RESEARCH LINE 5

EDUCATIONAL ASSESSMENT AND EDUCATIONAL MEASUREMENT

EDUCATIONAL ASSESSMENT AND EDUCATIONAL MEASUREMENT

The IPN Research Line 5 conducts projects in (1) Educational Assessment and in (2) Educational Measurement. Whereas educational assessment refers to collecting data to obtain information about variables such as students' learning outcomes and to gauge the results relative to particular benchmarks or standards, educational measurement refers to the statistical methods that are employed in educational assessment. Research Line 5 is thus dedicated to advancing educational assessment and investigating the results obtained as well as to studying methodological problems in empirical educational and psychological studies.

Important activities addressing educational assessment refer to test development, large-scale assessment like the Programme for International Student Assessment (PISA), and to empirical studies concerning the validity of competence tests. Of particular interest are tests assessing mathematics, science, and computer competences covering the lifespan. Based on models of competence structures developed in close cooperation with Research Line 2, researchers in Line 5 develop measurement models that form the basis for producing large item pools. A special challenge here is to provide items suitable to measure competences in various age groups (starting in kindergarten up to late adulthood). Consequently, the question arises whether the same construct can be assessed by means of a combination of common items and age-specific items over time. Complex statistical analyses based on item response theory (IRT) and structural equation modeling (SEM) are thus applied to data from empirical studies to test assumptions of measurement invariance over time. Further activities in educational assessment refer to the analysis of students' educational outcomes at the end of elementary as well as lower and upper secondary school. Large-scale assessments are used to acquire knowledge about the effectiveness of the German educational system. The IPN is part of the Centre for International Student Assessment (in German: Zentrum für Internationale Bildungsvergleichsstudien, ZIB) constituted by the TUM School of Education, the German Institute for International Educational Research (DIPF), and by the IPN. The ZIB conducted the PISA 2018 survey in Germany and published the PISA 2015 report on reading, mathematics, and science competences of 15-year-old students in Germany as compared with those of same-aged students from other countries. Researchers from the IPN were involved in the reporting of students' mathematics and science competences. Finally, studies in educational assessment address issues of validity. Over the past years tests for mathematics and sciences competences have been developed for different purposes. In the foreground are the international compar-

RESPONSIBLE FOR RESEARCH LINE 5:

Oliver Lüdtke (spokesperson),
Gabriel Nagy, Aiso Heinze
& Olaf Köller

ison of student competences (e.g., PISA), the examination of persons' competences over the lifespan as conducted by the National Educational Panel Study (NEPS), and monitoring the achievement of the educational standards in Germany. All tests are based on similar but to some extent different frameworks. From both a theoretical and an empirical point of view the question arises whether tests from these different assessments measure the same or different latent constructs.

The scope of the research conducted with regard to the topic of educational measurement also encompasses aspects relevant to large-scale studies. Classrooms form important developmental environments for students and research in teaching and learning is often interested in effects of context variables (e.g., teaching variables like classroom management) or composition characteristics (e.g., average socio-economic status of all classmates). One line of research at the IPN aims to develop multilevel modeling techniques that allow an appropriate modeling of context and composition effects in educational research. Another line of research focuses on new developments in the area of measurement models. Current research addresses models for assessing student achievement on the one hand and interest profiles on the other hand as well as models that allow in-depth examinations of the criterion validity of theoretical constructs. A third line of research deals with challenges typically encountered in the handling of missing data which represent a pervasive problem in educational research (e.g., students refuse to participate or provide only incomplete information). The current research develops strategies for dealing with missing data in multilevel designs (e.g., students nested within classroom) and provides software solutions that help applied researchers to implement the proposed strategies. Finally, a fourth line of research focuses on the estimation of causal effects with non-experimental data. Evidence-based educational research is interested in assessing the effects of targeted interventions on educational outcomes. In many cases, controlled experiments in which classes or students are randomly assigned to different treatments (e.g., low vs. high teaching quality) are not possible and researchers have to rely on observational data to test the effect of potential interventions. Current research at the IPN evaluates statistical procedures (e.g., propensity score matching) and designs (e.g., regression discontinuity design) that were proposed for drawing causal inferences from non-experimental data.

1 Educational Assessment

Core activities with regard to this research topic refer to international large-scale assessments (ILSA) such as PISA and the Trends in International Mathematics and Science Study (TIMSS). The main indicators of educational systems' outcomes provided by these studies are aggregations of student level data (domain-specific competences); they signify what students can do with what they have learned in school. PISA is a triennial cross-national assessment of 15-year-old students' key competences in reading, mathematics, and science. TIMSS on the other hand, assesses competences in terms of mathematical literacy and scientific literacy of fourth- and eighth-grade students every four years. Such comparisons with other countries allow gauging competences of students in Germany, and thus, evaluating the German educational system on both the elementary and the secondary level.

A further core project is NEPS which started in 2008 and examines competences of persons from different age cohorts in longitudinal studies covering the entire lifespan. The investigated age cohorts range from infancy to late adulthood. NEPS is funded by the Leibniz Association and is conducted by a research and infrastructure network under the auspices of the Leibniz Institute for Educational Trajectories (Leibniz-Institut für Bildungsverläufe) at the University of Bamberg. The IPN is one of several partners in this network. NEPS collects data in relevant areas of the German educational system so that central problems of empirical educational research can be addressed by analyzing the data. Within NEPS, the IPN develops and validates instruments to assess competences in mathematical literacy, scientific literacy, and computer literacy (ICT literacy). Competences in these domains will be assessed consistently and coherently over the lifespan so that persons' cumulative development can be reconstructed across educational stages. NEPS has a multi-cohort-sequence design. Particular challenges in NEPS refer to test mode changes from paper pencil tests to computer-based tests.

One major aim of ILSA is to monitor changes in student performance over time. A set of common items is repeatedly administered in each assessment and linking methods are used to align the results from the different assessments on a common scale to accomplish this task. In the following, the results of two research projects are reported that deal with methodological challenges when using linking methods to provide trend estimates in performance across assessments. First, PISA is conducted every three years to provide information on the effectiveness of educational systems. It is necessary to keep the test conditions and statistical methods in all PISA assessments as constant as possible to ensure valid trend estimation. However, several changes were implemented (test mode changed from paper pencil to computer tests, scaling models were changed) in PISA 2015. In the current research project, we investigated the effects of these changes on trend estimation in PISA using German data from all PISA cycles (2000–2015). Second, a framework is presented to discuss linking errors in ILSA. In this framework different components of linking errors are distinguished which can be used to quantify the uncertainty of national trend estimates.

1.1 Estimating trends in German PISA data 2015

Background

Initial reports on findings from PISA 2015 present a substantial decrease of science competences in Germany. While in 2012 on average German 15 year-old students reached $M(2012) = 524$ points on the international scale, their mean score in 2015 was only $M(2015) = 509$. Partially in line with this finding is the international difference (OECD mean) of minus 8 points between 2012 and 2015. The negative change between the two PISA time points raises the question whether this really represents decreasing science competences or whether the many changes in the test design of PISA 2015 might have caused differences in competences. The current research investigates two of the major changes in 2015 in detail:

1. PISA 2015 was the first computer-based assessment of students, while previous PISA assessments used paper-pencil tests. It is again an important research question whether the test medium moderates the estimation of country means and their trends (so-called mode effect).
2. Item calibration based on item response theory (IRT) has changed from the one-parameter model (1PL model) only estimating item difficulties, to a two-parameter model (2PL model) estimating item difficulties plus item discrimination parameters. The question addressed is whether the change of the IRT-model moderates means and trends of the participating countries.

