Large-scale Assessments in Education

EDITORIAL Open Access

The national educational panel study (NEPS) and methodological innovations in longitudinal large-scale assessments

Tanja Kutscher[1*†], Marie-Ann Sengewald[1†], Timo Gnambs[1], Claus H. Carstensen[2] and Christian Aßmann[1,2]

†Tanja Kutscher and Marie-Ann Sengewald shared first authorship.

*Correspondence:
Tanja Kutscher
tanja.kutscher@lifbi.de
[1]Leibniz Institute for Educational Trajectories, Bamberg, Germany
[2]University of Bamberg, Bamberg, Germany

Abstract
This editorial introduces a special issue of Large-Scale Assessments in Education (LSAE) that addresses key challenges in analyzing longitudinal data from large-scale studies. These challenges include ensuring fair measurement across time, developing common metrics, and correcting for measurement errors. The special issue highlights recent methodological innovations, particularly for studies like the National Education Panel Study (NEPS), providing approaches for improving the accuracy and robustness of longitudinal educational research. The papers in this issue present advances in methods for estimating trends, incorporating background information, and analyzing longitudinal relationships between constructs. Innovative approaches such as Bayesian modeling for borrowing historical information, continuous-time models for capturing developmental trends, and plausible value estimation provide practical solutions for researchers working with complex longitudinal data. In addition, the issue presents new software tools that facilitate the implementation of these advanced methodologies. Together, these papers contribute to both the theory and practice of educational assessment and provide valuable insights for those working with longitudinal data in national and international panel studies.

Keywords Longitudinal large-scale assessment, National Educational Panel Study, Bayesian historical borrowing, Plausible value estimation, Continuous-time models, Educational trend analysis

Like other longitudinal large-scale assessments in education (LSAE), the National Education Panel Study, or NEPS (Blossfeld et al., 2011), needs to address different methodological challenges to ensure the integrity of the data – such as the implementation of fair measurements across time and respondents, along with the construction of common metrics, as well as strategies to control for item and unit nonresponse (e.g., Pohl & Carstensen, 2013; von Maurice et al., 2017). Next to the complexity and sheer size of the database, further challenges arise for valid results in subsequent research, for instance, how to accurately model the relationships of different constructs and different time

points, along with corrections for measurement error, and the selection of relevant background information (e.g., Rutkowski et al., 2010).

Current methodological innovations for longitudinal LSAE address strategies for modeling longitudinal data with respect to different subsequent research questions. In addition, methodological research evaluates the performance of different analysis strategies in simulation studies and facilitates the implementation of methodological innovations. The aim of this special issue is to contribute to the advancement of methodological practices in longitudinal LSAEs. This special issue includes seven studies that make important contributions in this regard.

Longitudinal NEPS Data

As a national LSAE, the NEPS collects valuable data on education quality and performance development within the German system, similar to LSAE in other countries. Examples of such studies include the U.S. Early Childhood Longitudinal Studies Program (ECLS; National Center for Education Statistics [NCES], 2018) and the Australian National Assessment Program—Literacy and Numeracy (NAPLAN, Australian Curriculum, Assessment and Reporting Authority, 2023). International LSAE allow for comparisons across education systems, such as the Program for International Student Assessment (PISA, Organization for Economic Cooperation and Development [OECD], 2022), the Progress in International Reading Literacy Study (PIRLS, von Davier et al., 2022), and the Trends in International Mathematics and Science Study (TIMSS, Martin et al., 2020). Despite different orientations, the LSAE share common topics and standards for the assessments (e.g., Cresswell et al., 2015; Hernández-Torrano & Courtney, 2021).

Using a multicohort sequence design, the NEPS considers educational outcomes, processes, and decisions spanning from early childhood education in kindergarten and school, to vocational training, university studies and even after leaving the education system (e.g., Blossfeld & Roßbach, 2019). This comprehensive approach currently encompasses seven distinct cohort samples, including newborns, kindergarten children, secondary school children in 5th grade (2010 and 2022) and 9th grade, first-year undergraduate students, and adults. The surveys include competence tests and interviews with target persons and – at least for the younger cohorts – also interviews with parents and educators (see also NEPS Data Portfolio). Accordingly, extensive background data is available from different measurement occasions, next to comprehensive assessments of educationally relevant, domain-specific functional competencies (i.e., reading competence, listening comprehension, mathematical competence, scientific and information and communication technology literacy). A coherent and comparable assessment of the competencies across the different age and respondent groups in education (e.g., school types, status groups, educational careers and occupations) requires considerable effort for the construction of the test instruments as well as in the analysis of such data (e.g., Artelt et al., 2013).

