Methodological Challenges of International Student Assessment

Andreas Frey[1,2] & Johannes Hartig[3]

[1]Goethe University Frankfurt, Germany

[2]Centre for Educational Measurement (CEMO) at the University of Oslo, Norway

[3]Leibniz Institute for Research and Information in Education (DIPF), Germany

This is a postprint version of the following chapter:

Frey, A., & Hartig, J. (in press). Methodological challenges of international student assessment. In H. Harju-Luukkainen, N. McElvany, & J. Stang. Monitoring of Student Achievement in the 21st Century - European Policy Perspectives and Assessment Strategies. New York: Springer.

Author Notes

On behalf of all authors, the corresponding author states that there is no conflict of interest.

Correspondence should be addressed to: Andreas Frey, Goethe University Frankfurt, Theodor-W.-Adorno-Platz 6, 60323 Frankfurt, Germany. E-mail: frey@psych.uni-frankfurt.de

Abstract

International large-scale assessments are very successful. One key factor of this success is their rigorous methodological and psychometric basis. Because education systems worldwide are subject to rapid changes, international large-scale assessments need to evolve as well. We describe five current methodological challenges that should be addressed so that large-scale assessments can continue to provide highly useful information on educational outcomes in the future. First, new or changed constructs should be adopted and constructs with declining importance should be dropped from the assessments. Second, the heterogeneity of student performance within and between countries should be better accounted for. This can be achieved by completing the introduction of computerised adaptive testing into international large-scale assessments and making full use of computers to optimise the testing and scaling process. Third, more analytical effort should be invested in the measurement and modelling of context variables, mainly by applying latent variable modelling. Fourth, full compliance with open science standards should be an aim by improving the transparency and accessibility of data and methods. Fifth, longitudinal validation studies need to be included as integral parts of international large-scale assessments in order to provide justifications for test score interpretations and their intended uses.

Keywords: testing, large-scale assessments, item response theory, measurement, methodology

Methodological Challenges of International Student Assessment

Introduction

International large-scale assessments (ILSAs) are the pivotal type of study when it comes to comparing student competencies at an international level. ILSAs are considered to be "remarkably successful in comparing student performances and curricula" (van de Vijver, Jude, & Kuger, in press, p. 17) and "here to stay" (Singer, Braun, & Chudowsky, 2018, p. 77). One key factor in the success of ILSAs lies in their rigorous methodological and psychometric basis. All major ILSAs use current state-of-the-art methodology and, additionally, stimulate methodological innovations, which can then be directly applied at an international level. We currently do not see any serious problems limiting the feasibility of ILSAs. Nevertheless, because education systems worldwide are subject to rapid changes, ILSAs obviously need to evolve as well in order to maintain their capability of providing highly useful information. This applies in particular to the current period, which is characterised by extensive changes due to the introduction of digital technology in education. Against this background, it is necessary to set the course now so that future methodological challenges that can already be anticipated now can be met.

In this chapter, we describe five current methodological challenges of ILSAs and we outline pathways for how to tackle them. By discussing the challenges *adopting new constructs*, *consideration of performance heterogeneity, measurement and statistical modelling of context variables*, *transparency of data and methods*, and *validation of test score interpretations* we address issues concerned with the collection, measurement, analysis, and usage of ILSA data.

Our remarks are based on the assumption that the core objectives of ILSAs will continue to be benchmarking and monitoring in the future and that the descriptive character of ILSAs will thus be retained. This seems to be the most likely development track for ILSAs, although the desire to

derive causal conclusions from ILSAs has already been discussed (e.g., Kaplan & Kuger, 2016; Singer, Braun, & Chudowsky, 2018).

Adopting New Constructs

In today's fast-moving world, which is also characterised by the penetration of digital technology into all areas of life, the social relevance of educational constructs is also in flux. On the one hand, constructs change (e.g., reading competency due to the introduction of digital reading devices). On the other hand, new constructs come into focus and quickly acquire social relevance. For example, the competent use of computers and the Internet is without doubt an important prerequisite for successful participation in today's modern societies. Correspondingly, skills in using information and communication technology (ICT; International ICT Literacy Panel, 2002) have been under discussion as an educationally relevant construct for around 20 years now. The relevance of the construct for ILSAs is expressed by the fact that it has already been focused on in the International Computer and Information Literacy Study (ICILS; Fraillon, Schulz, & Ainley, 2013). It can be assumed that new constructs will continue to come into focus in the future. In this respect, there is already a trend towards constructs in which tasks have to be solved cooperatively by several persons (e.g., Herborn, Stadler, Mustafić, & Greiff, in press). The measurement of such constructs is associated with considerable challenges regarding test development, scoring, and scaling. The main challenge when incorporating such constructs into ILSAs will be to use digital technologies in such a way that the constructs are operationalised as realistically as possible while at the same time satisfying psychometric requirements.