Bearing in mind these two research questions, we present re-analyses of the German PISA data sets from 2000 to 2015. We focus on science, reading, and mathematics scores of German 15 year-old students and explore how changes in the IRT model as well as the change from paper-pencil to computer-based assessment affect the country mean. In addition, we use data from the German field trial for PISA 2015 conducted in spring 2014 to explore the mode effect. As students within schools were randomly assigned to two experimental conditions (paper-pencil vs. computer), the field test is an experimental study that allows a causal interpretation of a potential mode effect.

The German field trial in 2014

Robitzsch, A., Lüdtke, O., Köller, O., Kröhne, U., Goldhammer, F., & Heine, J.-H. (2017). Herausforderungen bei der Schätzung von Trends in Schulleistungsstudien. Eine Skalierung der deutschen PISA-Daten [Challenges in estimations of trends in large-scale assessments]. *Diagnostica, 63*(2), 148–165.

German students from 39 schools of the 2014 field trial were randomly assigned to two experimental conditions. While one group worked on paper-pencil tests (PBA group; *N* = 517), the other group worked on the same items presented on a computer (CBA group; *N* = 506). Table 1 provides findings of the field trial data based on IRT analyses applying the 1PL model. We start with results for all items of the field trial. There is clear evidence for mode effects in all three domains, that is, CBA items were on average more difficult than PBA items. Overall, with differences of 21, 14, and 16 points on the PISA scale, these effects are substantial and of practical importance. Differences between the three domains are not statistically significant, suggesting that the mode effect is of the same size across domains. The mean mode-effect of 17 points also reaches statistical significance.

Mode effects were also analyzed by the OECD. These analyses, however, were conducted with data including field test data of all countries and the OECD claimed that the quality of the field test data would not allow for country-specific analyses. Based on data from all participating countries of the field trial, the OECD identified items that were invariant (same item difficulties) with respect to the test mode (PBA vs. CBA). Using only these invariant items, we reran the mode-effect analyses of the German field trial data[1]. Assuming that no interaction effect mode by country exists, one would expect that all mode-effects in the German field trial data disappear

Table 1. Mode-effects on item difficulty; findings from the German field trial using the 1PL IRT model

	All Items							Invariant Items (OECD)				
Domain	*N*			*d*		PISA Scale			*d*		PISA Scale	
	PBA	CBA	*I*	*Est*	*SE*	*Est*	*SE*	*I*	*Est*	*SE*	*Est*	*SE*
Science	340	338	91	**–.23**	.08	**–21**	7.2	59	**–.17**	.08	**–15**	7.6
Math	345	340	68	**–.14**	.07	**–14**	6.1	48	–.09	.07	–8	6.3
Reading	349	334	87	–.13	.10	–16	8.9	36	–.06	.09	–7	10.3
Mean effect				**–.17**	.06	**–17**	5.6		–.11	.06	–10	6.1

Notes. N: Number of students; *I:* Number of items; PBA: Paper-based assessment; CBA: Computer-based assessment; *d:* Size of mode-effect CBA vs. PBA (negative effects indicate that CBA increase item difficulty); *Est:* Estimate; *SE:* Standard Error; PISA Scale: Mode-effect on the international PISA scale (originally *M* = 500; *SD* = 100); Bold parameters are statistically significant ($p < .05$).

[1] Information on these items was provided by the OECD in the file 2nd_PISA2015_Vol1_AnnexAI_Tab.xlsx (http://www.oecd.org/pisa/data/2015-technical-report; see Tables AI.1, AI.2, AI.4).

when restricting the analyses to these invariant items. As can be seen in Table 1, this assumption is not supported but mode effects still exist and decrease only slightly (and the effect for science remains statistically significant). These findings suggest that the OECD strategy of only using aggregated data from all countries cannot correct sufficiently for mode effects within countries, namely Germany. Furthermore, given that mode effects within countries could not be detected by means of the OECD analyses and assuming that country-specific mode effects could also have been substantial in the 2015 main study, these findings pose challenges to a valid interpretation of changes in country means between 2012 and 2015.

Trend estimation

Eight different analyses were carried out for the three PISA domains science, math, and reading to study mode effects and effects of the IRT model:

- **Model C1:** Concurrent calibration using the 1PL model without the German field trial data; item difficulties of all common items are set equal (mode effects are ignored)
- **Model C2:** Concurrent calibration using the 2PL model without the German field trial data; item difficulties of all common items are set equal (mode effects are ignored)
- **Model H1:** separate calibration (1PL) with item linking (based on anchor items); without the German field trial data (mode effects are ignored)
- **Model C1I:** Concurrent calibration using the 1PL model without the German field trial data; only item difficulties of invariant items (see Table 1) are set equal (effects from the international mode-effect study are taken into account)
- **Model C2I:** Concurrent calibration using the 2PL model without the German field trial data; only item difficulties of invariant items are set equal (effects from the international mode-effect study are taken into account)
- **Model H1I:** separate calibration (1PL) with item linking (based on invariant anchor items; effects from the international mode-effect study are taken into account)
- **Model C1F:** Concurrent calibration using the 1PL model including the German field trial data; item difficulties of invariant items are set equal and are adjusted for mode-effects found in the German field trial sample
- **Model H1F:** separate calibration (1PL) with item linking including the German field trial data; linking only for invariant anchor items; effects from the national mode-effect study are controlled)

For each of the analyses, the student sampling weights for PISA time points were included. Descriptive analysis for means and trend estimates were based on 20 plausible values. The starting point of the trend estimates in each domain was the PISA data collection in which it was major domain for the first time, i. e., 2000 for reading, 2003 for math, and 2006 for science. This procedure ensures that trend estimates are based on sufficient numbers of linking items.

Table 2 provides findings for science. Original results are those published in international PISA reports. All other scores come from our re-analyses of the German data sets. The models C1 und C2 reproduce the international findings (e. g., for PISA 2015: Original: 509 points for Germany; our re-analyses revealed 508 and 507 points, respectively). Restricting the analyses to items that were reported by the OECD to be invariant across test modes leads to slightly higher means for Germany (see models C1I und C2I: PISA 2015: 513 and 512 points, respectively); there is still a negative change from 2012 to 2015. Furthermore, the findings underline that both calibration models (1PL: models C1 and C1I vs. 2PL: models C2 and C2I) produce very similar estimates. This is also true for separated calibrations with link items (models H1 and H1I). Overall, all six models ignoring a country-specific mode effect strongly reproduce the international findings of PISA 2015 (Original), suggesting a negative trend in science from 2012 to 2015.

The picture changes when the mode-effect of the 2014 field trial in Germany is included (models C1F and H1F). Scores in 2015 increase and suggest a slight but not significant positive change in science. This finding is true for both methods (concurrent calibration and separate calibration). Due to the relatively small sample size in 2014 we did not run the analyses in a 2PL-framework. The main results are also summarized in Figure 1.