About this special issue

In longitudinal LSAE, methodological innovations aim at improving and evaluating the analytical routines and the approaches for subsequent substantive research addressing, for instance, group comparisons, prediction of future achievement, as well as inferring development and trends over time. Longitudinal data includes different methodological challenges such as measurement error and non-ignorable missing data, that can potentially bias research results. Strategies for adjusting for bias include incorporating available background information in the analysis, such as estimation of plausible values (PV) which is a special case of multiple imputation (e.g., Rubin, 1987). Furthermore, the specification of the temporal order is central in the longitudinal data analysis for accurately representing the theoretical processes of interest. This can be achieved through the consideration of discrete time intervals (e.g., Voelkle et al., 2018) or application of continuous-time dynamic models (e.g., van Montfort et al., 2018) offering insights into developmental trends over time. In addition to bias corrections and the correct representation of temporal processes, efficient and user-friendly strategies for analyzing extensive data are required. Sophisticated estimation strategies, like Bayesian approaches (e.g., van der Schoot et al., 2021) and automated analysis routines, for instance those available in the open-source software R (R Core Team, 2024), can help to combine various information sources and support the implementation of complex modeling strategies in practice.

This special issue presents seven studies contributing to the advancement of the methodological practices in LSAE. The studies use diverse methodological approaches and cover a range of topics, such as the incorporation of historical information within a Bayesian framework, the role of item selection in linking methods, the estimation of PV with customized background models, and the appropriate statistical modeling of longitudinal data. Multiple contributions conducted simulation studies for evaluating the performance of statistical methods under controlled conditions where the true results are known. Yet, the generated data in simulations typically does not match the complexity of empirical data. For this, illustration studies and tutorials are provided to showcase how the methods or software developments can be implemented in practice and what are the practical benefits or limitations. In addition, the applications of innovative methods in empirical studies with specific substantive research questions illustrate how to interpret the results and drawn substantive conclusions. The robustness of these conclusions in relation to the used methods is also a central question and provides insights for explaining the heterogeneity of empirical results (e.g., Nosek et al., 2022). Accordingly, we specified the type of the studies in this special issue and provide an overview in Table 1. The central aims of the seven studies can be summarized as follows.

The study conducted by Kaplan et al. (2023) investigates the potential benefits of incorporating information from previous cycles of longitudinal data within the Bayesian framework. The primary objective has been to assess the advantages of historical borrowing methods in estimating growth rates and predictive performance for the current cycle. The research systematically evaluates five different historical borrowing methods: three static approaches (complete pooling, Bayesian synthesis with aggregated data-dependent priors, and traditional power priors) and two dynamic methods (Bayesian dynamic borrowing and commensurate priors). The authors utilize data from two kindergarten cohorts of the U.S. ECLS, specifically the 2010–2011 and 1998–1999 cohorts, to scrutinize whether the inclusion of growth rate information from the 1998–1999

cohort can enhance the accuracy of estimating the growth rate in reading literacy for the 2010–2011 cohort, when employing a Bayesian multilevel growth curve model. Additionally, a simulation study has been conducted to assess the performance of Bayesian historical borrowing methods under various conditions, including different sample sizes and degrees of data heterogeneity (such as the similarity or dissimilarity between two cycles of data). The study's findings indicate that, in a single historical cycle, most methods performed similarly, except for pooling and power priors, which showed relatively poor performance under heterogeneous conditions. Based on previous research, the authors emphasize the efficacy of dynamic borrowing methods in leveraging information from previous cycles of longitudinal data. Therefore, it is recommended to use such methods when analyzing multiple cycles of longitudinal data to account for heterogeneity arising from cohort effects and changes in data collection strategies.

