Educational Assessment, Evaluation and Accountability (2021) 33:1–7
https://doi.org/10.1007/s11092-021-09354-y

# The integrity of educational outcome measures in international assessments

Rolf Strietholt [1,2,3] · Monica Rosén [3] · Olesya Gladushyna [2]

Published online: 21 January 2021


Among the most salient findings in the field of education are the huge differences in student achievement and in learning environments, as reported in international comparative studies. The results of the international studies, such as PISA (Programme for International Student Assessment) and TIMSS (Trends in International Mathematics and Science Study), are widely received by policymakers and academics, and their data are used in a variety of secondary analyses. However, for such international comparisons to be meaningful, the samples, as well as the educational outcome measures, must be comparable (e.g., Grek 2009; Hopfenbeck et al. 2017; Lenkeit et al. 2015; Lindblad et al. 2018; Meyer and Benavot 2013; Meyer et al. 2018; Strietholt and Scherer 2017).

This special issue focuses on the measures of educational outcomes achieved by students in international assessments and on the representativity of the samples. The overall aim is to scrutinize the strengths and the limitations of the current practice of comparative assessments. In nine contributions from international scholars, the quality of the data obtained from the original studies is evaluated, both conceptually and technically, in order to determine the integrity of these data for policy advice and research.

This special issue is organized into three sections. The first set of five studies concerns the comparability of the educational measures in different countries and over time (Edwin Cuellar Caicedo, Ivailo Partchev, Robert Zwiser, & Timo Bechger; Hüseyin H. Yıldırım; Andrés Christiansen & Rianne Janssen; Erika Majoros, Monica Rosén, Stefan Johansson, & Jan-Eric Gustafsson; Leah Natasha Glassow, Victoria Rolfe, & Kajsa Yang Hansen). This issue has been discussed as a measurement

✉ Rolf Strietholt
rolf.strietholt@tu–dortmund.de

[1] International Association for the Evaluation of Educational Achievement (IEA), Überseering 27, 22297 Hamburg, Germany

[2] Center for Research on Education and School Development, TU Dortmund University, Vogelpothsweg 78, 44227 Dortmund, Germany

[3] Department of Education and Special Education, University of Gothenburg, Box 300, 40530 Göteborg, Sweden

invariance in a confirmatory factor analysis (CFA) framework and as differential item functioning (DIF) in the IRT world. Incomparable measures are often described as posing a problem, but the authors demonstrate some constructive approaches to deal with this, such as DIF as an indicator of specific strengths or weaknesses in certain content areas, item positioning, culture, and so on.

The second section includes three papers that all focus on the concepts, measures and norms underlying educational outcome assessments (Andrés Strello, Rolf Strietholt, Isa Steinmann, & Charlotte Siepmann; Wangqiong Ye, Sigrid Blömeke, & Rolf Strietholt; Olesya Gladushyna, Rolf Strietholt, & Isa Steinmann). The authors of these papers observe a lack of conceptual clarity and provide evidence that different measures lead to different empirical findings. Researchers who intend to explore how to operationalize inequality, measure resilience or study math, reading and science should read these first three articles.

The third section of this special issue consists of only one study on non-response in the PISA (Jake Anders, Silvan Has, John Jerrim, Nikki Shure, & Laura Zieger). Non-response in surveys is a topic that receives too little attention by researchers and policymakers. Using Canada as an example, the authors of the final paper illustrate that the credibility of international assessments can be jeopardized by school and student non-response. The contribution is strongly recommended not only for those who initiate international studies, as well as for the national coordinators that administer the studies in their countries, but also for other stakeholders who use empirical data for monitoring and policy-making.

## 1 The papers in brief

### 1.1 Measures of equity and efficacy

One of the highly important topics in contemporary educational research is the inequality among students, which remains a major issue for many regions. In this regard, Andrés Strello, Rolf Strietholt, Isa Steinmann, and Charlotte Siepmann identify three types of inequality, namely dispersion inequality, social inequality and educational adequacy, and show the effect of early tracking on different types of inequality. Interestingly, the authors point out that "inequalities" usually connote injustice, but they also claim that dispersion inequality may be regarded as an acceptable outcome. Their conclusions, as well as their identification of these three types of inequality, contribute not only to further research but also to policy evaluation. For example, a key finding is that early tracking largely contributes to social inequality.