Applying the eight models to the German data in mathematics provides a very similar pattern of results. The models C1, C2 and H1 reproduce (with very small deviation) the German 2015 score reported by the OECD. When the German mode effect is taken into account (models C1F, H1F), the results suggest highly stable mean scores from 2012 to 2015 (see Figure 1). These findings are in contrast to the internation-

Table 2. Trend estimates in science; findings from IRT analyses

	Model	2006[†]	2009	2012	2015
Original		516	520	524	509
Without data from German field trial (all items)	C1	516	519	523	508
	C2	516	518	524	507
	H1	516	515	522	506
Without data from German field trial (only invariant items)	C1I	516	519	523	513
	C2I	516	519	524	513
	H1I	516	517	523	513
With data from German field trial (only invariant items)	C1F	516	520	524	528
	H1F	516	516	522	528

Notes. [†] Trend estimates anchored in PISA 2006.

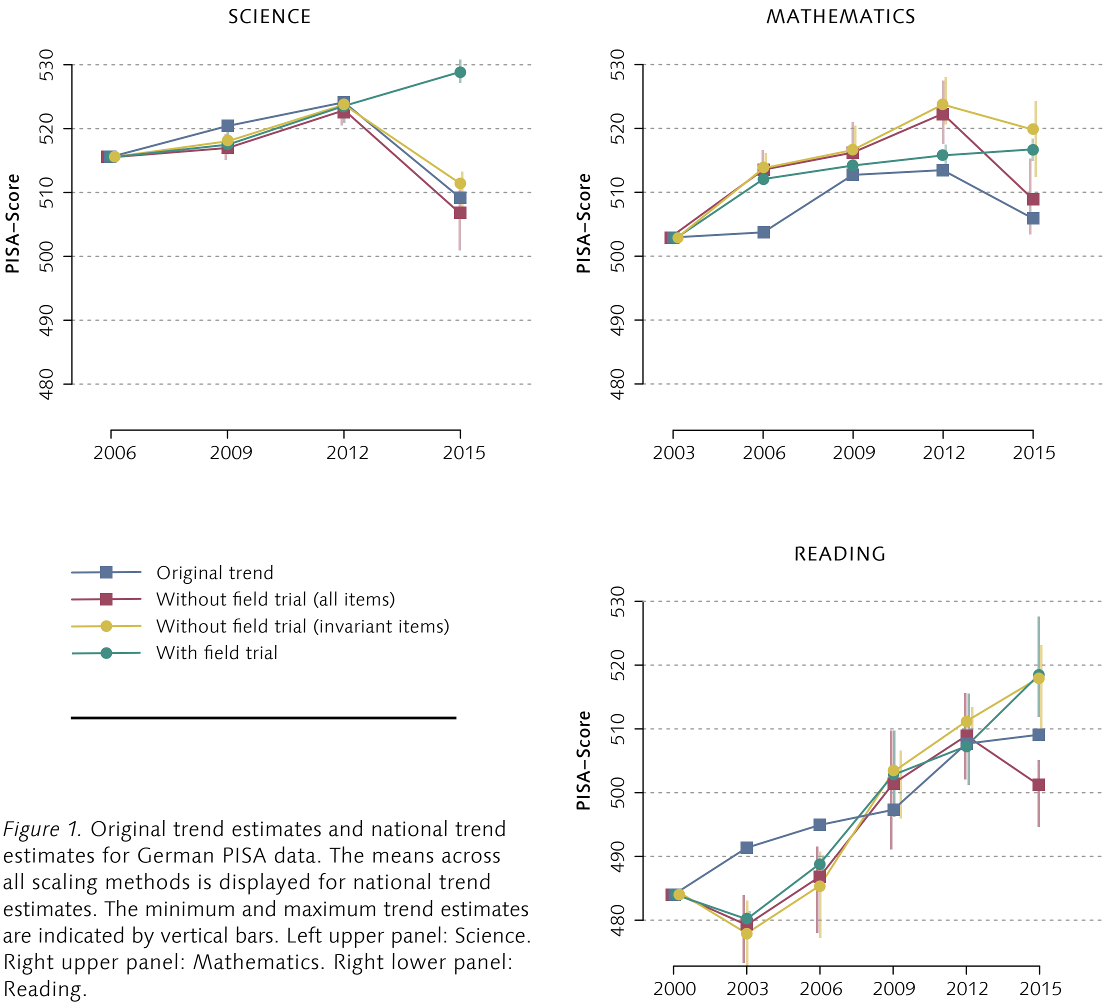

Figure 1. Original trend estimates and national trend estimates for German PISA data. The means across all scaling methods is displayed for national trend estimates. The minimum and maximum trend estimates are indicated by vertical bars. Left upper panel: Science. Right upper panel: Mathematics. Right lower panel: Reading.

ally reported finding of negative change in Germany from 2012 to 2015. Again, the calibration model (1PL vs. 2PL) does not moderate the trend estimates. When restricting the analyses to those items that were reported by the OECD to be invariant (models C1I, C2I, H1I), the results show no change in Germany between 2012 and 2015.

Compared to science and math, the findings in reading are slightly inconsistent but again underline that the choice of the calibration method (1PL vs. 2PL) does not substantially influence trend estimates (see Figure 1). Also in line with the other domains, reading scores increase when the mode-effect from the German field trial is included in the analyses. It is noteworthy, however, that restricting the analyses to German data leads to substantial differences in trend estimates compared to the internationally reported findings.

Discussion

All presented analyses were restricted to the German PISA data including national data from the 2014 field trial (marginal trend estimates). We did not explicitly strive to correct the international PISA trend results reported by the OECD. Our re-analyses of German data were carried out, however, to study the sensitivity of trend estimates under different assumptions with respect to the change of the calibration model (1PL vs. 2PL) and the test mode (PBA vs. CBA) changes from one PISA data collection to another. Findings of these analyses suggest that substantially lower scores of German students in PISA 2015 compared to 2012 could at least partly be explained by mode effects.

Overall, the German field trial data suggest that CBA increases item difficulties compared to PBA. We can only speculate about reasons for this finding. One explanation could be that German students lack experience in using computers in the school context. Findings of computer use in PISA 2015 support this explanation to some extent. The OECD analyzed mode-effects based on the complete 2014 data set, consisting of data sets from all participating countries. Only those item showing no mode effect were used for the trend estimation. Our analyses, however, suggest that these internationally invariant items could vary in individual countries. Findings of models C1I, C2I und H1I show that a mode-effect still occurs in Germany when calibration is restricted to these “invariant” items. Probably, the OECD procedure of eliminating mode-effects has not been fully successful in Germany.

It is also noteworthy that the OECD mean of all three domains decreased significantly between PISA 2012 and PISA 2015, the largest change (minus 8 points) is observed in science and the number of countries with negative changes is significantly larger than the number of countries with positive changes. These findings also suggest that even at the level of all OECD countries the strategy of adjusting for mode effects has not been completely successful. Further research (also including the field trial data from other countries) is necessary to better understand how the substantial changes in the study design affect students' performances on the achievement tests.