Fortunately, ILSAs are well equipped to identify and adopt new constructs because they have a differentiated international expert panel structure (van de Vijver, Jude, & Kuger, in press). ILSAs are likely to be faster and more flexible than education systems in this regard. However, most ILSAs still need to implement guidelines on when and how constructs will not be continued.

Consideration of Performance Heterogeneity

The heterogeneity of student performance within and between countries is a persisting challenge for ILSAs that has not yet been fully addressed. At the onset of ILSAs, only paper-based administration was possible. At that time, the practice of using test booklets with item difficulties roughly aligned with the assumed student performance distribution was the best option. However, due to the increased possibilities of using computers for testing purposes, performance heterogeneity can now be handled much better. Most importantly in this regard, computers can be used to adapt the difficulty of the presented items to the assumed or the measured individual performance level. This brings statistical and psychological improvements compared to assigning test booklets randomly.

The statistical advantages arise from the fact that the statistical information provided by the response to one single item is at its maximum when the difficulty parameter of this item is equal (in the Rasch model) or close to (in other item response theory [IRT] models) to the performance level of the test taker. Adapting the item difficulties of presented items to the performance level can therefore lead to substantial increases in measurement precision.

Regarding the psychological advantages, empirical evidence based on data from the Programme for International Student Assessment (PISA) suggests that the *ability-difficulty fit* (the difference between the ability level of a student and the average difficulty of the items this student worked on) affects test-taking effort and boredom/daydreaming significantly (Asseburg & Frey, 2013). Individuals whose ability level was much lower than the average difficulty of the items they had to answer reported lower levels of test-taking effort and higher levels of boredom/daydreaming compared to students whose ability level was equal to or higher than the average item difficulty. The adaptation of the difficulty level of the presented items to the individual student performance, and thus keeping the ability-difficulty fit constant across students, circumvents the potentially

performance-reducing effects of the test instrument itself that are mediated by variables such as test-taking effort or boredom/daydreaming. As a result, all tested students have the same opportunity to demonstrate their abilities.

The first steps towards adjusting the difficulty of the presented items to student abilities have already been taken. Starting with the 2009 assessment, PISA offered low-performing countries the option of including item clusters with easy items in the assessment (Organisation for Economic Co-operation and Development, 2012a). This mild adjustment of the average item difficulty, however, showed only marginal advantages regarding bias and the root mean square error (RMSE) of ability estimation in the simulation study of Rutkowski, Rutkowski, and Liaw (2018). The Progress in International Reading Literacy Study (PIRLS) uses a similar approach to that of PISA. Here, a less difficult version of the regular assessment, called PIRLS Literacy, is available for low-performing countries (Mullis & Martin, 2015).

Moving one step ahead, the adaptation of item difficulty to performance level is currently not only carried out on the level of countries but also on the level of individual students in PIAAC (Programme for the International Assessment of Adult Competencies) and PISA. In both ILSAs, multistage testing (MST) was used (for PIAAC, see Organisation for Economic Co-operation and Development, 2013a; for PISA, see Yamamoto, Shin, & Khorramdel, 2018). PIAAC used both, computer-based assessment and paper-based assessment. For computer-based assessment, MST was applied. The MST design included several layers of adaptation. Adaptation took previous student responses to cognitive items and background information (education level, native vs. non-native speaker) into account while controlling for item exposure rates and other constraints. Using this MST design proved to be 10-30% more efficient for the Literacy scale and 4-31% more efficient for the Numeracy scale compared to a nonadaptive linear test design (Yamamoto, Khorramdel, & Shin, 2018). In PISA, the efficiency gains compared to nonadaptive testing that

were achieved with the MST design were a bit smaller (4-7%) but still of a relevant magnitude (Yamamoto, Shin, & Khorramdel, 2018).