Robitzsch and Lüdtke (2023) investigate various approaches for estimating trends in International Large-Scale Assessments (ILSAs), specifically focusing on country means and standard deviations. The approaches distinguish three key factors: (1) the method of linking a country's performance to an international metric (indirect vs. direct), (2) the use of all items (both unique and common items) or only link items for linking, and (3) the assumption of item parameters as invariant (international item parameters), resulting from the concurrent scaling model, or noninvariant across countries (country-specific item parameters), resulting from the separate scaling model. The paper establishes that original trends and marginal trends correspond to indirect and direct linking approaches, respectively. The indirect linking approach estimates the trend for assessing a country's change between two measurement occasions by linking the country to the international metric at each measurement occasion (referred to as original trend estimates). Conversely, the direct linking approach links a specific country to the international metric only at the first measurement occasion and directly assesses the change in the country's mean or standard deviation (referred to as marginal trend estimates). Using simulation and analytical derivations, the authors demonstrate that direct linking and the use of link items outperformed alternative approaches, particularly when differential item functioning (DIF) was present. In the empirical application, trends in reading, mathematics, and science between two measurement occasions, PISA 2006 and PISA 2009, were evaluated. The results confirmed the efficacy of direct linking and the importance of link item selection for trend estimation, providing more accurate trend estimates and smaller linking errors in educational assessments compared to other approaches. Additionally, the study found that the choice of trend estimation methods had a greater impact on country standard deviations than on country means. Therefore, this work underscores the significance of careful methodological considerations when assessing trends in educational outcomes across different countries.

Heine and Robitzsch (2022) conduct a systematic examination of the impact of analytical decisions on cross-sectional and trend estimates in international large-scale educational assessments. The study focuses on the mathematical domain of the Programme for International Student Assessment (PISA) data from 2003 to 2012 and considers four crucial methodological factors: (1) the choice between concurrent or separate item calibration, (2) the selection of the calibration sample, for example, such as including either all participating countries or only OECD countries, (3) the inclusion of all items or only common items, and (4) the use of either the maximum likelihood estimator or

Table 1 Overview of the contributions in the special issue

Authors	Titel (Link)	Type of study	Data	Research focus	Innovative solutions for the analysis of longitudinal data	Relevant results
Kaplan, Chen, Lyu and Yavuz	Bayesian historical borrowing with longitudinal large-scale assessments (https://doi.org/10.1186/s40536-022-00140-w)	Simulation study and illustration study	ECLS kindergarten cohorts 2010–2011 and 1998–1999; Reading scores	Extension and evaluation of Bayesian historical borrowing methods as a technique to improve estimation accuracy by utilizing information across longitudinal studies, thereby accounting for heterogeneity that may be induced, for example, by cohort effects.	Include information from previous cycles to improve the accuracy of analysis of data from the current cycle of a longitudinal study	Most methods performed similarly in a historical cycle, except for pooling and power priors. The authors recommend using dynamic borrowing methods to use information from previous cycles of longitudinal data.
Robitzsch and Lüdtke	Comparing different trend estimation approaches in country means and standard deviations in international large-scale assessment studies (https://doi.org/10.1186/s40536-023-00176-6)	Simulation study and illustration study	PISA 2006 to 2009; Reading, mathematics, and science score	Effect of the type of methods used to estimate trends in ILSA on the accuracy of estimates of country means and standard deviations	Accounting for different methods of estimating trends in ILSA, e.g., direct vs. indirect linking, use of all items vs. only link item, international item parameters vs. country-specific item parameters	The direct linking approaches (e.g., for estimating the marginal trends) could result in more efficient trend estimates than indirect linking approaches (e.g., for estimating the original trend based on all items) if country DIF is present. As an alternative, there are indirect linking approaches that rely only on the link item. The choice of a trend estimation method could be more consequential for trend estimates in country standard deviations than for country means in the presence of uniform DIF.
Heine and Robitzsch	Evaluating the effects of analytical decisions in large-scale assessments: analyzing PISA mathematics 2003–2012 (https://doi.org/10.1186/s40536-022-00129-5)	Illustration study	PISA 2003 to 2012; Mathematics scores	Method variance due to analytical choices made during item calibration on country-specific cross-sectional and trend estimates	Comparisons of different analytical choices for country-specific cross-sectional and trend estimates	Significant influence of analytic choices on country means and related errors, potentially affecting the ranking of countries in mathematics literacy. The dominant role of item selection, both individually and in interaction with country-specific factors, in shaping cross-sectional and trend analyses