1.2 Trend estimation and linking errors in large-scale assessments

Background

The following section introduces a framework for discussing standard errors for trend estimates due to item selection (also denoted as linking errors) in international large-scale assessments (ILSA). In order to estimate national trends in student achievement, the results from different assessments need to be aligned so that the achievement scores can be directly compared. The general idea of linking procedures is to use a set of common items that are administered in more than one assessment in order to establish a common metric that makes it possible to compare the test results across different assessments. Figure 2 illustrates a typical linking design used in two assessments of an ILSA study. In both assessments, a set of I_0 common items is administered to a cohort of students. In addition, I_1 and I_2 unique items are presented in only one of the two assessments. One advantage of including unique items in an ILSA is that they can be made publically available for secondary analysis and also increase the transparency of what is being measured in an assessment.

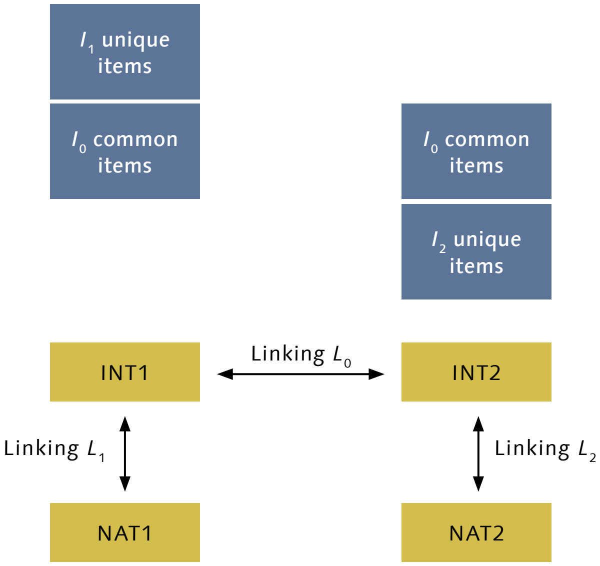

Figure 2. Linking steps in the estimation of original trends in international large-scale assessments. INT and NAT refer to the international and national scores, respectively, at time points 1 and 2.

Usually trend estimation for a country consists of three linkings in ILSA. At first, international item parameters are obtained by a separate calibration for each assessment. The item parameters of common items are used to link the two achievement scales on a common metric across time (see linking L0 in Figure 2). PISA applies the mean-mean linking method for transforming the item parameters from the 1PL scaling model to a common scale. Then, national achievement scores for each country are obtained by scaling the item responses in the first (linking L1) and second (linking L2) assessment. In the scaling models of both linkings, the item parameters are fixed to their international values. The outcome of the linkings L1 and L2 is a collection of country mean achievement scores for both assessments. The trend estimate for a country is defined by the difference between the country means in the two assessments (i.e., mean score at time 1 is subtracted from mean score at time 2) which is also denoted as the original trend estimate.

Linking error quantifies the degree to which a country's mean or trend estimate depends on the set of items that were included in the assessment. This dependency is manifested in three different aspects of item functioning. First, an item can function differently across assessments, which is known as item parameter drift (IPD). Second, an item can function differently (differential item functioning; DIF) across countries in one assessment, indicating that an item is relatively easier or more difficult for a specific country than at the international level. These cross-national differences have been studied extensively and termed country DIF. Third, country DIF can vary across assessments, which means that the relative difficulty changes across assessments (DIF × IPD). For example, changes in a country's curricula may affect the item difficulty of particular items. Furthermore, the magnitude of the linking error also depends on the linking design. A good illustration of the importance of the linking design can be seen in the PISA study with its distinction between major and minor domains (e.g., reading was the major domain in PISA 2009 and, therefore, a large number of reading items were used in PISA 2009, whereas, in PISA 2012, reading was a minor domain and a much smaller number of reading items were used). In this linking design, additional uncertainty in the linking is caused by the strong reduction in the number of items when a content area changes from being a major to a minor domain.

The linking errors calculated in the PISA study considered only the uncertainty of the linking at the international level (see linking L0 in Figure 2). Thereby, only the differential functioning of items across assessments (IPD) was taken into account when calculating the standard errors of trend estimates. In the operational procedure, PISA uses a variance components model for the international item parameters of the 1PL model that takes into account the IPD of testlet effects and item effects within a testlet. However, little is known about how the linking design (i. e., common and unique items) and the other variance components (i. e., DIF, DIF × IPD) affect the magnitude of linking errors. In the following we propose a variance components model for discussing linking errors in trend estimation of international assessments. The basic assumption is that the items administered in a study are a representative sample from a larger domain of items and the 1PL model is used as a scaling model for obtaining country means and trend estimates.

Variance components model

In the variance components model, the international item parameters for $\beta_{it}^{(int)}$ for item i ($i = 1, \ldots, I$) at time t ($t = 1, 2$) and the national item parameters $\beta_{itc}^{(nat)}$ at time t in country c ($c = 1, \ldots, C$) are decomposed into main item effects, country DIF effects, and IPD effects:

$$\begin{aligned} \beta_{it}^{(\text{int})} &= \mu_t^{(\text{int})} + \nu_i + \nu_{it} \\ \beta_{itc}^{(\text{nat})} &= \mu_{tc}^{(\text{nat})} + \nu_i + \nu_{ic} + \nu_{it} + \nu_{itc} \end{aligned} \qquad (1)$$

where $\mu_t^{(\text{int})}$ and $\mu_{tc}^{(\text{nat})}$ represent the average item difficulty at time point t for the international item parameters and the national item parameters in country c, respectively, and ν_i measures the effect of item i on item difficulty across countries and time points (i. e., the item's main effect on its difficulty independent of country and time point). The effect ν_{it} describes that the difficulty of item i changes across time (IPD effect). Differences in item difficulty across countries are reflected in the effect ν_{ic} of item i and country c across time points t (country DIF). Note that the effect ν_{ic} represents the country DIF that is stable across time. Country DIF that is specific to each assessment is captured by the effect ν_{itc} which is essentially a three-way interaction between item, country, and time point (i. e., DIF × IPD).

The model in Equation (1) assumes that the random effects ν_i, ν_{ic}, ν_{it}, and ν_{itc} are mean-centered and independent, with corresponding variances σ^2_{item}, $\sigma^2_{\text{DIF, }c}$, σ^2_{IPD}, and $\sigma^2_{\text{DIF}\times\text{IPD, }c}$. It should be emphasized that the DIF and DIF × IPD variances are allowed to vary across countries. The variance component σ^2_{item} describes the variability of item difficulties in the universe of items. The amount of DIF in a country *c* that is stable across time *t* is summarized in the variance $\sigma^2_{\text{DIF, }c}$. The variation of international item difficulties across time that is due to IPD is measured by σ^2_{IPD}. Finally, the country DIF in a country *c* can differ across assessments, which is reflected by the interaction $i \times t \times c$ and the corresponding variance component $\sigma^2_{\text{DIF}\times\text{IPD, }c}$.