The next obvious step would be to allow for adaptation on the smallest possible level in order to harvest efficiency gains in the best possible way. For many ILSAs, this level would be testlets and, for some, even single items. Such a fine-grained adaptation would require the implementation of computerised adaptive testing (CAT; e.g., van der Linden & Glas, 2000). Using CAT instead of MST seems to be an accessible option. Content balancing and item position effects (e.g., Debeer, Buchholz, Hartig, & Janssen, 2014; Nagy, Nagengast, Frey, Becker, & Rose, 2018) can well be accounted for in CAT, even though Yamamoto, Khorramdel and Shin (2018) mentioned them as reasons to use MST instead of CAT. Both can be addressed in CAT (in a statistically optimal sense), for example with the shadow-testing approach (van der Linden, & Reese, 1998). Of course, item position effects do not disappear by using CAT. However, controlling the position on which items are presented prevents systematic bias in ability estimates and group statistics.

Things are a bit different for the remaining major reason for favouring MST that is sometimes mentioned, namely that response revision is possible within blocks or stages. This is not possible in typical testlet- or item-level CAT systems. However, several solutions for response revision in CAT have recently been proposed. Although some of these are very promising (e.g., Cui, Liu, He, & Chen, 2018), the proposed methods have not yet reached operational status. In any case, it is worth developing these methods further because applying CAT in ILSAs makes further substantial improvements in measurement precision possible (e.g., Frey & Seitz, 2011).

With regard to the operational aspects of conducting and maintaining ILSAs, substantial improvements can also be achieved through CAT. For example, adding new items to the item pool, identifying drifted items, and linking with previous assessments can be considerably simplified by

adopting methods such as the continuous calibration strategy (CCS; Fink, Born, Frey, & Spoden, 2018). The CCS incorporates all these aspects in the sense of a self-learning system and makes use of technology in order to optimize the functioning of test instruments *across* cycles, and thus widening the view from one cycle to seeing ILSAs as a process. Adopting such a view and the corresponding methods, will make it easier to identify and to resolve issues that can be problematic for trend reporting. One example for such problems were the changes in the testing mode, the scaling method, and the type of some tasks incorporate in PISA 2015. The findings of Robitzsch et al. (2017) suggest that these changes could have biased the trend estimation for Germany. The CCS can help to introduce changes (such as new item types) more smoothly and more securely. The future success of ILSAs will largely depend on how well they adopt such methods to make optimal use of the enormous potential of computers. For computers to retain the status of essentially being used to imitate paper-based test procedures would hardly be viable for the future.

Measurement and Statistical Modelling of Context Variables

Traditionally, the main focus of ILSAs in education is on student achievement, and most of the assessment time is allotted to the assessment of *cognitive outcomes*, measured by means of achievement tests. In PISA 2015, for example, 120 minutes of assessment time were reserved for students to work on achievement tests, while they had 35 minutes to answer the student questionnaire. The variables assessed by questionnaire items (typically self-reports) are subsumed under *contextual information* (e.g., Organisation for Economic Co-operation and Development, 2017a) or *context assessment* (Kuger & Klieme, 2016). The amount of assessment time allotted to achievement tests compared to questionnaire variables stands in contrast to the number of constructs measured. While the 120 minutes of achievement testing in PISA 2015 were used to measure 10 literacy dimensions (reading, mathematics, and eight science scales), 32 derived variables (i.e., scales based on multiple items) in the student questionnaire were constructed from

student questionnaire data collected in 35 minutes (Organisation for Economic Co-operation and Development, 2017b). In addition to these derived variables, a number of student questionnaire variables are measured by individual items and, in addition to the student questionnaire, further context variables are assessed with context questionnaires for parents, teachers, and school principals.

The higher priority that is traditionally placed on achievement constructs is also visible in the methodology applied to achievement test data compared to questionnaire response data. The constructs assessed via achievement tests are modelled as latent variables with multidimensional IRT models. The multivariate distribution for the cognitive outcomes of the student population conditioned on a multitude of *background variables* from the context questionnaires is estimated using a latent linear regression model (Mislevy, 1991). Finally, *plausible values* (PVs) are used as estimates of student proficiency. They are essentially multiple imputations of the cognitive outcomes for each student. In contrast, for constructs assessed via questionnaire items, observed scores or point estimates such as the weighted likelihood estimate are used in further analyses without conditioning.