Table 1 (continued)

Authors	Titel (Link)	Type of study	Data	Research focus	Innovative solutions for the analysis of longitudinal data	Relevant results
Scharl and Zink	NEPSscaling: plausible value estimation for competence tests administered in the German National Educational Panel Study (https://doi.org/10.1186/s40536-022-00145-5)	Description of R package and tutorial	NEPS SC6 and NEPS SC3; Reading and mathematics scores	Accurate estimation of population effects	An automated estimation of plausible values with a customized specification of background variables using the NEPSscaling R package or graphical user interface NEPShiny	
Lohmann, Zitzmann, Voelkle and Hecht	A primer on continuous-time modeling in educational research: an exemplary application of a continuous-time latent curve model with structured residuals (CT-LCM-SR) to PISA Data (https://doi.org/10.1186/s40536-022-00126-8)	Illustration study with tutorial	PISA 2000 to 2018; Reading scores	Longitudinal data modeling with an appropriate statistical approach	Providing a continuous-time model to continuous-time latent curve model with structured residuals (CT-LCM-SR), allowing modeling of trends and process dynamics	An overall increase in average reading literacy scores across PISA countries, with considerable variation in the slope and direction of the trends. Countries also differed in their initial levels at the first measurement point in 2000. A significant continuous-time dynamic process of fluctuations around the trends. Countries' deviations from the trends disappeared, on average, after about five years.
Jindra, Sachse and Hecht	Dynamics between reading and math proficiency over time in secondary education – observational evidence from continuous time models (https://doi.org/10.1186/s40536-022-00136-6)	Empirical study	NEPS SC3; Reading and mathematics scores	Longitudinal data modeling with an appropriate statistical approach	Applying continuous-time models to examine the dynamics of two domains of competence	Reading literacy had a stronger effect on mathematics literacy than vice versa. Peak standardized cross-lagged effects occurred at an interval of about six-months.

Table 1 (continued)

Authors	Titel (Link)	Type of study	Data	Research focus	Innovative solutions for the analysis of longitudinal data	Relevant results
Sciffer, Perry and McConney	The substantiveness of socioeconomic school compositional effects in Australia: measurement error and the relationship with academic composition (https://doi.org/10.1186/s40536-022-00142-8)	Empirical study including simulation analyses	NAPLAN 2017; Scores in reading, writing, spelling, grammar, and numeracy	Estimation of bias due to measurement error by assessing composite measures such as school socioeconomic status or parental socioeconomic status; Examination of the relationships between school socioeconomic status and student achievement growth, considering the school's average prior academic achievement (so-called academic composition) as a mediator of this relationship.	Structural equation modeling and principal component analysis of appropriate approaches for assessing composite measures	No biased compositional effects in the dataset; Substantial effect of school socioeconomic status on student achievement growth in Australian schools; Academic composition mediated the relationship between socioeconomic school status and achievement growth.

the least square estimator for item calibration, as implemented in the R packages TAM (Robitzsch et al., 2024) and pairwise (Heine, 2023), respectively. The study reanalyzed 32 analytical scenarios and compared the results to those in the official OECD report. The findings underscore the substantial impact of analytical choices on country mean estimates and the amount of method variance. This impact results in wider confidence intervals of the countries' means compared to the official reports. The researchers find that item selection, both individually and in interaction with country, has a substantial impact on cross-sectional and trend estimates. They conclude that the analytical choices made during item calibration could introduce an additional source of method variance, which could potentially affect the ranking of countries in terms of their mathematical proficiency. To accurately interpret PISA rankings, it is crucial to consider the analytical options chosen and avoid overinterpreting small mean differences between countries.