Standard error of original trend estimates at the national level

The national trend for country *c* is given by the difference $\mu^{(\text{nat})}_{2c} - \mu^{(\text{nat})}_{1c}$. An estimate of this difference includes three sources of uncertainty: the error for L_0, which links the two assessments, and for L_1 and L_2, which link the country to the international scale at each time point (see Figure 2). More formally, the original trend estimate is given by $O = \hat{\mu}^{(\text{nat})}_{2c} - \hat{\mu}^{(\text{nat})}_{1c} = L_2 + L_0 - L_1$. For a linking design with I_0 common items and I_1 and I_2 unique items, the variance estimate for a trend estimate is calculated as follows:

$$Var(O) = \frac{2}{I_0}\sigma^2_{IPD} + \frac{I_1 + I_2}{(I_0+I_1)\,(I_0+I_2)}\sigma^2_{DIF,\,c} + \left[\frac{1}{I_0+I_1} + \frac{1}{I_0+I_2}\right]\sigma^2_{DIF\times IPD,\,c} \tag{2}$$

As can be seen, the variance of the trend estimate is a function of the DIF and IPD variance components and the linking design. The following observations can be made. First, the variance decreases with an increasing number of common items I_0. Second, using Equation 2 to calculate standard errors for national trend estimates will always result in larger standard error estimates than following the procedure that is currently used in PISA because it only contains IPD variance appearing in the first term $\frac{2}{I_0} \times \sigma^2_{IPD}$. Third, if the linking design only consists of common items (i. e., $I_1 + I_2 = 0$), country DIF variance does not impact the variability of trend estimates. This can be explained by the fact that country-specific DIF effects (ν_{ic} in Equation (1)) cancel each other out when only common items are administered in both assessments. However, country-specific IPD effects influence the uncertainty of trend estimates even in linking designs with only common items. Fourth, in the PISA study with its use of major and minor domains, only common items would be administered at the second time point, that is, I_1 would be much larger than I_0 and

$I_2 = 0$ (e.g., in PISA 2009 reading was the major domain, so in PISA 2012 only common items were administered for the assessment of reading). In this case, the country DIF effects do not cancel out across the two assessments and the contribution of the country DIF variance to the variance of the trend estimate is given by $I_1 / (I_0 + I_1) \times 1 / I_0 \times \sigma^2_{\text{DIF, c}}$. Fifth, if the number of unique items in both assessments is much larger than the number of common items (i.e., $I_1, I_2 \gg I_0$), the variance of the trend estimate is mostly determined by the IPD variance. In this condition, the PISA method can be expected to produce results similar to our proposed standard error formula.

Simulation study

In our simulation study, we compared the proposed formula for calculating standard errors of trend estimates with the method used in PISA, which only takes the linking errors at the international level into account. The simulation study was designed to resemble test designs used in previous PISA assessments. Two international data sets, each with 20 countries, were simulated. The main parameter of interest was the trend estimate for each country (i.e., the difference between the country mean of the second and the first assessment). To obtain realistic values for the data-generating model, we took the means and standard deviations for the reading achievement test of 20 selected countries that participated in PISA 2009 and 2012 and we transformed the values to the logit metric. The country means in both assessments showed considerable variation (PISA 2009: ranging from −.47 to .40, $M = .00$, SD = .20; PISA 2012: ranging from −.55 to .28, $M = .05$, SD = .20). In addition, the 20 coun-

Robitzsch, A., & Lüdtke, O. (2018). Linking errors in international large-scale assessments: Calculation of standard errors for trend estimation. *Assessment in Education: Principles, Policy & Practice.* Advance online publication.

tries in the data-generating model exhibited moderate variability in their true trends in achievement (ranging from –.17 to .27, M = .05, SD = .11). We implemented different linking designs in the simulation study. At each measurement point the total number of items was either 60 or 120. We specified six different linking designs, in which the number of common items I_0 and the number of unique items I_1 and I_2 at time 1 and 2 were manipulated. More specifically, the two linking designs A30 ($I_0 = 30$, $I_1 = 0$, $I_2 = 0$) and A60 ($I_0 = 60$, $I_1 = 0$, $I_2 = 0$) included only common items. The three linking designs B30a ($I_0 = 30$, $I_1 = 30$, $I_2 = 0$), B30b ($I_0 = 30$, $I_1 = 90$, $I_2 = 0$) and B60 ($I_0 = 60$, $I_1 = 60$, $I_2 = 0$) included additional unique items only at measurement point 1, and the linking design C60 ($I_0 = 60$, $I_1 = 60$, $I_2 = 60$) included unique items at measurement point 1 and 2. Following the test design of the original PISA studies items were arranged into 12 clusters, each consisting of 15 items (resulting in 180 items). The sample sizes of each country and each measurement point were set to n = 2 000.

Item difficulties for each country and each measurement point were generated according to the variance components model (see Equation 1) in each replication of the simulation. All DIF, IPD and DIF × IPD effects were assumed to be normally distributed and homogeneous across countries. This was in line with the preliminary analyses of the PISA 2009 and PISA 2012 assessments which revealed that the empirical DIF and IPD effects did not strongly deviate from normality. The IPD variance was set to 0, .025, or .05. The DIF variance was set to 0, .1, and .2. The DIF × IPD variance was set to 0 or .05. We determined the observed coverage of the 95 % confidence intervals. The results were aggregated across the 20 countries.

Table 3 shows the observed coverage for the trend estimates that were based on the operational PISA procedure and our proposed method as a function of the magnitude of the variance components and the linking design, for a sample size of n = 2 000. As expected, the PISA method only produced acceptable standard errors if one of the following conditions was fulfilled. First, if only IPD variance was present (and no DIF and DIF × IPD variances), the PISA standard error correctly reflected item parameter drift. Second, even in the case of country DIF, the PISA standard error for a trend estimate was accurate if only common items were administered (linking designs A30 or A60). In these conditions, the country-specific DIF effects cancelled each other out. In general, the coverage rates for the trend estimates provided by the PISA methods were consistently too low (M = 82.4 %, range = 51.8 % to 95.5 %); whereas, for our proposed standard error method they were acceptable (M = 93.8 %, range = 91.7 % to 95.0 %).

Table 3. Observed coverage of original trend estimates for a single country aggregated across countries as a function of IPD, DIF, and DIF × IPD variance for a sample size of n = 2 000

Variance			PISA method						Proposed method					
IPD	DIF	DIF×IPD	A30	A60	B30a	B30b	B60	C60	A30	A60	B30a	B30b	B60	C60
0	0	0	92.1	92.3	92.3	92.0	92.0	92.2	92.3	92.4	92.6	92.4	92.2	92.4
0	0	.05	**64.7**	**74.5**	**69.8**	**72.8**	**78.4**	**82.9**	92.5	93.1	93.0	93.3	93.1	93.3
0	.1	0	92.1	92.1	**73.5**	**66.9**	**81.3**	**81.3**	92.2	92.3	93.6	93.6	93.6	93.3
0	.1	.05	**64.9**	**73.5**	**61.1**	**59.9**	**71.1**	**74.8**	93.5	93.6	93.9	94.2	94.2	93.9
0	.2	0	91.7	91.5	**61.5**	**54.8**	**72.8**	**72.5**	92.0	91.7	94.0	94.5	94.0	94.0
0	.2	.05	**63.9**	**74.0**	**54.3**	**51.8**	**65.3**	**68.1**	93.4	93.9	94.5	94.5	94.6	94.6
.025	0	0	93.4	93.5	93.3	93.8	92.7	93.0	93.4	93.5	93.4	94.0	92.8	93.0
.025	0	.05	**80.5**	**82.6**	**82.9**	**84.5**	**85.3**	**87.6**	93.5	93.4	93.5	93.7	93.6	93.5
.025	.1	0	93.4	94.0	**85.3**	**83.4**	**87.8**	**87.8**	93.5	94.1	93.5	94.7	93.8	94.1
.025	.1	.05	**79.3**	**82.5**	**77.1**	**76.7**	**80.7**	**82.9**	93.8	93.6	94.2	94.7	94.3	94.2
.025	.2	0	93.1	92.6	**79.4**	**74.8**	**82.5**	**82.1**	93.2	92.7	94.4	95.0	94.3	94.3
.025	.2	.05	**79.4**	**82.1**	**72.4**	**69.6**	**76.2**	**78.2**	94.2	93.9	94.5	94.8	94.9	94.2
.05	0	0	93.3	94.2	94.1	94.4	93.7	94.5	93.3	94.2	94.1	94.4	93.7	94.6
.05	0	.05	**85.6**	**86.5**	**87.0**	**88.0**	**87.7**	**89.7**	94.0	93.8	93.7	93.5	93.7	93.8
.05	.1	0	93.7	93.3	**89.1**	**86.4**	**90.8**	**90.5**	93.8	93.4	94.1	94.2	94.6	94.8
.05	.1	.05	**85.0**	**86.4**	**83.0**	**82.2**	**83.9**	**86.8**	93.9	94.2	94.5	94.7	93.7	94.5
.05	.2	0	93.5	93.8	**84.7**	**81.0**	**85.7**	**85.9**	93.6	93.9	94.4	95.0	94.2	94.1
.05	.2	.05	**85.3**	**85.9**	**79.3**	**76.7**	**81.5**	**82.7**	94.5	94.1	94.9	94.9	94.9	94.6