In short, considerably more time, money, and analytical effort is invested in the measurement and modelling of cognitive outcomes than in that of context variables. However, the constructs assessed via questionnaires are receiving increasing attention. ILSAs use context variables not only as predictors of cognitive outcomes; several questionnaire variables are also treated as noncognitive educational outcomes (e.g., beliefs, motivation, and well-being) in their own right (e.g., Kuger & Klieme, 2016). The increasing importance of context variables in ILSAs is visible, for instance, by the fact that the PISA 2003 and 2006 assessment frameworks contained a single page on “the context questionnaires and their use” (Organisation for Economic Co-operation and Development, 2003, p. 18, and Organisation for Economic Co-operation and

Development, 2006, p. 14, respectively); in contrast, the frameworks for PISA 2009, 2012, and 2015 (Organisation for Economic Co-operation and Development, 2009, 2013b, 2017a, respectively), have whole book chapters on the context questionnaires' framework.

The imbalance between the efforts invested in modelling cognitive outcomes versus context variables can be regarded as problematic for at least two reasons. First, using observed scores or point estimates for context variables as predictor variables in the population models means that measurement error for those constructs is not taken into account and data are assumed to be complete (i.e., not missing; Rutkowski & Rutkowski, 2016). This is already a problem for the traditional focus on achievement using context variables merely as predictors, as missing data and measurement error can lead to biased estimates of the relationships between context variables and achievement (Rutkowski & Rutkowski, 2016). Second, although the methods applied to achievement test data aim to produce unbiased estimates of multivariate population and subgroup distributions that have been corrected for measurement error; this is not achieved with the observed scores or point estimates for questionnaire variables. Consequently, the reported results of analyses for noncognitive outcomes (e.g., group means and standard deviations or regression coefficients) are affected by measurement error. Those on cognitive outcomes are not (or if they are, at least to a smaller extent), as the analysis of PVs provides estimates based on a latent regression model. That again implies that differences between subgroups within education systems will be underestimated for noncognitive outcomes. The practical implications of this difference can be illustrated by looking at brief reports such as "PISA 2015 results in focus" (Organisation for Economic Co-operation and Development, 2018). The first major results table summarises achievement in science, reading, and mathematics by country, the second summarises results on students' science beliefs, engagement, and motivation. For both tables, significant differences between country means and the OECD average are highlighted. It is very

likely that the results on noncognitive outcomes would contain more statistically significant differences and also larger effect sizes if the same elaborated modelling and estimation approach used to obtain the achievement results were applied to the context questionnaire data.

Another possible shortcoming of modelling the context questionnaire data, which also applies to the achievement data, is that the multilevel structure of the data (students in schools and/or classrooms) is not taken fully into account. The estimation of the population model in PISA 2015, for instance, was conducted with a linear regression model that did not incorporate random effects on the school level (Organisation for Economic Co-operation and Development, 2017b). Depending on the strength of cluster effects, this could affect both the group means and standard errors obtained with the PVs based on the population model (Li, Oranje, & Jiang, 2009).

To summarise, a methodological challenge for ILSAs is to reduce the imbalance between the modelling approaches used for achievement test data and context questionnaire data. Ideally, all constructs should be modelled as latent variables in order to account for measurement error and missing data, and the hierarchical structure should be included in the model. Practically, this is not yet feasible given the large number of constructs measured. Nevertheless, it is important to keep the ideal in mind and try to move towards it step by step.

Transparency of Data and Methods

In recent years, the *open science* movement has become increasingly influential in science, politics, and society, and public awareness of the importance of the accessibility, transparency, and replicability of scientific results has grown substantially. Standards that need to be met in order to achieve an open science culture include the citation of data and materials, the transparency and accessibility of data, methods, and materials, the preregistration of studies, and the replication of research findings (Nosek et al., 2015). For ILSAs, the aims of open science are particularly important as they are explicitly designed to provide policy makers with the

information they need to make decisions on the configuration of education systems. Therefore, methods and results should be comprehensible and reproducible by independent experts. It has to be noted that ILSAs already meet some open science standards and have even been instrumental in setting some of these standards for the educational research community, particularly with respect to data and materials. The databases for ILSAs are made publicly available along with many of the instruments used in the assessment (except for the achievement test items for obvious reasons of test security and field trial data for most studies). Kuger, Klieme, Jude, and Kaplan (2016) even made all translated versions of the PISA 2015 context questionnaires openly available. Furthermore, some of the open science standards are not particularly relevant for ILSAs. Standards related to preregistration are not applicable to ILSAs, as they are descriptive in nature and do not aim to test hypotheses specified a priori.

