on_education
Journal for Research and Debate

challenges for the design of international assessments: sampling, measurement, and causality

Rolf Strietholt and Stefan Johansson

This paper examines critical design factors that influence data quality in educational research, using international large-scale educational assessments as an example. We will focus on statistical challenges related to sampling, measurement, and causality. While international assessments employ rigorous random sampling techniques, deviations such as exclusions and non-participation can introduce bias and affect representativeness. In terms of measurement, although these assessments excel in core domains, there is a growing call for broader assessment areas, such as environmental literacy and civic education. Additionally, concerns are emerging about the quality of context surveys. Causality remains a central concern, and despite the challenges posed by the cross-sectional design, combining data and applying sophisticated analytical methods can help address causal questions. Recognising the interconnectedness of sampling, measurement, and causality is essential for conducting robust research and informing evidence-based policies and practices.

Keywords: causality, educational measurement, international large-scale assessment, sampling, study design

Introduction

The concept of evidence-based policy and practice underscores the importance of relying on empirical data to guide and substantiate educational reforms instead of relying solely on ideological beliefs and advocacy. Evidence-based policy can only be effective when grounded in rigorous educational research, and such research requires high-quality data. The main purpose of the present paper is to discuss the quality of data using international large-scale assessments as an example.

In the field of educational research, the rise of international large-scale assessments such as PIRLS (Progress in International Reading Literacy Study), PISA (Programme for International Student Assessment), and TIMSS (Trends in International Mathematics and Science Study) generated extensive datasets. This data serves as a resource for educational research, and the findings of this research shape educational policy and practice (Hanushek & Woessmann, 2011; Johansson, 2016, 2020; Strietholt et al., 2019). In fact, the data produced by these studies gained strategic prominence in the national and international educational discourse, and they play a crucial role in guiding policy decisions within and across many countries (Baker & LeTendre, 2005; Grek, 2009, 2013; Ozga, 2012).

However, the utility of these data in education research that aims to generate meaningful evidence for policy and practice hinges on the quality of the data itself. In this paper, we delve into the

quality of data through the lens of three key issues related to study design: sampling, measurement, and causality. We discuss the data generated by international assessments based on these three aspects and examine their strengths and weaknesses. Robust empirical evidence can only be achieved when all three aspects converge. Even if a single component is inadequately addressed, the implications are substantial.

Sampling

Sampling is an important component of any rigorous study design. When a sample is representative, the knowledge acquired based on that sample can be generalised to a larger population. Conversely, results obtained from an unrepresentative sample hold limited value for generalisations. The gold standard for achieving this objective is the utilisation of random samples. Inference statistics rely on random samples to extrapolate research findings to the underlying populations.

The problems associated with non-randomly selected samples have been known for a long time. For example, the renowned statistician John Tukey is said to have responded to the Kinsey-Report published in 1948 by stating, "a random selection of three people would have been better than a group of 300 chosen by Mr. Kinsey." (Leonhardt, 2000, A19). Kinsey's research aimed to investigate the sexual behaviour of Americans, but for many, some of the findings appeared implausible. One potential explanation for this was a substantial bias in the sample, as a significant proportion of men included in the study were already incarcerated and had engaged in homosexual relationships. The crux of the issue with Kinsey's sample lay in the utilisation of a snowball sampling method.

A recent example illustrating the ongoing problems of non-representative samples is the so-called replication crisis, in which the results of many scientific studies are difficult or impossible to reproduce. Researchers conducted replications of roughly 100 studies in psychological science and found that only slightly more than one-third of the studies could be replicated (Open Science Collaboration, 2015). Unrepresentative samples in combination with publication bias (i.e., a tendency to publish significant results) are plausible explanations for the difficulty of replicating studies. Furthermore, Hedge and Nowell (1995, p. 41) point out that "[r]eviews and meta-analysis of data from nonrepresentative samples are not necessarily any more representative than the studies they are based on".

International assessments, on the contrary, employ sophisticated random sampling designs, which involve multi-stage plans where schools are initially randomly selected, followed by a random selection of students or classes. Furthermore, the samples include several thousand students per country, and due to these large sample sizes, very accurate estimates are possible. From a sampling perspective, the data from international assessments are arguably of very high quality. This claim is especially true when compared to many other studies in the field of educational research.

While we believe international assessments are of high quality from a sampling perspective, we also want to discuss two areas of concern. First, sophisticated sampling designs only guarantee quality when adhered to without deviation. For example, the PISA quality standards allow for the exclusion of up to 5% of students. These students can be those, for instance, who do not speak the test language or students with special educational needs. However, Sweden excluded more than 10% of the student samples from the PISA assessment in 2018, and this has faced heavy criticism (Andersson & Sandgren Massih, 2023).