Scharl and Zink (2022) introduce *NEPSscaling*, an R package that streamlines the estimation of plausible values (PVs) for competence tests within the National Educational Panel Study (NEPS). The resulting PVs can be utilized by NEPS data users to investigate population effects relevant to their research questions. *NEPSscaling* automates the PV estimation process, allowing users to concentrate on preparing background data specific to their research inquiries. PVs are estimated by following the scaling standards in NEPS and using an appropriate item response model. Missing values in the background data are handled using a nested multiple imputation scheme based on a classification and regression trees (CART) algorithm. The *NESPscaling* package provides a graphical user interface that simplifies the PV estimation process, making it accessible to researchers with minimal psychometric expertise and novice R users. The estimated PVs can be exported to statistical software, such as SPSS or Stata, for further analysis. These features enhance the accessibility of *NEPSscaling* to a wider audience. The package can estimate cross-sectional and longitudinally linked PVs for various competence assessments across NEPS cohorts, as demonstrated in the paper in two illustrated applications using R code and a graphical interface. The authors concluded that *NEPSscaling* is a valuable tool for NEPS data users, offering a simple and reliable method for automatic PV estimation. Compatible with a variety of statistical software and featuring a user-friendly design, *NEPSscaling* proves to be an invaluable resource for researchers conducting population-level analyses within the NEPS framework.

Lohmann et al. (2022) propose a continuous-time latent curve model with structured residuals (CT-LCM-SR) to address continuous development over time, considering both the dynamic process and trends in the data. The practical implementation of the CT-LCM-SR and its ability to separately estimate the trend and dynamic processes are demonstrated through an illustrative application using PISA reading literacy assessment data. The case study examined two main questions. (1) It analyzes the trends in mean PISA reading literacy scores from 2000 to 2018, revealing an overall increase with significant variations among countries in both slope and direction. (2) The authors examine the stability of education systems and identified a continuous-time dynamic process of fluctuations around trends. They found that temporary deviations tend to dissipate after approximately five years, as indicated by derived discrete-time autoregressive effects. The authors of the study conclude that the CT-LCM-SR produces parameters that are independent of specific time intervals and represents developmental processes on a continuous-time scale. The paper provides a tutorial and R code for specifying the

CT-LCM-SR model using the R package ctsem (Driver & Voelkle, 2018), along with an illustration of its application and interpretation. This tool is useful for modeling both trends and dynamic processes in educational research.

Jindra et al. (2022) examines the relationship between reading and mathematics literacy in children aged 10 to 19 using a continuous-time modeling approach. This innovative method enables the identification of peak effects and allows for the examination of the dynamics of these competence domains over time. The researchers utilize data from a large representative sample of German students in Starting Cohort 3 of the National Educational Panel Study (NEPS-SC3). The findings suggest that reading has a stronger impact on mathematics, indicating that mathematics literacy is a more persistent construct. Additionally, standardized cross-lagged effects peaked at approximately a six-month interval. These findings imply that interventions aimed at enhancing reading literacy could have a greater positive impact on mathematics literacy, as observed for mathematics. However, interventions targeting one domain may not yield enduring effects on the other. The authors emphasized the importance of using continuous-time models to capture the complex dynamics of educational constructs and identify peak effects.

Sciffer et al. (2022) address gaps in accurately assessing the effect of socioeconomic school composition on student achievement growth in Australian primary and secondary schools. The research also aims to clarify the relationships between socioeconomic school composition and academic composition using the dataset of the National Assessment Program – Literacy and Numeracy (NAPLAN) administered in 2017. Socioeconomic school composition is defined as the average socioeconomic status of a school, determined by parental educational and occupational status. The effects of socioeconomic school composition have been measured as the difference in academic achievement between students who have the same individual socioeconomic status but attend schools with different socioeconomic school compositions. Methodological concerns were raised regarding potential measurement errors in assessing the effects of socioeconomic school composition. To address this issue, the study compares two statistical approaches concerning their sensitivity in assessing measurement errors in indicators of socioeconomic status: multilevel residual-change regressions using composite measures provided by means of principal component analysis and multilevel residual-change structural equation models where composite measures are modelled as latent factors. The findings indicate that measurement error does not bias compositional effects in the dataset, providing confidence of using principal component analysis to develop reliable composite measures. Furthermore, the study utilizes multilevel path models to investigate whether academic composition acts as a mediator in the relationship between socioeconomic composition and achievement growth. Academic composition refers to a school's average prior achievement in academic domains, such as reading, writing, spelling, grammar, punctuation, and numeracy. The study finds that academic composition mediated the relationship between socioeconomic school composition and achievement growth. Noteworthy variations in achievement growth between schools with low and middle socioeconomic composition in Australia are highlighted, emphasizing the substantial impact of socioeconomic school composition on student achievement. The researchers conclude that educational reforms addressing both the academic and socioeconomic composition of schools are more likely to succeed. However, on reanalyzing

