Rohm *et al. Large-scale Assess Educ* (2021) 9:23
https://doi.org/10.1186/s40536-021-00116-2

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, $N(0, \infty)$, 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

I*A* 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