte Carlo Study,” *Journal of Educational and Behavioral Statistics*, 37, 31–56.

Maas, C. J., and J. J. Hox (2004), “Robustness Issues in Multilevel Regression Analysis,” *Statistica Neerlandica*, 58, 127–137.

Mahalanobis, P. C. (1946), “Recent Experiments in Statistical Sampling in the Indian Statistical Institute,” *Journal of the Royal Statistical Society. Series A*, 109, 325–378.

Moss, P. A., B. J. Girard, and L. C. Haniford (2006), “Validity in Educational Assessment,” *Review of Research in Education*, 30, 109–162.

Muthén, B., and T. Asparouhov (2012), “Bayesian SEM: A More Flexible Representation of Substantive Theory,” *Psychological Methods*, 17, 313–335.

———. (2016), “Multi-Dimensional, Multi-Level, and Multi-Timepoint Item Response Modeling,” in *Handbook of Item Response Theory*, eds. W. J. van der Linden, vol. 1, Chapter 8, pp. 527–539, Boca Raton: CRC Press.

Muthén, L. K., and B. O. Muthén (1998–2017), *Mplus User’s Guide* (8th ed.), Los Angeles, CA: Muthén and Muthén.

Neumann, I., C. Duchhardt, M. Grüßing, A. Heinze, E. Knopp, and T. Ehmke (2013), "Modeling and Assessing Mathematical Competence over the Lifespan," *Journal for Educational Research Online*, 5, 80.

OECD (2004), *Learning for Tomorrow's World – First Results from PISA 2003*, available at https://www.oecd.org/education/school/programmeforinternationalstudentassessmentpisa/34002216.pdf (last accessed June 18, 2020).

O'Muircheartaigh, C., and P. Campanelli (1998), "The Relative Impact of Interviewer Effects and Sample Design Effects on Survey Precision," *Journal of the Royal Statistical Society. Series A*, 161, 63–77.

O'Muircheartaigh, C., and P. Campanelli (1999), "A Multilevel Exploration of the Role of Interviewers in Survey-Nonresponse," *Journal of the Royal Statistical Society. Series A*, 162, 437–446.

Pickery, J., and G. Loosveldt (2004), "A Simultaneous Analysis of Interviewer Effects on Various Data Quality Indicators with Identification of Exceptional Interviewers," *Journal of Official Statistics*, 20, 77–89.

Pohl, S., and C. H. Carstensen (2013), "Scaling of Competence Tests in the National Educational Panel Study – Many Questions, Some Answers, and Further Challenges," *Journal for Educational Research Online*, 5, 189–216.

Rasbash, J., and H. Goldstein (1994), "Efficient Analysis of Mixed Hierarchical and Cross-Classified Random Structures Using a Multilevel Model," *Journal of Educational and Behavioral Statistics*, 19, 337–350.

Rasch, G. (1980), *Probabilistic Models for Some Intelligence and Attainment Tests*, Chicago, IL: University of Chicago Press.

Raudenbush, S. W. (1993), "A Crossed Random Effects Model for Unbalanced Data with Applications in Cross-Sectional and Longitudinal Research," *Journal of Educational Statistics*, 18, 321–349.

Raudenbush, S. W., and A. S. Bryk (2002), *Hierarchical Linear Models* (2nd ed.), Thousand Oaks, CA: Sage.

Rosenthal, R. (1967), "Covert Communication in the Psychological Experiment," *Psychological Bulletin*, 67, 356–367.

———. (1994), "Interpersonal Expectancy Effects: A 30-Year Perspective," *Current Directions in Psychological Science*, 3, 176–179.

———. (2002), "Covert Communication in Classrooms, Clinics, Courtrooms, and Cubicles," *American Psychologist*, 57, 839–849.

Schaeffer, N. C., J. Dykema, and D. W. Maynard (2010), "Interviewers and Interviewing," in *Handbook of Survey Research*, eds. P. V. Marsden and J. D. Wright, vol. 2, Chapter 4, pp. 437–479, Bingley, UK: Emerald.

Schnell, R., and F. Kreuter (2002), "Separating Interviewer and Sampling-point Effects," in *American Statistical Association Proceedings of the Section on Survey Research Methods*, pp. 3132–3133.

———. (2005), "Separating Interviewer and Sampling Point Effects," *Journal of Official Statistics*, 21, 389–410.

Shelton, J. T., E. M. Elliott, S. D. Eaves, and A. L. Exner (2009), "The Distracting Effects of a Ringing Cell Phone: An Investigation of the Laboratory and the Classroom Setting," *Journal of Environmental Psychology*, 29, 513–521.

Skrondal, A., and S. Rabe-Hesketh (2004), *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models*, Boca Raton, FL: Chapman & Hall/CRC.

———. (2009), "Prediction in Multilevel Generalized Linear Models," *Journal of the Royal Statistical Society. Series A*, 172, 659–687.

Tourangeau, R., and T. Yan (2007), "Sensitive Questions in Surveys," *Psychological Bulletin*, 133, 859–883.

Turner, M., P. Sturgis, D. Martin, and C. Skinner (2014), "Can Interviewer Personality, Attitudes and Experience Explain the Design Effect in Face-to-Face Surveys?," in *Improving Survey Methods: Lessons from Recent Research*, eds. U. Engel, B. Jann, P. Lynn, A. Scherpenzeel and P. Sturgis, Chapter 7, pp. 72–85, Abingdon: Routledge.

Van den Noortgate, W., P. De Boeck, and M. Meulders (2003), "Cross-Classification Multilevel Logistic Models in Psychometrics," *Journal of Educational and Behavioral Statistics*, 28, 369–386.

Vassallo, R., G. B. Durrant, and P. W. Smith (2017), "Separating Interviewer and Area Effects by Using a Cross-Classified Multilevel Logistic Model: Simulation Findings and Implications for Survey Design," *Journal of the Royal Statistical Society. Series A*, 180, 531–550.

Vassallo, R., G. B. Durrant, P. W. F. Smith, and H. Goldstein (2015), "Interviewer Effects on Non-Response Propensity in Longitudinal Surveys: A Multilevel Modelling Approach," *Journal of the Royal Statistical Society. Series A*, 178, 83–99.

Wang, S., H. Jiao, M. J. Young, T. Brooks, and J. Olson (2007), "A Meta-Analysis of Testing Mode Effects in Grade K-12 Mathematics Tests," *Educational and Psychological Measurement*, 67, 219–238.

West, B. T., F. Kreuter, and U. Jaenichen (2013), "Interviewer Effects in Face-to-Face Surveys: A Function of Sampling, Measurement Error or Nonresponse?," *Journal of Official Statistics*, 29, 277–297.

West, B. T., and K. Olson (2010), "How Much of Interviewer Variance Is Really Nonresponse Error Variance?," *Public Opinion Quarterly*, 74, 1004–1026.

White, H. (1982), "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1–25.

Disentangling Interviewer and Area Effects in Large Scale Educational Assessments using Cross-Classified Multilevel Item Response Models

Online-only supplementary material

List of Tables

List of Figures

Table S1. Summary statistics of selected variables by hierarchical level

Variable	***M*/ %**	***SD***	**Min.**	**Max.**	**Information on Recoding**	**Name in original Dataset**
Respondent Level (N = 5,139)						
Age	51.41	10.96	25	72	Grand-mean centred	tx29000
Gender (female)	0.51	-	0	1		t700001
Migration Background (yes)	0.17	-	0	1		t400500
Highest CASMIN					Recoded into 3 binary variables, reference category is 'secondary education'	tx28101
no degree or lower secondary education	0.19	-				
secondary education	0.33	-				
matriculation standard	0.19	-				
graduate degree	0.29	-				
Employment Status (unemployed)	0.20	-	0	1		tx29060
Cultural capital (Number of books)	4.11	1.33	1 (0 to 10 books)	6 (more than 500)	Grand-mean centred	t34005a
Political Area Size	4.17	1.78	1 (below 2000 inhabitants)	7 (more than 500k inhabitants)	Grand-mean centred	tx80103
Interviewer Level (N = 200)						
Gender (female)	0.40	-	0	1		tx80301
Age					Recoded into 2 binary variables, ref. categ. is ‚below 50 years'	tx80302
below 50 years	0.22	-				
50 to 65 years	0.58	-				
older than 65 year	0.20	-				
Educational Attainment					Recoded into 2 binary variables, ref. categ. is ‚lower secondary degree'	tx80303
lower secondary degree	0.14	-				
secondary education	0.31	-				
matriculation standard	0.55	-				
Work experience as interviewer					Recoded into 3 binary variables, ref. categ. is ‚up to 2 years'	tx80304
up to 2 years	0.15	-				
2 to 3 years	0.31	-				
4 to 5 years	0.25	-				
more than 5 years	0.29	-				

Table S2. Summary of area to interviewer distribution

Number of test administrations per interviewer	Number of areas (regional clusters/ strata) per interviewer								
	1	2	3	4	5	6	7	8	Sum
1	3								3
2	1								1
3	2								2
4	5			1					6
5	7								7
6	1	1							2
7	5	3							8
8	7	1	1						9
9	7	2							9
10	1		1						2
11	3	1							4
12	5	1							6
13	6		1						7
14	3	1	2						6
15	4	2							6
16	3								3
17	2	2			1				5
18	1	3	1						5
19	3	1							4
20	2		2						4
21	2	3							5
22	2	2							4
23	2	3	2	2					9
24	2	8	1						11
25	1	1	1						3
26	2	2			1				5
27			2						2
28	1	1		2					4
29		2							2
30			2						2
31		2	1						3
33	1			1					2
34		2	1	1					4
35		1	1						2
36				2					2
37		1							1
38			1						1
40			1						1
42				1		1			2
43				1					1
44			1						1
45	1								1
46				1	2			1	4
48				2					2
49			1						1
50			1						1
52				1					1
53				1					1
55							1		1
56			1					1	2
57				1					1
58				1					1
60	1			1					2
63				1					1
67				1					1
68			1		1				2
70					1				1
71					1				1
73					1				1
76						1			1
77					1		1		2
78			1						1
80				1					1
84							1		1
91					1				1
95					1				1
123						1			1
Sum	86	46	27	22	11	3	3	2	200

Table S3. Summary of interviewer to area distribution

Number of interviewer per area (strata)

Number of Interviews per Area

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	17	26	28	Sum
1	1																		1
7	1																		1
10	2																		2
12		1																	1
13	1																		1
14	2	1																	3
15	3	1	2																6
17	2	2	1																5
18		1																	1
19		1																	1
20	2	1																	3
21	1																		1
22	1		1																2
24					1														1
25		2	1																3
26	2	1																	3
28			1																1
29		1	1	1															3
30					1														1
31						1													1
33		1																	1
34				1															1
35				1															1
36			1																1
37		1																	1
38			1																1
39			2	1															3
40			1																1
42		1																	1
43		1		1															2
44				1	1														2
45			1																1
46		1	1																2
49			1																1
52				1	1														2
53				1															1
60				1															1
61							1												1
67						1													1
69						1													1
70					1														1
73				1		1													2
74							1												1
75						1													1
83				1															1
84												1							1
86									1										1
87									1										1
95							1												1
97														1					1
98								1											1
99										1									1
110												1							1
113									1										1
119											1								1
122									1										1
126											1								1
155													1						1
162														1					1
174															1				1
195																1			1
199													1						1
239																		1	1
360																	1		1
Sum	18	17	15	11	5	5	3	1	4	1	2	2	2	2	1	1	1	1	92

Table S4. Summary of interviewer to area distribution (German federal states)

Number of test administrations per interviewer	Number of visited German federal states per interviewer						
	1	2	3	4	5	6	Sum
1	3						3
2	1						1
3	2						2
4	5	1					6
5	7						7
6	1	1					2
7	7	1					8
8	9						9
9	8	1					9
10	1	1					2
11	3	1					4
12	5	1					6
13	6	1					7
14	4	2					6
15	6						6
16	3						3
17	4			1			5
18	2	2	1				5
19	4						4
20	2	1	1				4
21	5						5
22	3	1					4
23	5	2	2				9
24	10	1					11
25	1	1	1				3
26	2	3					5
27	1	1					2
28	2		2				4
29	1	1					2
30	2						2
31	2		1				3
33	1	1					2
34	2	1	1				4
35	1		1				2
36		1	1				2
37	1						1
38	1						1
40	1						1
42			2				2
43		1					1
44		1					1
45	1						1
46	1	1	1			1	4
48		2					2
49	1						1
50		1					1
52		1					1
53	1						1
55				1			1
56			2				2
57		1					1
58		1					1
60	1	1					2
63	1						1
67	1						1
68		2					2
70			1				1
71	1						1
73				1			1
76					1		1
77			2				2
78	1						1
80		1					1
84				1			1
91		1					1
95		1					1
123				1			1
Sum	133	41	19	5	1	1	200

As an example, for the highlighted rows in Tables S2 to S4 it is demonstrated how to read the presented information.

S2: It occurred two times, that 42 interviews were conducted per interviewer, one of these two interviewers worked in four areas and the other interviewer worked in six areas.

S3: It occurred two times that 46 interviews were realized per area and that in each of these two areas, the interviews were conducted in one of these regions by 2 different interviewers and in the other region by 3 different interviewers.

S4: Two interviewers visited 3 German Federal States and each interviewed 42 respondents. Furthermore, more than one-third of the interviewers worked in more than one German federal state (*n*= 75, 37.5 percent). Each interviewer worked on average in 1.52 German federal states (min = 1, max = 6, *SD* = 0.87).

Sensitivity of variance components to prior choice

Variance components in hierarchical models can be sensitive to the choice of priors (Gustafson, Hossain and MacNab 2006). We conducted sensitivity analyses for priors of the estimated latent factor variances of the first model. Setting an inverse gamma prior of *IG* (.001, .001) for the interviewer latent factor variance estimate, or for all three estimates of latent factor variances, did not change the results substantially compared to model 1 (see results in last column of the subsequent table S5). Using an inverse gamma prior of *IG* (1,1) led to an increase in estimated latent factor variances, compared to the variances obtained using the Mplus default setting of *IG* (-1,0). Surprisingly, using the *IG* (1,1) specification for all three latent factor variances increased the area variance to nearly the same amount as interviewer variance. Hence, the random area variance might be sensitive to the choice of the prior to some degree, whereas the random interviewer variance is rather robust. However, an inverse gamma prior specification is not recommended for near-zero variance parameters in hierarchical models (Gelman 2006). As the area variance parameter is close to zero in the default prior setting, this high value might show misspecification by using the alternative prior *IG* (1,1).

References:

Gelman, A. (2006) Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.*, 1, 515-534. doi:10.1214/06-BA117A. (Available online: http://www.stat.columbia.edu/~gelman/research/published/taumain.pdf)

Gustafson, P., Hossain, S., and MacNab, Y. (2006). Conservative Prior Distributions for Variance Parameters in Hierarchical Models. *The Canadian Journal of Statistics / La Revue Canadienne De Statistique*, 34(3), 377-390.

Table S5. Sensitivity of latent factor variance estimates to choice of prior

Parameter	**Interviewer latent factor variance only**		**All latent factor variances**		**Latent factor variances of Model 1**
	IG (1,1)	*IG* (.001, .001)	*IG* (1,1)	*IG* (.001, .001)	*IG* (-1, 0)
Interviewer Variance	0.056	0.029	0.057	0.029	0.030*
Area Variance	0.003*	0.004*	0.054	0.003	0.004*
Within Variance (first Level)	0.465*	0.422*	0.508	0.418	0.425*

*Parameters were estimated with an inverse gamma prior specification of *IG*(-1, 0).

Table S6. Random item effects for interviewer clusters (estimated with Mplus, Version 8)

Parameter	**M**	**SD**	**Item variance across interviewers**	**95% PPI**
Discrimination				
Item 1	0.848	0.041	0.041	(0.005, 0.104)
Item 2	0.994	0.048	0.021	(0.002, 0.086)
Item 3	0.835	0.045	0.067	(0.030, 0.125)
Item 4	0.784	0.044	0.020	(0.002, 0.068)
Item 5	1.206	0.054	0.059	(0.007, 0.154)
Item 6	1.292	0.061	0.070	(0.007, 0.179)
Item 7	0.869	0.047	0.100	(0.049, 0.177)
Item 8	0.846	0.048	0.059	(0.012, 0.131)
Item 9	1.053	0.049	0.038	(0.004, 0.113)
Item 10	0.622	0.041	0.079	(0.041, 0.138)
Item 11	1.159	0.048	0.030	(0.002, 0.101)
Item 12	0.944	0.061	0.102	(0.026, 0.213)
Item 13	1.325	0.060	0.093	(0.021, 0.202)
Item 14	0.800	0.043	0.045	(0.006, 0.107)
Item 15	0.594	0.049	0.149	(0.077, 0.255)
Item 16	1.329	0.063	0.108	(0.023, 0.236)
Item 17	0.670	0.036	0.036	(0.005, 0.091)
Item 18	1.021	0.056	0.091	(0.022, 0.196)
Item 19	1.063	0.050	0.036	(0.003, 0.115)
Item 20	1.167	0.086	0.019	(0.002, 0.097)
Item 21	1.459	0.075	0.182	(0.083, 0.335)
Threshold				
Item 1	-0.258	0.029	0.011	(0.002, 0.031)
Item 2	-1.049	0.038	0.018	(0.002, 0.053)
Item 3	0.939	0.038	0.058	(0.027, 0.102)
Item 4	-1.213	0.034	0.005	(0.001, 0.021)
Item 5	-0.484	0.036	0.006	(0.001, 0.022)
Item 6	0.535	0.042	0.042	(0.011, 0.086)
Item 7	-0.126	0.032	0.029	(0.008, 0.060)
Item 8	-1.117	0.034	0.006	(0.001, 0.026)
Item 9	-0.080	0.036	0.033	(0.010, 0.065)
Item 10	0.361	0.030	0.033	(0.011, 0.066)
Item 11	-0.185	0.038	0.047	(0.021, 0.084)
Item 12	0.912	0.046	0.035	(0.004, 0.089)
Item 13	-0.003	0.037	0.012	(0.002, 0.037)
Item 14	-0.638	0.032	0.029	(0.009, 0.057)
Item 15	0.834	0.033	0.032	(0.007, 0.071)
Item 16	-0.032	0.039	0.023	(0.003, 0.057)
Item 17	-0.008	0.030	0.032	(0.011, 0.061)
Item 18	-0.632	0.036	0.023	(0.004, 0.054)
Item 19	-0.466	0.038	0.039	(0.010, 0.084)
Item 20	-2.033	0.089	0.146	(0.056, 0.283)
Item 21	0.797	0.047	0.021	(0.002, 0.063)
Latent Trait Variance				
Observations	0.635	-		
Interviewer	0.068	0.012		

Note. *M* = posterior mean. *SD* = posterior standard deviation. Item variance across interviewers = the item-specific random effect variance across interviewers. PPI = posterior probability interval (2.5th and 97.5th percentile of the posterior distribution).

Table S7. Random item effects for interviewer clusters (estimated with R-package SIRT)

Parameter	**M**	**SD**	**Item variance across interviewers**	**95% PPI**
Discrimination				
Item 1	0.860	0.036	0.091	(0.057, 0.137)
Item 2	1.010	0.045	0.085	(0.053, 0.133)
Item 3	0.832	0.039	0.098	(0.064, 0.144)
Item 4	0.791	0.042	0.079	(0.050, 0.121)
Item 5	1.223	0.047	0.112	(0.068, 0.176)
Item 6	1.309	0.052	0.117	(0.068, 0.188)
Item 7	0.871	0.037	0.109	(0.072, 0.162)
Item 8	0.853	0.043	0.097	(0.060, 0.149)
Item 9	1.074	0.045	0.099	(0.060, 0.160)
Item 10	0.626	0.035	0.098	(0.063, 0.144)
Item 11	1.184	0.043	0.094	(0.059, 0.148)
Item 12	0.946	0.052	0.135	(0.081, 0.215)
Item 13	1.333	0.049	0.127	(0.077, 0.204)
Item 14	0.811	0.037	0.085	(0.053, 0.126)
Item 15	0.594	0.036	0.159	(0.100, 0.244)
Item 16	1.342	0.052	0.151	(0.086, 0.244)
Item 17	0.677	0.033	0.079	(0.051, 0.118)
Item 18	1.028	0.046	0.126	(0.076, 0.199)
Item 19	1.083	0.047	0.088	(0.055, 0.135)
Item 20	1.099	0.076	0.100	(0.057, 0.162)
Item 21	1.452	0.064	0.189	(0.109, 0.297)
Threshold				
Item 1	-0.301	0.022	0.048	(0.033, 0.067)
Item 2	-1.123	0.030	0.063	(0.042, 0.091)
Item 3	0.904	0.028	0.083	(0.058, 0.117)
Item 4	-1.274	0.029	0.048	(0.032, 0.071)
Item 5	-0.554	0.026	0.047	(0.032, 0.067)
Item 6	0.468	0.028	0.075	(0.050, 0.109)
Item 7	-0.174	0.023	0.062	(0.042, 0.089)
Item 8	-1.186	0.030	0.051	(0.034, 0.072)
Item 9	-0.134	0.025	0.064	(0.045, 0.088)
Item 10	0.331	0.023	0.064	(0.044, 0.089)
Item 11	-0.250	0.024	0.070	(0.048, 0.099)
Item 12	0.870	0.039	0.081	(0.052, 0.122)
Item 13	-0.078	0.024	0.054	(0.037, 0.077)
Item 14	-0.693	0.026	0.059	(0.042, 0.082)
Item 15	0.813	0.028	0.069	(0.046, 0.099)
Item 16	-0.102	0.025	0.062	(0.042, 0.089)
Item 17	-0.040	0.022	0.061	(0.042, 0.085)
Item 18	-0.694	0.028	0.062	(0.041, 0.089)
Item 19	-0.529	0.027	0.072	(0.048, 0.104)
Item 20	-2.078	0.077	0.162	(0.091, 0.260)
Item 21	0.729	0.034	0.070	(0.045, 0.102)
Latent Trait Variance				
Observations	0.635	0.011		
Interviewer	0.062	0.021		

Note. *M* = posterior mean. *SD* = posterior standard deviation. Item variance across interviewers = the item-specific random effect variance across interviewers. PPI = posterior probability interval (2.5th and 97.5th percentile of the posterior distribution).

Table S8. Absolute differences between values of Table S6 and Table S7

Parameter	M	SD	Item variance across interviewers	95% PPI
Discrimination				
Item 1	0.012	0.005	0.050	(0.052, 0.033)
Item 2	0.016	0.003	0.064	(0.051, 0.047)
Item 3	0.003	0.006	0.031	(0.034, 0.019)
Item 4	0.007	0.002	0.059	(0.048, 0.053)
Item 5	0.017	0.007	0.053	(0.061, 0.022)
Item 6	0.017	0.009	0.047	(0.061, 0.009)
Item 7	0.002	0.010	0.009	(0.023, 0.015)
Item 8	0.007	0.005	0.038	(0.048, 0.018)
Item 9	0.021	0.004	0.061	(0.056, 0.047)
Item 10	0.004	0.006	0.019	(0.022, 0.006)
Item 11	0.025	0.005	0.064	(0.057, 0.047)
Item 12	0.002	0.009	0.033	(0.055, 0.002)
Item 13	0.008	0.011	0.034	(0.056, 0.002)
Item 14	0.011	0.006	0.040	(0.047, 0.019)
Item 15	0.000	0.013	0.010	(0.023, 0.011)
Item 16	0.013	0.011	0.043	(0.063, 0.008)
Item 17	0.007	0.003	0.043	(0.046, 0.027)
Item 18	0.007	0.010	0.035	(0.054, 0.003)
Item 19	0.020	0.003	0.052	(0.052, 0.020)
Item 20	0.068	0.010	0.081	(0.055, 0.065)
Item 21	0.007	0.011	0.007	(0.026, 0.038)
Threshold				
Item 1	0.043	0.007	0.037	(0.031, 0.036)
Item 2	0.074	0.008	0.045	(0.040, 0.038)
Item 3	0.035	0.010	0.025	(0.031, 0.015)
Item 4	0.061	0.005	0.043	(0.031, 0.050)
Item 5	0.070	0.010	0.041	(0.031, 0.045)
Item 6	0.067	0.014	0.033	(0.039, 0.023)
Item 7	0.048	0.009	0.033	(0.034, 0.029)
Item 8	0.069	0.004	0.045	(0.033, 0.046)
Item 9	0.054	0.011	0.031	(0.035, 0.023)
Item 10	0.030	0.007	0.031	(0.033, 0.023)
Item 11	0.065	0.014	0.023	(0.027, 0.015)
Item 12	0.042	0.007	0.046	(0.048, 0.033)
Item 13	0.075	0.013	0.042	(0.035, 0.040)
Item 14	0.055	0.006	0.030	(0.033, 0.025)
Item 15	0.021	0.005	0.037	(0.039, 0.028)
Item 16	0.070	0.014	0.039	(0.039, 0.032)
Item 17	0.032	0.008	0.029	(0.031, 0.024)
Item 18	0.062	0.008	0.039	(0.037, 0.035)
Item 19	0.063	0.011	0.033	(0.038, 0.020)
Item 20	0.045	0.012	0.016	(0.035, 0.023)
Item 21	0.068	0.013	0.049	(0.043, 0.039)
Latent Trait Variance				
Observations	-	-		
Interviewer	0.006	0.009		

Table S9. Random item effects for area clusters (estimated with Mplus, Version 8)

Parameter	**M**	**SD**	**Item variance across areas**	**95% PPI**
Discrimination				
Item 1	0.865	0.043	0.025	(0.003, 0.082)
Item 2	1.018	0.051	0.017	(0.002, 0.068)
Item 3	0.833	0.048	0.045	(0.011, 0.108)
Item 4	0.809	0.047	0.019	(0.002, 0.072)
Item 5	1.212	0.054	0.039	(0.004, 0.118)
Item 6	1.303	0.059	0.033	(0.003, 0.119)
Item 7	0.828	0.051	0.074	(0.029, 0.150)
Item 8	0.850	0.048	0.025	(0.002, 0.082)
Item 9	1.070	0.049	0.019	(0.002, 0.077)
Item 10	0.593	0.048	0.074	(0.030, 0.144)
Item 11	1.170	0.049	0.024	(0.002, 0.093)
Item 12	0.938	0.059	0.023	(0.002, 0.098)
Item 13	1.366	0.068	0.087	(0.020, 0.207)
Item 14	0.789	0.043	0.025	(0.003, 0.074)
Item 15	0.512	0.053	0.101	(0.045, 0.196)
Item 16	1.329	0.058	0.031	(0.003, 0.108)
Item 17	0.682	0.040	0.030	(0.004, 0.081)
Item 18	1.025	0.056	0.047	(0.006, 0.128)
Item 19	1.063	0.055	0.039	(0.003, 0.117)
Item 20	1.130	0.090	0.062	(0.005, 0.202)
Item 21	1.475	0.094	0.267	(0.128, 0.515)
Threshold				
Item 1	-0.254	0.031	0.005	(0.001, 0.020)
Item 2	-1.036	0.039	0.011	(0.001, 0.038)
Item 3	0.904	0.043	0.063	(0.027, 0.118)
Item 4	-1.199	0.035	0.007	(0.001, 0.030)
Item 5	-0.469	0.040	0.017	(0.003, 0.046)
Item 6	0.577	0.048	0.036	(0.006, 0.091)
Item 7	-0.082	0.032	0.011	(0.001, 0.034)
Item 8	-1.097	0.037	0.012	(0.002, 0.037)
Item 9	-0.094	0.039	0.016	(0.002, 0.046)
Item 10	0.387	0.032	0.022	(0.005, 0.054)
Item 11	-0.145	0.040	0.019	(0.004, 0.046)
Item 12	0.912	0.048	0.025	(0.003, 0.072)
Item 13	0.009	0.040	0.007	(0.001, 0.027)
Item 14	-0.627	0.034	0.016	(0.003, 0.043)
Item 15	0.847	0.034	0.016	(0.002, 0.048)
Item 16	-0.033	0.042	0.013	(0.002, 0.041)
Item 17	-0.002	0.030	0.009	(0.001, 0.030)
Item 18	-0.623	0.041	0.025	(0.006, 0.058)
Item 19	-0.451	0.038	0.014	(0.002, 0.047)
Item 20	-1.956	0.081	0.053	(0.008, 0.146)
Item 21	0.812	0.053	0.023	(0.002, 0.080)
Latent Trait Variance				
Observations	0.632	-		
Area	0.035	0.010		

Note. *M* = posterior mean. *SD* = posterior standard deviation. Item variance across areas = the item-specific random effect variance across areas. PPI = posterior probability interval (2.5th and 97.5th percentile of the posterior distribution).

Table S10. Random item effects for area clusters (estimated with R-package SIRT)

Parameter	M	SD	Item variance across areas	95% PPI
Discrimination				
Item 1	0.875	0.039	0.085	(0.052, 0.132)
Item 2	1.042	0.048	0.084	(0.052, 0.133)
Item 3	0.833	0.041	0.094	(0.057, 0.144)
Item 4	0.823	0.045	0.086	(0.052, 0.141)
Item 5	1.210	0.048	0.099	(0.059, 0.158)
Item 6	1.320	0.052	0.101	(0.059, 0.171)
Item 7	0.836	0.039	0.102	(0.064, 0.154)
Item 8	0.852	0.044	0.081	(0.051, 0.128)
Item 9	1.096	0.046	0.090	(0.054, 0.145)
Item 10	0.599	0.036	0.108	(0.067, 0.165)
Item 11	1.177	0.045	0.091	(0.056, 0.147)
Item 12	0.972	0.055	0.107	(0.062, 0.172)
Item 13	1.375	0.051	0.123	(0.072, 0.203)
Item 14	0.787	0.037	0.079	(0.049, 0.125)
Item 15	0.508	0.038	0.133	(0.080, 0.202)
Item 16	1.355	0.053	0.102	(0.061, 0.165)
Item 17	0.678	0.034	0.080	(0.051, 0.127)
Item 18	1.041	0.047	0.106	(0.063, 0.171)
Item 19	1.085	0.047	0.094	(0.058, 0.151)
Item 20	1.083	0.075	0.132	(0.072, 0.231)
Item 21	1.452	0.064	0.257	(0.147, 0.410)
Threshold				
Item 1	-0.293	0.023	0.049	(0.033, 0.070)
Item 2	-1.104	0.032	0.061	(0.040, 0.089)
Item 3	0.865	0.027	0.093	(0.061, 0.141)
Item 4	-1.250	0.032	0.059	(0.039, 0.087)
Item 5	-0.537	0.027	0.063	(0.041, 0.092)
Item 6	0.525	0.030	0.082	(0.052, 0.127)
Item 7	-0.115	0.024	0.055	(0.036, 0.080)
Item 8	-1.158	0.031	0.057	(0.038, 0.084)
Item 9	-0.171	0.026	0.062	(0.041, 0.093)
Item 10	0.359	0.024	0.062	(0.040, 0.089)
Item 11	-0.198	0.025	0.060	(0.040, 0.086)
Item 12	0.881	0.039	0.080	(0.050, 0.125)
Item 13	-0.069	0.025	0.054	(0.036, 0.079)
Item 14	-0.681	0.025	0.060	(0.040, 0.086)
Item 15	0.841	0.029	0.066	(0.043, 0.098)
Item 16	-0.118	0.026	0.060	(0.040, 0.089)
Item 17	-0.022	0.024	0.057	(0.037, 0.081)
Item 18	-0.691	0.028	0.065	(0.042, 0.097)
Item 19	-0.517	0.029	0.066	(0.043, 0.097)
Item 20	-2.034	0.068	0.109	(0.064, 0.178)
Item 21	0.734	0.035	0.087	(0.054, 0.133)
Latent Trait Variance				
Observations	0.632	0.011		
Area	0.031	0.023		

Note. *M* = posterior mean. *SD* = posterior standard deviation. Item variance across areas = the item-specific random effect variance across areas. PPI = posterior probability interval (2.5th and 97.5th percentile of the posterior distribution).

Table S11. Absolute differences between values of Table S9 and Table S10

Parameter	M	SD	Item variance across areas	95% PPI
Discrimination				
Item 1	0.010	0.004	0.060	(0.049, 0.050)
Item 2	0.024	0.003	0.067	(0.050, 0.065)
Item 3	0.000	0.007	0.049	(0.046, 0.036)
Item 4	0.014	0.002	0.067	(0.050, 0.069)
Item 5	0.002	0.006	0.060	(0.055, 0.040)
Item 6	0.017	0.007	0.068	(0.056, 0.052)
Item 7	0.008	0.012	0.028	(0.035, 0.004)
Item 8	0.002	0.004	0.056	(0.049, 0.046)
Item 9	0.026	0.003	0.071	(0.052, 0.068)
Item 10	0.006	0.012	0.034	(0.037, 0.021)
Item 11	0.007	0.004	0.067	(0.054, 0.054)
Item 12	0.034	0.004	0.084	(0.060, 0.074)
Item 13	0.009	0.017	0.036	(0.052, 0.004)
Item 14	0.002	0.006	0.054	(0.046, 0.051)
Item 15	0.004	0.015	0.032	(0.035, 0.006)
Item 16	0.026	0.005	0.071	(0.058, 0.057)
Item 17	0.004	0.006	0.050	(0.047, 0.046)
Item 18	0.016	0.009	0.059	(0.057, 0.043)
Item 19	0.022	0.008	0.055	(0.055, 0.034)
Item 20	0.047	0.015	0.070	(0.067, 0.029)
Item 21	0.023	0.030	0.010	(0.019, 0.105)
Threshold				
Item 1	0.039	0.008	0.044	(0.032, 0.050)
Item 2	0.068	0.007	0.050	(0.039, 0.051)
Item 3	0.039	0.016	0.030	(0.034, 0.023)
Item 4	0.051	0.003	0.052	(0.038, 0.057)
Item 5	0.068	0.013	0.046	(0.038, 0.046)
Item 6	0.052	0.018	0.046	(0.046, 0.036)
Item 7	0.033	0.008	0.044	(0.035, 0.046)
Item 8	0.061	0.006	0.045	(0.036, 0.047)
Item 9	0.077	0.013	0.046	(0.039, 0.047)
Item 10	0.028	0.008	0.040	(0.035, 0.035)
Item 11	0.053	0.015	0.041	(0.036, 0.040)
Item 12	0.031	0.009	0.055	(0.047, 0.053)
Item 13	0.078	0.015	0.047	(0.035, 0.052)
Item 14	0.054	0.009	0.044	(0.037, 0.043)
Item 15	0.006	0.005	0.050	(0.041, 0.050)
Item 16	0.085	0.016	0.047	(0.038, 0.048)
Item 17	0.020	0.006	0.048	(0.036, 0.051)
Item 18	0.068	0.013	0.040	(0.036, 0.039)
Item 19	0.066	0.009	0.052	(0.041, 0.050)
Item 20	0.078	0.013	0.056	(0.056, 0.032)
Item 21	0.078	0.018	0.064	(0.052, 0.053)
Latent Trait Variance				
Observations	-	-		
Area	0.004	0.013		

Interpretation of random item effects for interviewer and area clusters

Two separate hierarchical item response models with random discrimination and threshold effects were estimated. Furthermore, the results obtained from using the software Mplus (Version 8) were confirmed by using the function *mcmc.2pno.ml* from the R-Package sirt (Robitzsch 2019). We obtained comparable results from both programs for discrimination and threshold parameters as well as their item variances across interviewers or across areas (Table S8 and Table S11). The average deviation across interviewers for item variances between both programs was *M* = 0.042, *SD* = 0.020 (*Min* = 0.007, *Max* = 0.081) for the discrimination parameter and for the threshold parameter item variances the difference was *M* = 0.036, *SD* = 0.008 (*Min* = 0.016, *Max* = 0.049). Across areas, the average deviation for item variances between programs was *M*= 0.055, *SD* = 0.018 (*Min* = 0.010, *Max* = 0.084) for the discrimination parameter and for the threshold parameter item variances, the average difference was *M* = 0.047, *SD* = 0.007 (*Min* = 0.030, *Max* = 0.064).