The exclusion of students may not necessarily jeopardise the representativeness of a sample if they are at random. However, studies suggest that the exclusion or non-response of students (or schools) in PISA is not random but rather correlates with students' social backgrounds and their performance (e.g., Anders et al., 2019; Durrant & Schnepf, 2012; Micklewright et al., 2012; Jerrim, 2021).

Another challenge in the field of sampling is defining meaningful target populations. An illustrative case is evident within the context of the PISA, where exclusively those 15-year-olds who remain enrolled within the formal school system are subjected to evaluation. While the majority, if not all, of the 15-year-old cohort, continues their educational pursuits in numerous countries of the Global North, such a scenario is not universally applicable. For instance, in Brazil, a mere 65% of 15-year-olds are still in the school system, as opposed to 80% in Argentina and 90% in Chile (OECD, 2019a). This underscores the fundamental question regarding the appropriateness of designating 15-year-olds within the educational system as the target population in question.

Lastly, it should be noted that students with special needs are often excluded from international assessments. Interestingly, significant variations exist in this regard among different countries, and the reasons for these differences are currently not well understood. Efforts are underway to make assessments more inclusive; however, in many cases, international assessments still rely on specific conceptions of 'typical' children.

Measurement

It is meaningful to distinguish between what should be conceptually measured (or not) and how it is operationalised. The extent to which measures capture what they are intended to measure is also reflected in the concepts of construct underrepresentation and construct irrelevant variance (Messick, 1989).

Central to international assessments are the achievement tests targeting students. The major studies prioritise three core domains–mathematics, science, and reading–while neglecting other important domains such as civic education or computer and information literacy (Gladushyna & Strietholt, 2023). Recognising that education encompasses a broader set of skills and knowledge is important. In times of environmental crisis, rapid technological advancements, and the rise of right-wing populism, environmental literacy, information and computer literacy, as well as civic education are gaining importance and are also essential in preparing students for active citizenship. Therefore, while the core domains remain significant, broadening the scope of international assessments is crucial.

International assessments have faced criticism for setting an agenda and potentially driving the global convergence of national curricula (e.g., Meyer et al., 2019). On this matter, we maintain a nuanced perspective. On the one hand, we concur that the emphasis on reading, mathematics, and natural sciences represents a narrowing of our broader understanding of education. Nevertheless, this does not imply that these three domains lack importance. Furthermore, Benavot et al. (1991) demonstrated that the range of subjects taught in primary and secondary schools has been virtually identical worldwide since the early 1900s, predating the establishment of the first international assessments. Similarly, Johansson and Strietholt (2019) did not find evidence for the global harmonisation of student achievement in mathematics.

Measurement is not only about what but also how the constructs of interest are measured. Complex

constructs like mathematics performance cannot be assessed using a single or a few items but require a large number of test items from various content areas (algebra, geometry, etc.). The tests used in international assessments are carefully developed and evaluated (Wagemaker, 2020). This involves various steps, including drafting comprehensive assessment frameworks, developing hundreds of test items to comprehensively cover the construct of interest, and meticulous translation into different test languages. The achievement tests encompass various item formats, including both multiple-choice and open-ended questions. Completing all these items would take several days. Therefore, each student works on a randomly distributed subset of items that can be completed within approximately one to two hours. Subsequently, all test items are linked to an achievement scale to measure the abilities of all students on a standard performance scale.

We posit that the achievement test employed in international assessments is at the forefront of contemporary measurement methodologies. Nevertheless, it is imperative to recognise that challenges endure in the interpretation of these test results. One of these challenges arises from the low-stakes nature of the tests for students, potentially giving rise to the hypothesis that some students may lack the motivation to perform at their best. Interestingly, evidence indicates significant international variability in student motivation within PISA (Eklöf & Hopfenbeck, 2019; Wise & DeMars, 2010). These findings imply that the level of students' seriousness may differ among countries, potentially threatening the comparability of the data.

International assessments not only collect achievement data but also gather supplementary information from students and schools about family backgrounds and teaching and learning contexts (Strietholt & Scherer, 2018). Compared to the achievement tests, the background questionnaires receive less emphasis. The survey time for students is typically 20 to 30 minutes, and constructs related to social background, socio-emotional skills or teaching quality are assessed with only a few items (Engzell, 2019; Strietholt & Strello, 2022). Complex constructs such as well-being or growth mindset are sometimes measured with just a single item (e.g., OECD, 2019b).