the data, Marks (2024) reported that the effects of socioeconomic status are rather minor when accounting for student-level prior achievement, with academic composition being a more relevant predictor of student achievement.

Conclusion

In conclusion, this special issue contributes significantly to advancing methodological practices in longitudinal LSAE by presenting eight studies that develop and employ diverse methodological approaches. The presented studies offer valuable insights into a variety of topics, including estimation accuracy using Bayesian strategies that incorporate information from previous time-points, the choice of trend estimation methods for fair country comparisons, plausible value estimation to account for measurement error, continuous-time models of longitudinal data, and compositional effects when investigating achievement growth. Through the use of simulation studies, illustration studies, tutorials, and the development of software, the contributions enrich the available modeling approaches for longitudinal LSAE, demonstrating their implementation and interpretation in practice.

Acknowledgements
The editors would like to express their gratitude to the Leibniz Institute for Educational Trajectories (LIfBi) in Bamberg (Germany) for organizing the 6th International Conference of the National Educational Panel Study (NEPS), which gave rise to the special issue.

Author contributions
TK: Conceptualization, Writing- Original draft preparation; MS: Conceptualization, Writing- Original draft preparation; TG: Conceptualization, Writing- Reviewing and Editing; CC: Writing- Reviewing and Editing, CA: Conceptualization, Writing- Reviewing and Editing; Supervision. All authors approved the final manuscript.

Funding
No funding was received for the preparation of the manuscript.

Data availability
Not applicable.

Declarations

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

Published online: 26 September 2024

References

Artelt, C., Weinert, S., & Carstensen, C. H. (2013). Assessing competencies across the lifespan within the German National Educational Panel Study (NEPS). *Journal for Educational Research Online*, *5*(2), 5–14. https://doi.org/10.25656/01:8422

Australian Curriculum, Assessment and Reporting Authority (2023). NAPLAN technical report for 2022, ACARA, Sydney. Retrieved from https://www.nap.edu.au/naplan

Blossfeld, H. P., & Roßbach, H. G. (Eds.). (2019). *Education as a Lifelong Process: The German National Educational Panel Study (NEPS)*, 2nd Edn. SpringerVS. https://doi.org/10.1007/978-3-658-23162-0

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

Cresswell, J., Schwantner, U., & Waters, C. (Eds.). (2015). A review of International large-scale assessments in education: Assessing component skills and collecting Contextual Data. PISA, The World Bank. https://doi.org/10.1787/9789264248373-en

Driver, C. C., & Voelkle, M. C. (2018). Hierarchical bayesian continuous time dynamic modeling. *Psychological Methods*, *23*(4), 774–799. https://doi.org/10.1037/met0000168

Heine, J. (2023). Pairwise: Rasch model parameters by pairwise algorithm [Computer software]. https://doi.org/10.32614/CRAN.package.pairwise

Heine, J. H., & Robitzsch, A. (2022). Evaluating the effects of analytical decisions in large-scale assessments: analyzing PISA mathematics 2003–2012. *Large-Scale Assessment in Education, 10,* 10. https://doi.org/10.1186/s40536-022-00129-5

Hernández-Torrano, D., & Courtney, M. G. R. (2021). Modern international large-scale assessment in education: An integrative review and mapping of the literature. *Large-scale Assessments in Education*, *9*(17). https://doi.org/10.1186/s40536-021-00109-1

Jindra, C., Sachse, K. A., & Hecht, M. (2022). Dynamics between reading and math proficiency over time in secondary education – observational evidence from continuous time models. *Large-Scale Assessment in Education, 10,* 1–19.https://doi.org/10.1186/s40536-022-00136-6