### 1.2 How would you measure academic resilience?

The issue of inequality in education is closely connected to academic resilience, which refers to students' capacity for high performance despite their disadvantaged background. Although most researchers using data from international large-scale assessments (ILSAs) define academic resilience with two criteria – student background and achievement – their conceptualisations and operationalisations vary substantially. In their systematic review, Wangqiong Ye, Sigrid Blömeke, and Rolf Strietholt identify

20 ILSA studies applying measures of socioeconomic status and achievement, different approaches to setting thresholds and consequently, different classifications of individual students as resilient or non-resilient. In their paper, they discuss the validity of these different definitions while showing how the classification of students as resilient depends heavily on the economic context where the students grow up. Moreover, significant interactions with gender and language background call for further research. The authors conclude that strong attention should be paid to the way that academic resilience is operationalized to avoid misfitting inferences. Additionally, they suggest using relative country-specific thresholds in defining students as disadvantaged or high achievers to overcome the risk of making the definition mainly dependent on the countries' developmental states.

### 1.3 Can students be good in all subjects?

When analysing student performance in ILSAs, in previous research, it was concluded that students varied in their overall performance levels across all subject domains but not in their individual performance profiles (e.g., Bergold et al. 2017; Wendt and Kasper 2016). In contrast, in their study, Olesya Gladushyna, Rolf Strietholt, and Isa Steinmann demonstrate that there are students who possess subject-specific strengths and weaknesses, and at the same time, there are those with similar scholastic performance across subject domains. The authors argue that traditional CFA and latent profile analysis (LPA) approaches have certain methodological limitations and propose using a factor mixture analysis (FMA) model to combine the advantages of both approaches. Indeed, the FMA has a better fit to the data than CFA and LPA models. The authors conclude that the choice of the methodology to analyse student performance is crucial because different methods lead to different results. The main finding of their study is that student performance is more than just general intelligence.

### 1.4 DIF and what it means for ILSA

The issue of comparability in measures of educational achievements across countries is the focus of the paper written by Edwin Cuellar Caicedo, Ivailo Partchev, Robert Zwiser, and Timo Bechger. They argue that measurement non-invariance (i.e., DIF) in ILSA should not only be regarded as a problem but as a potential source of interpretable information in the analysis of differences among educational systems. The authors propose methods to investigate and visualize measurement invariance when a large number of groups are involved (in their illustrative example, countries participating in PISA 2012), and they suggest a form of residual analysis after the dominant component has been removed. Their proposed multivariate techniques can be easily replicated. The analytical approach supports identifying biclusters of countries and items to reveal potentially interesting structures. This strategy connects the spirit of DIF analysis with classical methods of detecting DIF. Hence, their paper provides a methodological contribution, motivated by their belief that proper analysis of DIF may lead to more actionable insights in education.

### 1.5 Race for rankings or a wild-goose chase

Probably, the most popular outcome of international assessments constitutes country rankings. For these rankings to be meaningful, the parameters of test items must be

equivalent across participating countries. However, Hüseyin H. Yıldırım argues that it is very difficult, if not impossible, to reach item parameter equivalence in international assessments based on theories describing the culture and human-cognition relationship. It is a well-established finding that test items in international assessments may function differentially across countries. However, the general belief is that such problems may arise from only a few items and among a few countries. It is also assumed that these problems can be avoided if test items are adapted appropriately across countries. However, using the TIMSS 2015 data set, the results of Yıldırım's study show that this may not be the case. In international assessments, the non-equivalence of item parameters may be a general and inevitable consequence of cultural differences among countries, which calls for further research (see the previous contribution by Edwin Cuellar Caicedo and colleagues). From the comprehensive evidence presented, it is suggested that the current attention to country rankings in the international reports should be redirected to more informative and more useful outcomes to improve educational systems.

### 1.6 Learning by doing: Practice effect in language tests

In contrast to the assumptions made in standard measurement models used in large-scale assessments, student performance may change during the test administration. Andrés Christiansen and Rianne Janssen use an explanatory item response theory framework to analyse item position effects in the 2012 European Survey on Language Competences. Their analysis reveals consistent item position effects for listening but not for reading. More specifically, item difficulty decreases for a large subset of items along with item position, which is known as the practice effect. This practice effect differs among regions but is not related to the test administration mode. As the practice effects are substantial, it seems advisable to include them in the measurement model. Moreover, few educational measurement studies have been able to find practice effects; on the contrary, fatigue effects are commonly found throughout ILSAs. The authors contribute ideas for further research on position effects and their possible consequences for researchers' and policymakers' understanding of achievement scores in ILSAs.