The results for item variance across interviewer clusters (Table S6 and Table S7) and for item variance across area clusters (Table S9 and Table S10) are presented. Item discrimination and difficulty (threshold) parameters are depicted as well as the uncertainty which is given by the posterior standard deviation. Furthermore, random item effects at the interviewer level (Table S6 and Table S7) and area level (Table S9 and Table S10) with respective standard deviations are presented. In addition, 95% posterior probability intervals are given to evaluate significant deviations of the discrimination and threshold parameters. All estimated random effects at the interviewer level significantly deviate from zero when examining the 95% posterior probability interval. Nevertheless, it must be considered that variance estimates cannot become negative and in effect the probability interval will never include zero.

We assume that no strong violation of the measurement invariance is present. The share of variance in the latent trait across interviewers was 9.7 percent using Mplus (see last two rows of Table S6) and 8.9 percent using the sirt-package for estimation (see last two rows of Table S7). The average variances of item parameters among interviewers across all items was 0.051 (average of item variances in Table S6; average of discrimination item variances was 0.069; average of threshold item variances was 0.032). The share of variance in the latent trait across areas was 5.2 percent using Mplus (see the last two row of Table S9) and 4.7 percent using the sirt-package for estimation (see the last two rows of Table S10). The average variance of item parameters was 0.036 (average of item variances in Table S9; average of discrimination item variances was 0.053; average of threshold item variances was 0.020). Hence, we assume that mathematic competence was measured as a unidimensional construct among interviewers and areas.

References:

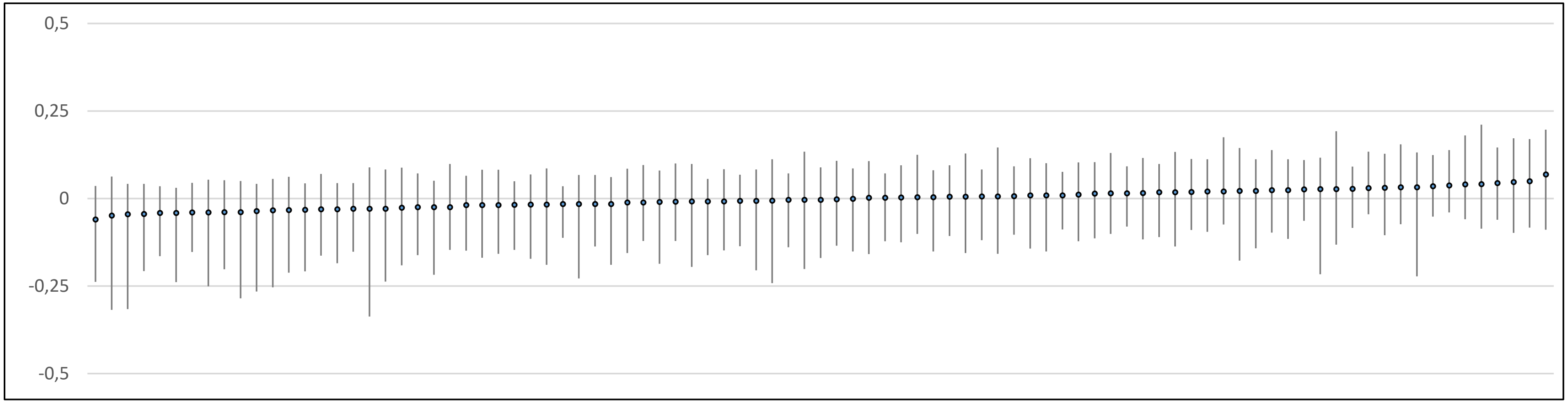
Robitzsch, A. (2019), *sirt: Supplementary Item Response Theory Models*. R package version 3.7-40, https://CRAN.R-project.org/package=sirt.

Table S12. Estimation results for the sample of interviewers having worked in at least two different regions (57 % of the interviewers)

	Model 1			**Model 2**			**Model 3**		
	M	*SD*	95% PPI	*M*	*SD*	95% PPI	*M*	*SD*	95% PPI
Fixed effects									
Age				-0.165	0.016	(-0.197, -0.134)	-0.165	0.016	(-0.195, -0.133)
Gender (ref. male)				-0.319	0.013	(-0.346, -0.293)	-0.318	0.014	(-0.345, -0.291)
Migration Background (ref. no)				-0.064	0.015	(-0.092, -0.035)	-0.064	0.014	(-0.093, -0.036)
Educational Attainment (ref. secondary education)									
no degree or lower sec. degree				-0.144	0.017	(-0.177, -0.111)	-0.144	0.017	(-0.178, -0.111)
matriculation standard				0.177	0.016	(0.145, 0.209)	0.177	0.016	(0.145, 0.208)
graduate degree				0.351	0.017	(0.318, 0.383)	0.351	0.017	(0.318, 0.383)
Employment status (ref. employed)				-0.056	0.015	(-0.087, -0.026)	-0.056	0.016	(-0.086, -0.026)
Cultural capital				0.160	0.017	(0.127, 0.193)	0.160	0.017	(0.127, 0.192)
Political Area Size				-0.043	0.021	(-0.084, -0.004)	-0.044	0.021	(-0.085, -0.003)
Interviewer Level Covariates									
Gender (ref. male)							-0.075	0.107	(-0.282, 0.141)
Age (ref. up to 49 years)									
50 to 65 years							-0.070	0.134	(-0.329, 0.194)
older than 65 years							0.103	0.136	(-0.166, 0.361)
Educational Attainment (ref. lower sec. degree)									
Secondary education							0.227	0.160	(-0.102, 0.521)
Matriculation standard							0.039	0.159	(-0.284, 0.338)
Work experience as interviewer (ref. up to two years)									
2 to 3 years							0.064	0.172	(-0.276, 0.394)
4 to 5 years							0.150	0.157	(-0.164, 0.454)
more than 5 years							0.000	0.170	(-0.330, 0.330)
Variance components of random effects									
Respondents	0.423	0.039	(0.355, 0.506)	0.243	0.024	(0.199, 0.290)	0.249	0.023	(0.208, 0.299)
Interviewers	0.029	0.007	(0.018, 0.045)	0.031	0.007	(0.021, 0.047)	0.031	0.007	(0.020, 0.048)
Areas	0.003	0.003	(0.000, 0.011)	0.001	0.001	(0.000, 0.006)	0.001	0.002	(0.000, 0.006)

Note. Standardized results are presented for fixed effects. *M* = posterior mean. *SD* = posterior standard deviation. PPI = posterior probability interval (2.5th and 97.5th percentile of the posterior distribution).

Figure S1. Residuals of area clusters with corresponding posterior probability interval (2.5th and 97.5th percentile of the posterior distribution).



Design Effect for Interviewer and Area Clusters

There are two main consequences resulting from interviewer effects on survey outcomes: first, an increased variance of a statistic and second, a reduction in effective sample size. The impact of the first consequence on the measurement of mathematic achievement was tested by indicating interviewer and area variance proportions based on variance component testing (Intraclass Correlation Coefficient, ICC). It showed that most variance is attributable to interviewer clusters and a much smaller amount to sampling clusters, which is a finding shared by previous surveys (Brunton-Smith *et al.*, 2012; Brunton-Smith *et al.*, 2016; Durrant *et al.*, 2010; Schnell and Kreuter, 2005). The second consequence stems from the overall increase of variance due to the high interviewer effects, as both lead to a decrease in effective sample size. For this reason, even small effects per interviewer can have an undue impact on the data quality, especially if the caseload per interviewer is high (Collins, 1980; Hox, 1994; Kish, 1965; Schaeffer *et al.*, 2010).

The amount of dependence of resulting competence estimates on the test administrator can furthermore be expressed by the design effect. By this, the average size of interviewers' caseloads is considered additional to the ICC:

$$\mathrm{D_{eff}} = 1 + (m - 1)\rho.$$

Thereby, m is the average number of test takers per interviewer and ρ is the ICC for all interviewers. Likewise, the design effect can be calculated for the area clusters, representing the effect of the two-stage sampling. Based on the intraclass correlation of Model 1 (see Table 1), the design effect for interviewer clusters amounts to 2.60 and to 1.44 for the area clusters. The design effect gives insight on how accurate the results are in comparison to a random sampling and at the same time it denotes how much larger the sample size must be to obtain the same precision in survey estimates (Schnell and Kreuter, 2005). Hence, it illustrates the increase in variance and also the decrease in effective sample size. For example, a design effect of 2 reduces the effective sample size by half (Schaeffer *et al.*, 2010). As the design effect has no unit of measurement, its values are comparable across different survey estimates.

Sensitivity of interviewer variance (ICC) to outlying interviewers

As interviewers with deviating residuals introduce variance to the estimation of latent mathematic competence, we tested how much the intraclass correlation reduces when first, the most outlying interviewer with respective respondents and second, all outlying interviewers with respective respondents are excluded from the analysis. The estimation of our null model without the most outlying interviewer resulted in a reduced interviewer variance of 3.5 percent (in comparison to 6.6 percent of variance in Model 1 of Table 1), whereas the variance attributable to the respondents nesting in areas slightly increased to 1.1 percent (in comparison to 0.8 percent of variance in Model 1 of Table 1). Estimating the null model without all 12 outlying interviewers resulted in a further reduction of interviewer variance. The interviewer clusters now account for 0.9 percent of variance, with area clusters showing likewise a variance of 0.9 percent.

Sensitivity of interviewer residuals to group size

For most of the obtained interviewer residuals from our estimated multilevel IRT analyses, shrinkage to the general mean is expectable. If an interviewer interviewed a high number of respondents, posterior means resemble practically the intercept of separate regression estimations for this interviewer. Hence, the identification of exceptional interviewers might depend on the group size (the number of respondents per interviewer), also termed sensitivity of interviewer residuals to group size (Pickery and Loosveldt, 2004). We tested if the amount of residual deviation per interviewer cluster is correlated with the number of test administrations per interviewer. The correlation coefficient (r = .074, p = .303) does not indicate that the amount of uncertainty on the interviewer level depends on the size of the clusters.

Large-scale Assessments in Education

RESEARCH Open Access

The achievement gap in reading competence: the effect of measurement non-invariance across school types

Theresa Rohm[1,2*], Claus H. Carstensen[2], Luise Fischer[1] and Timo Gnambs[1,3]

*Correspondence: theresa.rohm@uni-bamberg.de
[1] Leibniz Institute for Educational Trajectories, Wilhelmsplatz 3, 96047 Bamberg, Germany
Full list of author information is available at the end of the article

Abstract

Background: After elementary school, students in Germany are separated into different school tracks (i.e., school types) with the aim of creating homogeneous student groups in secondary school. Consequently, the development of students' reading achievement diverges across school types. Findings on this achievement gap have been criticized as depending on the quality of the administered measure. Therefore, the present study examined to what degree differential item functioning affects estimates of the achievement gap in reading competence.

Methods: Using data from the German National Educational Panel Study, reading competence was investigated across three timepoints during secondary school: in grades 5, 7, and 9 ($N = 7276$). First, using the invariance alignment method, measurement invariance across school types was tested. Then, multilevel structural equation models were used to examine whether a lack of measurement invariance between school types affected the results regarding reading development.

Results: Our analyses revealed some measurement non-invariant items that did not alter the patterns of competence development found among school types in the longitudinal modeling approach. However, misleading conclusions about the development of reading competence in different school types emerged when the hierarchical data structure (i.e., students being nested in schools) was not taken into account.

Conclusions: We assessed the relevance of measurement invariance and accounting for clustering in the context of longitudinal competence measurement. Even though differential item functioning between school types was found for each measurement occasion, taking these differences in item estimates into account did not alter the parallel pattern of reading competence development across German secondary school types. However, ignoring the clustered data structure of students being nested within schools led to an overestimation of the statistical significance of school type effects.

Keywords: Alignment method, Competence development, Measurement invariance, Multilevel item response theory, Multilevel structural equation modeling

Introduction

Evaluating measurement invariance is a premise for the meaningful interpretation of differences in latent constructs between groups or over time (Brown, 2006). By assessing measurement invariance, it is made certain that the observed changes present true change instead of differences in the interpretation of items. The present study investigates measurement invariance between secondary school types for student reading competence, which is the cornerstone of learning. Reading competences develop in secondary school from reading simple texts, retrieving information and making inference from what is explicitly stated, up to the level of being a fluent reader by reading longer and more complex texts and being able to infer from what is not explicitly stated in the text (Chall, 1983). In particular, students' reading competence is essential for the comprehension of educational content in secondary school (Edossa et al., 2019; O'Brien et al., 2001). Reading development is often investigated either from a school-level perspective or by focusing on individual-level differences. When taking a *school-level perspective* on reading competence growth within the German secondary school system, the high degree of segregation after the end of primary school must be considered. Most students are separated into different school tracks on the basis of their fourth-grade achievement level to obtain homogenous student groups in secondary school (Köller & Baumert, 2002). This homogenization based on proficiency levels is supposed to optimize teaching and education to account for students' preconditions, enhancing learning for all students (Baumert et al., 2006; Gamoran & Mare, 1989). Consequently, divergence in competence attainment already exists at the beginning of secondary school and might increase among the school tracks over the school years. Previous studies comparing reading competence development between different German secondary school types have presented ambiguous results by finding either a comparable increase in reading competence development (e.g., Retelsdorf & Möller, 2008; Schneider & Stefanek, 2004) or a widening gap between upper, middle, and lower academic school tracks (e.g., Pfost & Artelt, 2013) for the same schooling years. Increasing performance differences in reading over time are termed "Matthew effects", in the biblical analogy of rich getting richer and the poor getting poorer (e.g., Bast and Reitsma, 1998; Walberg & Tsai, 1983). This Matthew effect hypothesis was first used in the educational context by Stanovich (1986) to examine individual differences in reading competence development. Besides this widening pattern, as described by the Matthew effect phenomena, also parallel or compensatory patterns in reading development can be present. Parallel development is the case, when studied groups initially diverge in their reading competence and similarly increase over time. A compensatory pattern describes a reading competence development, where an initially diverging reading competence between groups converges over time.

Moreover, findings on the divergence in competence attainment have been criticized as being dependent on the quality of the measurement construct (Pfost et al., 2014; Protopapas et al., 2016). More precisely, the psychometric properties of the administered tests, such as the measurement (non-)invariance of items, can distort individual- or school-level differences. A core assumption of many measurement models pertains to comparable item functioning across groups, meaning that differences between item parameters are zero across groups, or in case of approximate measurement invariance, approximately zero. In practice, this often holds for only a subset of items and partial invariance can then be applied, where some item parameters (i.e., intercepts) are held constant across groups and others are allowed to be freely estimated (Van de Schoot et al., 2013). Using data from the German National Educational Panel Study (NEPS; Blossfeld et al., 2011), we focus on school-level differences in reading competence across three timepoints. We aim to examine the degree to which measurement non-invariance distorts comparisons of competence development across school types. We therefore compare a model that assumes partial measurement invariance across school types with a model that does not take differences in item estimates between school types into account. Finally, we demonstrate the need to account for clustering (i.e., students nested in schools) in longitudinal reading competence measurement when German secondary school types are compared.

School segregation and reading competence development

Ability tracking of students can take place within schools (e.g., differentiation through course assignment as, for example, in U.S. high schools) or between schools with a curricular differentiation between school types and with distinct learning certificates being offered by each school track, as is the German case (Heck et al., 2004; LeTendre et al., 2003; Oakes & Wells, 1996). The different kinds of curricula at each school type are tailored to the prerequisites of the students and provide different learning opportunities. German students are assigned to different school types based on primary school recommendations that take primary school performance during fourth grade into account, but factors such as support within the family are also considered (Cortina & Trommer, 2009; Pfost & Artelt, 2013; Retelsdorf et al., 2012). Nevertheless, this recommendation is not equally binding across German federal states, leaving room for parents to decide on their children's school track. Consequently, student achievement in secondary school is associated with the cognitive abilities of students but also with their social characteristics and family background (Baumert et al., 2006; Ditton et al., 2005). This explicit between-school tracking after fourth grade has consequences for students' achievement of reading competence in secondary school.

There might be several reasons why different trajectories of competence attainment are observed in the tracked secondary school system (Becker et al., 2006). First, students might already differ in their initial achievement and learning rates at the beginning of secondary school. This is related to curricular differentiation, as early separation aims to create homogenous student groups in terms of student proficiency levels and, in effect, enhances learning for all students by providing targeted learning opportunities (Baumert et al., 2003; Köller & Baumert, 2002; Retelsdorf & Möller, 2008). Hence, different learning rates are expected due to selection at the beginning of secondary school (Becker et al., 2006). Second, there are differences in learning and teaching methods among the school tracks, as learning settings are targeted towards students' preconditions. Differences among school types are related to cognitive activation, the amount of support from the teacher in problem solving and demands regarding students' accomplishments (Baumert et al., 2003). Third, composition effects due to the different socioeconomic and ethnic compositions of schools can shape student achievement. Not only belonging to a particular school type but also individual student characteristics determine student achievement. Moreover, the mixture of student characteristics might have decisive effects (Neumann et al., 2007). For example, average achievement rates and the characteristics of students' social backgrounds were found to have additional effects on competence attainment in secondary school (Baumert et al., 2006), beyond mere school track affiliation and individual characteristics. Hence, schools of the same school type were found to differ greatly from each other in their attainment levels and their social compositions (Baumert et al., 2003).

Findings from the cross-sectional *Programme for International Student Assessment* (PISA) studies, conducted on behalf of the OECD every three years since 2000, unanimously show large differences between school tracks in reading competence for German students in ninth grade (Baumert et al., 2001, 2003; Nagy et al., 2017; Naumann et al., 2010; Weis et al., 2016, 2020). Students in upper academic track schools have, on average, higher reading achievement scores than students in the middle and lower academic tracks. Reading competence is thereby highly correlated with other assessed competencies, such as mathematics and science, where these differences between school tracks hold as well.

A few studies have also examined between-school track differences in the development of reading competence in German secondary schools, with most studies focusing on fifth and seventh grade in selected German federal states (e.g., Bos et al., 2009; Lehmann & Lenkeit, 2008; Lehmann et al., 1999; Pfost & Artelt, 2013; Retelsdorf & Möller, 2008). While some studies reported parallel developments in reading competence from fifth to seventh grade between school types (Retelsdorf & Möller, 2008; Schneider & Stefanek, 2004), others found a widening gap (Pfost & Artelt, 2013; Pfost et al., 2010). A widening gap between school types was also found for other competence domains, such as mathematics (Baumert et al., 2003, 2006; Becker et al., 2006; Köller & Baumert, 2001), while parallel developments were rarely observed (Schneider & Stefanek, 2004).

In summary, there might be different school milieus created by the processes of selection into secondary school and formed by the social and ethnic origins of the students (Baumert et al., 2003). This has consequences for reading competence development during secondary school, which can follow a parallel, widening or compensatory pattern across school types. The cross-sectional PISA study regularly indicates large differences among German school types in ninth grade but does not offer insight into whether these differences already existed at the beginning of secondary school or how they developed throughout secondary school. In comparison, longitudinal studies have indicated a pattern in reading competence development through secondary school, but the studies conducted in the past were regionally limited and presented inconsistent findings on reading competence development among German secondary school types. In addition to differences in curricula, learning and teaching methods, students' social backgrounds, family support, and student composition, the manner in which competence development during secondary school is measured and analyzed might contribute to the observed pattern in reading competence development.

Measuring differences in reading development

A meaningful longitudinal comparison of reading competence between school types and across grades requires a scale with a common metric. To be more specific, the relationships between the latent trait score and each observed item should not depend on group membership. The interpretability of scales has been questioned due to scaling issues (Protopapas et al., 2016). While the item response theory (IRT) calibration is assumed to be theoretically invariant, it depends in practice on the sample, item fit, and equivalence of item properties (e.g., discrimination and difficulty) among test takers and compared groups. Hence, empirically discovered between-group differences might be confounded with the psychometric properties of the administered tests. For example, Pfost et al. (2014) concluded from a meta-analysis of 28 studies on Matthew effects in primary school (i.e., the longitudinally widening achievement gap between good and poor readers) that low measurement precision (e.g., constructs presenting floor or ceiling effects) is strongly linked with compensatory patterns in reading achievement. Consequently, measuring changes using reading competence scores might depend on the quality of the measurement. Regarding competence development in secondary school, measurement precision is enhanced through the consideration of measurement error, the consideration of the multilevel data structure, and measurement invariance across groups. A biased measurement model might result when measurement error or the multilevel data structure are ignored, while the presence of differential item functioning (DIF) can be evidence of test-internal item bias. Moreover, the presence of statistical item bias might also contribute to test unfairness and, thus, invalid systematic disadvantages for specific groups (Camilli, 2006).

Latent variable modeling for reading competence, such as latent change models (Raykov, 1999; Steyer et al., 2000), can be advantageous compared to using composite scores. When using composite scores representing latent competences, measurement error is ignored (Lüdtke et al., 2011). Hence, biased estimates might be obtained if the construct is represented by composite scores instead of a latent variable measured by multiple indicators and accounting for measurement error (Lüdtke et al., 2008). Investigating student competence growth in secondary school poses a further challenge, as the clustered structure of the data needs to be taken into account. This can for example be achieved using cluster robust standard error estimation methods or through hierarchical linear modeling (cf. McNeish et al., 2017). If the school is the primary sampling unit, students are nested within schools and classes. Ignoring this hierarchical structure during estimation might result in inaccurate standard errors and biased significance tests, as standard errors would be underestimated. In turn, the statistical significance of the effects would be overestimated (Finch & Bolin, 2017; Hox, 2002; Raudenbush & Bryk, 2002; Silva et al., 2019). As one solution, multilevel structural equation modeling (MSEM) takes the hierarchical structure of the data into account while allowing for the estimation of latent variables with dichotomous and ordered categorical indicators (Kaplan et al., 2009; Marsh et al., 2009; Rabe-Hesketh et al., 2007). Although explicitly modeling the multilevel structure (as compared to cluster robust standard error estimation) involves additional assumptions regarding the distribution of the random effects and the covariance structure of random effects, it allows for the partitioning of variance to different hierarchical levels and for cluster-specific inferences (McNeish et al., 2017).

Furthermore, regarding the longitudinal modeling of performance divergence, an interpretation of growth relies on the assumption that the same attributes are measured across all timepoints (Williamson et al., 1991) and that the administered instrument (e.g., reading competence test items) is measurement invariant across groups (Jöreskog, 1971; Schweig, 2014). The assumption of measurement invariance presupposes that all items discriminate comparably across groups as well as timepoints and are equally difficult, independent of group membership and measurement occasion. Hence, the item parameters of a measurement model have to be constant across groups, meaning that the probability of answering an item correctly should be the same for members of different groups and at different timepoints when they have equal ability levels (Holland & Wainer, 1993; Millsap & Everson, 1993). When an item parameter is not independent of group membership, DIF is present.

The aim of our study is to investigate the effects of measurement non-invariance among school types on the achievement gap in reading competence development in German secondary schools. Measurement invariance between secondary school types is investigated for each measurement occasion to test whether items are biased among the school types. Then, we embed detected DIF into the longitudinal estimation of reading competence development between school types. A model considering school-type-specific item discrimination and difficulty for items exhibiting non-invariance between school types is therefore compared to a model that does not consider these school-type specificities. To achieve measurement precision for this longitudinal competence measurement, we consider measurement error and the clustered data structure through multilevel latent variable modeling. Finally, we present the same models without consideration of the clustered data structure and compare school type effects on reading competence development.

It is our goal to investigate whether the longitudinal development of reading competence is sensitive to the consideration of measurement non-invariance between the analyzed groups and to the consideration of the clustered data structure. This has practical relevance for all studies on reading competence development, where comparisons between school types are of interest and where schools were the primary sampling unit. Such evaluations increase the certainty that observed changes between school types reflect true changes.

Method

Sample and procedure

The sample consisted of $N = 7276$ German secondary school students, repeatedly tested and interviewed in 2010 and 2011 (grade 5), 2012 and 2013 (grade 7), and 2014 and 2015 (grade 9) as part of the NEPS. Approximately half of the sample was female (48.08%), and 25.46% had a migration background (defined as either the student or at least one parent born abroad). Please note that migration background is unequally distributed across school types: 22.1% high school students, 26.9% middle secondary school students, 38.5% lower secondary school students, 31.2% comprehensive school students and 15.2% students from schools offering all tracks of secondary education except the high school track had a migration background. In fifth grade, the students' ages ranged from 9 to 15 years ($M = 11.17$, $SD = 0.54$). Students were tested within their class context through written questionnaires and achievement tests. For the first timepoint in grade 5, immediately after students were assigned to different school tracks, a representative sample of German secondary schools was drawn using a stratified multistage sampling design (Aßmann et al., 2011). First, schools that teach at the secondary level were randomly drawn, and second, two grade 5 classes were randomly selected within these schools. The five types of schools were distinguished and served as strata in the first step: high schools ("Gymnasium"), middle secondary schools ("Realschule"), lower secondary schools ("Hauptschule"), comprehensive schools ("Gesamtschule"), and schools offering all tracks of secondary education except the high school track ("Schule mit mehreren Bildungsgängen"). The schools were drawn proportional to their number of classes from these strata. Finally, all students of the selected classes for whom a positive parent's consent was obtained before panel participation were asked to take part in the study. At the second measurement timepoint in 2012 to 2013, when students attended grade 7, a refreshment sample was drawn due to German federal state-specific differences in the timing of the transition to lower secondary education ($N=2170$; 29.82% of the total sample). The sampling design of the refreshment sample resembles the sampling design of the original sample (Steinhauer & Zinn, 2016). The ninth-grade sample in 2014 and 2015 was taken at the third measurement timepoint and was a follow-up survey for the students from regular schools in both the original and the refreshment sample. Students were tested at their schools, but $N=1797$ students (24.70% of the total sample) had to be tested at least one measurement timepoint through an individual follow-up within their home context. In both cases, the competence assessments were conducted by a professional survey institute that sent test administrators to the participating schools or households. For an overview of the students being tested per measurement timepoint per school type, within the school or home context, as well as information on temporary and final sample attrition, see Table 1.

To group students into their corresponding school type, we used the information on the survey wave when the students were sampled (original sample in grade 5, refreshment sample in grade 7). Overall, most of the sampled students attended high schools ($N=3224$; 44.31%), 23.65% attended middle secondary schools ($N=1721$), 13.95% attended lower secondary schools ($N=1015$), 11.96% of students attended schools offering all tracks of secondary education except the high school track ($N=870$), and 6.13% attended comprehensive schools ($N=446$). Altogether, the students attended 299 different schools, with a median of 24 students per school. Further details on the survey and the data collection process are presented on the project website (http://www.neps-data.de/).

Instruments

During each assessment, reading competence was measured with a paper-based achievement test, including 32 items in fifth grade, 40 items in seventh grade administered in easy (27 items) and difficult (29 items) booklet versions, and 46 items in ninth grade administered in easy (30 items) and difficult (32 items) booklet versions. The items were specifically constructed for the administration of the NEPS, and each item was administered once (Krannich et al., 2017; Pohl et al., 2012; Scharl et al., 2017). Because memory effects might distort responses if items are repeatedly administered, the linking of the reading measurements in the NEPS is based on an anchor-group design (Fischer et al., 2016). With two independent link samples (one to link the grade 5 and grade 7 reading competence tests and the other to link the grade 7 with the grade 9 test), drawn from the

Table 1 Number of students per school type and per measurement occasion ($N = 7276$)

Type of school	N (%) tested overall	N (%) tested at all timepoints	N (%) tested grade 5	N (%) tested grade 7	N (%) tested grade 9
High school	3224 (44.31)	1457 (51.01)	2302 (47.12)	2909 (47.06)	2112 (46.18)
Refreshment sample				835	542
Tested in home context				176	706
Temporary attrition				176	586
Final attrition				22	142
Middle sec. school	1721 (23.65)	688 (24.12)	1114 (22.80)	1474 (23.84)	1096 (23.97)
Refreshment sample				554	340
Tested in home context				163	426
Temporary attrition				130	301
Final attrition				14	118
Lower sec. school	1015 (13.95)	293 (10.27)	698 (14.29)	706 (11.42)	501 (10.96)
Refreshment sample				278	164
Tested in home context				242	393
Temporary attrition				206	353
Final attrition				1	51
Schools offering all tracks of sec. education (except high school)	870 (11.96)	230 (8.06)	487 (9.97)	685 (11.08)	551 (12.05)
Refreshment sample				352	287
Tested in home context				146	206
Temporary attrition				98	189
Final attrition				1	40
Comprehensive school	446 (6.13)	184 (6.45)	284 (5.81)	408 (6.59)	313 (6.84)
Refreshment sample				123	98
Tested in home context				23	66
Temporary attrition				24	74
Final attrition				2	25

Absolute numbers are presented with percentages in parentheses. The percentages are to be read column wise

same population as the original sample, a mean/mean linking was performed (Loyd & Hoover, 1980). In addition, the unidimensionality of the tests, measurement invariance of the items regarding reading development over the grade levels, as well as for relevant sample characteristics (i.e., gender and migration background) was demonstrated (Fischer et al., 2016; Krannich et al., 2017; Pohl et al., 2012; Scharl et al., 2017). Marginal reliabilities were reported as good, with 0.81 in grade 5, 0.83 in grade 7, and 0.81 in grade 9.

Each test administered to the respondents consisted of five different text types (domains: information, instruction, advertising, commenting and literary text) with subsequent questions in either a simple or complex multiple-choice format or a matching response format. In addition, but unrelated to the five text types, the questions covered three types of cognitive requirements (finding information in the text, drawing text-related conclusions, and reflecting and assessing). To answer the respective question types, these cognitive processes needed to be activated. These dimensional concepts and question types are linked to the frameworks of other large-scale assessment studies, such as PISA (OECD, 2017) or the International Adult Literacy Survey (IALS/ALL; e.g., OECD & Statistics Canada 1995). Further details on the reading test construction and development are presented by Gehrer et al. (2003).

Statistical analysis

We adopted the multilevel structural equation modelling framework for the modeling of student reading competence development and fitted a two-level factor model with categorical indicators (Kamata & Vaughn, 2010) to the reading competence tests. Each of the three measurement occasions was modeled as a latent factor. Please note that MSEM is the more general framework to fitting multilevel item response theory models (Fox, 2010; Fox & Glas, 2001; Kamata & Vaughn, 2010; Lu et al., 2005; Muthén & Asparouhov, 2012), and therefore, each factor in our model resembles a unidimensional, two-parametric IRT model. The model setup was the same for the student and the school level and therefore discrimination parameters (i.e., item loadings) were constrained to be equal at the within- and between-level, while difficulty estimates (i.e., item thresholds) and item residual variances are measured on the between-level (i.e., school-level). School type variables were included as binary predictors of latent abilities at the school level.

The multilevel structural equation models for longitudinal competence measurement were estimated using Bayesian MCMC estimation methods in the Mplus software program (version 8.0, Muthén and Muthén 1998–2020). Two Markov chains were implemented for each parameter, and chain convergence was assessed using the potential scale reduction (PSR, Gelman & Rubin, 1992) criterion, where values below 1.10 indicate convergence (Gelman et al., 2004). Furthermore, successful convergence of the estimates was evaluated based on trace plots for each parameter. To determine whether the estimated models delivered reliable estimates, autocorrelation plots were investigated. The mean of the posterior distribution and the Bayesian 95% credibility interval were used to evaluate the model parameters. Using the Kolmogorov–Smirnov test, the hypothesis that both MCMC chains have an equal distribution was evaluated using 100 draws from each of the two chains per parameter. For all estimated models, the PSR criterion (i.e., Gelman and Rubin diagnostic) indicated that convergence was achieved, which was confirmed by a visual inspection of the trace plots for each model parameter.

Diffuse priors were used with a normal distribution with mean zero and infinite variance, N (0, ∞), for continuous indicators such as intercepts, loading parameters or regression slopes; normal distribution priors with mean zero and a variance of 5, N (0, 5), were used for categorical indicators; inverse-gamma priors IG (− 1, 0) were used for residual variances; and inverse-Wishart priors $IW\,(0, -4)$ for variances and covariances.