Kaplan, D., Chen, J., Lyu, W., & Yavuz, S. (2023). Bayesian historical borrowing with longitudinal large-scale assessments. *Large-Scale Assessment in Education, 11*(2), 1–30. https://doi.org/10.1186/s40536-022-00140-w

Lohmann, J. F., Zitzmann, S., Voelkle, M. C., & Hecht, M. (2022). A primer on continuous-time modeling in educational research: an exemplary application of a continuous-time latent curve model with structured residuals (CT-LCM-SR) to PISA Data. *Large-Scale Assessment in Education, 10,* 1–32. https://doi.org/10.1186/s40536-022-00126-8

Marks, G. N. (2024). No substantive effects of school socioeconomic composition on student achievement in Australia: A response to Sciffer, Perry and McConney. *Large-scale Assessment in Education*, *12*(8). https://doi.org/10.1186/s40536-024-00196-w

Martin, M. O., von Davier, M., & Mullis, I. V. S. (Eds.). (2020). Methods and Procedures: TIMSS 2019 technical report. Boston College, TIMSS & PIRLS International Study Center. Retrieved from https://timssandpirls.bc.edu/timss2019/methods/

NCES (2018). *Early Childhood Longitudinal Program (ECLS)* — Overview. Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Dept. of Education. Retrieved from https://nces.ed.gov/ecls/

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, Robustness, and reproducibility in Psychological Science. *Annual Review of Psychology*, *73*, 719–748. https://doi.org/10.1146/annurev-psych-020821-114157

OECD (2022). *PISA 2022 technical report*. Retrieved from https://www.oecd.org/pisa/

Pohl, S., & Carstensen, C. H. (2013). Scaling of competence tests in the National Educational Panel study – many questions, some answers, and further challenges. *Journal for Educational Research Online*, *5*, 189–216. https://doi.org/10.25656/01:8430

R Core Team (2024). R: A language and environment for statistical computing [Computer software]. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Robitzsch, A., Kiefer, T., & Wu, M. (2024). TAM: Test analysis modules [Computer software]. https://doi.org/10.32614/CRAN.package.TAM

Robitzsch, A., & Lüdtke, O. (2023). Comparing different trend estimation approaches in country means and standard deviations in international large-scale assessment studies. *Large-Scale Assessment in Education, 11*(26), 1–37. https://doi.org/10.1186/s40536-023-00176-6

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

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale Assessment Data: Issues in secondary analysis and reporting. *Educational Researcher*, *39*(2), 142–151. https://doi.org/10.3102/0013189X10363170

Scharl, A., & Zink, E. (2022). NEPSscaling: plausible value estimation for competence tests administered in the German National Educational Panel Study. *Large-scale Assessment in Education, 10*(1), 28. https://doi.org/10.1186/s40536-022-00145-5

Sciffer, M. G., Perry, L. B., & McConney, A. (2022). The substantiveness of socioeconomic school compositional effects in Australia: measurement error and the relationship with academic composition. *Large-scale Assessment in Education, 10*, 21. https://doi.org/10.1186/s40536-022-00142-8

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtnes, K., Tadesse, M. G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J., & Yau, C. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers*, *1*(1). https://doi.org/10.1038/s43586-020-00001-2

van Montfort, K., Oud, J. H. L., & Voelkle, M. C. (Eds.). (2018). *Continuous time modeling in the behavioral and related sciences*. Springer.

Voelkle, M. C., Gische, C., Driver, C. C., & Lindenberger, U. (2018). The role of time in the quest for understanding psychological mechanisms. *Multivariate Behavioral Research*, *53*(6), 782–805. https://doi.org/10.1080/00273171.2018.1496813

von Davier, M., Mullis, I. V. S., Fishbein, B., & Foy, P. (Eds.). (2022). Methods and Procedures: PIRLS 2021. Boston College, TIMSS & PIRLS International Study Center. Retrieved from https://pirls2021.org/.

von Maurice, J., Zinn, S., & Wolter, I. (2017). Large-scale assessments: Potentials and challenges in longitudinal designs. *Psychological Test and Assessment Modeling, 59*(1), 35–54. Retrieved from https://www.psychologie-aktuell.com/fileadmin/download/ptam/1-2017_20170323/03_Maurice.pdf

Publisher's note