Note. n = sample size. The different linking designs included I_0 common and I_1 and I_2 unique items and were specified as follows as follows: A30: I_0 = 30, I_1 = 0, I_2 = 0; A60: I_0 = 60, I_1 = 0, I_2 = 0; B30a: I_0 = 30, I_1 = 30, I_2 = 0; B30b: I_0 = 30, I_1 = 90, I_2 = 0; B60: I_0 = 60, I_1 = 60, I_2 = 0; C60: I_0 = 60, I_1 = 60, I_2 = 60. Coverage rates smaller than 91 % or larger than 98 % are printed in bold.

Discussion

National trend estimates are an important outcome of ILSA. We evaluated different methods for calculating standard errors for trend estimates and showed the method used in the PISA assessments until now underestimates the uncertainty of the linking errors. Our proposed variance components model for item difficulty parameters can be used to derive standard error formulas for original trend estimates. Furthermore, we showed how these formulas make it possible to clarify how the linking design (i. e., common and unique items) and the different variance components (i. e., DIF, IPD, DIF × IPD) affect the magnitude of linking errors.

Based on our findings, the following recommendations for the planning and analyzing of ILSA seem justified. First, the simulation results clearly indicate that the newly proposed standard errors for original trends were more accurate than the approaches currently used in PISA. Second, the analytical derivations that were based on the variance components model also revealed that country DIF affects the uncertainty of cross-national comparisons. The effect of country DIF on country means is particularly pronounced in cases with a smaller number of items, which is, for example, the case for minor domains in PISA. Therefore, it is important to increase the number of common items in order to obtain reliable country comparisons and trend estimates. Moreover, we believe that the calculation of standard errors for cross-sectional country means should also include uncertainty caused by country DIF.

IPN GROUP // Oliver Lüdtke, Alexander Robitzsch
FUNDED BY // Centre for International Student Assessment (ZIB)
DURATION // 2017 – 2018

2 Educational Measurement

The research conducted in Educational Measurement deal with methodological challenges in large-scale studies conducted at the IPN. Main areas of research encompass the development of multilevel modeling techniques, the refinement of measurement models using latent variable models, the treatment of missing data, and statistical procedures for causal inference from non-experimental data. In this report we limit our presentation to recent projects that were conducted in the area of missing data techniques.

In educational research, empirical data often contain missing values, for example, because participants fail to answer all items in a questionnaire or drop out before the end of a study. Selecting a proper treatment of missing data is very important because failing to do so can severely distort the conclusions obtained from the data. In the methodological literature, it is often argued that modern procedures such as multiple imputation (MI) are much better suited to treat missing data than traditional approaches that simply remove cases with missing data from the analysis (listwise deletion, LD).

Despite the growing popularity of MI, there still remain open questions about how best to apply MI when the data have a multilevel structure with for example individuals (e.g., students) nested within groups (e.g., classrooms). Not only can missing values occur at different levels, their analysis may also focus on different aspects of the multilevel structure, for example, the relations among student characteristics, the extent to which they vary across classrooms, or whether classroom-level variables can account for this variation. In the following projects, our goal was to investigate the application of MI in multilevel research.

2.1 Multiple imputation of incomplete multilevel data

Suppose that there exists a complete data set Y, two parts of which can be distinguished: the observed data Y_{obs} and the missing data Y_{mis}. The main principle of MI is to generate a set of plausible replacements for Y_{mis} on the basis of Y_{obs} and a statistical model (the imputation model) thereby creating multiple "filled-in" copies of the data. One crucial requirement of MI is that the imputation model has to match both the structure of the data and the research questions they are intended to address. Meeting this requirement can be challenging in multilevel research because of the variety in the structure of the data and the research questions to be answered therewith.

In multilevel data, variables measured for individuals can be separated into two parts, where each part varies only within or between groups, respectively. These parts can then be used to estimate the relations between variables at the level of both students and classrooms, respectively. In this context, the question of how the multilevel structure should be accommodated is very important.

Method

Lüdtke, O., Robitzsch, A., & Grund, S. (2017). Multiple imputation of missing data in multilevel designs: A comparison of different strategies. *Psychological Methods, 22*(1), 141–165.

We conducted a computer simulation study to evaluate the performance of several approaches to MI. Data were generated for two variables *X* and *Y* from a multilevel bivariate normal distribution, where *Y* contained missing data, and imputations were generated on the basis of (a) a multivariate normal distribution (NORM), (b) a set of dummy indicators, with one variable pertaining to each group (DI), and (c) a multivariate mixed-effects model (PAN). It is important to note that NORM ignores the multilevel structure, whereas DI represents the differences between groups with fixed effects, and PAN models the structure directly. Further, the simulation varied the number of groups ($J = 50, 150$), the group size ($n = 5, 15, 30$), the intraclass correlation of *X* and *Y* (ICC = .10, 150), the correlation at the group level ($\rho = .35, .60$), the missing data mechanism, and the amount of missing data.

Results

Figure 3 shows the estimated bias of the ICC of *Y* and the between-group regression coefficient of *X* on *Y* for selected conditions. It is easy to see that NORM tended to underestimate the ICC of *Y*, which in turn inflated the regression coefficient of *X* on *Y*. The reverse was true for DI, particularly in conditions with smaller groups and low ICCs. Only the use of PAN provided approximately unbiased estimates of the parameters across conditions. These results were also confirmed with analytical derivations.

Similar results were also obtained in a different study, which further contrasted the performance of MI with that of full-information maximum-likelihood (FIML). This study showed that, while FIML provided similar results to MI if the analyses was based on the latent covariate model but sometimes led to biased parameter estimates in models that use a manifest aggregation of covariates (e.g., observed class mean).