Model fit was assessed using the posterior predictive p-value (PPP), obtained through a fit statistic based on the likelihood-ratio χ^2 test of an H_0 model against an unrestricted H_1 model, as implemented in Mplus. A low PPP indicates poor fit, while an acceptable model fit starts with PPP > 0.05, and an excellent-fitting model has a PPP value of approximately 0.5 (Asparouhov & Muthén, 2010).

Differential item functioning was examined using the invariance alignment method (IA; Asparouhov & Muthén, 2014; Kim et al., 2017; Muthén & Asparouhov, 2014). These models were estimated with maximum likelihood estimation using numerical integration and taking the nested data structure into account through cluster robust estimation. One can choose between fixing one group or free estimation. As the fixed alignment was shown to slightly outperform the free alignment in a simulation study (Kim et al., 2017), we applied fixed alignment and ran several models fixing each of the five school types once. Item information for items exhibiting DIF between school types were then split to the respective non-aligning group versus the remaining student groups. Hence, new pseudo-items are introduced for the models that take school-type specific item properties into account.

In the multilevel structural equation models, for the students selected as part of the refreshment sample at the time of the second measurement, we treated their missing information from the first measurement occasion as missing completely at random (Rubin, 1987). Please note that student attrition from the seventh and ninth grade samples can be related to features of the sample, even though the multilevel SEM accounts for cases with missing values for the second and third measurement occasions. We fixed the latent factor intercept per assessment for seventh and ninth grade to the value of the respective link constant. The average changes in item difficulty to the original sample were computed from the link samples, and in that manner, an additive linking constant for the overall sample was obtained. Please note that this (additive) linking constant does not change the relations among school type effects per measurement occasion.

Furthermore, we applied weighted effect coding to the school type variables, which is preferred over effect coding, as the categorical variable school type has categories of different sizes (Sweeney & Ulveling, 1972; Te Grotenhuis et al., 2017). This procedure is advantageous for observational studies, as the data are not balanced, in contrast to data collected via experimental designs. First, we set the high school type as the reference category. Second, to obtain an estimate for this group, we re-estimated the model using middle secondary school as the reference category. Furthermore, we report the Cohen's (1969) d effect size per school type estimate. We calculated this effect size as the difference per value relative to the average of all other school type effects per measurement occasion and divided it by the square root of the factor variance (hence the standard deviation) per respective latent factor. For models where the multilevel structure was accounted for, the within- and between-level components of the respective factor variance were summed for the calculation of Cohen's d.

Data availability and analysis syntax

The data analyzed in this study and documentation are available at https://doi.org/10.5157/NEPS:SC3:9.0.0. Moreover, the syntax used to generate the reported results is provided in an online repository at https://osf.io/5ugwn/?view_only=327ba9ae72684d07be8b4e0c6e6f1684.

Results

We first tested for measurement invariance between school types and subsequently probed the sensitivity of school type comparisons when accounting for measurement non-invariance. In our analyses, sufficient convergence in the parameter estimation was indicated for all models through an investigation of the trace and autocorrelation plots. Furthermore, the PSR criterion fell below 1.10 for all parameters after 8000 iterations. Hence, appropriate posterior predictive quality for all parameters on the between and within levels was assumed.

DIF between school types

Measurement invariance of the reading competence test items across the school types was assessed using IA. Items with non-aligning, and hence measurement non-invariant, item parameters between these higher-level groups were found for each measurement occasion (see the third, sixth and last columns of Table 2). For the reading competence measurement in fifth grade, 11 out of the 32 administered items showed measurement non-invariance in either discrimination or threshold parameters across school types. Most non-invariance occurred for the lowest (lower secondary school) and the highest (high school) types. For 5 of the 11 non-invariant items, the school types with non-invariance were the same for both the discrimination and threshold parameters. In seventh grade, non-invariance across school types was found for 11 out of the 40 test items in either discrimination or threshold parameters. While non-invariance occurred six times in discrimination parameters, it occurred seven times in threshold parameters, and most non-invariance occurred for the high school type (10 out of the 11 non-invariant items). Applying the IA to the competence test administered in ninth grade showed non-invariance for 11 out of the 44 test items. Nearly all non-invariances were between the lowest and highest school types, and most item non-invariance in discrimination and threshold parameters occurred for the last test items.

Consequences of DIF for school type effects

Comparisons of competence development across school types were estimated using MSEM. Each timepoint was modeled as a latent factor, and the between-level component of each latent factor was regressed on the school type. Furthermore, the latent factors were correlated through this modeling approach, both at the within and between levels. Please note that the within- and between-level model setup was the same, and each factor was modeled with several categorical indicators. In Models 1a and 1b, no school-type specific item discrimination or item difficulty estimates were accounted for, while in Models 2a and 2b, school-type specific item discrimination

Table 2 Results from the invariance alignment method per measurement occasion

Grade 5 (N = 4885)		**Grade 7 (N = 6182)**		**Grade 9 (N = 4573)**	
Item	**IA**	**Item**	**IA**	**Item**	**IA**
Discrimination		*Discrimination*		*Discrimination*	
reg50110_c		reg70110_c		reg90610_c	
reg5012s_c		reg70120_c		reg90620_c	
reg50130_c		reg7013s_c	HS	reg9063s_c	
reg50140_c	LS	reg70140_c		reg90640_c	
reg50150_c		reg7015s_c		reg90660_c	
reg5016s_c		reg7016s_c		reg90670_c	
reg50170_c		reg70610_c		reg90680_c	
reg50210_c		reg70620_c	HS	reg90810_c	
reg50220_c		reg7063s_c		reg90820_c	
reg50230_c		reg70640_c		reg9083s_c	
reg50240_c		reg70650_c		reg90840_c	
reg50250_c		reg7066s_c		reg90850_c	
reg5026s_c		reg70210_c		reg90860_c	
reg50310_c		reg70220_c		reg90870_c	
reg50320_c		reg7023s_c		reg90210_c	
reg50330_c		reg7024s_c		reg90220_c	
reg50340_c		reg70250_c		reg90230_c	
reg50350_c		reg7026s_c		reg90250_c	
reg50360_c		reg70310_c		reg90710_c	
reg50370_c	MS, HS	reg70320_c		reg90720_c	
reg50410_c	LS	reg7033s_c	HS	reg90730_c	
reg5042s_c	LS	reg70340_c		reg9074s_c	
reg50430_c	LS	reg70350_c		reg90750_c	
reg50440_c	HS	reg70360_c		reg9091s_c	
reg50460_c	LS	reg70410_c		reg90920_c	
reg50510_c		reg70420_c		reg90930_c	
reg5052s_c		reg70430_c		reg90940_c	
reg50530_c		reg70440_c		reg90950_c	
reg50540_c		reg7045s_c		reg90960_c	
reg5055s_c		reg70460_c		reg9097s_c	
reg50560_c		reg7051s_c		reg90410_c	
reg50570_c		reg70520_c		reg90420_c	
		reg7053s_c		reg90430_c	
		reg7055s_c		reg90440_c	
		reg70560_c		reg90450_c	
		reg7071s_c	MS	reg90460_c	
		reg70720_c	HS	reg9047s_c	
		reg70730_c		reg90510_c	
		reg70740_c		reg90520_c	LS
		reg7075s_c	LS	reg90530_c	LS
				reg90540_c	
				reg90550_c	LS
				reg90560_c	HS
				reg90570_c	
Threshold		*Threshold*		*Threshold*	
reg50110_c		reg70110_c		reg90610_c	
reg5012s_c, cat.1	HS	reg70120_c		reg90620_c	

Table 2 (continued)

Grade 5 (N = 4885)		Grade 7 (N = 6182)		Grade 9 (N = 4573)	
Item	**IA**	**Item**	**IA**	**Item**	**IA**
reg5012s_c, cat.2		reg7013s_c, cat.1		reg9063s_c, cat.1	
reg50130_c		reg7013s_c, cat.2		reg9063s_c, cat.2	AT
reg50140_c	LS	reg70140_c		reg90640_c	
reg50150_c		reg7015s_c		reg90660_c	
reg5016s_c, cat.1		reg7016s_c, cat.1		reg90670_c	
reg5016s_c, cat.2		reg7016s_c, cat.2		reg90680_c	
reg5016s_c, cat.3		reg7016s_c, cat.3		reg90810_c	
reg5016s_c, cat.4		reg70610_c		reg90820_c	HS
reg5016s_c, cat.5	HS	reg70620_c		reg9083s_c	
reg50170_c	HS	reg7063s_c, cat.1		reg90840_c	
reg50210_c		reg7063s_c, cat.2		reg90850_c	
reg50220_c		reg70640_c		reg90860_c	
reg50230_c		reg70650_c		reg90870_c	
reg50240_c		reg7066s_c, cat.1		reg90210_c	
reg50250_c		reg7066s_c, cat.2		reg90220_c	
reg5026s_c		reg7066s_c, cat.3		reg90230_c	
reg50310_c		reg7066s_c, cat.4		reg90250_c	
reg50320_c		reg70210_c		reg90710_c	
reg50330_c		reg70220_c		reg90720_c	
reg50340_c		reg7023s_c, cat.1		reg90730_c	
reg50350_c		reg7023s_c, cat.2		reg9074s_c, cat.1	
reg50360_c		reg7024s_c, cat.1		reg9074s_c, cat.2	
reg50370_c	MS, HS	reg7024s_c, cat.2		reg9074s_c, cat.3	
reg50410_c	LS	reg70250_c		reg9074s_c, cat.4	LS
reg5042s_c, cat.1	LS	reg7026s_c, cat.1		reg90750_c	
reg5042s_c, cat.2	LS	reg7026s_c, cat.2		reg9091s_c, cat.1	
reg5042s_c, cat.3	LS	reg7026s_c, cat.3		reg9091s_c, cat.2	
reg50430_c	AT, HS	reg7026s_c, cat.4		reg90920_c	
reg50440_c	HS	reg70310_c		reg90930_c	
reg50460_c	HS	reg70320_c		reg90940_c	
reg50510_c		reg7033s_c, cat.1		reg90950_c	
reg5052s_c, cat.1		reg7033s_c, cat.2		reg90960_c	LS
reg5052s_c, cat.2		reg7033s_c, cat.3		reg9097s_c, cat.1	
reg5052s_c, cat.3		reg70340_c		reg9097s_c, cat.2	
reg50530_c		reg70350_c		reg9097s_c, cat.3	
reg50540_c	AT	reg70360_c		reg90410_c	
reg5055s_c, cat.1		reg70410_c	HS	reg90420_c	
reg5055s_c, cat.2		reg70420_c		reg90430_c	
reg5055s_c, cat.3		reg70430_c	HS	reg90440_c	
reg50560_c		reg70440_c		reg90450_c	HS
reg50570_c		reg7045s_c, cat.1	LS, HS	reg90460_c	
		reg7045s_c, cat.2		reg9047s_c, cat.1	
		reg7045s_c, cat.3		reg9047s_c, cat.2	
		reg70460_c		reg90510_c	
		reg7051s_c, cat.1		reg90520_c	LS, HS
		reg7051s_c, cat.2		reg90530_c	HS
		reg70520_c		reg90540_c	HS
		reg7053s_c, cat.1		reg90550_c	LS

Table 2 (continued)

Grade 5 (N = 4885)		Grade 7 (N = 6182)		Grade 9 (N = 4573)	
Item	**IA**	**Item**	**IA**	**Item**	**IA**
		reg7053s_c, cat.2		reg90560_c	HS
		reg7055s_c, cat.1		reg90570_c	LS, HS
		reg7055s_c, cat.2	HS		
		reg7055s_c, cat.3			
		reg70560_c	HS		
		reg7071s_c, cat.1			
		reg7071s_c, cat.2	HS		
		reg70720_c			
		reg70730_c	MS		
		reg70740_c			
		reg7075s_c, cat.1			
		reg7075s_c, cat.2			
		reg7075s_c, cat.3			
All grade 5	0.644	All grade 7	0.631	All grade 9	0.492

IA invariance alignment method for school types, presenting non-invariant groups (*HS* high schools, *MS* middle secondary schools, *CS* comprehensive schools, *AT* schools offering all school tracks except high school, *LS* lower secondary schools). *All* Average Invariance Index: Average R^2 across all parameters ranging from 0 (indicating full non-invariance) to 1 (indicating perfect scalar invariance)

and item difficulty estimates were taken into account for items exhibiting DIF. The amount of variance attributable to the school type (intraclass correlation) was high in both of these longitudinal models and amounted to 43.0% (Model 1a)/42.4% (Model 2a) in grade 5, 40.3% (Model 1a)/40.6% (Model 2a) in grade 7 and 43.4% (Model 1a)/43.3% (Model 2a) in grade 9. After including the school type covariates (Model 1b and Model 2b), the amount of variance in the school-level random effects was reduced by approximately two-thirds for each school-level factor, while the amount of variance in the student-level random effects remained nearly the same.

The development of reading competence from fifth to ninth grade appeared to be almost parallel between school types. The results of the first model (see Model 1b in Table 3) present quite similar differences in reading competence between school types at each measurement occasion. The highest reading competence is achieved by students attending high schools, followed by middle secondary schools, comprehensive schools and schools offering all school tracks except high school. Students in lower secondary schools had the lowest achievement at all timepoints. As the 95 percent posterior probability intervals overlap between the middle secondary school type, the comprehensive school type and the type of schools offering all school tracks except high school (see Model 1b and Model 2b in Table 3), three distinct groups of school types, as defined by reading competence achievement, remain. Furthermore, the comparison of competence development from fifth to ninth grade across these school types was quite stable. The Cohen's *d* effect size per school type estimate and per estimated model are presented in Table 4 and support this finding. A large positive effect relative to the average reading competence of the other school types is found for high school students across all grades. A large negative effect is found across all grades for lower secondary school students relative to the other school types. The other three school types have overall small effect sizes across all grades relative to the averages of the other school types.

Table 3 Results of multilevel structural equation models for longitudinal competence measurement (N = 7276)

	Model 1a			**Model 1b**			**Model 2a**			**Model 2b**		
	M	**SD**	**95% PPI**	**M**	**SD**	**95% PPI**	**M**	**SD**	**95% PPI**	**M**	**SD**	**95% PPI**
Fixed effects: School level covariates												
Grade 5												
HS				0.865	0.031	(0.802, 0.920)				0.851	0.032	(0.784, 0.909)
MS				− 0.174	0.039	(− 0.252, − 0.099)				− 0.158	0.040	(− 0.238, − 0.080)
CS				− 0.207	0.039	(− 0.283, − 0.129)				− 0.201	0.040	(− 0.279, − 0.123)
AT				− 0.206	0.041	(− 0.286, − 0.124)				− 0.210	0.041	(− 0.291, − 0.130)
LS				− 0.689	0.030	(− 0.743, − 0.624)				− 0.692	0.030	(− 0.746, − 0.627)
Grade 7												
HS				0.844	0.029	(0.783, 0.897)				0.848	0.029	(0.788, 0.902)
MS				− 0.206	0.036	(− 0.278, − 0.137)				− 0.209	0.036	(− 0.277, − 0.138)
CS				− 0.119	0.035	(− 0.188, − 0.049)				− 0.120	0.035	(− 0.187, − 0.050)
AT				− 0.174	0.038	(− 0.247, − 0.100)				− 0.179	0.038	(− 0.252, − 0.102)
LS				− 0.712	0.027	(− 0.762, − 0.656)				− 0.710	0.027	(− 0.762, − 0.655)
Grade 9												
HS				0.861	0.031	(0.797, 0.917)				0.856	0.031	(0.790, 0.912)
MS				− 0.214	0.039	(− 0.288, − 0.138)				− 0.206	0.038	(− 0.282, − 0.134)
CS				− 0.111	0.038	(− 0.185, − 0.037)				− 0.110	0.038	(− 0.183, − 0.037)
AT				− 0.221	0.039	(− 0.298, − 0.142)				− 0.232	0.040	(− 0.308, − 0.152)
LS				− 0.682	0.030	(− 0.735, − 0.617)				− 0.676	0.032	(− 0.734, − 0.608)
Variance components of random effects Student level												
Grade 5	0.383	0.047	(0.301, 0.491)	0.404	0.056	(0.319, 0.551)	0.353	0.042	(0.286, 0.445)	0.359	0.044	(0.289, 0.458)
Grade 7	0.154	0.019	(0.120, 0.192)	0.160	0.016	(0.135, 0.196)	0.151	0.016	(0.125, 0.187)	0.153	0.012	(0.131, 0.176)
Grade 9	0.277	0.030	(0.223, 0.340)	0.281	0.029	(0.227, 0.338)	0.265	0.025	(0.217, 0.313)	0.275	0.021	(0.237, 0.320)

Table 3 (continued)

	Model 1a			Model 1b				Model 2a			Model 2b		
	M	SD	95% PPI	M	SD		95% PPI	M	SD	95% PPI	M	SD	95% PPI
School Level													
Grade 5	0.289	0.044	(0.217, 0.388)	0.078	0.014		(0.057, 0.111)	0.260	0.039	(0.196, 0.349)	0.071	0.012	(0.052, 0.097)
Grade 7	0.104	0.016	(0.079, 0.139)	0.030	0.004		(0.023, 0.040)	0.103	0.014	(0.080, 0.135)	0.029	0.004	(0.023, 0.037)
Grade 9	0.212	0.030	(0.162, 0.280)	0.060	0.009		(0.045, 0.080)	0.202	0.027	(0.157, 0.262)	0.061	0.009	(0.047, 0.080)
Correlations and Covariances between latent factors	Correlation		Covariance	Correlation		Covariance		Correlation		Covariance	Correlation		Covariance
Student Level													
Grade 5 with Grade 7	0.644		0.172	0.632		0.179		0.631		0.139	0.634		0.139
Grade 5 with Grade 9	0.649		0.223	0.652		0.232		0.662		0.210	0.650		0.185
Grade 7 with Grade 9	0.723		0.163	0.722		0.160		0.722		0.156	0.721		0.131
School Level													
Grade 5 with Grade 7	0.000		0.000	0.001		0.000		0.002		0.000	0.002		0.000
Grade 5 with Grade 9	0.000		0.000	0.001		0.000		0.002		0.000	0.003		0.001
Grade 7 with Grade 9	0.001		0.000	0.002		0.000		0.003		0.001	0.000		0.000
PPP	0.027			0.019				0.179			0.196		

Note. Standardized results are presented for fixed effects. *M* = posterior mean. *SD* = posterior standard deviation, *PPI* posterior probability interval (2.5 and 97.5 percentile of the posterior distribution). School level covariates: *HS* high schools, *MS* middle secondary schools, *CS* comprehensive schools, *AT* schools offering all school tracks except high school, *LS* lower secondary schools, *PPP* posterior predictive p value. In Models 1a and 1b, no school-type specific item discrimination or item difficulty estimates were accounted for. In Models 2a and 2b, school-type specific item discrimination and item difficulty estimates were taken into account for items exhibiting DIF. Correlations and covariances were reported based on sample statistics

Table 4 Effect sizes (Cohen's *d*) for school type covariates per estimated model

Type of School	Model 1b	Model 2b	Model 3b	Model 4b
Grade 5				
HS	1.26	1.23	1.22	1.21
MS	0.11	0.13	0.08	0.10
CS	− 0.31	− 0.29	− 0.13	− 0.11
AT	− 0.08	− 0.10	− 0.13	− 0.15
LS	− 0.94	− 0.95	− 1.01	− 1.02
Grade 7				
HS	1.16	1.16	1.12	1.13
MS	− 0.01	− 0.01	− 0.04	− 0.04
CS	− 0.06	− 0.06	0.06	0.06
AT	− 0.06	− 0.07	− 0.08	− 0.08
LS	− 1.02	− 1.01	− 1.06	− 1.06
Grade 9				
HS	1.21	1.21	1.15	1.15
MS	− 0.02	0.00	− 0.04	− 0.03
CS	− 0.01	0.00	0.15	0.16
AT	− 0.17	− 0.20	− 0.20	− 0.23
LS	− 0.98	− 0.97	− 1.03	− 1.02

School level covariates: *HS* high schools, *MS* middle secondary schools, *CS* comprehensive schools, *AT* schools offering all school tracks except high school, *LS* lower secondary schools. Cohen's *d* effect size: calculated as the difference per value from the average of all other school type effects and divided by the square root of the factor variance per respective latent factor. The average of all school type effects can differ slightly from zero due to effect coding and model re-estimation using the reference group to obtain a reference group estimate. In Models 1b and 3b, no school-type specific item discrimination or item difficulty estimates were accounted for. In Model 2b and 4b, school-type specific item discrimination and item difficulty estimates were taken into account for items exhibiting DIF. The multilevel data structure was taken into account for estimation of Models 1b and 2b but not for Models 3b and 4b

The results of the second model (see Model 2b in Table 3) show similar differences between the school types when compared to the former model. Additionally, effect sizes are similar between the two models. Hence, differences in the development of reading competence across school types are parallel, and this pattern is robust to the discovered school-type specific DIF of item discrimination and difficulty estimates. With regard to model fit, only two models (Models 2a and 2b) showed an acceptable fit with PPP > 0.05 when school type-specific item discrimination and item difficulty estimates for items exhibiting DIF were accounted for. Furthermore, single-level regression analyses with cluster robust standard error estimation using the robust maximum likelihood (MLR) estimator were performed to investigate if the findings were robust to the application of an alternative estimation method for hierarchical data. Please note that result tables for these analyses are presented in the Additional file 1. The main findings remain unaltered, as a parallel pattern of reading competence development between the school types was found, as well as three distinct school type groups.

Consequences when ignoring clustering effects

Finally, we estimated the same models without accounting for the clustered data structure (see Table 5). In comparison to the previous models, Model 3a and Model

Table 5 Results of structural equation models for longitudinal competence measurement ignoring clustered structure (N = 7276)

	Model 3a			Model 3b			Model 4a			Model 4b		
	M	SD	95% PPI	M	SD	95% PPI	M	SD	95% PPI	M	SD	95% PPI
School Type Covariates												
Grade 5												
HS				0.627	0.014	(0.600, 0.653)				0.614	0.015	(0.587, 0.644)
MS				− 0.129	0.014	(− 0.156, − 0.101)				− 0.117	0.015	(− 0.145, − 0.089)
CS				− 0.098	0.013	(− 0.124, − 0.073)				− 0.094	0.014	(− 0.119, − 0.066)
AT				− 0.138	0.014	(− 0.165, − 0.109)				− 0.140	0.015	(− 0.170, − 0.112)
LS				− 0.404	0.012	(− 0.427, − 0.380)				− 0.403	0.013	(− 0.428, − 0.378)
Grade 7												
HS				0.609	0.012	(0.584, 0.633)				0.615	0.013	(0.589, 0.640)
MS				− 0.156	0.012	(− 0.180, − 0.131)				− 0.158	0.013	(− 0.184, − 0.133)
CS				− 0.053	0.012	(− 0.076, − 0.030)				− 0.054	0.012	(− 0.078, − 0.031)
AT				− 0.111	0.013	(− 0.136, − 0.086)				− 0.113	0.013	(− 0.138, − 0.088)
LS				− 0.406	0.012	(− 0.428, − 0.382)				− 0.406	0.012	(− 0.429, − 0.382)
Grade 9												
HS				0.631	0.014	(0.604, 0.658)				0.628	0.014	(0.599, 0.656)
MS				− 0.162	0.014	(− v0.188, − 0.135)				− 0.155	0.014	(− 0.184, − 0.128)
CS				− 0.043	0.013	(− 0.068, − 0.018)				− 0.041	0.013	(− 0.066, − 0.015)
AT				− 0.148	0.014	(− 0.174, − 0.120)				− 0.155	0.014	(− 0.183, − 0.127)
LS				− 0.398	0.013	(− 0.425, − 0.372)				− 0.396	0.013	(− 0.421, − 0.370)
Variance components of the random effects												
Grade 5	0.542	0.072	(0.401, 0.689)	0.391	0.050	(0.291, 0.490)	0.490	0.069	(0.368, 0.666)	0.337	0.046	(0.255, 0.438)
Grade 7	0.126	0.015	(0.103, 0.157)	0.094	0.012	(0.074, 0.119)	0.123	0.013	(0.101, 0.152)	0.098	0.014	(0.078, 0.133)
Grade 9	0.307	0.045	(0.216, 0.390)	0.221	0.028	(0.174, 0.282)	0.255	0.035	(0.202, 0.340)	0.204	0.029	(0.154, 0.266)

Table 5 (continued)

	Model 3a			Model 3b			Model 4a			Model 4b		
	M	**SD**	**95% PPI**	**M**	**SD**	**95% PPI**	**M**	**SD**	**95% PPI**	**M**	**SD**	**95% PPI**
Correlations and Covariances between latent factors	Correlation		Covariance	Correlation		Covariance	Correlation		Covariance	Correlation		Covariance
Grade 5 with Grade 7	0.754		0.188	0.753		0.194	0.743		0.178	0.751		0.192
Grade 5 with Grade 9	0.771		0.281	0.767		0.304	0.766		0.242	0.766		0.275
Grade 7 with Grade 9	0.810		0.168	0.808		0.155	0.807		0.134	0.809		0.142
PPP	0.005			0.003			0.117			0.045		

Standardized results are presented for fixed effects. *M* posterior mean, *SD* posterior standard deviation, *PPI* posterior probability interval (2.5 and 97.5 percentile of the posterior distribution). School type covariates: *HS* high schools, *MS* middle secondary schools, *CS* comprehensive schools, *AT* schools offering all school tracks except high school, *LS* lower secondary schools, *PPP* posterior predictive p value. In Models 3a and 3b, no school-type specific item discrimination or item difficulty estimates were accounted for. In Models 4a and 4b, school-type specific item discrimination and item difficulty estimates were taken into account for items exhibiting DIF

Correlations and covariances were reported based on sample statistics

4a show that in seventh and ninth grade the comprehensive school type performed significantly better than the middle secondary schools and schools offering all school tracks except high school.

Additionally, we replicated the analyses of longitudinal reading competence development using point estimates of student reading competence. The point estimates are the linked weighted maximum likelihood estimates (WLE; Warm, 1989) as provided by NEPS and we performed linear growth modelling with and without cluster robust standard error estimation. Results are presented in Additional file 1: Tables S3–S5. As before, these results support our main findings on the pattern of competence development between German secondary school types and the three distinct school type groups. When it was not accounted for the clustered data structure, the misleading finding resulted that the comprehensive schools performed significantly better in seventh and ninth grade than middle secondary schools and schools offering all school tracks except high school.

Discussion

We evaluated measurement invariance between German secondary school types and tested the sensitivity of longitudinal comparisons to the found measurement non-invariance. Differences in reading competence between German secondary school types from fifth to ninth grade were investigated, while reading competence was modeled as a latent variable with measurement error taken into account. Multilevel modeling was employed to account for the clustered data structure, and measurement invariance between school types was assessed. Based on our results, partial invariance between school types is assumed (i.e., more than half of the items were measurement invariant/ free of DIF; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000).

The results on the longitudinal estimation of reading competence revealed a parallel pattern between German secondary school types, and that pattern remained when school-type-specific item estimates were included for items exhibiting DIF. Nevertheless, estimations of the same models without consideration of the clustered data structure led to misleading assumptions about the pattern of longitudinal reading competence development. In these models, students attending the comprehensive school type are estimated to be significantly better in seventh and ninth grade than students attending the middle secondary school type and those attending schools offering all school tracks except high school. For research focusing on school type comparisons of latent competence, we emphasize the use of hierarchical modeling when a nested data structure is present.

Furthermore, although we recommend the assessment of measurement invariance, it is not (or not only) a statistical question whether an item induces bias for group comparisons. Rather, procedures for measurement invariance testing are at best part of the test development process, including expert reviews on items exhibiting DIF (Camilli, 1993). Items that are measurement non-invariant and judged to be associated with construct irrelevant factors are revised or replaced throughout the test development process. Robitzsch and Lüdtke (2020) provide a thoughtful discussion on the reasoning behind (partial) measurement invariance for group comparison under construct relevant DIF and DIF caused by construct irrelevant factors.

Information about the amount of item bias for a developed test is also useful to quantify the uncertainty in group comparisons, which is analogous to the report of linking errors in longitudinal large-scale assessments (cf. Robitzsch & Lüdtke, 2020). While the assumption of exact item parameter invariance across groups is quite strict, we presented a method to assess the less strict approach of partial measurement invariance. Even when a measured construct is only partially invariant, comparisons of school types can be valid. Nevertheless, no statistical method alone can define construct validity without further theoretical reasoning and expert evaluation. As demonstrated in this study, the sensitivity of longitudinal reading competence development to partial measurement invariance between school types can be assessed.

Implications for research on the achievement gap in reading competence

Studies on reading competence development have presented either parallel development (e.g., Retelsdorf & Möller, 2008; Schneider & Stefanek, 2004) or a widening gap (e.g., Pfost & Artelt, 2013) among secondary school types. In these studies, samples were drawn from different regions (i.e., German federal states), and different methods of statistical analysis were used. We argued that group differences, such as school type effects, can be distorted by measurement non-invariance of test items. As these previous studies have not reported analyses of measurement invariance such as DIF, it is unknown whether the differences found relate to the psychometric properties of the administered tests. With our analyses, we found no indication that the pattern of competence development is affected by DIF. As a prerequisite for group-mean comparisons, studies should present evidence of measurement invariance between investigated groups and in the longitudinal case, across measurement occasions, or refer to the respective sources where these analyses are presented. Also, to enhance comparability of results across studies on reading competence development, researchers should discuss if the construct has the same meaning for all groups and over all measurement occasions. On a further note, the previous analyses were regionally limited and considered only one or two German federal states. In comparison, the sample we used is representative on a national level, and we encourage future research to strive to include more regions. Please note that the clustered data structure was always accounted for in previous analyses on reading competence development through cluster robust maximum likelihood estimation. When the focus is on regression coefficients and variance partitioning or inference on the cluster-level is not of interest, researchers need to make less assumptions of their data when choosing the cluster robust maximum likelihood estimation approach, as compared to hierarchical linear modeling (McNeish et al., 2017; Stapleton et al., 2016). As mentioned before, inaccurate standard errors and biased significance tests can result when hierarchical structures are ignored during estimation (Hox, 2002; Raudenbush & Bryk, 2002). As a result, standard errors are underestimated and the confidence intervals are narrower than they actually are, and effects become statistically significant more easily. As our results showed, ignoring the clustered data structure can result in misleading conclusions about the pattern of longitudinal reading competence development in comparisons of German secondary school types.

Limitations

One focus of our study was to investigate the consequences for longitudinal measurements of latent competence when partial invariance is taken into account in the estimation model. It was assumed that the psychometric properties of the scale and the underlying relationship among variables can be affected when some items are non-invariant and thus unfair between school types. With the NEPS study design for reading competence measurement, this assumption cannot be entirely tested, as for each measurement occasion, a completely new set of items is administered to circumvent memory effects. The three measurement occasions are linked through a mean/mean linking approach based on an anchor-group design (Fischer et al., 2016, 2019). Hence, a unique linking constant is assumed to hold for all school types. The computation of the linking constant relies on the assumption that items are invariant across all groups under investigation (e.g., school types). Due to data restrictions, as the data from the additional linking studies are not published by NEPS, we could not investigate the effect of item non-invariance across school types on the computation of linking constants. Therefore, we cannot test the assumption that the scale score metric, upon which the linking constant is computed, holds across measurement occasions for the school clusters and the school types under study. Overall, we assume that high effort was invested in the item and test construction for the NEPS. However, we can conclude that the longitudinal competence measurement is quite robust against the findings presented here regarding measurement non-invariance between school types, as the same measurement instruments are used to create the linking constants. Whenever possible, we encourage researchers to additionally assess measurement invariance across repeated measurements.

On a more general note, and looking beyond issues of statistical modeling, the available information on school types for our study is not exhaustive, as the German secondary school system is very complex and offers several options for students regarding schooling trajectories. A detailed variable on secondary school types and an identification of students who change school types between measurement occasions is desired but difficult to provide for longitudinal analyses (Bayer et al., 2014). As we use the school type information that generated the strata for the sampling of students, this information is constant over measurement occasions, but the comparability for later measurement timepoints (e.g., ninth grade) is rather limited.

Conclusion

In summary, it was assumed that school-level differences in measurement constructs may impact the longitudinal measurement of reading competence development. Therefore, we assessed measurement invariance between school types. Differences in item estimates between school types were found for each of the three measurement occasions. Nevertheless, taking these differences in item discrimination and difficulty estimates into account did not alter the parallel pattern of reading competence development when comparing German secondary school types from fifth to ninth grade. Furthermore, the necessity of taking the hierarchical data structure into account when comparing competence development across the school types was demonstrated. Ignoring the fact that students are nested within schools by sampling design in the estimation led to an overestimation of the statistical significance of the effects for the comprehensive school type in seventh and ninth grade.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40536-021-00116-2.

Additional file 1: Table S1. Results of models for longitudinal competence measurement (N= 7276) with cluster robust standard error estimation. **Table S2**. Effect sizes (Cohen's d) for school type covariates per estimated model. **Table S3**. Results of models for longitudinal competence development using WLEs (N= 7276) with cluster robust standard error estimation. **Table S4**. Results of models for longitudinal competence development using WLEs (N= 7276) without cluster robust standard error estimation. **Table S5**. Effect sizes (Cohen's d) for school type covariates per estimated model.

Acknowledgements

The authors would like to thank David Kaplan for helpful suggestions on the analysis of the data. We would also like to thank Marie-Ann Sengewald for consultation on latent variable modelling.

Authors' contributions

TR analyzed and interpreted the data used in this study. TR conducted the literature review and drafted significant parts of the manuscript. CHC, LF and TG substantially revised the manuscript and provided substantial input regarding the statistical analyses. All authors read and approved the final manuscript.

Funding

This research project was partially funded by the Deutsche Forschungsgemeinschaft (DFG; http://www.dfg.de) within Priority Programme 1646 entitled "A Bayesian model framework for analyzing data from longitudinal large-scale assessments" under Grant No. CA 289/8–2 (awarded to Claus H. Carstensen). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Availability of data and materials

The data analyzed in this study and documentation are available at doi: https://doi.org/10.5157/NEPS:SC3:9.0.0. Moreover, the syntax used to generate the reported results is provided in an online repository at https://osf.io/5ugwn/?view_only=327ba9ae72684d07be8b4e0c6e6f1684.