### 1.7 Tracking half-century trends in mathematics achievement

In a series of studies on the relation between education and economic growth, Hanushek and Woessmann (see, e.g., 2011, 2012, 2015) based their cognitive outcomes on achievement measures from large-scale assessments by calculating standardized scores for all countries on all assessments. They used the US National Assessment of Educational Progress (NAEP) to link various ILSAs to the same scale using the mean-sigma method. Their approach is based on the assumption that the samples within educational systems are comparable across studies and over time. However, Erika Majoros, Monica Rosén, Stefan Johansson, and Jan-Eric Gustafsson make no such assumptions; instead, their analysis takes into account some variations in both the indicators of mathematics achievement and the comparability of the samples from the participating countries over time. The authors apply a more rigorous linking approach based on the item response theory, where the trait score estimates and their corresponding standard errors are independent of population distributions (Embretson and Reise

2000; Strietholt and Rosén 2016). Thus, they are able to link mathematics achievement using the population of eighth-grade students from the four countries (England, Israel, Japan and the United States) that participated in all assessments from 1964 to 2015, thereby achieving comparable scores over a 50-year period. Their study contributes not only a more well-founded trend scale but also a valuable time perspective. Both should be used in further research to include more countries and to better address issues of stability and change in educational achievement.

### 1.8 Which countries have the happiest teachers?

Research related to the "characteristics" dimension of teacher quality (including self-efficacy, job satisfaction and perceptions of work environments) has proven this factor's inconclusive or weak relation to student achievement (Goe 2007; Nilsen and Gustafsson 2016). Using data from TIMSS 2015 and multiple group confirmatory factor analysis (MGCFA) with an alignment optimization approach outlined by Asparouhov and Muthen (2014), Leah Natasha Glassow, Victoria Rolfe, and Kajsa Yang Hansen investigate teacher-related characteristics and perceptions of work contexts across countries using the newly constructed latent means of mathematics teacher job satisfaction, self-efficacy, perceptions of school academic climate, perceptions of school conditions and resources, and perceptions of school safety and organization. Particularly interesting results are found for teacher job satisfaction and self-efficacy, where clear geographical patterns emerge in some cases. Teacher job satisfaction and mathematics teacher self-efficacy tend to be higher in East and Southeast Asian countries, such as Japan, Singapore, Chinese Taipei and Hong Kong, and lower in Middle Eastern countries at the bottom of the achievement rankings, such as Qatar, Oman, UAE, Lebanon and Kuwait. Ultimately, in this paper, the authors demonstrate an approach to how educational researchers can tackle previously unanswerable substantive questions through new methodological advancements. Future research can make use of the newly constructed means for further secondary analysis or build on this research to examine teacher characteristic means across subgroups of students within countries.

### 1.9 Muddy waters: Non-participation in international assessments

One of the leading ILSAs in education is the PISA, which claims to put robust measures in place to ensure that the final sample from each participating nation is a true representation of its 15-year-old population. Jake Anders, Silvan Has, John Jerrim, Nikki Shure, and Laura Zieger provide a case study of one "educational superpower" country (Canada), discussing how various issues with the quality of its PISA 2015 data bring into question its status as one of the highest-performing educational systems worldwide. The authors point out how various biases can emerge in the PISA sample and show how the Canadian PISA data fail to meet some of the key quality criteria set by the Organization for Economic Co-operation and Development (OECD; such as vastly exceeding the number of permissible student exclusions). The authors thus conclude their paper by offering some constructive suggestions on how this element of the PISA study could be improved in the future. It should be noted that Canada is not the only country that has not met quality standards for samples; for example, the USA has not met the standards in any PISA cycle so far. Nevertheless, the data from Canada,

the USA and other countries with high non-response rates are being used repeatedly, cited and used for far-reaching recommendations ("superpower" and so forth). Even before the publication of this special issue (i.e., after the first publication online), there has been heated debate involving the authors, another stakeholder and the editors. We look forward to continuing such discussions to expand our knowledge of the integrity of the PISA and other international studies.

## 2 Concluding remarks

The expansion of ILSAs provides many advantages for monitoring and evaluating educational progress. Such assessments produce a great amount of data, which enable researchers to address many unanswered questions and thereby also contribute to innovative solutions for improving education. However, secondary analyses must be thoughtful, and some areas, such as measurements of inequality and resilience, suffer from their lack of conceptual clarity. The international comparability of the measures is another contentious issue. As discussed in the papers of this special issue, a new understanding of what comparability means is needed. The analysis of comparability implies that using partly new multivariate approaches for addressing assumptions about measurement invariance can reveal substantive differences in subdimensions. Studies that take into account complex patterns in the data are desirable to learn more about country-specific strengths and weaknesses. Dealing with measurement invariance and DIF calls for the development of flexible approaches to deal with non-comparable measures in international assessments. The alignment method has both advantages and disadvantages. Testing for equality among thousands of parameters is unrealistic. Sample comparability is also a serious issue. One of the great achievements of ILSAs is the development of rigorous quality standards; for example, both the International Association for the Evaluation of Educational Achievement (IEA) and the OECD have developed rigorous (and much needed) standards, and it is important not to sacrifice them.