Nonetheless, there is room for improvement in ILSAs when it comes to meeting open science standards with respect to the transparency and accessibility of methods. Because ILSAs are often used as a reference by researchers working on smaller studies, it is particularly important that their methods are well documented and well founded. This is currently not always the case. For example, in PISA 2015, the root mean square deviation (RMSD) fit statistic was used to identify differential item functioning. RMSD values larger than 0.15 were interpreted to be indicators of deviations for achievement test items and values larger than 0.30 for context questionnaire items (Organisation for Economic Co-operation and Development, 2017b). However, no rationale or evidence (e.g., based on simulation studies) was provided for these cut-off criteria. Another issue regarding documentation, which has been criticised by Rutkowski and Rutkowski (2016), is that the technical reports of several ILSAs are made available much later than the publication of the corresponding results, making it difficult for independent researchers to critically evaluate the methods of an ILSA in time. Finally, the transparency of methods

should, in the case of statistical analyses, include the publication of the code used in the analyses (Nosek et al., 2015).

Most code used for ILSA analyses has not been published in recent ILSAs, and proprietary software is often used that is not easily accessible to other researchers (e.g., the mdltm software [von Davier, 2005] used in PISA 2015). A noteworthy and encouraging exception is the documentation of an Austrian national large-scale assessment, whose researchers published their methods as a comprehensive book (Breit & Schreiner, 2016) including the R code used for the analyses, and in which the routines used for reporting were made available as an R package (Robitzsch & Oberwimmer, 2018). In summary, there are possibilities for ILSAs to better meet open science standards, namely, the timely publication of technical documentations, the use of freely available software, and the publication of the analysis code used for the analyses and for processing the results.

Validation of Test Score Interpretations

According to the Standards for Educational and Psychological Testing, validity is considered "the most fundamental consideration in developing tests and evaluating tests" (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014, p. 11). Measured against this central position of validity for assessment, the efforts made in ILSAs to provide evidence for the desired test score interpretations and the assumed effects on the future lives of the tested students are very sparse. This is surprising as several ILSAs formulate rather strong claims about the importance of the test scores for later life chances and individual success in society in general. As an example, the objective of PISA is defined as: "PISA assesses the extent to which 15-year-old students, near the end of their compulsory education, have acquired key knowledge and skills that are essential for

full participation in modern societies." Empirical evidence is needed in order to support such test score interpretations.

Some such evidence was published based on the Youth in Transition Survey (YITS; Motte, Hanqing, Zhang, & Bussière, 2008; Statistics Canada, 2011) for Canada. YITS was set up to examine major transitions in young people's lives, with respect to education, training, and work. One cohort of YITS participated in PISA 2000 and a subsample of that cohort took the PISA reading test in 2009 again. The results show that PISA test results can indeed be connected to desirable outcomes. For example, a positive relationship between reading performance in PISA 2000 and the probability of attending a university at the age of 24 was reported (Organisation for Economic Co-operation and Development, 2012b). Nevertheless, the study contained only a limited number of indictors for a *full participation in modern societies* and was conducted in Canada only. Another example for a study capable to provide some validity evidence is the TREE study (Transitions in Youth and Young Adulthood; Scharenberg, Hupka-Brunner, Meyer, & Bergman, 2016). The study uses the Swiss PISA 2000 sample for several follow-ups. The results show, for example, that reading performance in PISA is predictive for educational attainment ten years later. A substantial proportion of the students with low reading competency in PISA 2000 with regards to reading did not complete an upper secondary educational program (below proficiency level 1: 37 %; proficiency level 1: 19 %) compared to only 4 % on each of the proficieny levels 4 and 5 (Scharenberg et al., 2016).

As the claims made by ILSA are ambitious, more support from more countries is needed to justify the intended test score interpretations and uses in a broad sense. Because the validity of test score interpretations and uses can change over time (Kane, 2013), collecting validity evidence at one point in time only is not sufficient.