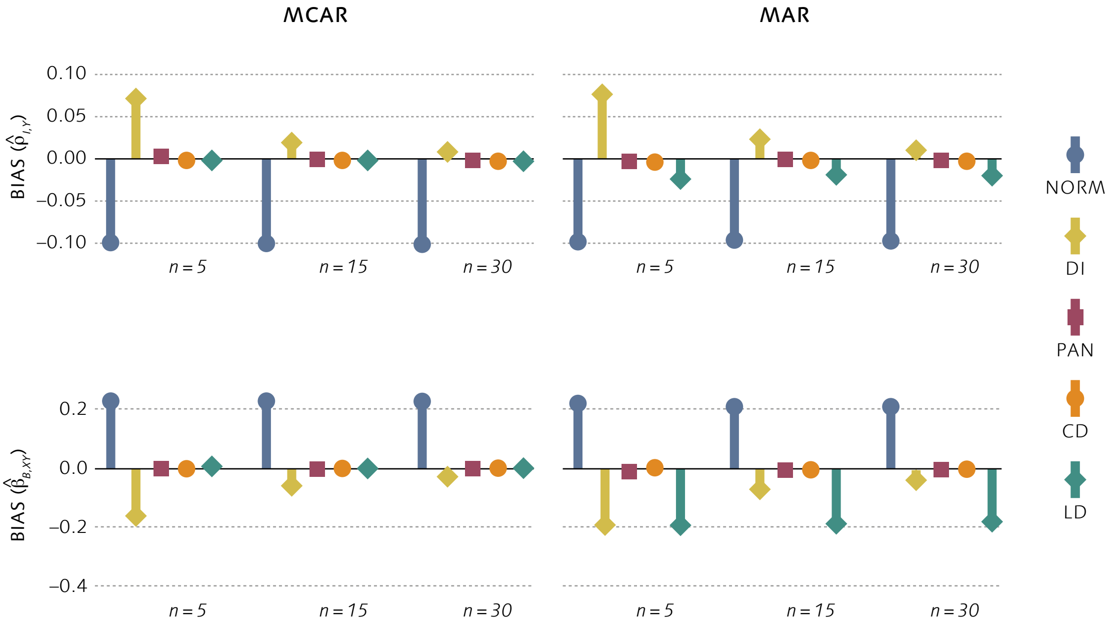

Figure 3. Bias for the intraclass correlation (ICC) of *Y* ($\rho_{I,Y}$) and the between-group regression coefficient of *X* on *Y* ($\beta_{B,XY}$) for varying group size (*n*) and missing data mechanisms (MCAR, MAR) in conditions with a large number of groups (*J* = 150), large ICCs (.30), and 25 % missing data. NORM = multivariate normal imputation; DI = dummy-indicator approach; PAN = two-level imputation; CD = complete data; LD = listwise deletion.

2.2 Multiple imputation at the classroom level

The previous study investigated the imputation of missing data at the level of individuals (e.g., students). However, little is known about how to conduct MI when missing data occur at the group level (e.g., classrooms or teachers) and how individual-level variables should best be included. In an additional study, we investigated the theoretical and practical properties of different approaches to MI for missing data at the group level. The two approaches considered in this study were joint modeling (JM) and the fully conditional specification (FCS) of MI, which make different use of individual-level variables. Whereas current implementations of FCS use the observed group means, JM uses latent means (i.e., random effects). In this context, we developed an alternative sampling procedure for the FCS approach that used latent means (FCS-LAT), similar to JM.

Grund, S., Lüdtke, O., & Robitzsch, A. (2018a). Multiple imputation of missing data at level 2: A comparison of fully conditional and joint modeling in multilevel designs. *Journal of Educational and Behavioral Statistics, 43*, 316–353.

Method

To evaluate these procedures, we generated data for two variables Y and Z from a bivariate distribution, where Z was measured at the group level and contained missing data. The simulation varied the number of groups (J = 30, ...1 000), the group size (n = 5, 20), the intraclass correlation of (ICC = .10, .30), the missing data mechanism, and the amount of missing data. In addition, the study distinguished conditions with balanced groups (i. e., all of the same size) und unbalanced groups (i. e., groups of different size).

Results

The results are illustrated in Table 4, which shows the estimated bias of the covariance of Y and Z for selected conditions with balanced and unbalanced groups. In the balanced conditions, both FCS and JM provided unbiased estimates of the covariance of Y and Z. However, this was no longer the case in unbalanced conditions. Only FCS-LAT and JM provided unbiased results across all conditions, although the bias was fairly small, even with FCS, unless the group sizes were extremely imbalanced. These results were confirmed with analytical derivations showing that the FCS approach (with observed means) provides unbiased estimates of the covariance only with balanced groups, with the bias depending on the distribution of the group sizes, the ICC of Y, and the amount of missing data.

Table 4. Bias (in %) for the covariance of Y and Z in balanced and unbalanced data (uniform and bimodally distributed) for large number of groups (J = 1 000), a small ICC of Y (.10) and 20 % missing data (MAR)

	Balanced			uniform (± 40 %)			uniform (± 80 %)			bimodal (± 80 %)		
	FCS	FCS-LAT	JM	FCS	FCS-LAT	JM	FCS	FCS-LAT	JM	FCS	FCS-LAT	JM
$\bar{n}$ = 5	0.3	0.4	−0.9	−0.9	−0.1	−1.5	−3.2	0.2	−1.2	−7.1	0.0	−1.0
$\bar{n}$ = 20	−0.3	−0.2	−0.9	−0.1	0.0	−0.6	−0.9	−0.1	−0.8	−2.8	0.3	−0.6

Note. $\bar{n}$ = average group size; FCS = two-level FCS with manifest group means; FCS-LAT = two-level FCS with latent group means; JM = joint modeling.

2.3 Multiple imputation with interaction effects and random slopes

In multilevel research, the interest is often not only in estimating the relations among variables within and between groups but also in whether or not they vary across groups (i. e., random slopes) or as a function of other moderating variables (e. g., cross-level interactions, CLIs). Not much is known about how to conduct MI if the predictor variables associated with random slopes or CLIs contain missing data due to the presence of nonlinear effects in these models. The present study evaluated a number of approaches that are currently available in standard software.

Grund, S., Lüdtke, O., & Robitzsch, A. (2018b). Multiple imputation of missing data for multilevel models: Simulations and recommendations. *Organizational Research Methods, 21,* 111–149.

Method

The data were simulated according to a multilevel model with an outcome variable *Y*, a predictor variable *X* with a random slope, and a group-level predictor variable *Z*, with or without CLI. We varied the number of groups (J = 50, ...,500), the group size (n = 5, 10), the ICC of *Y* (ICC = .10, .20, .50), the size of the CLI (.0, .20), and the missing data mechanism. In addition, this study featured conditions without random slopes and CLIs, with missing data at the group level, and missing data in categorical variables similar to the simulations above. To accommodate models with random slopes and CLIs, we compared FCS ignoring the multilevel structure, multilevel FCS with "passive imputation" of the CLI but ignoring random slopes (FCS-CLI/no RS), FCS with random slopes and "passive imputation" of the CLI (FCS-CLI/RS), and FIML.

Results

When missing data occurred only in the outcome *Y*, both FCS-CLI/RS and FIML provided unbiased parameter estimates. By contrast, FCS resulted in biased estimates, and FCS-CLI/no RS—while unbiased—provided confidence intervals that were too narrow. When missing data occurred in *X*, FIML could no longer be used, and none of the MI procedures provided perfect results. Under FCS-CLI/no RS and FCS-CLI/RS, the bias was mostly restricted to the CLI, whereas main effects were estimated with little bias. Under FCS, both the CLI and main effects were biased.