This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort Grade 5, doi: https://doi.org/10.5157/NEPS:SC3:9.0.0. From 2008 to 2013, NEPS data was collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

Declarations

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Leibniz Institute for Educational Trajectories, Wilhelmsplatz 3, 96047 Bamberg, Germany. [2]University of Bamberg, Bamberg, Germany. [3]Johannes Kepler University Linz, Linz, Austria.

Received: 7 December 2020 Accepted: 15 October 2021

Published online: 28 October 2021

References

Asparouhov, T., & Muthén, B. (2010). Bayesian analysis using Mplus: Technical implementation (Mplus Technical Report). http://statmodel.com/download/Bayes3.pdf. Accessed 12 November 2020.

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(4), 495–508. https://doi.org/10.1080/10705511.2014.919210

Aßmann, C., Steinhauer, H. W., Kiesl, H., Koch, S., Schönberger, B., Müller-Kuller, A., Rohwer, G., Rässler, S., & Blossfeld, H.-P. (2011). 4 Sampling designs of the National Educational Panel Study: Challenges and solutions. *Zeitschrift Für Erziehungswissenschaft, 14*(S2), 51–65. https://doi.org/10.1007/s11618-011-0181-8

Bast, J., & Reitsma, P. (1998). Analyzing the development of individual differences in terms of Matthew effects in reading: Results from a Dutch longitudinal study. *Developmental Psychology, 34*(6), 1373–1399. https://doi.org/10.1037/0012-1649.34.6.1373

Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Stanat, P., Tillmann, K.-J., & Weiß, M. (2001). *PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich*. Leske + Budrich. https://doi.org/10.1007/978-3-322-83412-6

Baumert, J., Stanat, P., & Watermann, R. (2006). Schulstruktur und die Entstehung differenzieller Lern- und Entwicklungsmilieus. In J. Baumert, P. Stanat, & R. Watermann (Eds.), *Herkunftsbedingte Disparitäten im Bildungssystem* (pp. 95–188). VS Verlag für Sozialwissenschaften.

Baumert, J., Trautwein, U., & Artelt, C. (2003). Schulumwelten—institutionelle Bedingungen des Lehrens und Lernens. In J. Baumert, C. Artelt, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, K.-J. Tillmann, & M. Weiß (Eds.), *PISA 2000. Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland* (pp. 261–331). Leske u. Budrich.

Bayer, M., Goßmann, F., & Bela, D. (2014). NEPS technical report: Generated school type variable t723080_g1 in Starting Cohorts 3 and 4 (NEPS Working Paper No. 46). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study. https://www.neps-data.de/Portals/0/Working%20Papers/WP_XLVI.pdf. Accessed 12 November 2020.

Becker, M., Lüdtke, O., Trautwein, U., & Baumert, J. (2006). Leistungszuwachs in Mathematik. *Zeitschrift Für Pädagogische Psychologie, 20*(4), 233–242. https://doi.org/10.1024/1010-0652.20.4.233

Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (Eds.), (2011). Education as a lifelong process: The German National Educational Panel Study (NEPS) [Special Issue]. *Zeitschrift für Erziehungswissenschaft*, 14.

Bos, W., Bonsen, M., & Gröhlich, C. (2009). *KESS 7 Kompetenzen und Einstellungen von Schülerinnen und Schülern an Hamburger Schulen zu Beginn der Jahrgangsstufe 7. HANSE—Hamburger Schriften zur Qualität im Bildungswesen* (Vol. 5). Waxmann.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. Guilford Press.

Camilli, G. (1993). The case against item bias detection techniques based on internal criteria: Do item bias procedures obscure test fairness issues? In P. W. Holland & H. Wainer (Eds.), *Differential item functioning: Theory and practice* (pp. 397–417). Erlbaum.

Camilli, G. (2006). Test fairness. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 221–256). American Council on Education and Praeger.

Chall, J. S. (1983). *Stages of reading development*. McGraw-Hill.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. Academic Press.

Cortina, K. S., & Trommer, L. (2009). *Bildungswege und Bildungsbiographien in der Sekundarstufe I. Das Bildungswesen in der Bundesrepublik Deutschland: Strukturen und Entwicklungen im Überblick*. Waxmann.

Ditton, H., Krüsken, J., & Schauenberg, M. (2005). Bildungsungleichheit—der Beitrag von Familie und Schule. *Zeitschrift Für Erziehungswissenschaft, 8*(2), 285–304. https://doi.org/10.1007/s11618-005-0138-x

Edossa, A. K., Neuenhaus, N., Artelt, C., Lingel, K., & Schneider, W. (2019). Developmental relationship between declarative metacognitive knowledge and reading comprehension during secondary school. *European Journal of Psychology of Education, 34*(2), 397–416. https://doi.org/10.1007/s10212-018-0393-x

Finch, W. H., & Bolin, J. E. (2017). *Multilevel Modeling using Mplus*. Chapman and Hall—CRC.

Fischer, L., Gnambs, T., Rohm, T., & Carstensen, C. H. (2019). Longitudinal linking of Rasch-model-scaled competence tests in large-scale assessments: A comparison and evaluation of different linking methods and anchoring designs based on two tests on mathematical competence administered in grades 5 and 7. *Psychological Test and Assessment Modeling, 61*, 37–64.

Fischer, L., Rohm, T., Gnambs, T., & Carstensen, C. H. (2016). Linking the data of the competence tests (NEPS Survey Paper No. 1). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study. https://www.lifbi.de/Portals/0/Survey%20Papers/SP_I.pdf. Accessed 12 November 2020.

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. Springer.

Fox, J.-P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using gibbs sampling. *Psychometrika, 66*, 271–288.

Gamoran, A., & Mare, R. D. (1989). Secondary school tracking and educational inequality: Compensation, reinforcement, or neutrality? *American Journal of Sociology, 94*(5), 1146–1183. https://doi.org/10.1086/229114

Gehrer, K., Zimmermann, S., Artelt, C., & Weinert, S. (2003). NEPS framework for assessing reading competence and results from an adult pilot study. *Journal for Educational Research Online, 5*, 50–79.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Chapman & Hall.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple Sequences. *Statistical Science, 7*, 457–472.

Heck, R. H., Price, C. L., & Thomas, S. L. (2004). Tracks as emergent structures: A network analysis of student differentiation in a high school. *American Journal of Education, 110*(4), 321–353. https://doi.org/10.1086/422789

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Routledge. https://doi.org/10.4324/9780203357811

Hox, J. J. (2002). *Multilevel analysis: Techniques and applications. Quantitative methodology series*. Erlbaum.

Jak, S., & Jorgensen, T. (2017). Relating measurement invariance, cross-level invariance, and multilevel reliability. *Frontiers in Psychology, 8*, 1640. https://doi.org/10.3389/fpsyg.2017.01640

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*(4), 409–426. https://doi.org/10.1007/BF02291366

Kamata, A., & Vaughn, B. K. (2010). Multilevel IRT modeling. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 41–57). Routledge.

Kaplan, D., Kim, J.-S., & Kim, S.-Y. (2009). Multilevel latent variable modeling: Current research and recent developments. In R. E. Millsap & A. Maydeu-Olivares (Eds.), *The Sage handbook of quantitative methods in psychology* (pp. 592–612). Sage Publications Ltd. https://doi.org/10.4135/9780857020994.n24

Kim, E., Cao, C., Wang, Y., & Nguyen, D. (2017). Measurement invariance testing with many groups: A comparison of five approaches. *Structural Equation Modeling: A Multidisciplinary Journal*. https://doi.org/10.1080/10705511.2017.1304822

Köller, O., & Baumert, J. (2001). Leistungsgruppierungen in der Sekundarstufe I. Ihre Konsequenzen für die Mathematikleistung und das mathematische Selbstkonzept der Begabung. *Zeitschrift Für Pädagogische Psychologie, 15*, 99–110. https://doi.org/10.1024//1010-0652.15.2.99

Köller, O., & Baumert, J. (2002). Entwicklung von Schulleistungen. In R. Oerter & L. Montada (Eds.), *Entwicklungspsychologie* (pp. 735–768). Beltz/PVU.

Krannich, M., Jost, O., Rohm, T., Koller, I., Carstensen, C. H., Fischer, L., & Gnambs, T. (2017). NEPS Technical report for reading—scaling results of starting cohort 3 for grade 7 (NEPS Survey Paper No. 14). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study. https://www.neps-data.de/Portals/0/Survey%20Papers/SP_XIV.pdf. Accessed 12 November 2020.

Lehmann, R., Gänsfuß, R., & Peek, R. (1999). *Aspekte der Lernausgangslage und der Lernentwicklung von Schülerinnen und Schülern an Hamburger Schulen: Klassenstufe 7; Bericht über die Untersuchung im September 1999*. Hamburg: Behörde für Schule, Jugend und Berufsbildung, Amt für Schule.

Lehmann, R. H., & Lenkeit, J. (2008). *ELEMENT. Erhebung zum Lese- und Mathematikverständnis. Entwicklungen in den Jahrgangsstufen 4 bis 6 in Berlin*. Berlin: Senatsverwaltung für Bildung, Jugend und Sport.

LeTendre, G. K., Hofer, B. K., & Shimizu, H. (2003). What Is tracking? Cultural expectations in the United States, Germany, and Japan. *American Educational Research Journal, 40*(1), 43–89. https://doi.org/10.3102/00028312040001043

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17*, 179–193.

Lu, I. R. R., Thomas, D. R., & Zumbo, B. D. (2005). Embedding IRT in structural equation models: A comparison with regression based on IRT scores. *Structural Equation Modeling: A Multidisciplinary Journal, 12*(2), 263–277. https://doi.org/10.1207/s15328007sem1202_5

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods, 13*, 203–229.

Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2x2 taxonomy of multilevel latent contextual model: Accuracy-bias trade-offs in full and partial error correction models. *Psychological Methods, 16*, 444–467.

Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B., & Nagengast, B. (2009). Doubly-latent models of school contextual effects: Integrating multilevel and structural equation approaches to control measurement and sampling error. *Multivariate Behavioral Research, 44*, 764–802.

McNeish, D., Stapleton, L. M., & Silverman, R. D. (2017). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods, 22*(1), 114–140. https://doi.org/10.1037/met0000078

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*(4), 297–334. https://doi.org/10.1177/014662169301700401

Muthén, B., & Asparouhov, T. (2012). Bayesian SEM: A more flexible representation of substantive theory. *Psychological Methods, 17*, 313–335.

Muthén, B., & Asparouhov, T. (2014). IRT studies of many groups: The alignment method. *Frontiers in Psychology, 5*, 978. https://doi.org/10.3389/fpsyg.2014.00978

Muthén, L.K. and Muthén, B.O. (1998–2020). *Mplus User's Guide* (8th ed.), Los Angeles, CA: Muthén and Muthén.

Nagy, G., Retelsdorf, J., Goldhammer, F., Schiepe-Tiska, A., & Lüdtke, O. (2017). Veränderungen der Lesekompetenz von der 9. zur 10. Klasse: Differenzielle Entwicklungen in Abhängigkeit der Schulform, des Geschlechts und des soziodemografischen Hintergrunds? *Zeitschrift Für Erziehungswissenschaft, 20*(S2), 177–203. https://doi.org/10.1007/s11618-017-0747-1

Naumann, J., Artelt, C., Schneider, W. & Stanat, P. (2010). Lesekompetenz von PISA 2000 bis PISA 2009. In E. Klieme, C. Artelt, J. Hartig, N. Jude, O. Köller, M. Prenzel (Eds.), *PISA 2009. Bilanz nach einem Jahrzehnt*. Münster: Waxmann. https://www.pedocs.de/volltexte/2011/3526/pdf/DIPF_PISA_ISBN_2450_PDFX_1b_D_A.pdf. Accessed 12 November 2020.

Neumann, M., Schnyder, I., Trautwein, U., Niggli, A., Lüdtke, O., & Cathomas, R. (2007). Schulformen als differenzielle Lernmilieus. *Zeitschrift Für Erziehungswissenschaft, 10*(3), 399–420. https://doi.org/10.1007/s11618-007-0043-6

O'Brien, D. G., Moje, E. B., & Stewart, R. A. (2001). Exploring the context of secondary literacy: Literacy in people's everyday school lives. In E. B. Moje & D. G. O'Brien (Eds.), *Constructions of literacy: Studies of teaching and learning in and out of secondary classrooms* (pp. 27–48). Erlbaum.

Oakes, J., & Wells, A. S. (1996). *Beyond the technicalities of school reform: Policy lessons from detracking schools*. UCLA Graduate School of Education & Information Studies.

OECD. (2017). *PISA 2015 assessment and analytical framework: science, reading, mathematic, financial literacy and collaborative problem solving*. OECD Publishing. https://doi.org/10.1787/9789264281820-en

OECD & Statistics Canada. (1995). *Literacy, economy and society: Results of the first international adult literacy survey*. OECD Publishing.

Pfost, M., & Artelt, C. (2013). Reading literacy development in secondary school and the effect of differential institutional learning environments. In M. Pfost, C. Artelt, & S. Weinert (Eds.), *The development of reading literacy from early childhood to adolescence empirical findings from the Bamberg BiKS longitudinal studies* (pp. 229–278). Bamberg: University of Bamberg Press.

Pfost, M., Hattie, J., Dörfler, T., & Artelt, C. (2014). Individual differences in reading development: A review of 25 years of empirical research on Matthew effects in reading. *Review of Educational Research, 84*(2), 203–244. https://doi.org/10.3102/0034654313509492

Pfost, M., Karing, C., Lorenz, C., & Artelt, C. (2010). Schereneffekte im ein- und mehrgliedrigen Schulsystem: Differenzielle Entwicklung sprachlicher Kompetenzen am Übergang von der Grund- in die weiterführende Schule? *Zeitschrift Für Pädagogische Psychologie, 24*(3–4), 259–272. https://doi.org/10.1024/1010-0652/a000025

Pohl, S., Haberkorn, K., Hardt, K., & Wiegand, E. (2012). NEPS technical report for reading—scaling results of starting cohort 3 in fifth grade (NEPS Working Paper No. 15). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

Protopapas, A., Parrila, R., & Simos, P. G. (2016). In Search of Matthew effects in reading. *Journal of Learning Disabilities, 49*(5), 499–514. https://doi.org/10.1177/0022219414559974

Rabe-Hesketh, S., Skrondal, A., & Zheng, X. (2007). Multilevel Structural Equation Modeling. In S.-Y. Lee (Ed.), *Handbook of Latent Variable and Related Models* (pp. 209–227). Elsevier.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods. Advanced quantitative techniques in the social sciences*, (Vol. 1). Thousand Oaks, CA.: Sage Publ.

Raykov, T. (1999). Are simple change scores obsolete? An approach to studying correlates and predictors of change. *Applied Psychological Measurement, 23*(2), 120–126. https://doi.org/10.1177/01466219922031248

Retelsdorf, J., Becker, M., Köller, O., & Möller, J. (2012). Reading development in a tracked school system: A longitudinal study over 3 years using propensity score matching. *The British Journal of Educational Psychology, 82*(4), 647–671. https://doi.org/10.1111/j.2044-8279.2011.02051.x

Retelsdorf, J., & Möller, J. (2008). Entwicklungen von Lesekompetenz und Lesemotivation: Schereneffekte in der Sekundarstufe? *Zeitschrift Für Entwicklungspsychologie Und Pädagogische Psychologie, 40*(4), 179–188. https://doi.org/10.1026/0049-8637.40.4.179

Robitzsch, A., & Lüdtke, O. (2020). A review of different scaling approaches under full invariance, partial invariance, and noninvariance for cross-sectional country comparisons in large-scale assessments. *Psychological Test and Assessment Modeling*, 62(2), 233–279. https://www.psychologie-aktuell.com/fileadmin/Redaktion/Journale/ptam-2020-2/03_Robitzsch.pdf

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley. https://doi.org/10.1002/9780470316696

Scharl, A., Fischer, L., Gnambs, T., & Rohm, T. (2017). NEPS Technical report for reading: scaling results of starting cohort 3 for grade 9 (NEPS Survey Paper No. 20). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study. https://www.neps-data.de/Portals/0/Survey%20Papers/SP_XX.pdf. Accessed 12 November 2020.

Schneider, W., & Stefanek, J. (2004). Entwicklungsveränderungen allgemeiner kognitiver Fähigkeiten und schulbezogener Fertigkeiten im Kindes- und Jugendalter. *Zeitschrift Für Entwicklungspsychologie Und Pädagogische Psychologie, 36*(3), 147–159. https://doi.org/10.1026/0049-8637.36.3.147

Schweig, J. (2014). Cross-level measurement invariance in school and classroom environment surveys: Implications for policy and practice. *Educational Evaluation and Policy Analysis, 36*(3), 259–280. https://doi.org/10.3102/0162373713509880

Silva, C., Bosancianu, B. C. M., & Littvay, L. (2019). *Multilevel Structural Equation Modeling*. Sage.

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21*(4), 360–407. https://doi.org/10.1598/RRQ.21.4.1

Stapleton, L. M., McNeish, D. M., & Yang, J. S. (2016). Multilevel and single-level models for measured and latent variables when data are clustered. *Educational Psychologist, 51*(3–4), 317–330. https://doi.org/10.1080/00461520.2016.1207178

Steenkamp, J. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, 25*, 78–90. https://doi.org/10.1086/209528

Steinhauer, H. W. & Zinn, S. (2016). NEPS technical report for weighting: Weighting the sample of starting cohort 3 of the national educational panel study (Waves 1 to 3) (NEPS Working Paper No. 63). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study. https://www.neps-data.de/Portals/0/Working%20Papers/WP_LXIII.pdf. Accessed 12 November 2020.

Steyer, R., Partchev, I., & Shanahan, M. J. (2000). Modeling True Intraindividual Change in Structural Equation Models: The Case of Poverty and Children's Psychosocial Adjustment. In T. D. Little, K. U. Schnabel, & J. Baumert (Eds.), *Modeling longitudinal and multilevel data: Practical issues, applied approaches and specific examples* (pp. 109–26). Mahwah, N.J.: Lawrence Erlbaum Associates. https://www.metheval.uni-jena.de/materialien/publikationen/steyer_et_al.pdf. Accessed 12 November 2020.

Sweeney, R. E., & Ulveling, E. F. (1972). A Transformation for simplifying the interpretation of coefficients of binary variables in regression analysis. *The American Statistician, 26*(5), 30–32. https://doi.org/10.2307/2683780

Te Grotenhuis, M., Pelzer, B., Eisinga, R., Nieuwenhuis, R., Schmidt-Catran, A., & Konig, R. (2017). When size matters: Advantages of weighted effect coding in observational studies. *International Journal of Public Health, 62*(1), 163–167. https://doi.org/10.1007/s00038-016-0901-1

van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthén, B. (2013). Facing off with Scylla and Charybdis: A comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in Psychology, 4*, 770. https://doi.org/10.3389/fpsyg.2013.00770

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4–70. https://doi.org/10.1177/109442810031002

Walberg, H. J., & Tsai, S.-L. (1983). Matthew effects in education. *American Educational Research Journal, 20*(3), 359–373. https://doi.org/10.2307/1162605

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

Weis, M., Doroganova, A., Hahnel, C., Becker-Mrotzek, M., Lindauer, T., Artelt, C., & Reiss, K. (2020). Aktueller Stand der Lesekompetenz in PISA 2018. In K. Reiss, M. Weis & A Schiepe-Tiska (Hrsg). *Schulmanagement Handbuch* (pp. 9–19). München: Cornelsen. https://www.pisa.tum.de/fileadmin/w00bgi/www/_my_direct_uploads/PISA_Bericht_2018_.pdf. Accessed 12 November 2020.

Weis, M., Zehner, F., Sälzer, C., Strohmeier, A., Artelt, C., & Pfost, M. (2016). Lesekompetenz in PISA 2015: Ergebnisse, Veränderungen und Perspektiven. In K. Reiss, C. Sälzer, A. Schiepe-Tiska, E. Klieme & O. Köller (Eds.), *PISA 2015—Eine Studie zwischen Kontinuität und Innovation* (pp. 249–283). Münster: Waxmann.

Williamson, G. L., Appelbaum, M., & Epanchin, A. (1991). Longitudinal analyses of academic achievement. *Journal of Educational Measurement, 28*(1), 61–76. https://doi.org/10.1111/j.1745-3984.1991.tb00344.x

Publisher's Note

The achievement gap in reading competence: The effect of measurement non-invariance across school types

Supplementary material

List of Tables

Table S1. Results of models for longitudinal competence measurement (N= 7,276) with cluster robust standard error estimation.

	Model S1a			**Model S1b**			**Model S2a**			**Model S2b**		
	Estim.	S.E.	95% CI	Estim.	S.E.	95% CI	Estim.	S.E.	95% CI	Estim.	S.E.	95% CI
School Covariates												
Grade 5												
HS				0.632	0.027	(0.578, 0.685)				0.618	0.028	(0.563, 0.674)
MS				-0.134	0.024	(-0.180, -0.088)				-0.123	0.024	(-0.171, -0.075)
CS				-0.097	0.039	(-0.173, -0.021)				-0.094	0.038	(-0.169, -0.018)
AT				-0.140	0.028	(-0.195, -0.086)				-0.141	0.027	(-0.195, -0.088)
LS				-0.400	0.018	(-0.436, -0.364)				-0.399	0.019	(-0.437, -0.361)
Grade 7												
HS				0.611	0.025	(0.563, 0.660)				0.617	0.025	(0.568, 0.665)
MS				-0.159	0.023	(-0.204, -0.115)				-0.162	0.023	(-0.206, -0.117)
CS				-0.054	0.029	(-0.111, 0.002)				-0.056	0.028	(-0.111, 0.000)
AT				-0.113	0.022	(-0.155, -0.070)				-0.114	0.022	(-0.157, -0.072)
LS				-0.402	0.016	(-0.433, -0.371)				-0.403	0.016	(-0.434, -0.372)
Grade 9												
HS				0.631	0.028	(0.577, 0.685)				0.628	0.027	(0.574, 0.682)
MS				-0.165	0.023	(-0.209, -0.120)				-0.159	0.023	(-0.204, -0.113)
CS				-0.045	0.031	(-0.105, 0.015)				-0.042	0.031	(-0.103, 0.018)
AT				-0.148	0.026	(-0.199, -0.097)				-0.154	0.027	(-0.207, -0.102)
LS				-0.393	0.019	(-0.430, -0.355)				-0.392	0.020	(-0.432, -0.352)
Variance components of random effects												
Grade 5	3.042	0.419	(2.221, 3.862)	2.054	0.286	(1.494, 2.614)	2.989	0.425	(2.157, 3.822)	2.046	0.294	(1.468, 2.623)
Grade 7	0.191	0.049	(0.094, 0.288)	0.137	0.037	(0.065, 0.210)	0.188	0.070	(0.051, 0.325)	0.136	0.048	(0.042, 0.230)
Grade 9	1.029	0.161	(0.714, 1.344)	0.729	0.115	(0.503, 0.955)	1.051	0.170	(0.719, 1.384)	0.750	0.120	(0.514, 0.985)
Correlations and covariances between latent factors	Correlation		Covariance	Correlation		Covariance	Correlation		Covariance	Correlation		Covariance
Grade 5 with Grade 7	0.893		0.534	0.889		0.528	0.889		0.518	0.888		0.521
Grade 5 with Grade 9	0.915		1.227	0.910		1.215	0.915		1.221	0.911		1.223
Grade 7 with Grade 9	0.922		0.322	0.922		0.327	0.922		0.322	0.923		0.331
Model Fit Information												
AIC	470006			467556			474113			471748		
BIC	471916			469547			476807			474526		

Note. Standardized results are presented for fixed effects. Estim. = Estimated model parameters from robust maximum likelihood estimation. S.E. = Standard error. CI = Confidence interval. School type covariates: HS = high schools, MS = middle secondary schools, CS = comprehensive schools, AT = schools offering all school tracks except high school, LS = lower secondary schools. In Models S1a and S1b, no school-type specific item discrimination or item difficulty estimates were accounted for. In Models S2a and S2b, school-type specific item discrimination and item difficulty estimates were taken into account for items exhibiting DIF.
AIC = Akaike information criterion, BIC = Bayes information criterion. Correlations and covariances were reported based on sample statistics.

Table S2. Effect sizes (Cohen's *d*) for school type covariates per estimated model.

Type of School		Model S1b	Model S2b
Grade 5			
	HS	1.23	1.20
	MS	0.06	0.08
	CS	-0.12	-0.11
	AT	-0.14	-0.15
	LS	-1.00	-1.00
Grade 7			
	HS	1.14	1.16
	MS	-0.05	-0.05
	CS	0.05	0.05
	AT	-0.09	-0.09
	LS	-1.05	-1.05
Grade 9			
	HS	1.18	1.18
	MS	-0.06	-0.04
	CS	0.13	0.15
	AT	-0.21	-0.24
	LS	-1.02	-1.01

Note. School level covariates: HS = high schools, MS = middle secondary schools, CS = comprehensive schools, AT = schools offering all school tracks except high school, LS = lower secondary schools. Cohen's *d* effect size: calculated as the difference per value from the average of all other school type effects and divided by the square root of the variance per respective latent factor. The average of all school type effects can differ slightly from zero due to effect coding and model re-estimation using the reference group to obtain a reference group estimate. The clustered data structure was taken into account in both models. In Models S1b, school-type specific item discrimination or item difficulty estimates were accounted for. In Model S2b school-type specific item discrimination and item difficulty estimates were taken into account for items exhibiting DIF.

Table S3. Results of models for longitudinal competence development using WLEs (N= 7, 276) with cluster robust standard error estimation.

	Model S3a			**Model S3b**		
	Estim.	S.E.	95% CI	Estim.	S.E.	95% CI
Intercept	0.040	0.041	(-0.040, 0.121)	0.045	0.030	(-0.015, 0.104)
Competence Growth	1.479	0.087	(1.309, 1.649)	1.782	0.130	(1.527, 2.038)
School Effects						
Grade 5						
HS				0.546	0.032	(0.464, 0.607)
MS				-0.109	0.021	(-0.151, -0.067)
CS				-0.083	0.038	(-0.158, -0.009)
AT				-0.114	0.027	(-0.167, -0.062)
LS				-0.360	0.023	(-0.406, -0.315)
Grade 7						
HS				0.556	0.030	(0.497, 0.615)
MS				-0.139	0.022	(-0.183, -0.095)
CS				-0.048	0.028	(-0.102, 0.006)
AT				-0.097	0.022	(-0.141, -0.053)
LS				-0.378	0.022	(-0.422, -0.334)
Grade 9						
HS				0.550	0.032	(0.487, 0.613)
MS				-0.142	0.021	(-0.183, -0.101)
CS				-0.034	0.027	(-0.086, 0.019)
AT				-0.129	0.026	(-0.181, -0.077)
LS				-0.348	0.024	(-0.396, -0.300)
Variance components						
Grade 5	0.306	0.045	(0.219, 0.394)	0.416	0.038	(0.341, 0.491)
Grade 7	0.863	0.030	(0.805, 0.922)	0.797	0.027	(0.744, 0.851)
Grade 9	0.157	0.037	(0.084, 0.229)	0.258	0.033	(0.193, 0.323)
Model Fit Information						
AIC	46684			44330		
BIC	46739			44468		
RMSEA	0.067			0.104		
CFI	0.993			0.982		
TLI	0.978			0.736		
SRMR	0.026			0.012		

Note. Standardized results are presented for fixed effects. Estim. = Estimated model parameters from robust maximum likelihood estimation. S.E. = Standard error. CI = Confidence interval. School type covariates: HS = high schools, MS = middle secondary schools, CS = comprehensive schools, AT = schools offering all school tracks except high school, LS = lower secondary schools. AIC = Akaike information criterion, BIC = Bayes information criterion. RMSEA = Root mean square error of approximation. CFI = Comparative fit index. TLI = Tucker–Lewis index. SRMR = Standardized root mean square residual.

Table S4. Results of models for longitudinal competence development using WLEs (N= 7, 276) without cluster robust standard error estimation.

	Model S4a			**Model S4b**		
	Estim.	SD	95% CI	Estim.	SD	95% CI
Intercept	0.040	0.015	(0.012, 0.069)	0.045	0.016	(0.013, 0.076)
Competence Growth	1.479	0.081	(1.321, 1.638)	1.782	0.126	(1.536, 2.029)
School Effects						
Grade 5						
HS				0.546	0.014	(0.519, 0.573)
MS				-0.088	0.014	(-0.115, -0.060)
CS				-0.064	0.014	(-0.090, -0.037)
AT				-0.050	0.015	(-0.079, -0.022)
LS				-0.310	0.013	(-0.336, -0.285)
Grade 7						
HS				0.556	0.013	(0.531, 0.581)
MS				-0.140	0.012	(-0.163, -0.118)
CS				-0.047	0.011	(-0.068, -0.026)
AT				-0.109	0.012	(-0.132, -0.086)
LS				-0.400	0.011	(-0.422, -0.378)
Grade 9						
HS				0.550	0.014	(0.522, 0.578)
MS				-0.109	0.014	(-0.137, -0.082)
CS				-0.011	0.013	(-0.036, 0.014)
AT				-0.116	0.014	(-0.142, -0.089)
LS				-0.359	0.014	(-0.386, -0.331)
Variance components						
Grade 5	0.306	0.039	(0.230, 0.383)	0.416	0.035	(0.347, 0.485)
Grade 7	0.863	0.022	(0.820, 0.906)	0.797	0.019	(0.759, 0.835)
Grade 9	0.157	0.033	(0.091, 0.222)	0.258	0.030	(0.199, 0.316)
Model Fit Information						
AIC	46684			44330		
BIC	46739			44468		
RMSEA	0.099			0.093		
CFI	0.984			0.991		
TLI	0.951			0.861		
SRMR	0.026			0.012		

Note. Standardized results are presented for fixed effects. Estim. = Estimated model parameters from robust maximum likelihood estimation. S.E. = Standard error. CI = Confidence interval. School type covariates: HS = high schools, MS = middle secondary schools, CS = comprehensive schools, AT = schools offering all school tracks except high school, LS = lower secondary schools. AIC = Akaike information criterion, BIC = Bayes information criterion. RMSEA = Root mean square error of approximation. CFI = Comparative fit index. TLI = Tucker–Lewis index. SRMR = Standardized root mean square residual.

Table S5. Effect sizes (Cohen's *d*) for school type covariates per estimated model.

Type of School	Model S3b	Model S4b
Grade 5		
HS	1.02	1.02
MS	0.07	0.07
CS	-0.10	-0.10
AT	-0.09	-0.09
LS	-0.87	-0.87
Grade 7		
HS	0.96	0.96
MS	-0.03	-0.03
CS	0.05	0.05
AT	-0.05	-0.05
LS	-0.92	-0.92
Grade 9		
HS	0.94	0.94
MS	-0.04	-0.04
CS	0.13	0.13
AT	-0.17	-0.17
LS	-0.83	-0.83

Note. School level covariates: HS = high schools, MS = middle secondary schools, CS = comprehensive schools, AT = schools offering all school tracks except high school, LS = lower secondary schools. Cohen's *d* effect size: calculated as the difference per value from the average of all other school type effects and divided by the square root of the variance per respective indicator (i.e., reading competence per measurement occasion). The average of all school type effects can differ slightly from zero due to effect coding and model re-estimation using the reference group to obtain a reference group estimate. The clustered data structure was taken into account for estimation of Model 3b but not for Model 4b.

Original Article

Comparing Perceptual Speed Between Educational Contexts

The Case of Students With Special Educational Needs

Timo Gnambs, Anna Scharl, and Theresa Rohm

Leibniz Institute for Educational Trajectories, Bamberg, Germany

Abstract. Perceptual speed is a basic component of cognitive functioning that allows people to efficiently process novel visual stimuli and quickly react to them. In educational studies, tests measuring perceptual speed are frequently developed using students from regular schools without considering students with special educational needs. Therefore, it is unclear whether these instruments allow valid comparisons between different school tracks. The present study on N = 3,312 students from the National Educational Panel Study evaluated differential item functioning (DIF) of a short test of perceptual speed between four school tracks in Germany (special, basic, intermediate, and upper secondary schools). Bayesian Rasch Poisson counts modeling identified negligible DIF that did not systematically disadvantage specific students. Moreover, the test reliabilities were comparable between school tracks. These results highlight that perceptual speed can be comparably measured in special schools, thus enabling educational researchers to study schooling effects in the German educational system.

Keywords: perceptual speed, differential item functioning, large-scale assessment, Poisson counts models, Bayesian analyses

Educational research frequently involves comparisons between different school tracks, for example, to evaluate the effect of different curricula or instructional approaches on academic achievement. To properly address these research questions, the administered measures must be comparable across contexts. Otherwise, comparisons are unfair and might lead to biased conclusions. In practice, systematic differences in the measurement properties of cognitive tests might be expected if tests are administered in contexts they have not been developed or validated for. Many commercially available tests and even custom-designed tests administered in educational large-scale assessments are developed using students attending regular schools. However, if these tests are also administered to students with cognitive impairments (e.g., with special educational needs [SENs]), the measured constructs might differ to some degree, for example, because these students process instructions or item contents differently than regular students (Nusser & Weinert, 2017; Pohl et al., 2016; Südkamp et al., 2015). Before comparing a measured cognitive ability between different educational contexts, measurement invariance must be demonstrated. Therefore, the present study evaluates differential item functioning (DIF) in a test for figural perceptual speed. Of particular interest are students with SENs in the area of learning who attend various special schools in Germany (cf. Heydrich et al., 2013). Using a latent variable approach in a Bayesian framework, we examined whether the test allows for fair comparisons between different school tracks.