In order to justify the intended test score interpretations of ILSAs and the sometimes far-reaching decisions made by educational policy makers based on those interpretations, the future challenge is to implement longitudinal validation studies as an integral study part for all countries. Otherwise, ILSAs are based on unsupported claims; this does not do justice to the effort and costs invested, nor does it sufficiently support the conclusions derived.

Conclusions

This chapter outlined some of the challenges of ILSAs, as well as possible approaches on how to meet them. These relate to the following aspects: inclusion of new constructs, consideration of proficiency differences, measurement and statistical modelling of context variables, transparency of data and methods, and validation. The issues covered in this chapter are by no means exhaustive, e.g. we did not address challenges related to adaptation and translation of instruments and dealing with measurement invariance violations (e.g., Caro, Sandoval-Hernández, & Lüdtke, 2014; Rutkowski & Rutkowski, 2013). Due to the high methodological standards that have already been met, the ever-growing interest, the increasing use of data-based decision making, and the international character of ILSAs, it is likely that the discussed challenges can be met well.

ILSAs will probably continue to initiate and stimulate methodological developments in the future, to develop and examine these developments scientifically, and to apply successful approaches. Thus, they provide very good opportunities for addressing one of the Achilles' heels of basic methodological-statistical research: Mastering the leap from basic psychometric research to actual application.

References

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

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

Breit, S., & Schreiner, C. (Eds.) (2016). *Large-Scale Assessment mit R: Methodische Grundlagen der österreichischen Bildungsstandard-Überprüfung [Large-scale assessment with R: Methodological foundations of the Austrian educational standard evaluation]*. Vienna: Facultas.

Caro, D. H., Sandoval-Hernández, A., & Lüdtke, O. (2014). Cultural, social, and economic capital constructs in international assessments: An evaluation using exploratory structural equation modeling. *School Effectiveness and School Improvement*, *25*, 433–450.

Cui, Z., Liu, C., He, Y., & Chen, H. (2018) Evaluation of a new method for providing full review opportunities in computerized adaptive testing—computerized adaptive testing with salt. *Journal of Educational Measurement*, *55*, 582–594.

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

Fink, A., Born, S., Frey, A., & Spoden, C. (2018). A continuous calibration strategy for computerized adaptive testing. *Psychological Test and Assessment Modeling*, *60*, 327–346.

Fraillon, J., Schulz, W., & Ainley, J. (2013). *International computer and information literacy study assessment framework*. Amsterdam, the Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

Frey, A., & Seitz, N. N. (2011). Hypothetical use of multidimensional adaptive testing for the assessment of student achievement in PISA. *Educational and Psychological Measurement*, *71*, 503–522.

Herborn, K., Stadler, M., Mustafić, M., & Greiff, S. (in press). The assessment of collaborative problem solving in PISA 2015: Can computer agents replace humans? *Computers in Human Behavior*.

International ICT Literacy Panel (2002). *Digital transformation: A framework for ICT literacy*. Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/ICTREPORT.pdf

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*, 1–73.

Kaplan, D., & Kuger, S. (2016). The methodology of PISA: Past, present, and future. In S. Kuger, E. Klieme, N. Jude, & D. Kaplan (Eds.), *Assessing Contexts of Learning* (pp 53–73). New York: Springer.

Kuger, S., & Klieme, E. (2016). Dimensions of context assessment. In S. Kuger, E. Klieme, N. Jude, & D. Kaplan (Eds.), *Assessing Contexts of Learning* (pp 3–37). New York: Springer.

Kuger, S., Klieme, E., Jude, N., & Kaplan, D. (Eds.) (2016), *Assessing contexts of learning*. New York: Springer.

Li, D., Oranje, A., & Jiang, Y. (2009). On the estimation of hierarchical latent regression models for large-scale assessments. *Journal of Educational and Behavioral Statistics, 34,* 433–463.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*, 177–196.

Motte, A., Hanqing, Q., Zhang, Y., & Bussière, P. (2008). The youth in transition survey: Following Canadian youth through time. In Ross Finnie, R., Mueller, R. E., Sweetman, A., & Usher, A. (Eds.), *Who goes? Who stays? What matters? Accessing and persisting in post-secondary education in Canada*. (pp 63–75). Montréal: Mc-Gill, Queen's University Press.