To conclude, this special issue addresses a number of prominent and sensitive topics related to ILSAs in education. The presented papers encourage thought-provoking discussions and prompt future research to continue seeking new perspectives and elaborating on relevant methods to evaluate educational systems. We welcome follow-up papers to engage in scientific debates on the issues raised, and we hope that this discussion will spur research on the integrity of educational measures in ILSAs and bring positive changes to the world of education.

**Funding** Open Access funding enabled and organized by Projekt DEAL.

## References


Asparouhov, T., & Muthen, B. (2014). Multiple group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal, 21*, 495–508.

Bergold, S., Wendt, H., Kasper, D., & Steinmayr, R. (2017). Academic competencies: Their interrelatedness and gender differences at their high end. *Journal of Educational Psychology, 109*(3), 439–449. https://doi.org/10.1037/edu0000140.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory*. Mahwah: Psychology Press.

Goe, L. (2007). *The link between teacher quality and student outcomes: A research synthesis*. Washington, DC: National center for teacher quality.

Grek, S. (2009). Governing by numbers: The PISA 'effect' in Europe. *Journal of Education Policy, 24*(1), 23–37.

Hanushek, E. A., & Woessmann, L. (2011). The economics of international differences in educational achievement. In E. A. Hanushek, S. Machin, & L. Woessmann (Eds.), *Handbook of the economics of education* (Vol. 3, pp. 89–200). Amsterdam: Elsevier.

Hanushek, E. A., & Woessmann, L. (2012). Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. *Journal of Economic Growth, 17*(4), 267–321. https://doi.org/10.1007/s10887-012-9081-x.

Hanushek, E. A., & Woessmann, L. (2015). *The knowledge capital of nations: Education and the economics of growth. CESifo book series*. Cambridge: The MIT Press.

Hopfenbeck, T. N., Lenkeit, J., El Masri, Y., Cantrell, K., Ryan, J., & Baird, J.-A. (2017). Lessons learned from PISA: A systematic review of peer-reviewed articles on the Programme for international student assessment. *Scandinavian Journal of Educational Research, 62*(3), 333–353. https://doi.org/10.1080/00313831.2016.1258726.

Lenkeit, J., Chan, J., Hopfenbeck, T. N., & Baird, J.-A. (2015). A review of the representation of PIRLS related research in scientific journals. *Educational Research Review, 16*, 102–115. https://doi.org/10.1016/j.edurev.2015.10.002.

Lindblad, S., Pettersson, D., & Popkewitz, T. S. (2018). *Numbers, education and the making of society: International assessments and its expertise*. London: Routledge.

Meyer, H. D., & Benavot, A. O. (Eds.). (2013). *PISA, power, policy. The emergence of global educational governance*. Oxford: Oxford Studies in Comparative Education.

Meyer, H. D., Strietholt, R., & Epstein, D. Y. (2018). Three models of global education quality and the emerging democratic deficit in global education governance. In M. Akiba & G. K. LeTendre (Eds.), *Routledge international handbook of teacher quality and policy*. New York: Routledge.

Nilsen, T., & Gustafsson, J. E. (2016). *Teacher quality, instructional quality, and student outcomes: Evidence across countries, cohorts, and time*. Cham: Springer International Publishing.

Strietholt, R., & Rosén, M. (2016). Linking large-scale Reading assessments: Measuring international trends over 40 years. *Measurement: Interdisciplinary Research and Perspectives, 14*(1), 1–26. https://doi.org/10.1080/15366367.2015.1112711.

Strietholt, R., & Scherer, R. (2017). The contribution of international large-scale assessments to educational research: Combining individual and institutional data sources. *Scandinavian Journal of Educational Research*, 1–18. https://doi.org/10.1080/00313831.2016.1258729.

Wendt, H., & Kasper, D. (2016). Subject-specific strength and weaknesses of fourth-grade students in Europe: A comparative latent profile analysis of multidimensional proficiency patterns based on PIRLS/TIMSS combined 2011. *Large-scale Assessments in Education, 4*(14), 14. https://doi.org/10.1186/s40536-016-0026-2.