The quality of background questionnaires has been the subject of extensive academic discussion (Hooper, 2022; Rutkowski & Rutkowski, 2010). We contend that the contextual factors that influence student achievement, as assessed through background questionnaires, do presently not receive commensurate attention compared to achievement tests. It is pertinent to note that this has not been a historical constant. For instance, the first TIMSS study conducted in 1995 represents the comprehensive evaluation of learning conditions compared to all following international assessments. To achieve a more profound comprehension of the variability in student performance, it is essential to scrutinise the relationships with contextual variables. Consequently, we advocate for a heightened emphasis on measuring these variables in future research endeavours.

Causality

Causal questions are central in educational research: Why do some students, classes, schools, or countries outperform others? The question of the determinants of student achievement is pertinent when considering the huge variation in student performance both between and within countries (Strietholt et al., 2014a, 2014b).

It is extremely challenging to draw causal inferences based on data from international assessments such as PIRLS, PISA, and TIMSS. The primary issue is that these studies have a cross-sectional design, making it difficult to distinguish between correlations and causation. For example, Mullis et

al. (2012) argue that an early start in pre-primary education is crucial for developing children's reading achievement, as evidenced by the higher performance of students in PIRLS at the end of primary school who have had more pre-primary education experience. However, an equally plausible interpretation is the presence of selection effects, where more privileged children may start preschool earlier than those from disadvantaged backgrounds. Another plausible explanation is that wealthier countries have more advanced early childhood education systems. It is crucial to remember that correlation does not imply causation.

However, it is possible to combine data from international school performance studies to address well-known issues in cross-sectional analyses. Just recently, researchers Joshua Angrist and Guido Imbens were awarded the Nobel Memorial Prize in Economic Sciences for their methodological contributions to the analysis of causal relationships. Strietholt et al. (2020) applied such an approach to the data from international assessments to investigate whether increases in national-level preschool enrollment rates lead to improvements in student achievement. For this purpose, they combine data from different years to examine the relationship between preschool attendance and academic performance at the national level and over time. This approach incorporates a longitudinal component at the national level, offering similar methodological advantages to those found in longitudinal studies where individuals are repeatedly observed: the ability to control for prior achievement.

There are further examples of how data from cross-sectional international assessment studies can be combined or supplemented with external data to strengthen designs that make causal inference more credible, such as difference-in-differences designs, instrument variable regression, regression discontinuity analyses, and matching approaches (e.g., Cordero et al., 2018; Hogrebe & Strietholt, 2016; Kennedy & Strietholt, 2023; Schlotter et al., 2011; Steinmann & Olsen, 2022; Strello et al., 2021)

Beyond creatively utilising and combining existing cross-sectional data, there are ongoing initiatives aimed at altering the designs of these studies themselves. For instance, amidst the COVID-19 pandemic, adaptations were made to the teacher survey component of ICILS, resulting in the same teachers being surveyed in 2020 as they were in 2018 (Strietholt et al., 2021). Similarly, some countries have extended the current TIMSS cycle, involving the re-testing of the same students in 2024 following the regular assessment in 2023. In both instances, these international assessments incorporate a longitudinal component at the individual level.

Conclusion

In conclusion, international large-scale assessments in education have become instrumental in shaping educational policies and practices, emphasising the significance of evidence-based decision-making over ideological perspectives. This paper has explored the critical aspects of data quality in international assessments, focusing on sampling, measurement, and causality.

Sampling plays a pivotal role in the credibility of research findings. International assessments excel in employing sophisticated random sampling techniques, providing high-quality data compared to many other educational and psychological studies. Nevertheless, it is crucial to acknowledge the potential deviations from these designs, as the exclusion of students may introduce bias and challenge representativeness, further emphasising the need for vigilance in adherence to sampling standards.

Measurement is another pivotal aspect, especially concerning achievement tests. While international assessments excel in assessing core domains, there is a growing call to expand the assessment scope to include essential areas such as environmental literacy and civic education. The construction and evaluation of achievement tests in these assessments are commendable. Still, the low-stakes nature of the tests can affect student motivation, leading to variations in seriousness among countries.

Causality remains the central concern in educational research, attempting to explain why some entities outperform others and which policies and practices support student learning. It is indeed the area where there is the most pressure for action: Educators, policy-makers, and other education stakeholders must formulate clear causal questions, and researchers must employ rigorous analytical strategies that allow for robust causal inferences. While the cross-sectional design of international assessments poses challenges for making causal inferences, there are opportunities to utilise their data or integrate it with other sources to employ sophisticated analytical approaches that offer substantial evidence for addressing causal questions.