Mullis, I. V. S. & Martin, M. O. (Eds.). (2015). *PIRLS 2016 assessment framework*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Nagy, G., Nagengast, B., Frey, A., Becker, M., & Rose, N. (2018). A multilevel study of position effects in PISA achievement tests: Student- and school-level predictors in the German tracked school system. *Assessment in Education: Principles, Policy & Practice*. Advance online publication.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... Yarkoni, T. (2015). Promoting an open research culture. *Science, 348,* 1422–1425.

Organisation for Economic Co-operation and Development (2006). *Assessing scientific, reading and mathematical literacy. A framework for PISA 2006.* Paris: OECD Publishing.

Organisation for Economic Co-operation and Development (2009). *PISA 2009 assessment framework. Key competencies in reading, mathematics and science.* Paris: OECD Publishing.

Organisation for Economic Co-operation and Development (2012a). *PISA 2009 technical report*. Paris: OECD Publishing.

Organisation for Economic Co-operation and Development (2012b). *Learning beyond fifteen: Ten years after PISA*. Paris: OECD Publishing.

Organisation for Economic Co-operation and Development (2013a). *Technical report of the Survey of Adult Skills (PIAAC)*. Paris: OECD Publishing.

Organisation for Economic Co-operation and Development (2013b). *PISA 2012 assessment and analytical framework mathematics, reading, science, problem solving and financial literacy.* Paris: OECD Publishing.

Organisation for Economic Co-operation and Development (2017a). *PISA 2015 assessment and analytical framework science, reading, mathematic, financial literacy and collaborative problem solving.* Paris: OECD Publishing.

Organisation for Economic Co-operation and Development (2017b). *PISA 2015 technical report*. Paris: OECD Publishing.

Organisation for Economic Co-operation and Development (2018). *PISA 2015 results in focus.* URL: https://www.oecd.org/pisa/pisa-2015-results-in-focus.pdf

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: A calibration of the German PISA data]. *Diagnostica*, *63*, 148–165.

Robitzsch, A., & Oberwimmer, K. (2018). *BIFIE survey: Tools for survey statistics in educational assessment. R package version 3.0-14.* URL: https://CRAN.R-project.org/package=BIFIEsurvey

Rutkowski, D., & Rutkowski, L. (2013). Measuring socioeconomic background in PISA: One size might not fit all. *Research in Comparative and International Education*, *8*, 259–278.

Rutkowski, L., & Rutkowski, D. (2016). A call for a more measured approach to reporting and interpreting PISA results. *Educational Researcher*, *45*, 252–257.

Rutkowski, D., Rutkowski, L., & Liaw (2018). Measuring widening proficiency differences in international assessments: Are current approaches enough? *Educational Measurement: Issues and Practice, 37*(4), 40–48.

Scharenberg, K., Hupka-Brunner, S., Meyer, T., & Bergman, M. M. (Eds.) (2016). *Transitions in Youth and Young Adulthood: Results from the Swiss TREE Panel Study*. Volume 2. Zürich: Seismo.

Singer, J. D., Braun, H. I., & Chudowsky, N. (2018). *International education assessments. Cautions, conundrums, and common sense*. Washington: National Academy of Education.

Statistics Canada (2011). *Youth in transition survey (YITS). Cohort A - 25-year-olds, cycle 6. User guide*. Ottawa: Statistics Canada.

van de Vijver, F. J. R., Jude, N., & Kuger, N. (in press). Challenges in international large-scale educational surveys. In B. Denman, L. E. Suter, & E. Smith (Eds.), *The SAGE Handbook of Comparative Studies in Education.* Thousand Oaks: Sage.

van der Linden, W. J., & Glas, C. A. (Eds.) (2000). *Computerized adaptive testing: Theory and practice.* Dordrecht: Kluwer Academic Publishers.

van der Linden, W. J., & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measuremen*t, *22*, 259–270.

von Davier, M. (2005). *A general diagnostic model applied to language testing data (Research Report No. RR-05-16).* Princeton, NJ: Educational Testing Service.

Yamamoto, K., Khorramdel, L., & Shin, H. J. (2018). Introducing multistage adaptive testing into international large-scale assessments designs using the example of PIAAC. *Psychological Test and Assessment Modeling*, *61*, 347–368.

Yamamoto, K. Shin, H. J., & Khorramdel, L. (2018). Multistage adaptive testing design in international large-scale assessments. *Educational Measurement: Issues and Practice*, *37*(4), 6–27.