Theoretical Background

Perceptual Speed as a Facet of Processing Speed

Processing speed is a component of cognitive functioning and represents the ability to quickly identify, discriminate, and decide about visual, auditory, or kinesthetic sensory information of different types of complexity (Holdnack, 2019). Measures of processing speed indicate how efficiently an individual can perform basic tasks in early stages of information processing. Slow processing speed may make a cognitive task (e.g., solving a math problem) more difficult, whereas higher speed can support thinking and learning. In the three-stratum model of cognitive abilities (Carroll, 1993), processing speed represents one of the eight broad abilities of which perceptual speed is one narrow facet. Perceptual speed indicates the automaticity and efficiency of processing novel visual information and the speed of decision-making. It is typically assessed with speeded tests that require the quick identification of specific targets from a set of stimuli. In these tests, perceptual speed is quantified either as the time until all targets are identified

https://doi.org/10.1027/2698-1866/a000013

or as the number of correctly identified targets per time. Perceptual speed is routinely assessed in various contexts because of its ability to predict various real-life outcomes. For example, in occupational and educational settings, it was associated with job (Mount et al., 2008) and school performance (Rindermann & Neubauer, 2004). Moreover, meta-analytic evidence highlighted the importance of processing speed for learning because children and adults with mathematical difficulties or reading disorders typically show substantially impaired speed performance compared to healthy comparison groups (Kudo et al., 2015; Peng et al., 2018). Thus, individual differences in perceptual speed can have a profound impact on learning outcomes and academic success.

The Educational System in Germany

Germany has a tiered system of educational tracking that separates children at the age of about 10 years by ability into different school tracks. Low-achieving students attend *Hauptschule* (basic secondary school) and receive simplified educational training up to the ninth school grade, while students in *Realschule* (intermediate secondary school) receive extended education combined with more practical elements that may lead to an apprenticeship after tenth grade. High-achieving students attend *Gymnasium* (upper secondary school) and receive more advanced instructions in the same academic subjects and qualify for university entrance after the twelfth grade. Exceptions are students with cognitive difficulties and, therefore, SENs, for example, in the area of learning (SEN-L). These have problems in comprehending complex and abstract information which frequently leads to performance difficulties in regular schools (e.g., Nusser & Weinert, 2017). Moreover, SEN-L as compared to regular students shows frequently impaired cognitive performance in various domains such as reasoning abilities (Gnambs & Nusser, 2019), different components of working memory (Pickering & Gathercole, 2004), or reading competencies (Pohl et al., 2016). Therefore, students with SEN-L typically attend *Förderschule* (special school) that provides training and support targeted at the difficulties of these students.

Comparison of Cognitive Abilities Between Educational Contexts

Evaluating the impact of different educational contexts on student outcomes requires valid and fair assessments of cognitive abilities in the examined contexts. Otherwise, test scores may not be comparable. However, cognitive tests are typically developed using samples from regular schools that rarely include SEN-L students. But those students might have more difficulties in appropriately understanding standard test instructions and testing procedures (Nusser & Weinert, 2017). Similarly, they might interpret item contents differently or adopt less effective task solution strategies. All of this can contribute to differential test functioning between school tracks and result in incomparable measurements.

The few studies that evaluated the measurement properties of cognitive tests among SEN-L students concluded that comparative analyses are difficult or even impossible because the administered tests seemed to measure different constructs in different educational contexts (Bolt & Ysseldyke, 2008; Pohl et al., 2016; Südkamp et al., 2015). For example, a test for mathematical competence showed substantial DIF between groups of students with different SENs (Bolt & Ysseldyke, 2008). Similarly, Südkamp et al. (2015) showed that a valid comparison of reading competencies between students from regular and special schools was impossible because of substantial rates of missing responses, low item discrimination, and an inferior test reliability among SEN-L students. In contrast, Gnambs and Nusser (2019) reported that a short instrument measuring reasoning abilities exhibited comparable measurement properties among students from special and regular schools. Thus, the matter of measurement invariance across educational contexts seems to be test-specific and needs to be explored for each test setting anew.

Present Study

The comparison of cognitive abilities across educational contexts requires comparable measurements in the studied settings. Therefore, the present study examines DIF of a test measuring figural perceptual speed between students from four educational tracks in Germany. Of particular interest are students with SEN-L who attend special schools because these were not considered during the development of the instrument (Lang et al., 2014). If the learning difficulties of students in special schools affect how they process and respond to the test items, comparisons between school tracks might be distorted. Like other frequently employed tests of cognitive speed (cf. Schmitz & Wilhelm, 2019), the present study evaluated an economical instrument that can be administered in less than 2 minutes and, thus, is ideally suited for educational large-scale studies where assessment times are costly. The test was limited to figural item material and, thus, is intended as a quick screening instrument to study population effects rather than a broad measure of mental speed for precise individual assessments.

Materials and Methods

Sample and Procedure

The *National Educational Panel Study* (NEPS; Blossfeld et al., 2011) is a multicohort, large-scale assessment of student characteristics and educational outcomes in Germany. We draw on starting cohort 4 of the NEPS that included 11,580 students. To reduce confounding effects from systematic differences in the students' background characteristics, the students were matched across school tracks. While the matching worked well for sex, age, and migration background, it could not improve the distribution of the cultural capital indicator (for details, see Electronic Supplementary Material 1 [ESM 1]). Thus, the present study examined a matched sample of N = 3,312 (45% girls) students from ninth grade attending 396 special (n = 901), basic (n = 818), intermediate (n = 789), and upper secondary schools (n = 804). Students attending specialized school types such as comprehensive schools (*Gesamtschule*) were not considered because of smaller sample sizes in these groups. Their M_{age} was 15.9 years (SD = 0.6). All students were tested in small groups at their respective institutions by experienced supervisors who received a priori training to guarantee standardized assessment conditions.

Measure

Figural perceptual speed was measured with three items from the *Bilder-Zeichen-Test* (Lang et al., 2014). For each item, a set of nine target stimuli including a figure and a corresponding number was presented on the top of the page (see Figure 1). Beneath the targets, 31 figures had to be matched to their target within 30 seconds by noting the number corresponding to the respective target stimulus. The number of correctly matched figures represented the item score (see left plot in Figure 2). The density plots of the z-standardized test scores in Figure 2 showed a distribution with two modes for each school track. Despite a time constraint of 30 seconds, the test exhibited ceiling effects for a non-negligible part of the sample. For methodological reasons (see below), we thus decided to reverse-code the items to indicate the number of errors (instead of correct responses).

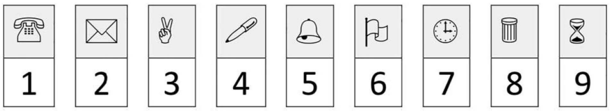


Figure 1. Target stimuli of an example item for the perceptual speed test (Lang et al., 2014, p. 9). © Leibniz Institute for Educational Trajectories (LIfBi). Reproduced with permission.

Statistical Analyses

Item Response Model

The total error scores of each item were modeled using the Rasch (1960) Poisson counts model parameterized as a generalized linear mixed-effects model (GLMM; Fox, 2010). In this approach, item difficulty parameters are represented by fixed effects, while person abilities are

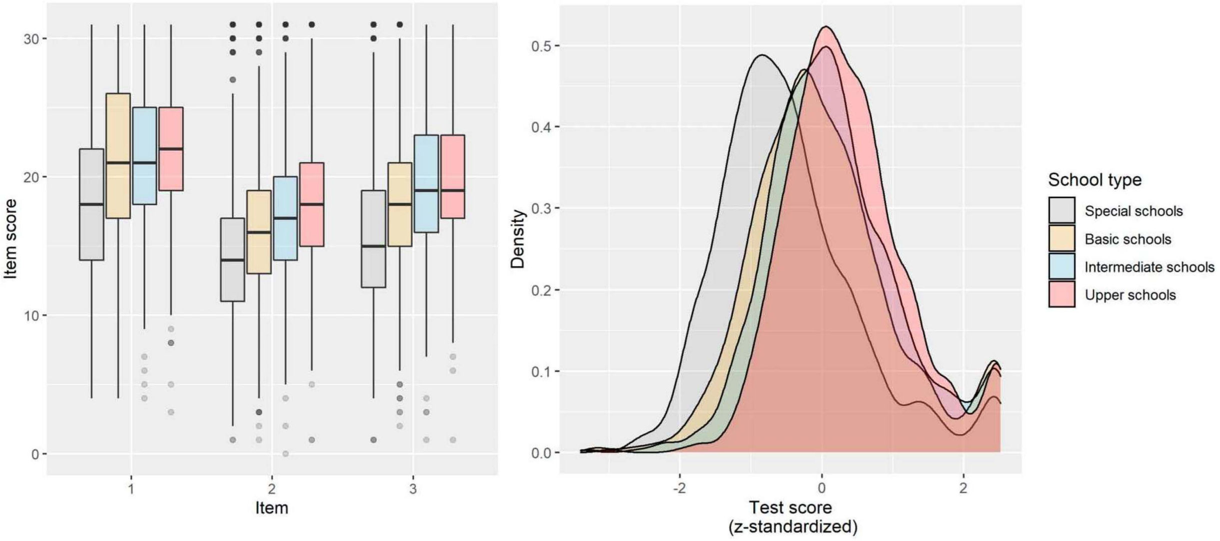


Figure 2. Item and test score distributions by school track. Box plots of item scores are given on the left, while kernel density estimates of the z-standardized test score distributions are given on the right. Descriptive statistics for these distributions are reported in ESM 1.

given by a random effect. Because students were nested in different schools, we included an additional random school effect. To account for the second mode of the response distribution (see Figure 2), the model was extended by a zero-inflation process that accounts for the excess zeros of the error scores (see Lambert, 1992). As a robustness check, we compared several response distributions: (a) a Poisson distribution (without zero-inflation), (b) a zero-inflated Poisson counts distribution, (c) a zero-inflated Poisson-lognormal distribution, and (d) a zero-inflated Poisson-Gamma distribution. All but the Poisson-Gamma distribution were truncated at 31, the maximum number of errors in the administered items, because software constraints prevented us from truncating the Poisson-Gamma distribution. The dispersion parameter φ was calculated from (d) following Doebler and Holling (2016) to check if it substantially deviated from 1 (as implied by the Poisson counts model). Model comparisons were based on the Watanabe–Akaike information criterion and leave-one-out cross-validation (LOO; Vehtari et al., 2017) for which lower values indicate better fit. Overlapping confidence intervals for these indices were interpreted as comparable fit. The reliability of the scale along the proficiency distribution was estimated following Baghaei and Doebler (2019).

Differential Item Functioning

In the GLMM framework, DIF is represented by significant cross-level interactions between the item parameters and some grouping variable (i.e., school track) after accounting for the main effects (Bürkner, 2020). In our application, both the probability for excess zeros and the rate parameter of the Poisson counts distribution can be regressed on covariates and, thus, allow for the examination of DIF. Therefore, we compared different models: (a) a model that included only main effects of school track and assumes no DIF, (b) a model with school track-specific DIF for the zero-inflation and item difficulty parameters, and (c) a model that additionally included school-level random item effects in both parts of the model. The latter allowed for distinguishing further differences stemming from the individual school's context (Hartig et al., 2020). The formal model specifications are given in ESM 1.

DIF effects were evaluated using 95% credible intervals (CrIs); effects that did not include 0 indicated DIF. The meaningfulness of the identified DIF was evaluated using a Cohen's *d*-like measure by standardizing the difference in item difficulties between two school tracks at their pooled population variances. Following the Educational Testing Service (Holland & Wainer, 1993), absolute values up to 0.25 (i.e., a quarter of a SD) were considered negligible, while values exceeding 0.50 were viewed as substantial DIF.

Bayesian Model Estimation

All models were estimated using the *R* package ***brms*** (Bürkner, 2017) that provides an interface to Stan's (Carpenter et al., 2017) implementation of a Hamiltonian Monte Carlo algorithm. Fixed effects (e.g., the item difficulties and school track main effects) were modeled with weakly informative priors whose density mostly covered the commonly expected parameter space of the effects (see ESM 1 for details). For the DIF models, 8,000 posterior draws were obtained for each parameter. Convergence diagnostics for the focal models are reported in ESM 1. The posterior distributions of the parameters are summarized using the median and a 95% highest density interval to determine whether the parameter was different from zero.

Open Practices

The anonymized data including information on the assessment procedure are available after registration at https://doi.org/10.5157/NEPS:SC4:10.0.0. Moreover, the R code including the analysis results is available at https://osf.io/yfecp.

Results

The distributions of the *z*-standardized number correct scores for all school tracks substantially overlapped (see Figure 2). However, the mode for SEN-L students was markedly shifted to the left, while students in upper secondary schools had, on average, the highest test scores. First, we fitted four different item response models to the data without acknowledging differences between school tracks. Model comparisons (see Table 1) showed that acknowledging a zero-inflation process improved model fit compared to the ordinary Poisson counts model. Because differences between the selected response distributions were negligible, and little overdispersion was observed (φ = 1.02, 95% CrI [1.01, 1.03]), we proceeded with the zero-inflated Poisson counts distribution model. The item difficulty parameters (on the log-scale) were 2.33, 95% CrI [2.31, 2.36], 2.65, 95% CrI [2.62, 2.67], and 2.52, 95% CrI [2.49, 2.54] for the three items which corresponded to error scores (on a scale from 0 to 31) of 10.28, 14.12, and 12.37, respectively. Thus, the first item was slightly easier than the other two. For the zero-inflation process, the item parameters (on the logit-scale) were −4.13, −6.19, and −5.42. Hence, the probability of observing a ceiling effect was only about 1% for item 1 and less than half this size for the remaining items. Given the negligible size of the zero-inflation parameters, we focus on DIF for the rate parameters of the Poisson process indicating the item difficulties.

Table 1. Model comparisons of estimated item response models

Response models	WAIC	SE_{WAIC}	Δ_{WAIC}	$SE_{\Delta WAIC}$	LOO	SE_{LOO}	Δ_{LOO}	$SE_{\Delta LOO}$
Models without DIF and different response distributions								
Poisson counts distribution without zero-inflation	60,370	293	4,777	251	60,912	301	5,007	255
Zero-inflated Poisson counts distribution	55,593	213	—	—	55,923	216	18	10
Zero-inflated Poisson-lognormal distribution	55,595	213	2	5	55,920	215	15	9
Zero-inflated Poisson-Gamma distribution	55,608	212	15	7	55,905	214	—	—
Zero-inflated Poisson counts distribution models with DIF								
Model 1: Model without DIF	55,582	214	859	60	55,925	217	702	61
Model 2: DIF model with school track-specific DIF	55,504	214	781	57	55,859	218	636	59
Model 3: DIF model with school-level random items effects	54,723	198	—	—	55,224	202	—	—

Note. DIF = differential item functioning, LOO = leave-one-out cross-validation, WAIC = Watanabe–Akaike information criterion, Δ = difference to the lowest WAIC/LOO.

Next, we compared three models that acknowledged differences between school tracks (see Table 1). Model comparisons showed the best fit for the most complex model with fixed-effects DIF for the item parameters and, additionally, random group DIF across schools for each item. The estimated model parameters for the Poisson process are summarized in Table 2. The item difficulty parameters for the three perceptual speed items were similar to the previously reported results, with item 1 being the easiest and item 2 the most difficult. This pattern was rather robust and emerged for all four school tracks.

As expected, we observed mean differences (on the log scale) in the latent proficiencies between school tracks with respondents in regular schools exhibiting substantially larger perceptual speed compared to students in special schools, β = −.16, 95% CrI [−0.22, −0.10] for basic secondary schools, β = −.26, 95% CrI [−0.32, −0.19] for intermediate secondary schools, and β = −.29, 95% CrI [−0.35, −0.23] for upper secondary schools. These mean differences corresponded to standardized effect sizes Δ of −0.62, −0.99, and −1.16, respectively. Since we modeled error scores, the negative effects indicate fewer errors in, for example, upper secondary schools than in special schools and, thus, a higher proficiency. In contrast, the variances of the proficiency distributions did not differ markedly between school tracks with *SD*s of 0.23, 0.29, 0.27, and 0.24, respectively (see Table 2).

Moreover, the DIF model contained evidence for DIF for SEN-L students. Using item 3 as anchor item, item 1 was easier in basic schools than in special schools (β = −.07, 95% CrI [−0.14, −0.01]), while item 2 was more difficulty in basic schools (β = .05, 95% CrI [0.01, 0.09]) and intermediate schools (β = .07, 95% CrI [0.03, −0.11]). These DIF effects corresponded to about 0.93 to 1.17 percentage changes in mean scores and standardized ESs Δ of −0.27, 0.19, and 0.28, respectively. Following rules of thumb for the interpretation of these ESs, two of them can be considered as exhibiting moderate DIF. For the remaining school tracks, no substantial DIF was observed (see Table 2). Recently, it has been argued (Hartig et al., 2020) that, in addition to average DIF across fixed groups (i.e., school tracks), random group DIF should be studied to determine whether relevant differences in item difficulties exist between schools. The respective random school effects for item 2, σ = .03, 90% CrI [0.00, 0.06], and item 3, σ = .05, 90% CrI [0.01, 0.08], were rather small. Moreover, the posterior probabilities that these variances equaled 0 were 98% and 88%, thus giving weak evidence for random school DIF. In contrast, for item 1, the respective effect was substantially larger, σ = .17, 90% CrI [0.15, 0.19], with a posterior probability of no variance of 0%, which suggests differences in the difficulty of item 1 between schools. These results also replicated in sensitivity analyses using unmatched samples across school types (see ESM 1).

Finally, we explored the reliability of the administered test in the four school tracks. The average reliability coefficients were .74 in special schools, .78 in basic secondary schools, .75 in intermediate secondary schools, and .73 in upper secondary schools. These model-based reliability estimates were slightly smaller than traditional omega reliabilities of .80, .80, .82, and .76, respectively. Moreover, the reliabilities along the proficiency scale are given in Figure 3. Although students with lower ability exhibited slightly lower reliabilities, the reliability estimates were reasonably high for a range of abilities and, typically, exceeded .70. Importantly, reliabilities did not differ substantially between school tracks.

Discussion

Measurement is a cornerstone of educational and psychological research. If tests are not comparable across

Table 2. Summary of model parameters for DIF effects

Parameters	*Mdn*	*MAD*	*LL CrI*	*UL CrI*	e^{Mdn}
Fixed effects					
Item 1	**2.53**	0.03	2.47	2.58	12.50
Item 2	**2.79**	0.02	2.74	2.83	16.22
Item 3	**2.69**	0.02	2.64	2.73	14.71
School track (ref. cat.: Special schools)					
BS	**−0.16**	0.03	−0.22	−0.10	0.85
IS	**−0.26**	0.03	−0.32	−0.19	0.77
US	**−0.29**	0.03	−0.35	−0.23	0.75
DIF effects (ref. cat.: Special schools, item 3)					
Item 1 × BS	**−0.07**	0.03	−0.14	−0.00	0.93
Item 2 × BS	**0.05**	0.02	0.01	0.09	1.05
Item 1 × IS	−0.01	0.03	−0.08	0.05	0.99
Item 2 × IS	**0.07**	0.02	0.03	0.11	1.07
Item 1 × US	−0.04	0.03	−0.10	0.03	0.96
Item 2 × US	0.03	0.02	−0.01	0.07	1.03
Random effects (*SD*)					
Students					
Special schools	**0.23**	0.01	0.21	0.25	
BS	**0.29**	0.01	0.27	0.31	
IS	**0.27**	0.01	0.25	0.30	
US	**0.24**	0.01	0.22	0.26	
Schools					
Special schools	**0.18**	0.02	0.15	0.22	
BS	**0.13**	0.02	0.09	0.16	
IS	**0.16**	0.02	0.12	0.20	
US	**0.12**	0.02	0.09	0.15	
Items in schools					
Item 1	**0.17**	0.01	0.15	0.19	
Item 2	0.03	0.02	0.00	0.07	
Item 3	**0.05**	0.02	0.01	0.08	

Note. BS = basic schools, CrI = credibility interval, DIF = differential item functioning, IS = intermediate school, *LL CrI* = lower limit of the 95% CrI, *MAD* = median absolute deviation of the posterior distribution, *Mdn* = median of the posterior distribution, *UL CrI* = upper limit of the 95% CrI; US = upper schools. e^{Mdn} = exponentiated parameter estimate indicating the expected percentage change for a unit change in the predictor. Values in boldface indicate fixed effects for which zero is not contained in the CrI or the posterior probabilities of a variance of 0 exceeds 90%. Full results are given in ESM 1.

relevant groups, DIF can bias cross-group comparisons and lead to inappropriate conclusions. Particularly, in representative large-scale assessments, it is important to show that the administered measures can be used to compare individuals across different contexts. Therefore, the present study evaluated a short test measuring figural perceptual speed. We examined whether the test allows valid comparisons between different school tracks in the German educational system. These analyses revealed only modest DIF effects between special and regular schools. Importantly, the direction of DIF did not systematically disadvantage a specific school track: While item 1 was more difficult for SEN-L students, item 2 was easier for them. Thus, it is unlikely that differences in the measurement properties of the test would systematically bias school track comparisons in substantial research. Interestingly, item-specific random school DIF was more pronounced (cf. Hartig et al., 2020). Particularly for the first item, notable differences in the estimated item difficulties were identified across schools. This might be a sign of problems in the standardization of the test procedure. Despite elaborated test protocols and extensive training of the test administrators, test instructions or organization of the test setting (e.g., the

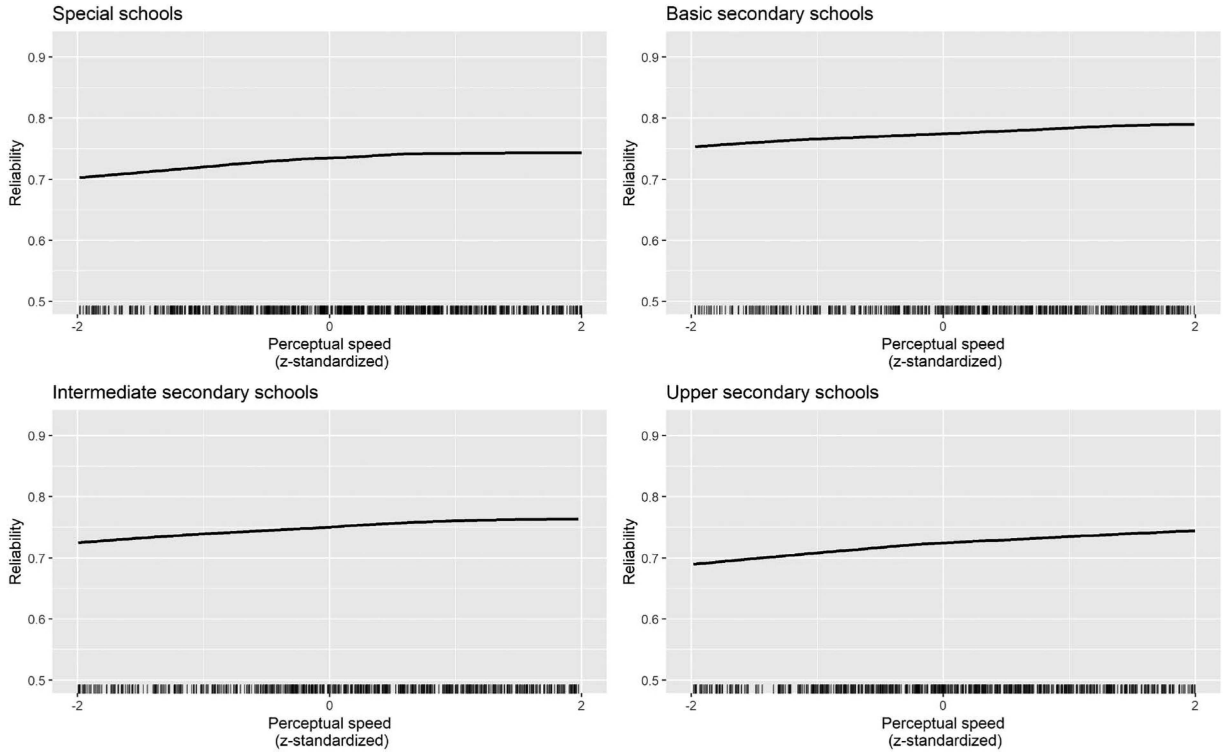


Figure 3. Reliability by school track across the proficiency scale.

seating of students, promoting test engagement, and adherence to time limits) might have varied between schools to some degree and, thus, affected student performance at the beginning of the test. This underscores the importance of standardized assessment conditions in educational large-scale assessments for comparable cognitive measures across different educational contexts. Finally, the test exhibited acceptable levels of reliabilities across different levels on the proficiency scale. Importantly, no substantial differences in the measurement precision were observed for SEN-L students.

Despite the encouraging findings, the generalizability of our results might be limited by solely relying on figural item material. Given the higher prevalence of dyslexia and dyscalculia among SEN-L students (e.g., Van der Veen et al., 2010), more pronounced DIF effects might be observed for instruments with numeric or verbal item material. Furthermore, even though a three items scale might be considered short, each item of the figural speed test provided counts data and was, thus, substantially more informative than binary correct/incorrect responses in typical power tests. Moreover, for unidimensional cognitive scales, test shortening does not systematically impair criterion validities or mean-group comparisons (Heene et al., 2014). Finally, recent advancements in the modeling of counts data have suggested alternative modeling strategies that involve more realistic assumptions for empirical data (e.g., Forthman et al., 2020). Although software constraints prevented us from exploring these modeling strategies, we have little reason to believe that this would have substantially affected our findings, as model comparisons did not suggest substantial overdispersion in our data.

In summary, the reported results demonstrate that the administered test of perceptual speed can be validly used for school track comparisons in educational large-scale assessments. These results fall in line with recent research (e.g., Gnambs & Nusser, 2019), showing that some measures originally developed for students in regular schools can be comparably administered to SEN-L students. However, we want to emphasize that these results do not render comparable analyses for other cognitive measures obsolete. For other tests requiring higher reading abilities or higher-order thinking, or adopting more complex response formats, comparisons between special and regular schools might be more challenging (cf. Bolt & Ysseldyke, 2008; Pohl et al., 2016; Südkamp et al., 2015). Future research is also encouraged to extend comparable analyses to other school types such as comprehensive or Waldorf schools. This would strengthen comparative educational research for schools with substantially different curricula and pedagogical concepts.

Electronic Supplementary Material

The electronic supplementary material is available with the online version of the article at https://doi.org/10.1027/2698-1866/a000013

ESM 1. The ESM includes information on (a) the matching of samples across school types, (b) descriptive statistics, (c) formal model specifications, (d) the Bayesian estimation and statistical software, (e) the parameter estimates, and (f) sensitivity analyses.

References

Bürkner, P. (2020). *Bayesian item response modeling in R with brms and Stan*. arXiv Preprint. https://arxiv.org/abs/1905.09501

Bürkner, P. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Baghaei, P., & Doebler, P. (2019). Introduction to the Rasch Poisson counts model: An R tutorial. *Psychological Reports*, *122*(5), 1967–1994. https://doi.org/10.1177/0033294118797577

Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (2011). Education as a lifelong process – The German National Educational Panel Study (NEPS). *Zeitschrift für Erziehungswissenschaft*, *14*, 19–34. https://doi.org/10.1007/s11618-011-0179-2

Bolt, S. E., & Ysseldyke, J. (2008). Accommodating students with disabilities in large-scale testing: A comparison of differential item functioning (DIF) identified across disability types. *Journal of Psychoeducational Assessment*, *26*(2), 121–138, https://doi.org/10.1177/0734282907307703

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), https://doi.org/10.18637/jss.v076.i01

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press.

Doebler, A., & Holling, H. (2016). A processing speed test based on rule-based item generation: An analysis with the Rasch Poisson counts model. *Learning and Individual Differences*, *52*, 121–128. https://doi.org/10.1016/j.lindif.2015.01.013

Forthmann, B., Gühne, D., & Doebler, P. (2020). Revisiting dispersion in count data item response theory models: The Conway–Maxwell–Poisson counts model. *British Journal of Mathematical and Statistical Psychology*, *73*(S1), 32–50. https://doi.org/10.1111/bmsp.12184

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. Springer.

Gnambs, T., & Nusser, L. (2019). The longitudinal measurement of reasoning abilities in students with special educational needs. *Frontiers in Psychology*, *10*, 232. https://doi.org/10.3389/fpsyg.2019.00232

Gnambs, T., Scharl, A., & Rohm, T., (2021). Comparing perceptual speed between educational contexts [Data set]. https://osf.io/yfecp/

Hartig, J., Köhler, C., & Naumann, A. (2020). Using a multilevel random item Rasch model to examine item difficulty variance between random groups. *Psychological Test and Assessment Modeling*, *62*(1), 11–27.

Heene, M., Bollmann, S., & Bühner, M. (2014). Much ado about nothing, or much to do about something? Effects of scale shortening on criterion validity and mean differences. *Journal of Individual Differences*, *35*(4), 245–249. https://doi.org/10.1027/1614-0001/a000146

Heydrich, J., Weinert, S., Nusser, L., Artelt, C., & Carstensen, C. H. (2013). Including students with special educational needs into large-scale assessments of competencies: Challenges and approaches with the German National Educational Panel Study (NEPS). *Journal of Educational Research Online*, *5*(2), 217–240.

Holdnack, J. A., Prifitera, A., Weiss, L. G., & Saklofske, D. H. (2019). WISC-V and the personalized assessment approach. In L. G., Weiss, D. H., Saklofske, J. A., Holdnack, & A., Prifitera (Eds.), *WISC-VE clinical use and interpretation* (pp. 447–488). Academic Press. https://doi.org/10.1016/B978-0-12-815744-2.00013-6

Holland, P. W., & Wainer, H. E. (1993). *Differential item functioning*. Erlbaum.

Kudo, M. F., Lussier, C. M., & Swanson, H. L. (2015). Reading disabilities in children: A selective meta-analysis of the cognitive literature. *Research in Developmental Disabilities*, *40*, 51–62, https://doi.org/10.1016/j.ridd.2015.01.002

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, *34*(1), 1–14. https://doi.org/10.2307/1269547

Lang, F. R., Kamin, S., Rohr, M., Stünkel, C., & Williger, B. (2014). *Erfassung der fluiden kognitiven Leistungsfähigkeit über die Lebensspanne im Rahmen des Nationalen Bildungspanels* [Assessment of fluid cognitive abilities across the life span in the National Educational Panel Study] (NEPS Working Paper No. 43). Leibniz-Institute for Educational Trajectories.

Mount, M. K., Oh, I.-S., & Burns, M. (2008). Incremental validity of perceptual speed and accuracy over general mental ability. *Personnel Psychology*, *61*(1), 113–139. https://doi.org/10.1111/j.1744-6570.2008.00107.x

Nusser, L., & Weinert, S. (2017). Appropriate test-taking instructions for students with special educational needs. *Journal of Cognitive Education and Psychology*, *16*(3), 227–240. https://doi.org/10.1891/1945-8959.16.3.227

Peng, P., Wang, C., & Namkung, J. (2018). Understanding the cognition related to mathematics difficulties: A meta-analysis on the cognitive deficit profiles and the bottleneck theory. *Review of Educational Research*, *88*(3), 434–476. https://doi.org/10.3102/0034654317753350

Pickering, S. J., & Gathercole, S. E. (2004). Distinctive working memory profiles in children with special educational needs. *Educational Psychology*, *24*(3), 393–408. https://doi.org/10.1080/0144341042000211715

Pohl, S., Südkamp, A., Hardt, K., Carstensen, C. H., & Weinert, S. (2016). Testing students with special educational needs in large-scale assessments – Psychometric properties of test scores and associations with test taking behavior. *Frontiers in Psychology*, *7*, 154. https://doi.org/10.3389/fpsyg.2016.00154

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. University of Chicago Press.

Rindermann, H., & Neubauer, A. (2004). Processing speed, intelligence, creativity, and school performance: Testing of causal hypotheses using structural equation models. *Intelligence*, *32*(6), 573–589. https://doi.org/10.1016/j.intell.2004.06.005

Südkamp, A., Pohl, S., & Weinert, S. (2015). Competence assessment of students with special educational needs—Identification of appropriate testing accommodations. *Frontline Learning Research*, *3*(2), 1–26. https://doi.org/10.14786/flr.v3i2.130

Schmitz, F., & Wilhelm, O. (2019). Mene mene tekel upharsin: Clerical speed and elementary cognitive speed are different by virtue of test mode only. *Journal of Intelligence*, *7*(3), 16. https://doi.org/10.3390/jintelligence7030016

Van der Veen, I., Smeets, E., & Derriks, M. (2010). Children with special educational needs in the Netherlands: Number, characteristics and school career. *Educational Research*, *52*(1), 15–43. https://doi.org/10.1080/00131881003588147

Vehtari, A., Gelman, A. & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

History
Received February 19, 2021
Revision received June 30, 2021
Accepted July 1, 2021
Published online August 27, 2021

Section: Educational Psychology

Acknowledgments
This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort Grade 9, doi:10.5157/NEPS:SC4:10.0.0. From 2008 to 2013, NEPS data was collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

Conflict of Interest
The authors have no conflicts of interest to disclose.

Open Data
The anonymized data including information on the assessment procedure are available after registration at https://doi.org/10.5157/NEPS:SC4:10.0.0. Moreover, the R code including the analysis results is available at https://osf.io/yfecp (Gnambs et al., 2021).

Funding
The publication of this article was funded by the Open Access Fund of the Leibniz Association.

ORCID
Timo Gnambs
https://orcid.org/0000-0002-6984-1276
Anna Scharl
https://orcid.org/0000-0003-0081-1893
Theresa Rohm
https://orcid.org/0000-0001-9203-327X

Timo Gnambs
Leibniz Institute for Educational Trajectories
Wilhelmsplatz 3
96047 Bamberg
Germany
timo.gnambs@lifbi.de

Supplemental Material for

Comparing Perceptual Speed between Educational Contexts: The Case of Students with Special Educational Needs

Timo Gnambs, Anna Scharl, & Theresa Rohm

published in

Psychological Test Adaptation and Development

https://doi.org/10.1027/2698-1866/a000013

Matching of Samples across School Types

The study aims at examining context effects of different school types. Because students do not choose schools at random, the choice of a specific school type may depend on various background characteristics. Thus, school types are confounded with various student attributes. For example, substantially more girls attend upper secondary schools as compared to basic secondary schools, while the proportion of students with a migration background is typically higher in special and basic schools. To be able to draw causal conclusions on schooling effects, students were matched on gender, age (in years), migration background, and the number of books at home (as a proxy for cultural capital; cf. Sieben & Lechner, 2019) across school types. Because the number of missing values on these background variables was small (i.e., less than 3%), students with missing values on these variables were excluded from the matching procedure. To account for the nested sample structure resulting from students attending the same schools, we performed multilevel matching (Pimental et al., 2018). The goal of this procedure is to simultaneously identify schools and students that are comparable on the covariates. Special schools were used as the treatment group in the matching procedure to which the other school types were matched.