These results indicated that, while procedures with "passive inputation" can be used relatively well to accommodate random slopes and CLIs, more sophisticated methods are needed to treat missing data in multilevel models with nonlinear effects. Because so-called "substantive model compatible" procedures have been recommended for this purpose, we conducted an additional simulation in the course of this study. Although not yet available in standard software, these procedures provided promising results and led to unbiased estimates of all model parameters. This study was accompanied by a detailed list of recommendations for the treatment of missing data in various settings, including worked and reproducible examples in statistical software.

2.4 Discussion

In multiple studies, we have considered a number of practical problems that arise due to missing data in multilevel research. This included a variety of multilevel models with random intercepts and slopes, contextual effects, nonlinear effects, and missing data at different levels and in different types of variables. Overall, the studies provide evidence that missing data in multilevel research can be effectively treated with different approaches to multilevel MI (i. e., FCS and JM). This is in contrast to procedures that neglect the multilevel structure overall or certain aspects of the multilevel structure in particular (e. g., effects within and between groups, random slopes, CLIs). However, the studies also illustrate potential for the future: Despite the utility of MI, there are still open questions regarding, for example, the treatment of missing data in models with random slopes and nonlinear effects. Further questions aim toward situations in which variables are measured with error, for example, in large-scale assessment studies, or concern the way in which statistical inferences can be drawn under MI.

In addition to the evaluation of these procedures in different studies, one additional part of this project was the development of a software package ("mitml") for the statistical software R. This package aims to make multilevel MI more accessible with a convenient interface for specifying the imputation model and by providing automated procedures for analyses and statistical inferences under MI. The use of this package for multilevel MI is illustrated in a separate article.

Grund, S., Robitzsch, A., & Lüdtke, O. (2017). *mitml: Tools for multiple imputation in multilevel modeling (Version 0.3-5)*. Abgerufen von http://CRAN.R-project.org/package=mitml

IPN GROUP // Simon Grund, Oliver Lüdtke, Alexander Robitzsch
FUNDED BY // Centre for International Student Assessment (ZIB)
DURATION // 2014–2018

Perspectives for Research Line 5

In the area of educational assessment, test development for NEPS is an ongoing project with a fixed agenda until 2022. Developing NEPS tests and moving from paper-pencil to computer-based tests, carries many opportunities for substantive and methodological research, for example questions concerning the validity of the tests or questions exploring mode effects. These research questions will be addressed during the upcoming research period. Furthermore, the research using eye-tracking technology to better understand students' information processing while working on a test will continue.

The IPN – together with the ZIB partners – will be responsible for the analyses and documentation of the national results of PISA 2018.

The next TIMSS data collection will take place in 2019. Researchers at IPN have reviewed the science items and are responsible for analyzing the national science data for the national TIMSS report which will be published in 2020.

In a totally new line of research the potential of automatic scoring methods in essay writing will be explored. In this line, one important question is how text length influences proficiency scores provided by human and machine ratings.

In the area of educational measurement one major goal is to develop a common statistical framework which integrates latent variable models, multilevel modeling techniques and the treatment of missing data. This will be accompanied by the further refinement of statistical software packages that allow researchers a flexible and user-friendly application of the proposed statistical procedures.

Projects in Research Line 5:

Project // Homepage	**Methodological issues in longitudinal educational studies (MILES) //** **www.ipn.uni-kiel.de/de/forschung/projektliste/miles**
Funded by	German Research Foundation (DFG)
Term	2017–2019
Departments involved	Educational Research and Educational Psychology, Educational Measurement
Staff (IPN)	Karin Guill, Olaf Köller, Oliver Lüdtke, Gabriel Nagy
Cooperation partners	German Institute for International Educational Research (DIPF) // Kiel University // Technical University Dortmund // University of Hamburg // University of Tübingen

Project // Homepage	**International computer and information literacy study (ICILS 2018) //** **www.ipn.uni-kiel.de/de/forschung/projektliste/icils**
Funded by	Federal Ministry of Education and Research (BMBF) and Federal States
Term	2016–2020
Departments involved	Educational Research and Educational Psychology, Educational Measurement
Staff (IPN)	Jan Marten Ihme, Martin Senkbeil
Cooperation partners	German Institute for International Educational Research (DIPF) // University of Paderborn

Project // Homepage	**Trends in international mathematics and science study (TIMSS) //** **www.ipn.uni-kiel.de/de/forschung/projektliste/timss-grundschule**
Funded by	Federal Ministry of Education and Research (BMBF) and Federal States
Term	2017–2020
Departments involved	Educational Research and Educational Psychology, Chemistry Education
Staff (IPN)	Olaf Köller, Mirjam Steffensky
Cooperation partners	German Institute for International Educational Research (DIPF) // Technical University of Dortmund // Universität Hamburg

Project	**Contextual effects in educational research**
Funded by	Centre for International Student Assessment (ZIB)
Term	2014–ongoing
Departments involved	Educational Research and Educational Psychology, Educational Measurement
Staff (IPN)	Simon Grund, Oliver Lüdtke, Alexander Robitzsch, Steffen Zitzmann
Cooperation partners	

Project // Homepage	**Programme for international student assessment (PISA) //** **www.ipn.uni-kiel.de/de/forschung/projektliste/pisa-2015**
Funded by	Centre for International Student Assessment (ZIB)
Term	2017–2022
Departments involved	Educational Research and Educational Psychology, Educational Measurement, Mathematics Education
Staff (IPN)	Aiso Heinze, Olaf Köller, Oliver Lüdtke, Gabriel Nagy, Alexander Robitzsch
Cooperation partners	German Institute for International Educational Research (DIPF) // Technical University Munich

Project // Homepage	**National educational panel study (NEPS) //** **www.ipn.uni-kiel.de/de/forschung/projektliste/neps**
Funded by	Leibniz Institute for Educational Trajectories
Term	2017–2022
Departments involved	Educational Research and Educational Psychology, Educational Measurement, Mathematics Education
Staff (IPN)	Inga Hahn, Aiso Heinze, Jan Marten Ihme, Jana Kähler, Olaf Köller, Martin Senkbeil, Helene Wagner
Cooperation partners	Leibniz-Institute for Educational Trajectories and NEPS Network

Project // Homepage	**Cognition in educational assessment (COGEA) //** **www.ipn.uni-kiel.de/en/research/projectlist/cogea**
Funded by	
Term	2012–ongoing
Departments involved	Educational Research and Educational Psychology, Educational Measurement
Staff (IPN)	Marlit Annalena Lindner, Steffani Saß, Benjamin Strobel
Cooperation partners	Leibniz-Institut für Wissensmedien

Project	**Item position effects**
Funded by	
Term	2012–ongoing
Departments involved	Educational Research and Educational Psychology, Educational Measurement
Staff (IPN)	Marit List, Gabriel Nagy
Cooperation partners	

Project	**Psychometric measurement models**
Funded by	
Term	2014–ongoing
Departments involved	Educational Research and Educational Psychology, Educational Measurement
Staff (IPN)	Oliver Lüdtke, Gabriel Nagy, Alexander Robitzsch
Cooperation partners	