The matching resulted in a total sample size of $N = 3,313$. Table S1 shows the original and matched samples' characteristics broken down by school type. Overall, the matching procedure worked well for gender, age, and migration background. In contrast, the matching was unable to substantially improve the distribution of the cultural capital variable because the available data did not allow an exact match.

Table S1 also reports *z*-standardized mean scores of a short test measuring reasoning abilities (Lang et al., 2014) to emphasize differences in general cognitive abilities between school types. As expected, the students in special schools had substantially lower reasoning abilities as compared to students in basic schools ($d = -0.77$), intermediate schools ($d = -1.36$), and upper schools ($d = 1.81$). These differences were negligibly affected by the matching procedure, thus, emphasizing the robustness of the cognitive differences between school types.

Table S1

Characteristics of the Original and Matched Samples by School Type

	Original Sample				Matched Sample			
Variable	SS	BS	IS	US	SS	BS	IS	US
N	911	2913	3002	4754	901	818	789	804
n	9	17	29	33	9	8	8	8
Number of schools	99	158	103	147	99	99	99	99
Female (%)	0.44	0.44	0.49	0.55	0.44	0.44	0.45	0.45
Age (*M*)	15.98	15.89	15.62	15.37	15.98	15.97	15.84	15.68
(*SD*)	0.64	0.69	0.59	0.50	0.64	0.64	0.61	0.58
Migration (%)	0.28	0.38	0.23	0.18	0.28	0.28	0.27	0.25
Cultural capital (*M*)	2.38	2.96	3.71	4.56	2.40	2.52	2.96	3.57
(*SD*)	1.34	1.40	1.34	1.23	1.34	1.25	1.24	1.24
Reasoning (*M*)	-1.29	-0.52	0.07	0.52	-0.84	-0.17	0.38	0.75
(*SD*)	0.87	0.87	0.84	0.79	0.80	0.81	0.81	0.76

Note. N = Total sample size, *n* = Median number of students per school, SS = Special schools, BS = Basic secondary schools, IS = Intermediate secondary schools, US = Upper secondary schools.

Descriptive Statistics for Test Items by School Type

The means and standard deviations of the three perceptual speed items by school type are summarized in Table S2. The means for the three item scores increased across school types as would be expected of the ability levels within the German ability-tiered school system.

Table S2

Means, Standard Deviations, and Correlations for Item Scores by School Type

	Means (Standard deviations)		
	Item 1	Item 2	Item 3
Special schools	18.84 (6.24)	14.43 (5.42)	15.98 (5.50)
Basic secondary schools	21.63 (5.94)	16.41 (5.68)	18.56 (5.81)
Intermediate secondary schools	21.51 (5.22)	17.33 (5.34)	19.58 (5.66)
Upper secondary schools	22.37 (4.84)	18.51 (5.04)	20.34 (5.40)
	Correlations		
	Items 1 / 2	Items 1 / 3	Items 2 / 3
Special schools	.55	.48	.69
Basic secondary schools	.53	.51	.65
Intermediate secondary schools	.57	.48	.72
Upper secondary schools	.41	.41	.64

Note. The reported statistics are based on the number correct scores (and not the error scores).

Formal Specifications of the Generalized Linear Mixed-Effects Models

The three perceptual speed items indicated the number of errors (i.e., missed targets) and, thus, can be described by a Poisson distribution (Baghaei & Doebler, 2019; Doebler & Holling, 2015) that expresses the probability to observe an error score of y_{vi} on item *i* for respondent *v* as

$$P(Y_{vi} = y_{vi}) = e^{-\lambda_{vi}} \lambda_{vi}{}^{y_{vi}} / y_{vi}!. \qquad [1]$$

Here, λ_{vi} represents the expected number of errors of person *v* on item *i* which is a function of a person's ability θ_v and an item's difficulty β_i (Rasch, 1960). Applying a log-link allows expressing λ_{vi} as

$$\log(\lambda_{vi}) = \log(\theta_v) + \log(\beta_i). \qquad [2]$$

In this way, the item response model can be specified as a generalized linear mixed-effects model (GLMM; Fox, 2010; Van den Noortgate & De Boeck, 2005) with fixed effects representing the item difficulties and a random effect for the person abilities:

$$\log\left[E\left(Y \middle| I_{i \in \{1,2,3\}}, \theta\right)\right] = \sum_i (\beta_i \cdot I_i) + U_\theta \qquad [3]$$

In [3], I_i represents three dummy-coded variables that take the value 1 if the score refers to item *i* and 0 otherwise. Because the model does not include an intercept, item difficulty parameters β_i can be estimated for each item. As described above, the response variable *Y* is assumed to follow a Poisson distribution, while a normal distribution $N(0, \sigma_\theta)$ is assumed for the random person effect U_θ.

To account for the bimodality of the response distribution (see Figure 2), the model in [3] can be extended by a zero-inflation process (Lambert, 1992). In zero-inflation models, two processes are assumed: The excess number of zeros in the data is modeled by a binomial distribution with the zero-inflation probability *z*, while a second random process occurring with probability (1 - *z*) describes the count data as in [1].

$$(1)\ P_z(Y=0) = z + (1-z)\cdot P(Y=0) \quad \text{if } y = 0 \qquad [4]$$

$$(2)\ P_z(Y=y) = (1-z)\cdot P(Y=y) \qquad \text{if } y > 0$$

Using a logit-link function for the zero-inflation process enables the estimation of the zero-inflated process as a GLMM:

$$\text{logit}\left[E\left(Y_z \middle| I_{i\,\in\{1,2,3\}}, \theta_z\right)\right] = \sum_i (\gamma_i \cdot I_i) + U_{\theta_z} \qquad [5]$$

In [5], θ_z represents a student's ability to show a ceiling effect. As described above, the response variable Y_z is assumed to follow a binomial distribution, while a normal distribution is assumed for the random person effect U_{θ_z}. However, the random effects for the zero-inflation process and the Poisson count process are not independent but follow a joint normal distribution with

$$\begin{bmatrix} \mathrm{U}_\theta \\ U_{\theta_z} \end{bmatrix} = N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}; \begin{bmatrix} \sigma_{\theta,\theta} & \sigma_{\theta,\theta_z} \\ \sigma_{\theta_z,\theta} & \sigma_{\theta_{z,z}} \end{bmatrix}\right). \qquad [6]$$

Because students are nested in different schools, these dependencies can be acknowledged by including respective random school effects s:

$$(1)\ \text{logit}\left[E\left(Y_z \middle| I_{i\,\in\{1,2,3\}}, \theta_z, s_z\right)\right] = \sum_i (\gamma_i \cdot I_i) + U_{\theta_z} + U_{s_z} \qquad [7]$$

$$(2)\ \log\left[E\left(Y \middle| I_{i\,\in\{1,2,3\}}, \theta, s\right)\right] = \sum_i (\beta_i \cdot I_i) + U_\theta + U_s$$

Again, the random effects for the zero-inflation process and the Poisson count process are assumed to follow a joint distribution with

$$\begin{bmatrix} U_\theta \\ U_{\theta_z} \\ U_s \\ U_{s_z} \end{bmatrix} = N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}; \begin{bmatrix} \sigma_{\theta,\theta} & \sigma_{\theta,\theta_z} & 0 & 0 \\ \sigma_{\theta_z,\theta} & \sigma_{\theta_z,\theta_z} & 0 & 0 \\ 0 & 0 & \sigma_{s,s} & \sigma_{s,s_z} \\ 0 & 0 & \sigma_{s_z,s} & \sigma_{s_z,s_z} \end{bmatrix}\right). \qquad [8]$$

Finally, differential item functioning can be described as cross-level interactions between the item parameters I_i in [7] and the school type (Bürkner, 2019; Van den Noortgate & De Boeck, 2005):

$$(1)\ \text{logit}\left[E\left(Y_z \middle| I_{i\,\in\{1,2,3\}}, T_{t\,\in\{SS,BS,IS,US\}}, \theta_z, s_z\right)\right] = \qquad [9]$$

$$\textstyle\sum_i\ (\gamma_i \cdot I_i) + \sum_{t\,\notin\{SS\}}(\gamma_t \cdot T_t) + \sum_{i\,\notin\{3\}} \sum_{t\,\notin\{SS\}}\left(\gamma_{i,t} \cdot I_i \cdot T_t\right) +$$

$$\textstyle\sum_t\ \left(U_{\theta_z,t} \cdot T_t\right) + \sum_t\ \left(U_{s_z,t} \cdot T_t\right)$$

$$(2)\ \log\left[E\left(Y \middle| I_{i\,\in\{1,2,3\}}, T_{t\,\in\{SS,BS,IS,US\}}, \theta, s\right)\right] =$$

$$\textstyle\sum_i\ (\beta_i \cdot I_i) + \sum_{t\,\notin\{SS\}}(\beta_t \cdot T_t) + \sum_{i\,\notin\{3\}} \sum_{t\,\notin\{SS\}}\left(\beta_{i,t} \cdot I_i \cdot T_t\right) +$$

$$\textstyle\sum_t\ \left(U_{\theta,t} \cdot T_t\right) + \sum_t\ \left(U_{s,t} \cdot T_t\right)$$

In [9], T_t represents dummy-coded variables that take the value 1 if the score was obtained in school type t and 0 otherwise. Here, the school types are represented as basic (BS), intermediate (IS), and upper (US) secondary schools. Special schools were used as reference category, while item 3 was used as an anchor item. Again, the random effects are jointly normally distributed. However, to allow for differently homogeneous ability distributions in the four school types, different variance components are estimated in each school type, thus, resulting in different variance-covariance matrices in [8] for each school type. The model in [9] can be flexibly extended to include, for example, item-specific random school effects $U_{s,i}$ or even interactive random school effects $U_{s,t,i}$.

As described in the main manuscript, we evaluated three different models to examine DIF:

1. The baseline model included main effects for the school type without DIF.

$$(1)\ \text{logit}\left[E\left(Y_z \middle| I_{i\,\in\{1,2,3\}}, T_{t\,\in\{SS,BS,IS,US\}}, \theta_z, s_z\right)\right] = \qquad [10]$$

$$\textstyle\sum_i\ (\gamma_i \cdot I_i) + \sum_{t\,\notin\{SS\}}(\gamma_t \cdot T_t) +$$

$$\textstyle\sum_t\ \left(U_{\theta_z,t} \cdot T_t\right) + \sum_t\ \left(U_{s_z,t} \cdot T_t\right)$$

$$(2)\ \log\left[E\left(Y \middle| I_{i\,\in\{1,2,3\}}, T_{t\,\in\{SS,BS,IS,US\}}, \theta, s\right)\right] =$$

$$\sum_i (\beta_i \cdot I_i) + \sum_{t \notin \{SS\}} (\beta_t \cdot T_t) +$$

$$\sum_t \left(U_{\theta,t} \cdot T_t\right) + \sum_t \left(U_{s,t} \cdot T_t\right)$$

The corresponding random effects followed a diagonal-block structure with the block structure in school type *t* given in [11].

$$\begin{bmatrix} U_{\theta,t} \\ U_{\theta_{z,t}} \\ U_{s,t} \\ U_{s_{z,t}} \end{bmatrix} = N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}; \begin{bmatrix} \sigma_{(\theta,t),(\theta,t)} & \sigma_{(\theta,t),(\theta_{z,t})} & 0 & 0 \\ \sigma_{(\theta_{z,t}),(\theta,t)} & \sigma_{(\theta_{z,t}),(\theta_{z,t})} & 0 & 0 \\ 0 & 0 & \sigma_{(s,t),(s,t)} & \sigma_{(s,t),(s_z,t)} \\ 0 & 0 & \sigma_{(s_z,t),(s,t)} & \sigma_{(s_z,t),(s_z,t)} \end{bmatrix}\right). \qquad [11]$$

2. The second model included DIF for the item difficulty and zero-inflation parameters.

(1) $\text{logit}\left[E\left(Y_z \middle| I_{i \in \{1,2,3\}}, T_{t \in \{SS,BS,IS,US\}}, \theta_z, s_z\right)\right] =$ [12]

$$\sum_i (\gamma_i \cdot I_i) + \sum_{t \notin \{SS\}} (\gamma_t \cdot T_t) + \sum_{i \notin \{3\}} \sum_{t \notin \{SS\}} \left(\gamma_{i,t} \cdot I_i \cdot T_t\right) +$$

$$\sum_t \left(U_{\theta_z,t} \cdot T_t\right) + \sum_t \left(U_{s_z,t} \cdot T_t\right)$$

(2) $\log\left[E\left(Y \middle| I_{i \in \{1,2,3\}}, T_{t \in \{SS,BS,IS,US\}}, \theta, s\right)\right] =$

$$\sum_i (\beta_i \cdot I_i) + \sum_{t \notin \{SS\}} (\beta_t \cdot T_t) + \sum_{i \notin \{3\}} \sum_{t \notin \{SS\}} \left(\beta_{i,t} \cdot I_i \cdot T_t\right) +$$

$$\sum_t \left(U_{\theta,t} \cdot T_t\right) + \sum_t \left(U_{s,t} \cdot T_t\right)$$

The corresponding random effects structure in school type *t* is given in [11].

3. The third model additionally added item-specific random school effects $U_{s,i}$.

(1) $\text{logit}\left[E\left(Y_z \middle| I_{i \in \{1,2,3\}}, T_{t \in \{SS,BS,IS,US\}}, \theta_z, s\right)\right] =$ [13]

$$\sum_i (\gamma_i \cdot I_i) + \sum_{t \notin \{SS\}} (\gamma_t \cdot T_t) + \sum_{i \notin \{3\}} \sum_{t \notin \{SS\}} \left(\gamma_{i,t} \cdot I_i \cdot T_t\right) +$$

$$\sum_t \left(U_{\theta_z,t} \cdot T_t\right) + \sum_t \left(U_{s_z,t} \cdot T_t\right) + \sum_i \left(U_{s_z,i} \cdot I_i\right)$$

(2) $\log\left[E\left(Y \middle| I_{i \in \{1,2,3\}}, T_{t \in \{SS,BS,IS,US\}}, \theta, s\right)\right] =$

$$\sum_i (\beta_i \cdot I_i) + \sum_{t \notin \{SS\}} (\beta_t \cdot T_t) + \sum_{i \notin \{3\}} \sum_{t \notin \{SS\}} \left(\beta_{i,t} \cdot I_i \cdot T_t\right) +$$

$$\sum_t \left(U_{\theta,t} \cdot T_t\right) + \sum_t \left(U_{s,t} \cdot T_t\right) + \sum_t \left(U_{s,i} \cdot I_i\right)$$

The corresponding random effects followed a diagonal-block structure with the block structure in school type *t* given in [11] and an additional diagonal-block for the random item effects.

Bayesian Model Specifications

All models were estimated using the *R* package *brms* (Bürkner, 2017) that provides an interface to Stan's (Carpenter et al., 2017) implementation of a Hamiltonian Monte Carlo algorithm, the adaptive no-U-turn sampler (Hoffman & Gelman, 2014). We chose weakly informative prior distributions for the models to focus the exploration of the parameters space on plausible values without restricting it. Fixed effects (e.g., the item difficulties and school type main effects) were modeled with normally distributed priors of *N*(0, 4) that are based on recommendations from the item response theory literature and cover the expected range of the latent proficiency distribution (e.g., Patz & Junker, 1999). Random effects were modeled with zero-truncated normal distributions of *N*(0, 2) to allow for larger variances as typically observed in item response modeling (i.e., values around 1) and accommodate the expected context effects of the German ability-tiered school system. Correlations between the model parameters were estimated using the LKJ prior distribution which allows easier sampling from the distribution of correlation matrices (Lewandowski, Kurowicka, & Joe, 2009). We used the default prior distribution of the *brms* package for the standard deviation of the zero-inflation parameter, a *t*-distribution with *t*(3, 0, 2.5) because we did not have any prior expectations regarding that parameter The basic models without school track specific effects were run with two chains for 4,000 iterations of which 2,000 were discarded as warm-up iterations. In total, this resulted in 4,000 posterior draws for each parameter. For the more complex models with school track specific effects, we used four chains for 4,000 iterations with 2,000 burnin samples that resulted in a total of 8,000 posterior draws for each parameter.

The posterior distributions of the parameters are summarized using the median and the median absolute deviation. Moreover, a 95% credibility interval was used to determine whether the parameter was different from zero.

Statistical Software

All analyses were conducted on a 64-bit Windows 10 machine using *R* (version 3.5.3, R Core Team, 2020) and the following packages: tidyverse (version 1.3.0, Wickham et al., 2019), tidyr (version 1.0.2, Wickham, 2020), TAM (version 3.5-19, Robitzsch, Kiefer, & Wu, 2020), sjlabelled (version 1.1.1, Lüdecke, 2020), rstan (version 2.19.2, Stan Development Team, 2020), rmarkdown (version 2.0, Xie, Dervieux, & Riederer, 2020), psych (version 1.8.12, Revelle, 2020), optmatch (version 0.9-13, Hansen & Klopfer, 2006), MBESS (version 4.8.0, Kelley, 2020), matchMulti (version 1.1.7, Keele, Pimentel, & Rosenbaum, 2018), lavaan (version 0.6-7, Rosseel, 2012), labelled (version 2.2.1, Larmarange, 2020), knitr (version 1.25, Xie, 2020), here (version 0.1, Müller, 2020), haven (version 2.2.0, Wickham & Miller, 2020), ggplot2 (version 3.2.1, Wickham, 2016), dplyr (version 1.0.2, Wickham, Francois, Henry, & Müller, 2020), cowplot (version 1.0.0, Wilke, 2020), brms (version 2.13.5, Bürkner, 2018), Rcpp (version 1.0.3, Eddelbuettel, 2013), gtools (version 3.8.2, Warnes, Bolker, & Lumley, 2020), broom.mixed (version 0.2.6, Bolker & Robinson, 2020), DescTools (version 0.99.41, Signorell et al., 2021), and bayestestR (version 0.4.0, Makowski, Ben-Shachar, & Lüdecke, 2019) as well as possible dependencies.

Convergence Diagnostics

We checked several diagnostic statistics and plots to ensure the convergence of our models to a stationary distribution. The mixing behavior of our chains was assessed in two ways: visually by examining the trace plots and by considering the potential scale reduction factor ($\hat{R}$). Trace plots with chains that overlap to a large degree indicate good mixing behaviors, while a $\hat{R}$ below 1.01 indicates convergence (Gelman et al., 2013). Moreover, the effect of autocorrelations within chains was evaluated visually using autocorrelation plots and the effective sample size (ESS). Because MCMC samples are dependent, the amount of correlations between subsequent draws can affect the precision of the estimates. The ESS of the posterior draws represents the number of independent samples with the same precision as the autocorrelated MCMC samples. Vehtari et al. (2021) suggest that an ESS greater than 100 times the number of chains (in our case: 100 x 4 = 400) is sufficient for stable estimates. Finally, we inspected the kernel density plots of each parameter.

Zero-Inflated Poisson Count Model without School Track-Specific Effects (see [13])

The trace plots (see Figure A) show good mixing behavior since the chains overlap well and only some spikes of the individual chains can be discerned. This finding is corroborated by the potential scale reduction factors (see Figure B) which range around 1 for all parameters and, thus, indicate good convergence (Gelman et al., 2013). The autocorrelation plots (see Figures C to E) show a steep decline of autocorrelations for most of the parameters and low coefficients by a lag of 5. Moreover, the ESS of the posterior draws were greater than 1,000 for all parameters. Thus, autocorrelations did not bias the results and provided reliable parameter estimates. Finally, the kernel density plots showed symmetric distributions for all parameters (see Figure F). These results support the convergence of the Bayesian estimation model and allow interpreting the estimated model parameters.

Figure A

Trace Plots for Zero-Inflated Poisson Counts Model

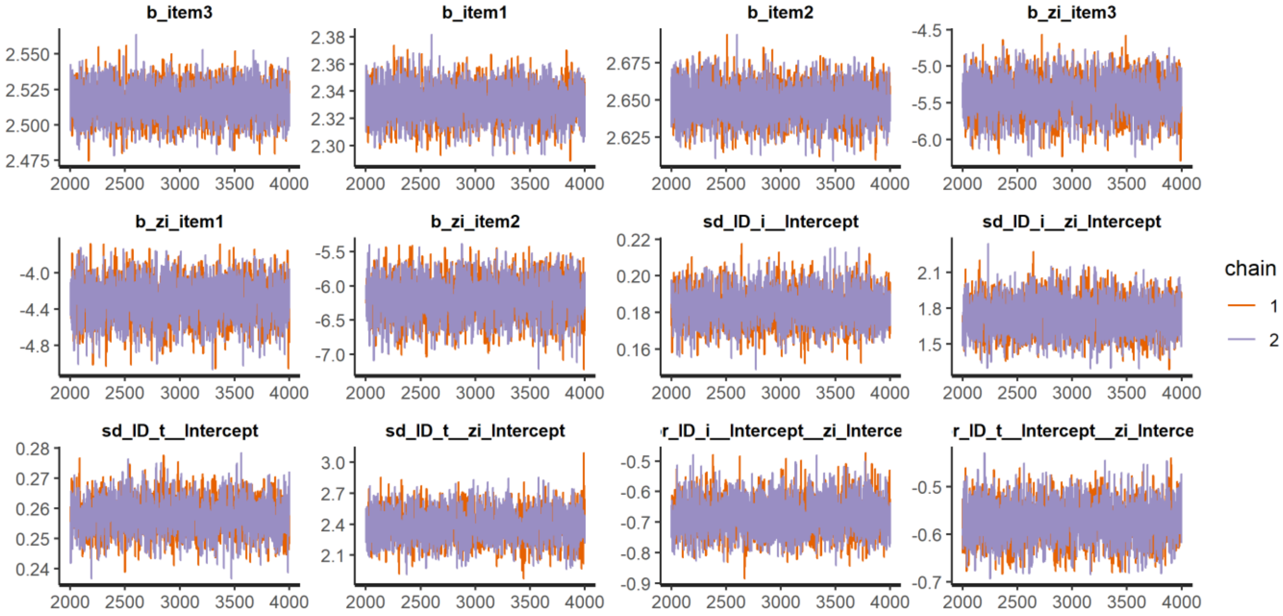


Figure B

Potential Scale Reduction Factors for Zero-Inflated Poisson Counts Model

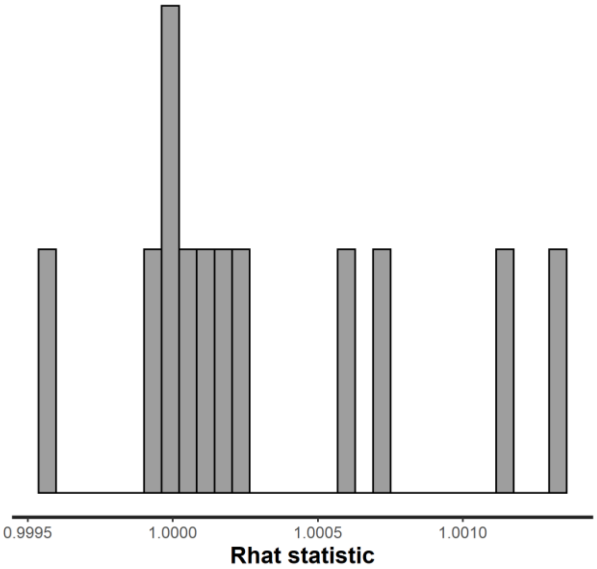


Figure C

Autocorrelation Plots for Zero-Inflated Poisson Counts Model and a Lag of up to 20 for each Chain (Part 1).

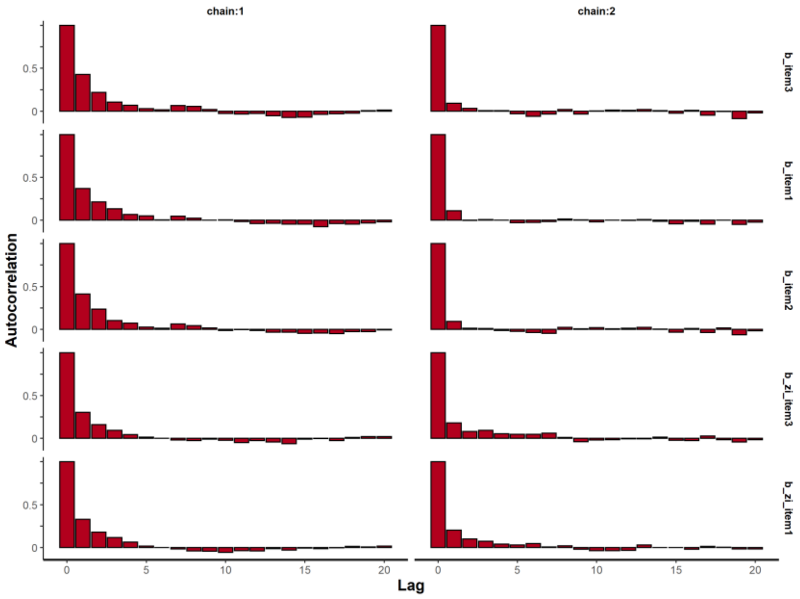


Figure D

Autocorrelation Plots for Zero-Inflated Poisson Counts Model and a Lag of up to 20 for each Chain (Part 2).

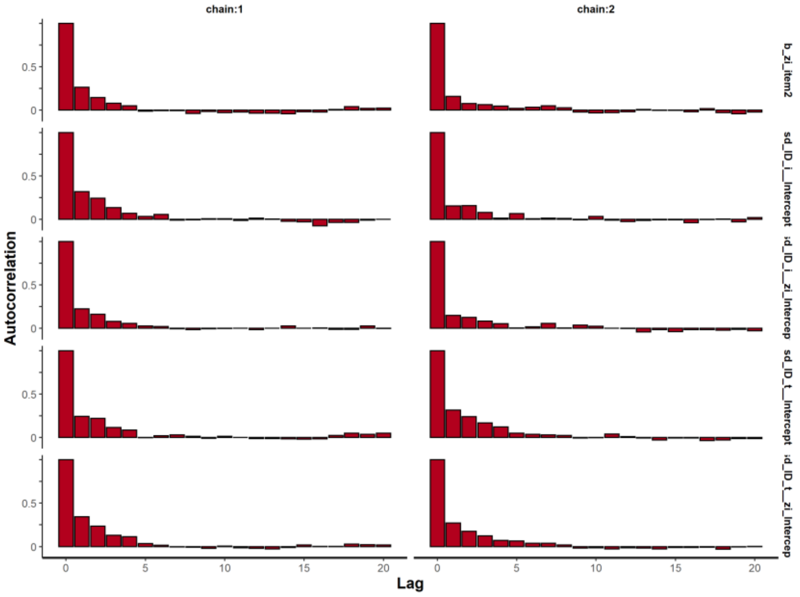


Figure E

Autocorrelation Plots for Zero-Inflated Poisson Counts Model and a Lag of up to 20 for each Chain (Part 3).

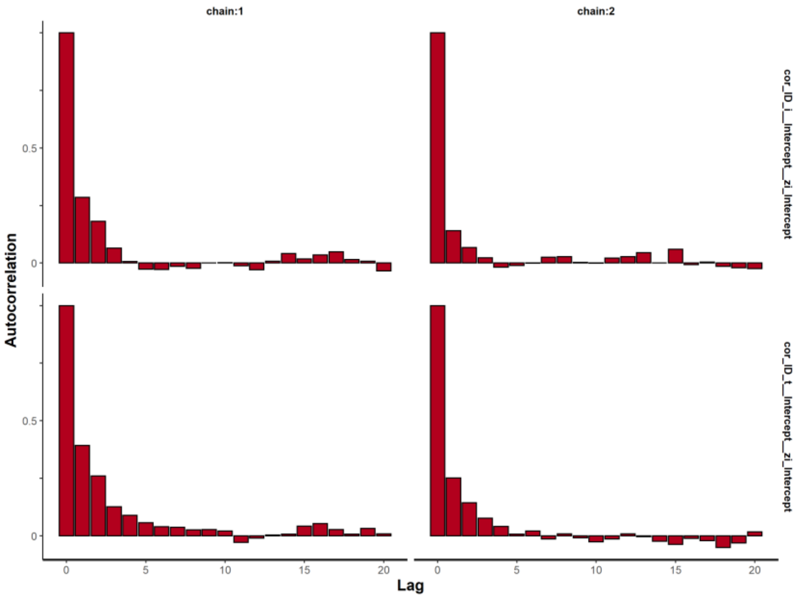
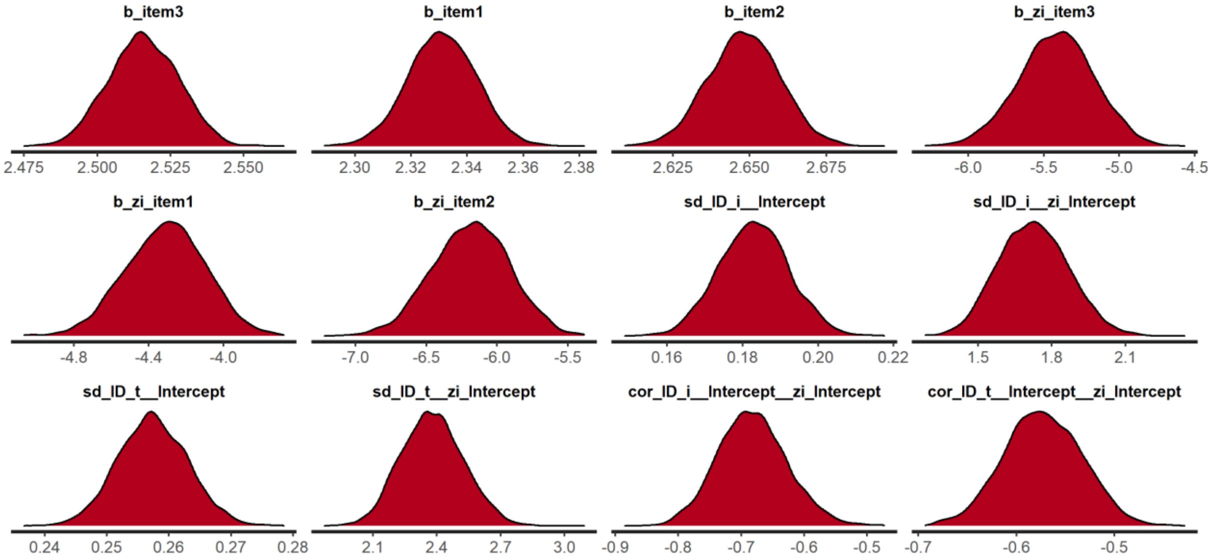


Figure F

Kernel Density Plots for Zero-Inflated Poisson Counts Model



Zero-inflated Poisson Count Model with School Track-Specific DIF and Item-Specific Random School Effects (see [13])

The trace plots (see Figures G to M) show good mixing behavior since the chains overlap well and only some spikes of the individual chains can be discerned. This finding is corroborated by the potential scale reduction factors (see Figure N) which range around 1 for most parameters. Only some parameters related to the item-specific random school variances exhibited larger values around 1.01 or slightly above. However, overall, these results indicate sufficient convergence (Gelman et al., 2013). The autocorrelation plots (see Figures O to Y) show a steep decline of autocorrelations for most of the parameters and low coefficients by a lag of 10. Moreover, the ESS of the posterior draws were greater than 400 for all parameters. Thus, autocorrelations did not bias the results and provided reliable parameter estimates. Finally, the kernel density plots showed symmetric distributions for most parameters (see Figures Z to FF). Only the correlation between random effects in intermediate schools and the random school effect for item 2 seems somewhat skewed. A large part of the density groups at the edge of the parameter space. Since diagnostics of autocorrelation and ESS give confidence in sufficiently large independent posterior samples, the skewness reflects a high correlation among the parameters (in intermediate schools) and a random school variance close to zero (for item 2), rather than problems in the exploration of the parameter space. Overall, these results support the convergence of the Bayesian estimation model and allow interpreting the estimated model parameters.

Figure G

Trace Plots for DIF Model (Part 1)

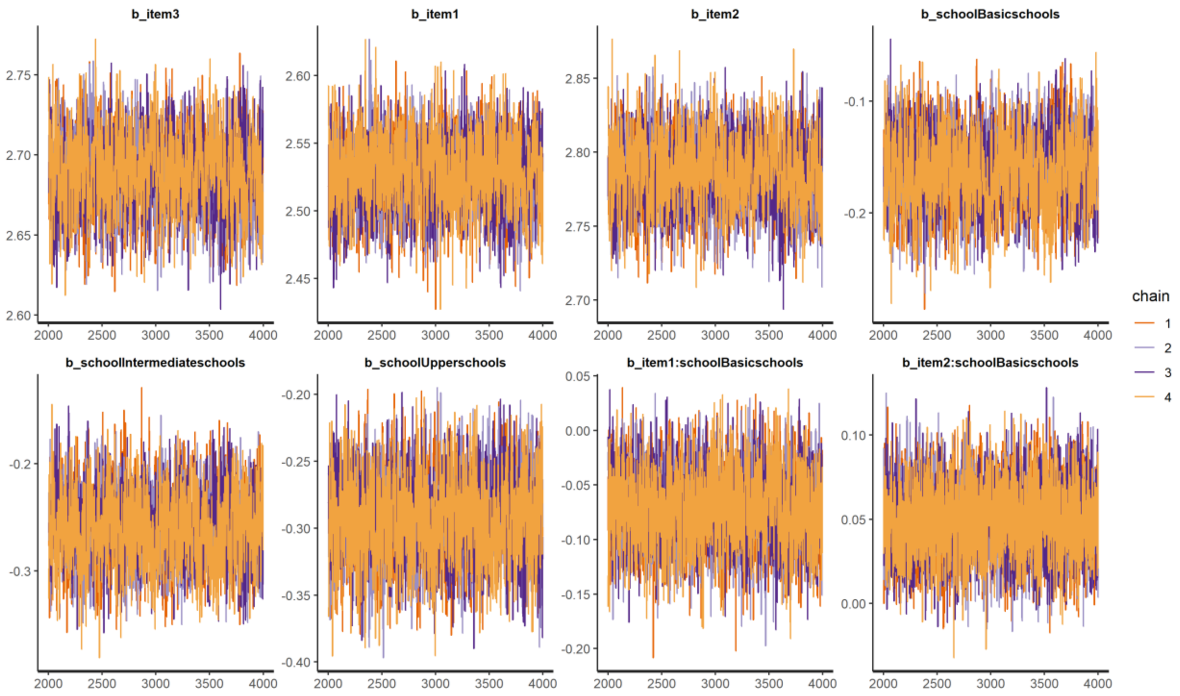


Figure H

Trace Plots for DIF Model (Part 2).

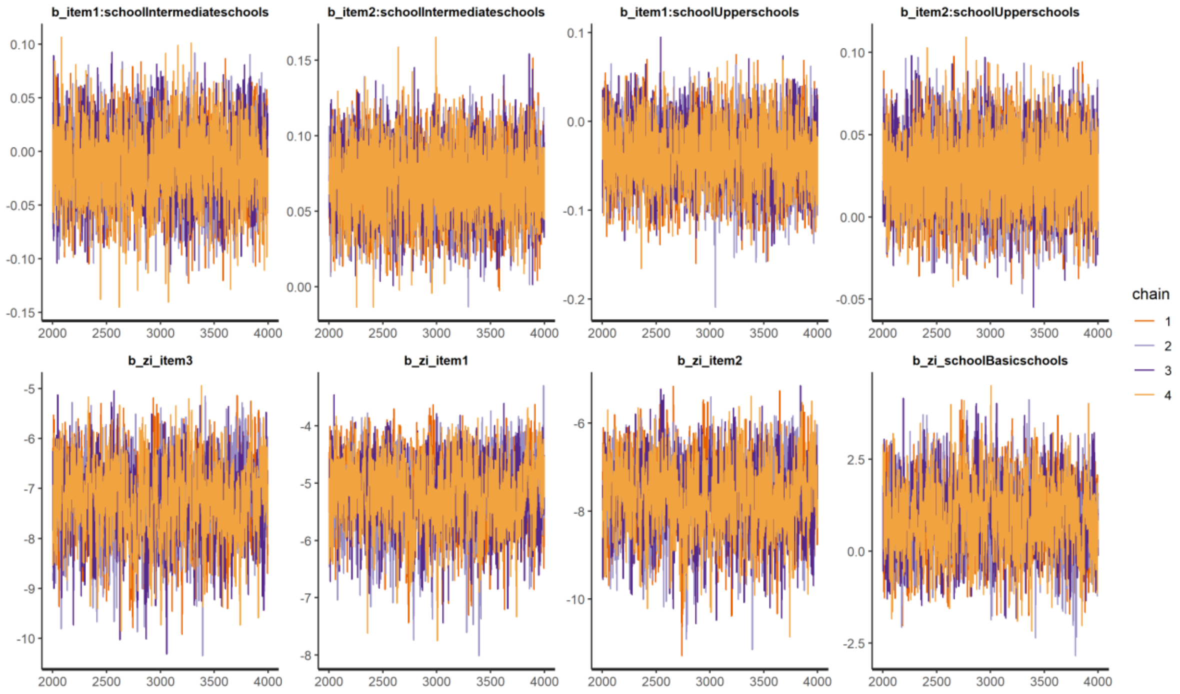


Figure I

Trace Plots for DIF Model (Part 3)

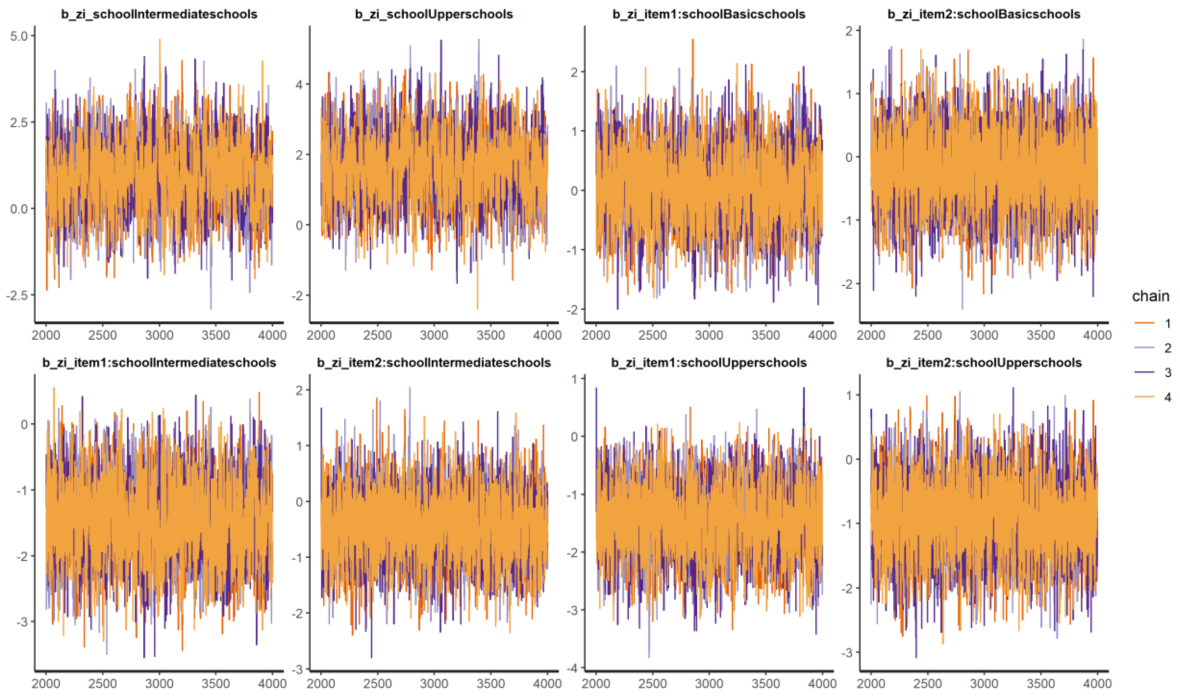


Figure J

Trace Plots for DIF Model (Part 4)

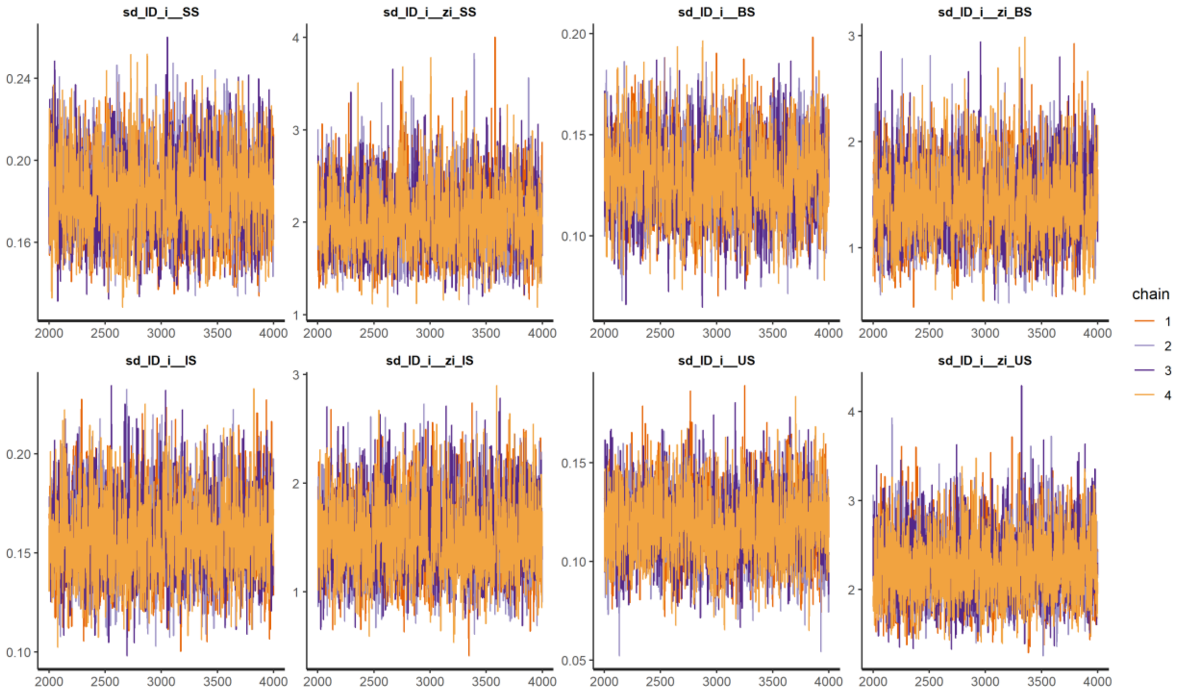


Figure K

Trace Plots for DIF Model (Part 5)

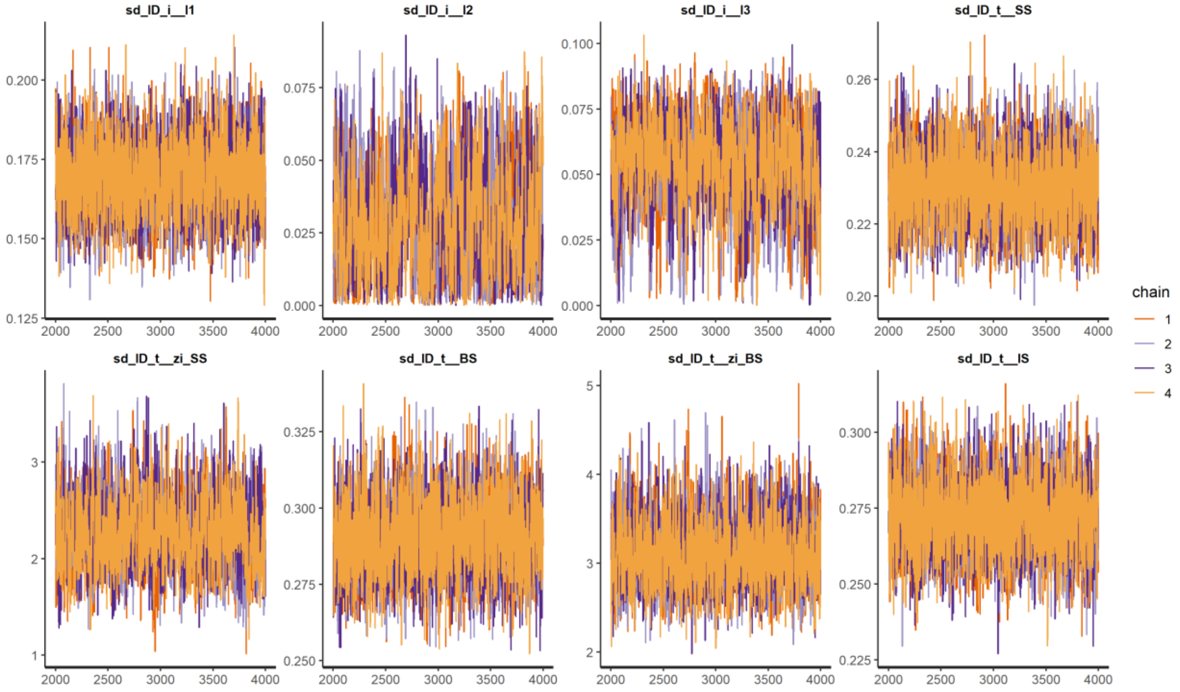


Figure L

Trace Plots for DIF Model (Part 6)

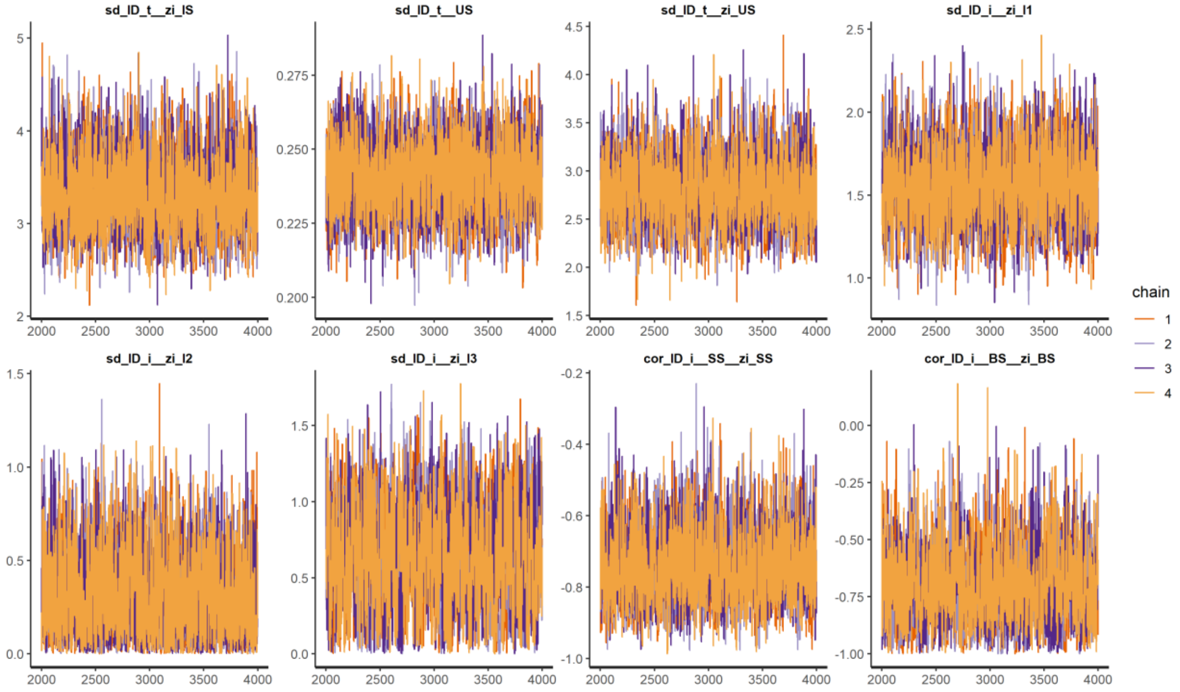


Figure M

Trace Plots for DIF Model (Part 7)

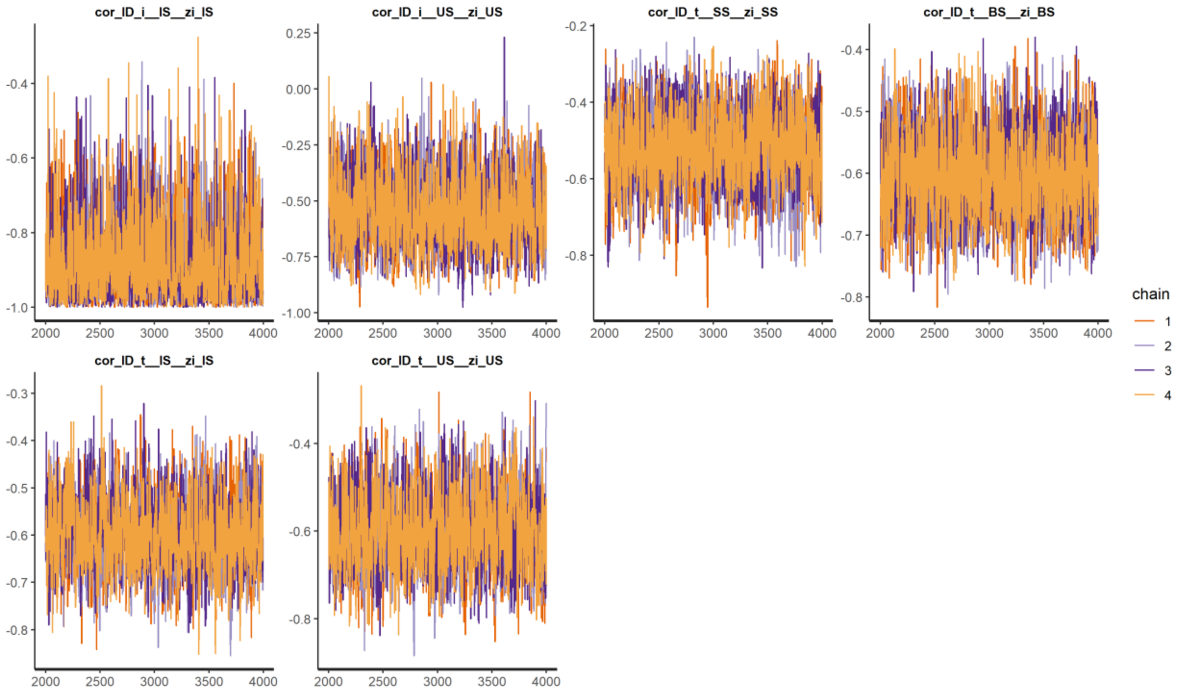


Figure N

Potential Scale Reduction Factors for DIF Model

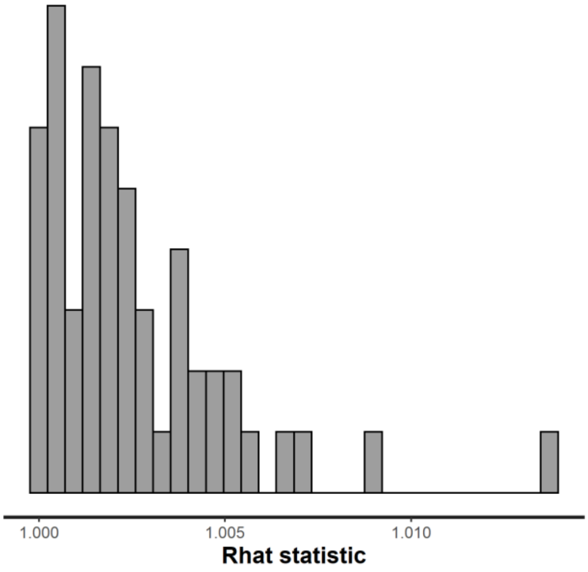


Figure O

Autocorrelation Plots for DIF Model and a Lag of up to 20 for each Chain (Part 1)

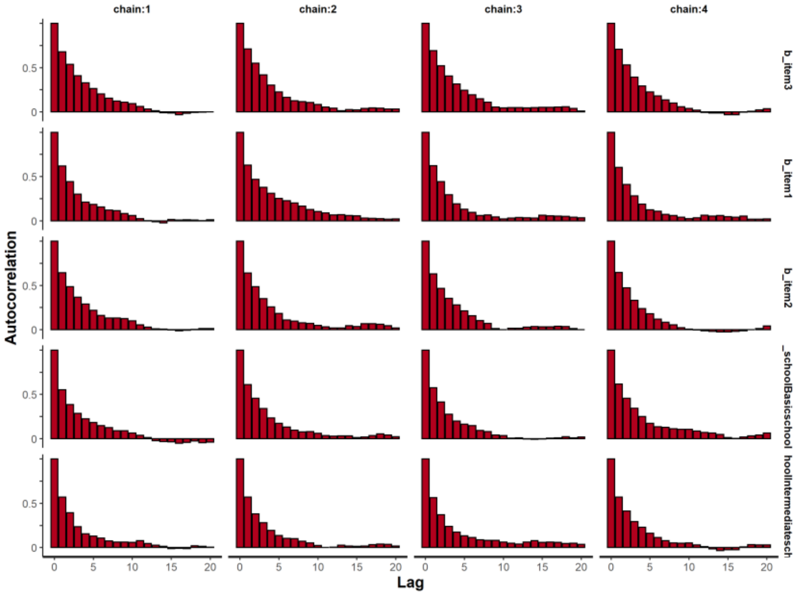


Figure P

Autocorrelation Plots for DIF Model and a Lag of up to 20 for each Chain (Part 2)

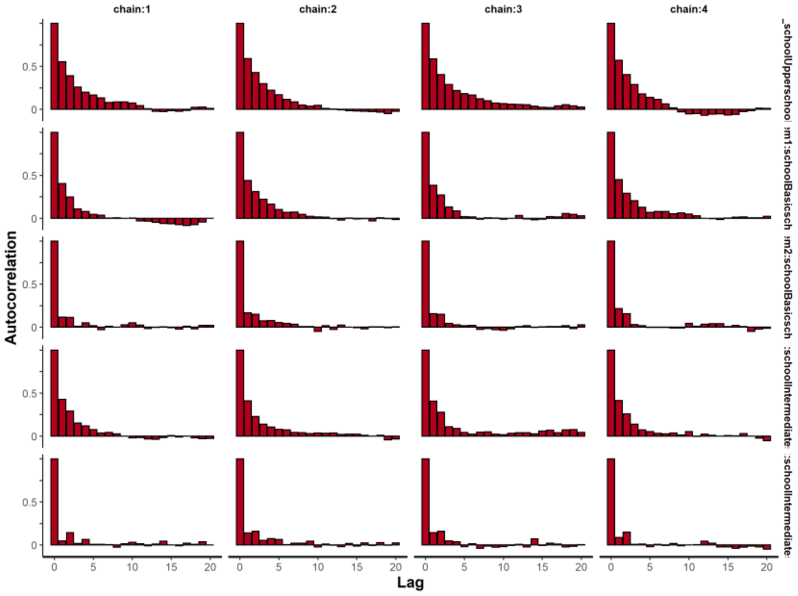


Figure Q

Autocorrelation Plots for DIF Model and a Lag of up to 20 for each Chain (Part 3)

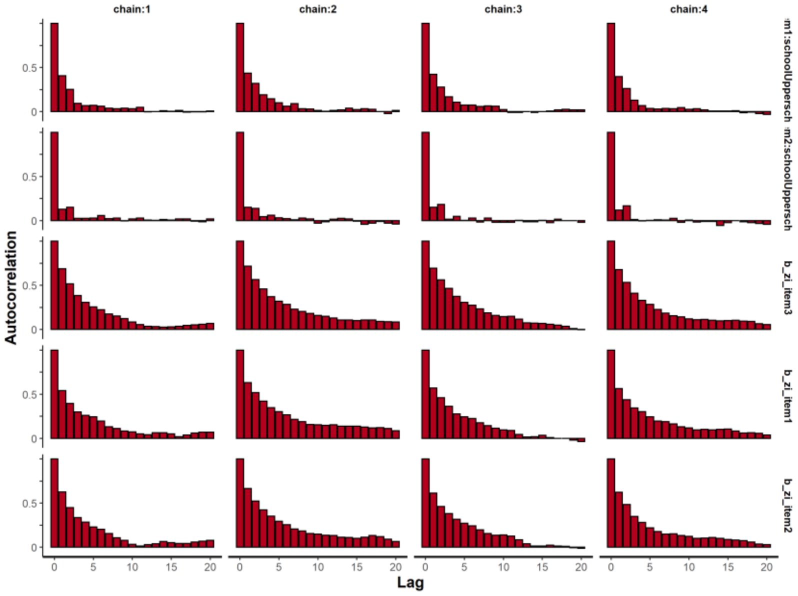


Figure R

Autocorrelation Plots for DIF Model and a Lag of up to 20 for each Chain (Part 4)

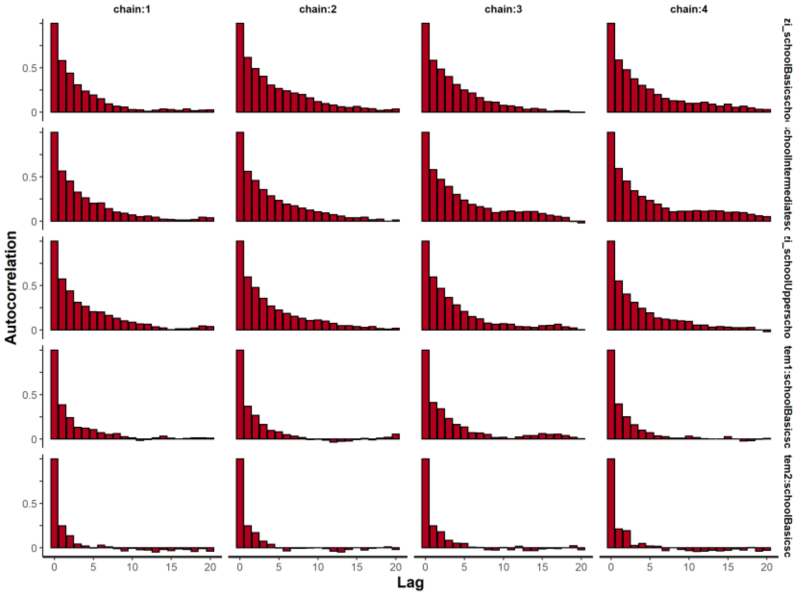


Figure S

Autocorrelation Plots for DIF Model and a Lag of up to 20 for each Chain (Part 5)

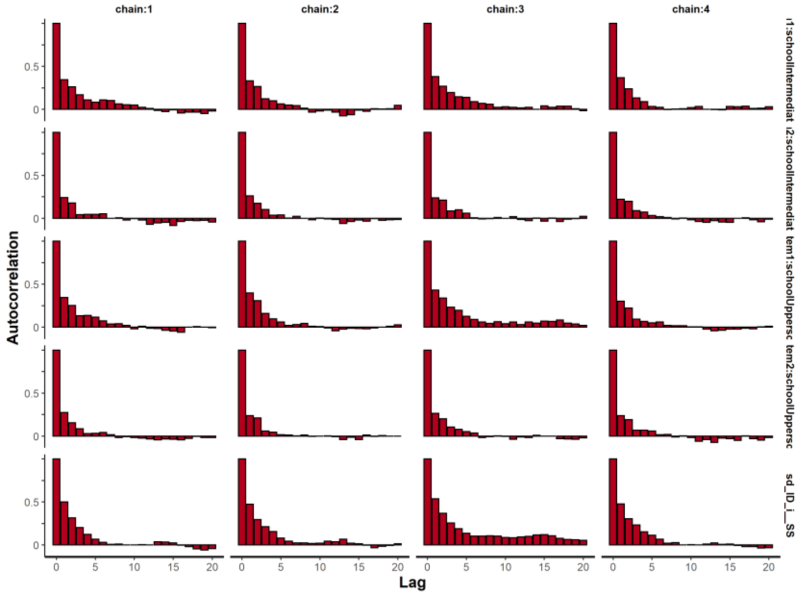


Figure T

Autocorrelation Plots for DIF Model and a Lag of up to 20 for each Chain (Part 6)

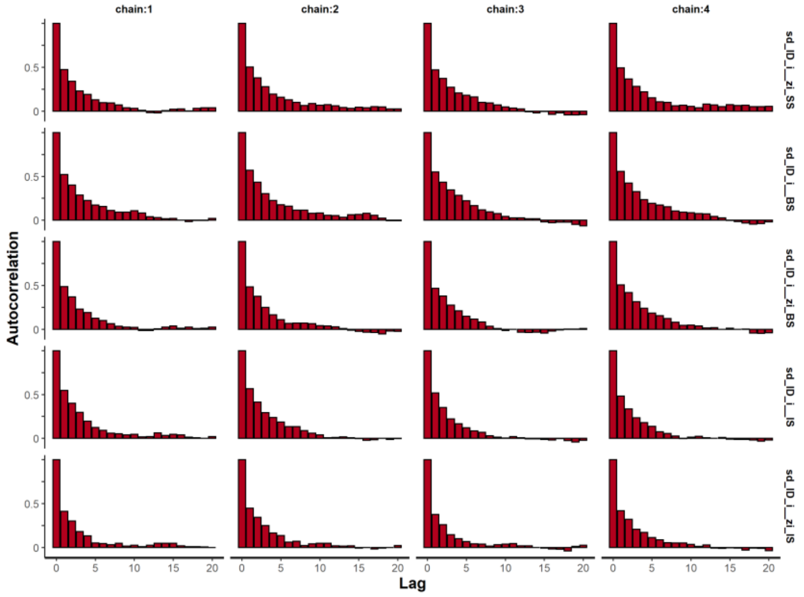


Figure U

Autocorrelation Plots for DIF Model and a Lag of up to 20 for each Chain (part 7)

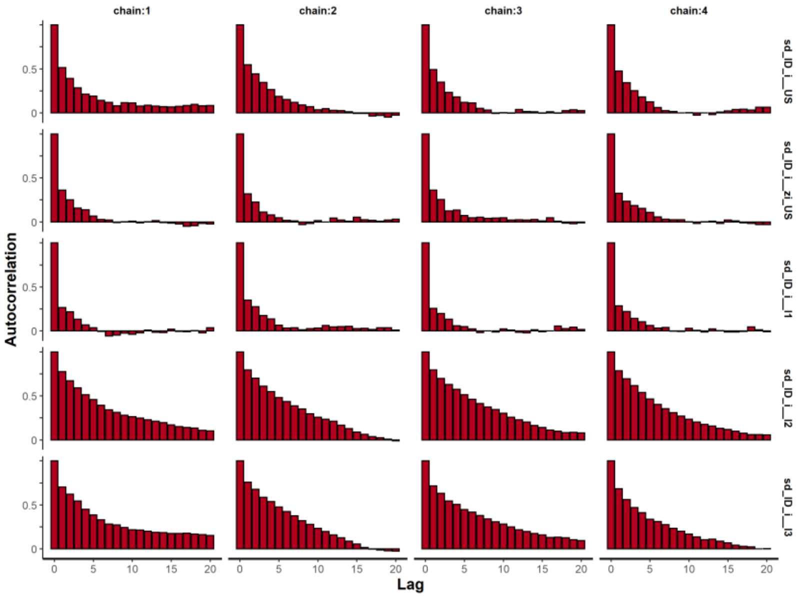


Figure V

Autocorrelation Plots for DIF Model and a Lag of up to 20 for each Chain (part 8)

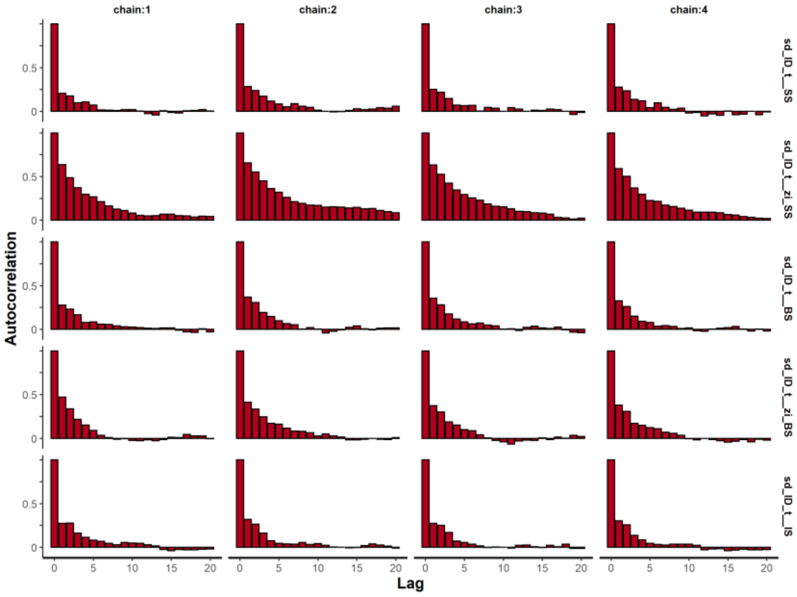


Figure W

Autocorrelation Plots for DIF Model and a Lag of up to 20 for each Chain (Part 9)

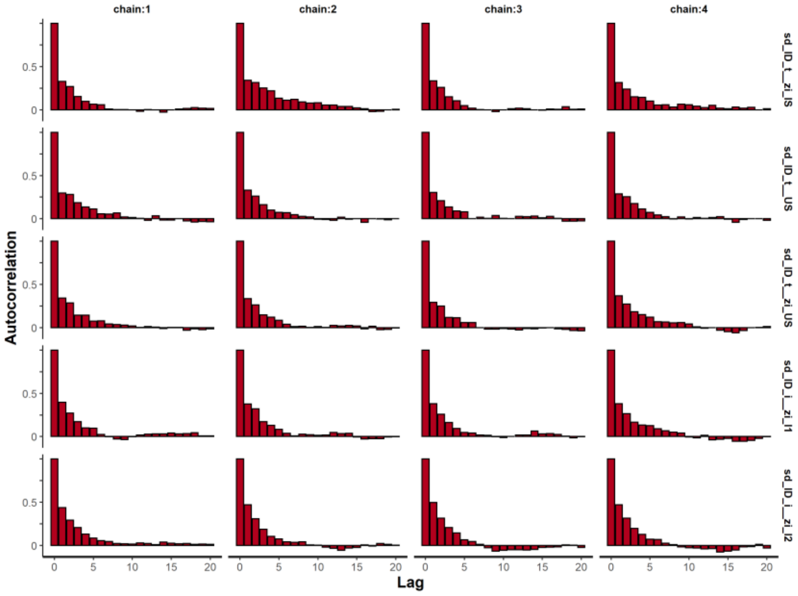


Figure X

Autocorrelation Plots for DIF Model and a Lag of up to 20 for each Chain (Part 10)

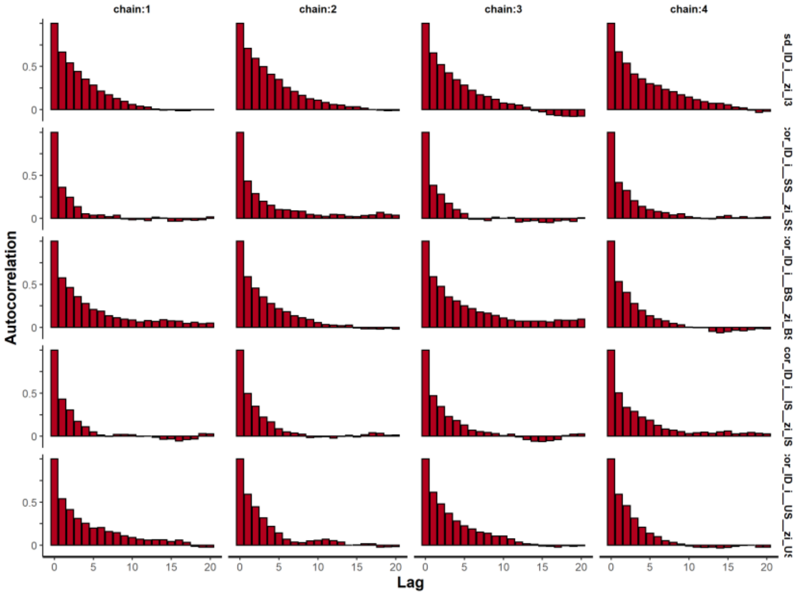


Figure Y

Autocorrelation Plots for DIF Model and a Lag of up to 20 for each Chain (Part 11)

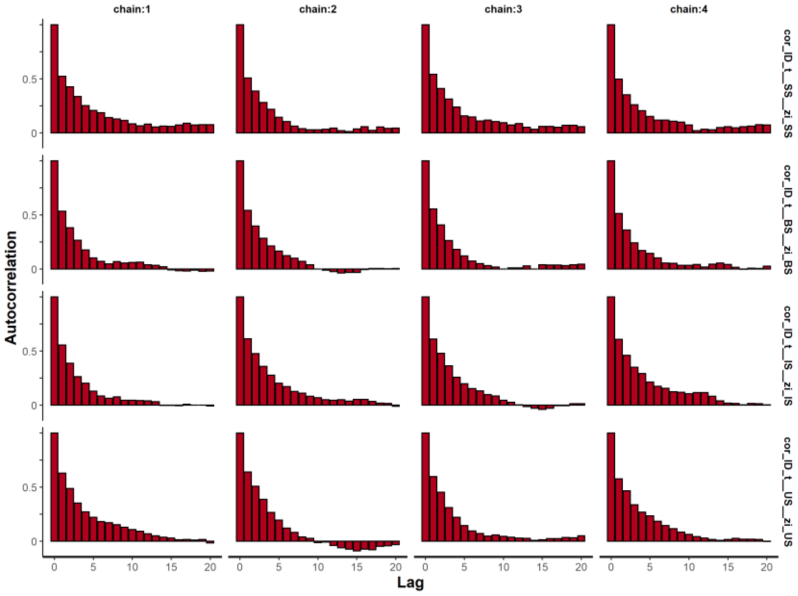


Figure Z

Kernel Density Plots for DIF Model (Part 1)

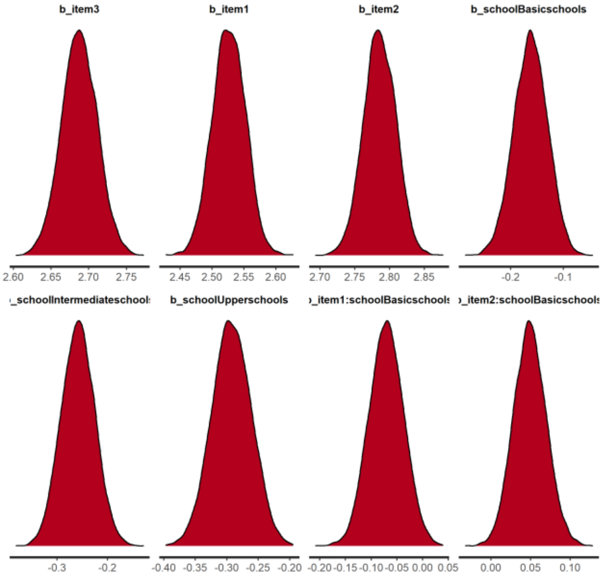


Figure AA

Kernel Density Plots for DIF Model (Part 2)

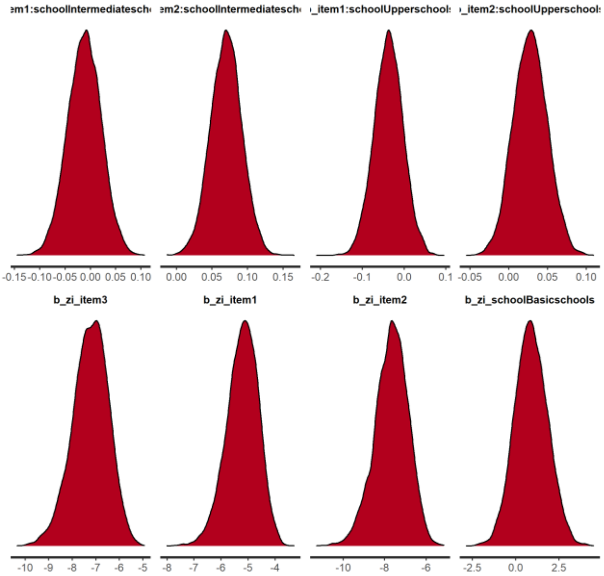


Figure BB

Kernel Density Plots for DIF Model (Part 3)

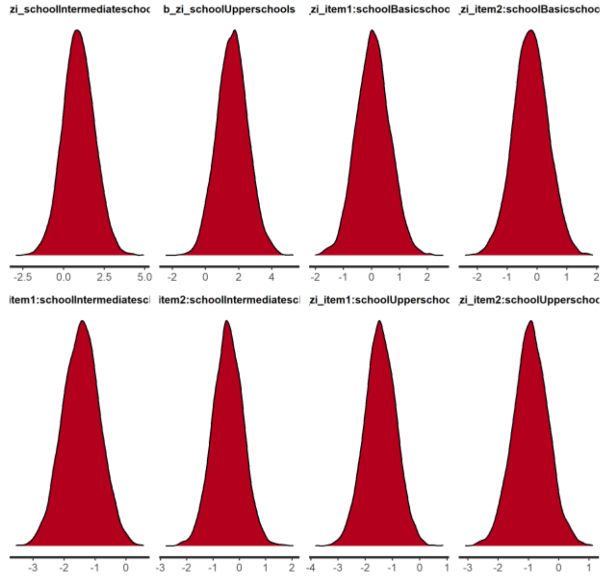


Figure CC

Kernel Density Plots for DIF Model (Part 4)

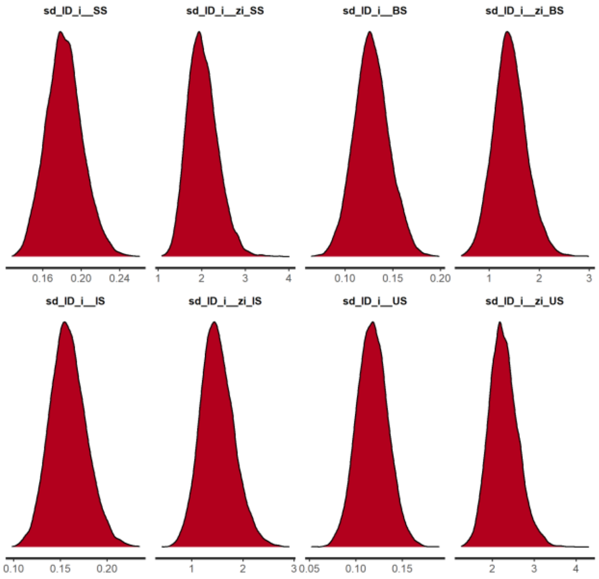


Figure DD

Kernel Density Plots for DIF Model (Part 5)

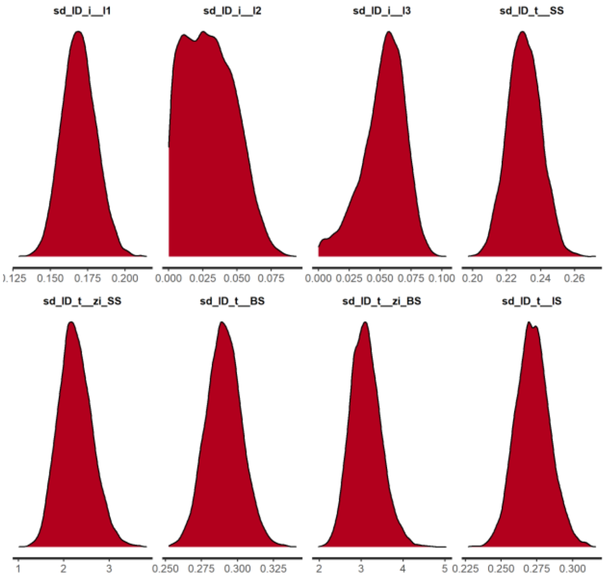


Figure EE

Kernel Density Plots for DIF Model (Part 6)

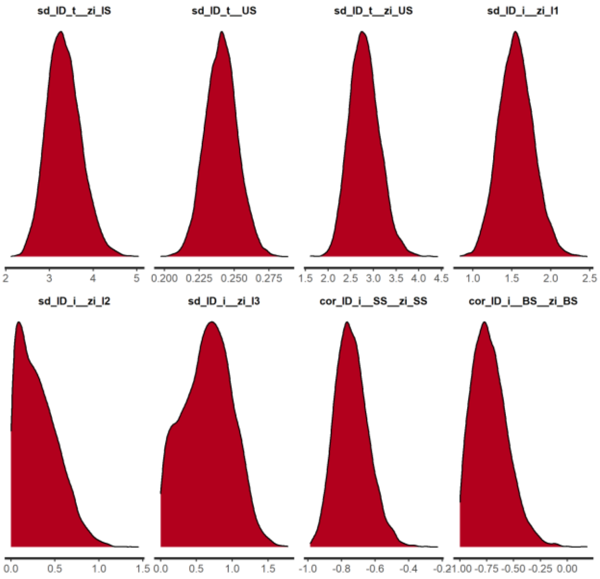
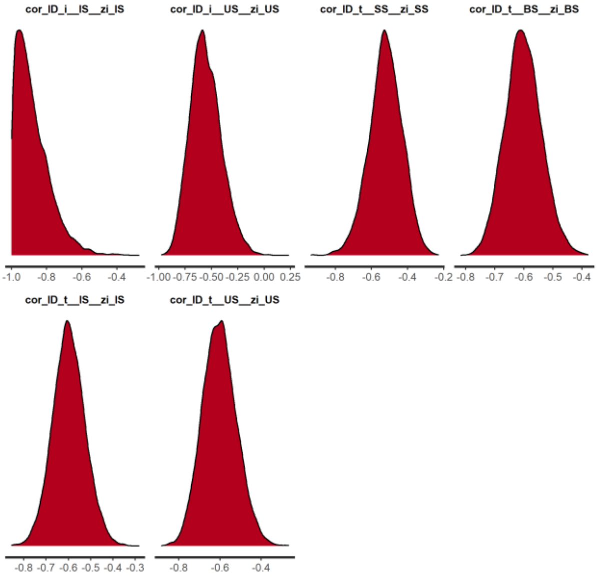


Figure FF

Kernel Density Plots for DIF Model (Part 7)



Parameter Estimates of Zero-Inflated Poisson-Counts Models

	Model without DIF						Model with DIF					
	Mdn	*MAD*	*l.CrI*	*u.CrI*	*ESS*	$\hat{R}$	*Mdn*	*MAD*	*l.CrI*	*u.CrI*	*ESS*	$\hat{R}$
Zero-Inflation Process												
Fixed Effects												
Item 1 (γ_1)	**-4.31**	0.22	-5.89	-4.91	1319	1.00	**-5.17**	0.57	-6.40	-4.12	917	1.01
Item 2 (γ_2)	**-6.18**	0.27	-6.69	-5.63	1400	1.00	**-7.63**	0.81	-9.32	-6.22	1116	1.00
Item 3 (γ_3)	**-5.41**	0.25	-6.75	-5.66	1235	1.00	**-7.19**	0.76	-8.74	-5.83	955	1.00
School type (ref. cat.: SS)												
Basic schools (γ_{BS})							0.90	0.94	-0.85	2.74	1223	1.00
Intermediate schools (γ_{IS})							0.92	0.96	-0.83	2.92	1225	1.00
Upper schools (γ_{US})							1.60	0.88	-0.24	3.38	1110	1.00
DIF effects (ref. cat.: SS, item 3)												
Item 1 x BS ($\gamma_{1,BS}$)							0.07	0.60	-1.17	1.16	2023	1.00
Item 2 x BS ($\gamma_{2,BS}$)							-0.21	0.55	-1.29	0.95	3541	1.00
Item 1 x IS ($\gamma_{1,IS}$)							**-1.46**	0.59	-2.61	-0.32	2176	1.00
Item 2 x IS ($\gamma_{2,IS}$)							-0.43	0.56	-1.62	0.65	3414	1.00
Item 1 x US ($\gamma_{1,US}$)							**-1.45**	0.58	-2.58	-0.32	2119	1.00
Item 2 x US ($\gamma_{2,US}$)							-0.93	0.55	-2.02	0.20	3775	1.00
Random Effects (*SD*)												
Students												
$\sigma_{(\theta_z),(\theta_z)}$	**2.38**	0.15	2.10	2.69	1020	1.01						
$\sigma_{(\theta_z,SS),(\theta_z,SS)}$							**2.24**	0.34	1.59	2.99	972	1.01
$\sigma_{(\theta_z,BS),(\theta_z,BS)}$							**3.08**	0.37	2.43	3.85	2013	1.00
$\sigma_{(\theta_z,IS),(\theta_z,IS)}$							**3.31**	0.38	2.57	4.10	2224	1.00
$\sigma_{(\theta_z,US),(\theta_z,US)}$							**2.79**	0.33	2.17	3.48	2153	1.00
Schools												
$\sigma_{(s_z),(s_z)}$	**1.72**	0.14	1.44	2.01	1367	1.00						
$\sigma_{(s_z,SS),(s_z,SS)}$							**1.99**	0.34	1.38	2.71	1521	1.00

$\sigma_{(s_z,BS),(s_z,BS)}$							**1.40**	0.32	0.83	2.10	1655	1.00
$\sigma_{(s_z,IS),(s_z,IS)}$							**1.47**	0.30	0.87	2.10	1933	1.00
$\sigma_{(s_z,US),(s_z,US)}$							**2.25**	0.33	1.61	2.98	2329	1.00
Items in schools												
$\sigma_{(s_z,1),(s_z,1)}$							**1.55**	0.22	1.15	2.02	2297	1.00
$\sigma_{(s_z,2),(s_z,2)}$							**0.27**	0.23	0.00	0.74	2059	1.00
$\sigma_{(s_z,3),(s_z,3)}$							**0.68**	0.36	0.00	1.21	1020	1.00
Poisson process												
Fixed Effects												
Item 1 (β_1)	**2.33**	0.01	2.31	2.36	1280	1.00	**2.53**	0.03	2.47	2.58	1220	1.01
Item 2 (β_2)	**2.65**	0.01	2.62	2.67	1235	1.01	**2.79**	0.02	2.74	2.83	1287	1.01
Item 3 (β_3)	**2.52**	0.01	2.49	2.54	1290	1.00	**2.69**	0.02	2.64	2.73	1066	1.01
School type (ref. cat.: SS)												
Basic schools (β_{BS})							**-0.16**	0.03	-0.22	-0.10	1386	1.00
Intermediate schools (β_{IS})							**-0.26**	0.03	-0.32	-0.19	1448	1.00
Upper schools (β_{US})							**-0.29**	0.03	-0.35	-0.23	1580	1.00
DIF effects (ref. cat.: SS, item 3)												
Item 1 x BS ($\beta_{1,BS}$)							**-0.07**	0.03	-0.14	-0.00	2379	1.00
Item 2 x BS ($\beta_{2,BS}$)							**0.05**	0.02	0.01	0.09	4377	1.00
Item 1 x IS ($\beta_{1,IS}$)							-0.01	0.03	-0.08	0.05	2356	1.00
Item 2 x IS ($\beta_{2,IS}$)							**0.07**	0.02	0.03	0.11	4687	1.00
Item 1 x US ($\beta_{1,US}$)							-0.04	0.03	-0.10	0.03	2232	1.00
Item 2 x US ($\beta_{2,US}$)							0.03	0.02	-0.01	0.07	4338	1.00
Random Effects (*SD*)												
Students												
$\sigma_{(\theta),(\theta)}$	**0.26**	0.01	0.25	0.27	1603	1.00						
$\sigma_{(\theta,SS),(\theta,SS)}$							**0.23**	0.01	0.21	0.25	2547	1.00

$\sigma_{(\theta,BS),(\theta,BS)}$							**0.29**	0.01	0.27	0.31	2377	1.00
$\sigma_{(\theta,IS),(\theta,IS)}$							**0.27**	0.01	0.25	0.30	2394	1.00
$\sigma_{(\theta,US),(\theta,US)}$							**0.24**	0.01	0.22	0.26	2624	1.00
Schools												
$\sigma_{(s),(s)}$	**0.18**	0.01	0.16	0.20	1174	1.00						
$\sigma_{(s,SS),(s,SS)}$							**0.18**	0.02	0.15	0.22	2008	1.00
$\sigma_{(s,BS),(s,BS)}$							**0.13**	0.02	0.09	0.16	1279	1.00
$\sigma_{(s,IS),(s,IS)}$							**0.16**	0.02	0.12	0.20	1798	1.00
$\sigma_{(s,US),(s,US)}$							**0.12**	0.02	0.09	0.15	1723	1.00
Items in schools												
$\sigma_{(s,1),(s,1)}$							**0.17**	0.01	0.15	0.19	3061	1.00
$\sigma_{(s,2),(s,2)}$							0.03	0.02	0.00	0.06	411	1.01
$\sigma_{(s,3),(s,3)}$							**0.05**	0.02	0.01	0.08	548	1.01
Correlations between Random Effects												
$\sigma_{(\theta),(\theta_z)}$	**-0.57**	0.04	-0.65	-0.49	985	1.00						
$\sigma_{(\theta,SS),(\theta_z,SS)}$							**-0.52**	0.09	-0.70	-0.34	1472	1.00
$\sigma_{(\theta,BS),(\theta_z,BS)}$							**-0.60**	0.06	-0.73	-0.48	1397	1.00
$\sigma_{(\theta,IS),(\theta_z,IS)}$							**-0.60**	0.08	-0.74	-0.46	1275	1.00
$\sigma_{(\theta,US),(\theta_z,US)}$							**-0.60**	0.08	-0.76	-0.44	1031	1.00
$\sigma_{(s),(s_z)}$	**-0.68**	0.06	-0.79	-0.57	1501	1.00						
$\sigma_{(s,SS),(s_z,SS)}$							**-0.75**	0.10	-0.93	-0.55	2000	1.00
$\sigma_{(s,BS),(s_z,BS)}$							**-0.73**	0.16	-1.00	-0.42	1196	1.00
$\sigma_{(s,IS),(s_z,IS)}$							**-0.90**	0.09	-1.00	-0.69	1523	1.00
$\sigma_{(s,US),(s_z,US)}$							**-0.55**	0.15	-0.82	-0.24	1368	1.00

Note. Mdn = Median of posterior distribution, *MAD* = Median absolute deviation of posterior distribution, *l.CrI* = Lower boundary of the 95% credibility interval, *u.CrI* = Upper boundary of the 95% credibility interval, ESS = Effective sample size, $\hat{R}$ = Potential scale reduction factor, SS = Special schools, BS = Basic secondary schools, IS = Intermediate secondary schools, US = Upper secondary schools. Values in boldface indicate fixed effects for which zero is not contained in the credible interval or the posterior probabilities of a variance of 0 exceeds 90%. Parameter names correspond to equations [11] and [13].

Sensitivity Analyses

It could be argued that specific background characteristics of students are defining attributes of specific school types. For example, it is well known that girls are more likely to attend upper secondary schools, while boys are more likely to attend basic schools. By matching our analyzed samples on sociodemographic characteristics (i.e., sex, age, migration background, cultural capital), we examined hypothetical effects while controlling for these sociodemographic characteristics. This allowed us to investigate whether specific psychological profiles of students with SEN-L or specific context conditions associated with different school tracks introduced DIF. Because applied researchers might also be interested in knowing whether the observed samples including their specific sociodemographic patterns exhibit DIF, we replicated our primary analyses for a random sample of schools from each school track. Because the available sample sizes for the three school tracks were highly unbalanced with substantially more students attending, for example, upper school as compared to special schools, we tried to create roughly equally sized samples for each school track. Therefore, we conducted a two-step sampling approach. First, we randomly drew 99 schools (i.e., the number of schools available for special schools) from basic, intermediate, and upper schools. In the second step, we randomly drew a proportion of students from each of these schools to create a total sample size in each school track that matched the sample size available for special schools. Basic information on the original sample and the thus created random sample are given in Table S3. The summary highlights that the random samples exhibited similar differences in the background characteristics as the original sample. For example, 55% of students in upper schools were girls, while this was the case for only 44% of students in special schools. Thus, the random samples allowed us to examine DIF across school types while maintaining sociodemographic differences across school types.

Table S3

Characteristics of the Original and Random Samples by School Type

	Original Sample				Random Sample			
Variable	SS	BS	IS	US	SS	BS	IS	US
N	911	2913	3002	4754	911	981	968	956
n	9	17	29	33	9	9	10	10
Number of schools	99	158	103	147	99	99	99	99
Female (%)	0.44	0.44	0.49	0.55	0.44	0.45	0.48	0.55
Age (*M*)	15.98	15.89	15.62	15.37	15.98	15.90	15.60	15.40
(*SD*)	0.64	0.69	0.59	0.50	0.64	0.71	0.56	0.51
Migration (%)	0.28	0.38	0.23	0.18	0.28	0.40	0.23	0.20
Cultural capital (*M*)	2.38	2.96	3.71	4.56	2.38	2.89	3.70	4.54
(*SD*)	1.34	1.40	1.34	1.23	1.34	1.39	1.48	1.26
Reasoning (*M*)	-1.29	-0.52	0.07	0.52	-1.75	-0.72	0.07	0.63
(*SD*)	0.87	0.87	0.84	0.79	1.14	1.10	1.07	0.96

Note. N = Total sample size, *n* = Median number of students per school, SS = Special schools, BS = Basic secondary schools, IS = Intermediate secondary schools, US = Upper secondary schools.

The parameter estimates of the DIF model for the unmatched samples are given in Table S4. For a large part, these results replicated the findings from the matched samples. The estimated item difficulties were equivalent and did not change. Also, standardized mean differences Δ between school tracks of -0.68, -0.95, and -1.15 were comparable to the respective results found for the matched samples. Using item 3 as an anchor item, item 1 was again easier in basic schools as compared to special schools, while item 2 was more difficult in basic schools and intermediate schools. Importantly, the respective standardized effects Δ of -0.20, 0.13, 0.23 were slightly smaller as compared to the DIF effects reported for the matched samples and, thus, did not reach our threshold for non-negligible DIF. Moreover, only the credible interval for the DIF effect for intermediate schools did not contain 0. Finally, we found comparable reliabilities in the four school tracks as compared to the matched samples (.74, .78, .75, and .73). In conclusion, these results give even greater confidence in using the administered test for perceptual speed in special schools because even the unmatched sample found no pronounced DIF that might bias comparisons across school tracks.

Table S4

Parameter Estimates of Zero-Inflated Poisson-Counts Models for Random Sample

	Mdn	*MAD*	*l.CrI*	*u.CrI*	*ESS*	$\hat{R}$
			Zero-Inflation Process			
Fixed Effects						
Item 1 (γ_1)	**-5.09**	0.57	-6.26	-4.07	837	1.00
Item 2 (γ_2)	**-7.54**	0.76	-9.16	-6.20	995	1.00
Item 3 (γ_3)	**-7.05**	0.72	-8.42	-5.68	978	1.00
School type (ref. cat.: SS)						
Basic schools (γ_{BS})	0.64	0.89	-1.10	2.38	1013	1.00
Intermediate schools (γ_{IS})	1.30	0.87	-4.06	2.93	880	1.00
Upper schools (γ_{US})	1.20	0.87	-5.54	2.93	932	1.00
DIF effects (ref. cat.: SS, item 3)						
Item 1 x BS ($\gamma_{1,BS}$)	-0.01	0.55	-1.03	1.09	1997	1.00
Item 2 x BS ($\gamma_{2,BS}$)	0.43	0.51	-5.12	1.43	3494	1.00
Item 1 x IS ($\gamma_{1,IS}$)	**-1.30**	0.53	-2.43	-0.35	2333	1.00
Item 2 x IS ($\gamma_{2,IS}$)	-0.23	0.50	-1.21	0.75	3644	1.00
Item 1 x US ($\gamma_{1,US}$)	**-1.13**	0.53	-2.19	-0.09	2131	1.00
Item 2 x US ($\gamma_{2,US}$)	-0.24	0.50	-1.18	0.79	3712	1.00
Random Effects (*SD*)						
Students						
$\sigma_{(\theta_z),(\theta_z)}$						
$\sigma_{(\theta_z,SS),(\theta_z,SS)}$	**2.21**	0.35	1.56	2.89	1427	1.01
$\sigma_{(\theta_z,BS),(\theta_z,BS)}$	**3.00**	0.30	2.42	3.61	2067	1.00
$\sigma_{(\theta_z,IS),(\theta_z,IS)}$	**2.74**	0.30	2.17	3.37	1603	1.00
$\sigma_{(\theta_z,US),(\theta_z,US)}$	**2.77**	0.32	2.17	3.44	2004	1.00
Schools						
$\sigma_{(s_z),(s_z)}$						
$\sigma_{(s_z,SS),(s_z,SS)}$	**1.97**	0.33	1.37	2.68	1009	1.00
$\sigma_{(s_z,BS),(s_z,BS)}$	**1.95**	0.31	1.39	2.62	2011	1.00
$\sigma_{(s_z,IS),(s_z,IS)}$	**1.51**	0.29	0.94	2.08	1733	1.00
$\sigma_{(s_z,US),(s_z,US)}$	**1.65**	0.28	1.11	2.27	1464	1.00
Items in schools						
$\sigma_{(s_z,1),(s_z,1)}$	**1.39**	0.20	1.01	1.81	2313	1.00
$\sigma_{(s_z,2),(s_z,2)}$	0.15	0.13	0.00	0.44	2925	1.00
$\sigma_{(s_z,3),(s_z,3)}$	0.40	0.31	0.00	0.91	1191	1.00
			Poisson Process			
Fixed Effects						
Item 1 (β_1)	**2.53**	0.03	2.48	2.73	1157	1.00
Item 2 (β_2)	**2.79**	0.02	2.74	2.58	1049	1.00

Item 3 (β_3)	**2.69**	0.02	2.64	2.83	928	1.00
School type (ref. cat.: SS)						
Basic schools (β_{BS})	**-0.18**	0.03	-0.24	-0.11	1221	1.00
Intermediate schools (β_{IS})	**-0.25**	0.03	-0.31	-0.19	1131	1.00
Upper schools (β_{US})	**-0.28**	0.03	-0.34	-0.22	1190	1.00
DIF effects (ref. cat.: SS, item 3)						
Item 1 x BS ($\beta_{1,BS}$)	-0.05	0.03	-0.12	0.01	2261	1.00
Item 2 x BS ($\beta_{2,BS}$)	0.04	0.02	-0.01	0.08	3383	1.00
Item 1 x IS ($\beta_{1,IS}$)	-0.00	0.03	-0.07	0.06	2389	1.00
Item 2 x IS ($\beta_{2,IS}$)	**0.06**	0.02	0.01	0.10	3085	1.00
Item 1 x US ($\beta_{1,US}$)	-0.03	0.03	-0.10	0.03	2161	1.00
Item 2 x US ($\beta_{2,US}$)	0.04	0.02	-0.01	0.08	3441	1.00
Random Effects (*SD*)						
Students						
$\sigma_{(\theta),(\theta)}$						
$\sigma_{(\theta,SS),(\theta,SS)}$	**0.23**	0.01	0.21	0.25	1831	1.00
$\sigma_{(\theta,BS),(\theta,BS)}$	**0.29**	0.01	0.26	0.31	1225	1.00
$\sigma_{(\theta,IS),(\theta,IS)}$	**0.26**	0.01	0.24	0.28	1758	1.00
$\sigma_{(\theta,US),(\theta,US)}$	**0.24**	0.01	0.22	0.26	1695	1.00
Schools						
$\sigma_{(s),(s)}$						
$\sigma_{(s,SS),(s,SS)}$	**0.18**	0.02	0.15	0.22	2729	1.00
$\sigma_{(s,BS),(s,BS)}$	**0.14**	0.02	0.10	0.17	2038	1.00
$\sigma_{(s,IS),(s,IS)}$	**0.14**	0.02	0.11	0.18	1867	1.00
$\sigma_{(s,US),(s,US)}$	**0.12**	0.02	0.09	0.15	2884	1.00
Items in schools						
$\sigma_{(s,1),(s,1)}$	**0.16**	0.01	0.14	0.18	2594	1.00
$\sigma_{(s,2),(s,2)}$	0.02	0.02	0.00	0.06	504	1.01
$\sigma_{(s,3),(s,3)}$	**0.08**	0.01	0.06	0.10	1045	1.00
	Correlations between Random Effects					
$\sigma_{(\theta),(\theta_z)}$						
$\sigma_{(\theta,SS),(\theta_z,SS)}$	**-0.55**	0.09	-0.72	-0.37	1831	1.00
$\sigma_{(\theta,BS),(\theta_z,BS)}$	**-0.63**	0.06	-0.74	-0.52	1225	1.00
$\sigma_{(\theta,IS),(\theta_z,IS)}$	**-0.68**	0.07	-0.81	-0.54	1758	1.00
$\sigma_{(\theta,US),(\theta_z,US)}$	**-0.70**	0.08	-0.85	-0.54	1695	1.00
$\sigma_{(s),(s_z)}$						
$\sigma_{(s,SS),(s_z,SS)}$	**-0.72**	0.10	-0.90	-0.51	2729	1.00
$\sigma_{(s,BS),(s_z,BS)}$	**-0.73**	0.12	-0.94	-0.49	2038	1.00
$\sigma_{(s,IS),(s_z,IS)}$	**-0.73**	0.12	-0.98	-0.49	1867	1.00
$\sigma_{(s,US),(s_z,US)}$	**-0.47**	0.17	-0.77	-0.13	2884	1.00

Note. Mdn = Median of posterior distribution, *MAD* = Median absolute deviation of posterior distribution, *l.CrI* = Lower boundary of the 95% credibility interval, *u.CrI* = Upper boundary of the 95% credibility interval, ESS = Effective sample size, $\hat{R}$ = Potential scale reduction factor, SS = Special schools, BS = Basic secondary schools, IS = Intermediate secondary schools, US = Upper secondary schools. Values in boldface indicate fixed effects for which zero is not contained in the credible interval or the posterior probabilities of a variance of 0 exceeds 90%. Parameter names correspond to equation [13].

References

Bolker, B., & Robinson (2020). broom.mixed: Tidying methods for mixed models (Version 0.2.6) [Computer software].

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal, 10*(1), 395-411. https://doi.org/10.32614/RJ-2018-017

Eddelbuettel, D. (2013). *Seamless R and C++ integration with Rcpp*. Springer Publishing. https://doi.org/10.1007%2F978-1-4614-6868-4

Hansen, B. B., & Klopfer, S. O. (2006). Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics*, *15*(3), 609-627. https://doi.org/10.1198/106186006X137047

Keele, L., Pimentel, S., & Rosenbaum, P. (2018). *matchMulti: Optimal multilevel matching using a network algorithm (Version 1.1.7) [Computer software]*.

Kelley, K. (2020). *MBESS: The MBESS R package (Version 4.8.0) [Computer software]*.

Larmarange, J. (2020). *labelled: Manipulating labelled data (Version 2.7.0) [Computer software]*.

Lüdecke D (2020). *sjlabelled: Labelled data utility functions* (Version 1.1.7) [Computer software]. https://doi.org/10.5281/zenodo.1249215

Makowski, D., Ben-Shachar, M., & Lüdecke, D. (2019). bayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework. *Journal of Open Source Software, 4*(40), 1541. https://doi.org/10.21105/joss.01541

Müller, K. (2020). *here: A simpler way to find your files (Version 1.0.0) [Computer software]*.

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov Chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24*(2), 146–178. https://doi.org/10.2307/1165199

Pimentel, S. D., Page, L. C., Lenard, M., & Keele, L. (2018). Optimal multilevel matching using network flows: An application to a summer reading intervention. *The Annals of Applied Statistics*, *12*(3), 1479-1505. https://doi.org/10.1214/17-AOAS1118

R Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Revelle, W. (2020) *psych: Procedures for personality and psychological research (Version 2.0.9) [Computer software].*

Robitzsch, A., Kiefer, T., & Wu, M. (2020). *TAM: Test analysis modules (Version 3.5-19) [Computer software]*.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1-36. URL http://www.jstatsoft.org/v48/i02/

Sieben, S., & Lechner, C. M. (2019). Measuring cultural capital through the number of books in the household. *Measurement Instruments for the Social Sciences*, *1*(1), 1. https://doi.org/10.1186/s42409-018-0006-0

Signorell, A. et al. (2021). DescTools: Tools for descriptive statistics (Version 0.99.41) [Computer software].

Stan Development Team (2020*). RStan: The R interface to Stan* (Version 2.21.2) [Computer software].

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of MCMC. *Bayesian Analysis*. *Advance online publication*. https://doi.org/10.1214/20-BA1221

Warnes, G. R., Bolker, B., & Lumley, T. (2020). gtools: Various R programming tools (Version 3.8.2) [Computer software].

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag.

Wickham et al. (2019). Welcome to the tidyverse. *Journal of Open Source Software, 4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H. (2020). *tidyr: Tidy messy data* (Version 1.1.2) [Computer software].

Wickham, H., François, R., Henry, L., & Müller, K. (2020). *dplyr: A grammar of data manipulation* (Version 1.0.2) [Computer software].

Wickham, H., & Miller, E. (2020). *haven: Import and export 'SPSS', 'Stata' and 'SAS' files* (Version 2.3.1) [Computer software].

Wilke, C.O. (2020). *cowplot: Streamlined plot theme and plot annotations for 'ggplot2'* (Version 1.1.0) [Computer software].

Xie, Y. (2020). *knitr: A general-purpose package for dynamic report generation in R* (Version 1.30) [Computer software].

Xie, Y., Dervieux, C., & Riederer, E. (2020). *R Markdown Cookbook*. Chapman and Hall/CRC.