In large-scale educational research, it is essential to acknowledge the interconnectedness of sampling, measurement, and causality. Neglecting any one of these components diminishes the overall value of a study. This principle holds true not only for international assessments but for all research endeavours aiming to make high-level inferences and generalisable conclusions (e.g., Ercikan & Roth, 2006; Gustafsson, 2008). A comprehensive approach that simultaneously addresses sampling, measurement, and causality is paramount. This helps ensure that research outcomes possess the potency to drive positive transformations in education on a global scale.

Acknowledgment

This work was financially supported by Riksbankens Jubileumsfond under grant number P20-0095.

References

Anders, J., Has, S., Jerrim, J., Shure, N., & Zieger, L. (2020). Is Canada really an education superpower? The impact of non-participation on results from PISA 2015. *Educational Assessment, Evaluation and Accountability*, *33*(1), 229–249. https://doi.org/10.1007/s11092-020-09329-5

Andersson, C., & Sandgren Massih, S. (2023). PISA 2018: Did Sweden exclude students according to the rules? *Assessment in Education: Principles, Policy & Practice*, *30*(1), 33–52. https://doi.org/10.1080/0969594X.2023.2189566

Baker, D., & LeTendre, G. (2005). *National differences, global similarities: World culture and the future of schooling*. Stanford University Press.

Benavot, A., Cha, Y., Kamens, D. H., Meyer, J. W., & Wong, S. (1991). Knowledge for the masses: World models and national curricula, 1920-1986. *American Sociological Review*, *56*(1), 85. https://doi.org/10.2307/2095675

Cordero, J. M., Cristóbal, V., & Santín, D. (2018). Causal inference on education policies: A survey of empirical studies using PISA, TIMSS and PIRLS. *Journal of Economic Surveys*, *32*(3), 878–915.

https://doi.org/10.1111/joes.12217

Durrant, G. B., & Schnepf, S. V. (2017). Which schools and pupils respond to educational achievement surveys? A focus on the English Programme for International Student Assessment sample. *Journal of the Royal Statistical Society*, *181*(4), 1057–1075.
https://doi.org/10.1111/rssa.12337

Eklöf, H., & Hopfenbeck, T. (2019). Self-reported effort and motivation in the PISA test. In B. Maddox (Ed.), *International large-scale assessments in education: Insider research perspectives* (pp. 121–136). Bloomsbury Academic.

Engzell, P. (2019). What do books in the home proxy for? A cautionary tale. *Sociological Methods & Research, 50*(4), 1487–1514.
https://doi.org/10.1177/0049124119826143

Ercikan, K., & Roth, W.-M. (2006). What good is polarizing research into qualitative and quantitative? *Educational Researcher*, *35*(5), 14–23.

Gladushyna, O., & Strietholt, R. (2023). Measuring education: Do we need a plethora of assessment studies or just a single score? *International Journal of Educational Research Open*, *5*.
https://doi.org/10.1016/j.ijedro.2023.100281

Grek, S. (2009). Governing by numbers: The PISA 'effect' in Europe. *Journal of Education Policy*, *24*(1), 23–37.
https://doi.org/10.1080/02680930802412669

Grek, S. (2013). Expert moves: International comparative testing and the rise of expertocracy. *Journal of Education Policy*, *28*(5), 695–709.
https://doi.org/10.1080/02680939.2012.758825

Gustafsson, J.-E. (2008). Effects of international comparative studies on educational quality on the quality of educational research. *European Educational Research Journal*, *7*(1), 1–17.
https://doi.org/10.2304/eerj.2008.7.1.1

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

Hedges, L. V., & Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science, 269*(5220), 41–45.
https://doi.org/10.1126/science.7604277

Hogrebe, N., & Strietholt, R. (2016). Does non-participation in preschool affect children's reading achievement? International evidence from propensity score analyses. *Large-scale Assessments in Education*, *4*(1), 1–22.
https://doi.org/10.1186/s40536-016-0017-3

Hooper, M. (2022). Dilemmas in developing context questionnaires for international large-scale assessments. In T. Nilsen, A. Stancel-Piątak, & J.-E. Gustafsson (Eds.), *International handbook of*

comparative large-scale studies in education: Perspectives, methods and findings (pp. 721–747). Springer International Publishing.
https://doi.org/10.1007/978-3-030-38298-8_29-1

Johansson, S. (2016). International large-scale assessments: What uses, what consequences? *Educational Research*, *58*(2), 139–148.
https://doi.org/10.1080/00131881.2016.1165559

Johansson, S. (2020). Analysing the (mis)use and consequences of international large-scale assessments. In J. Zajda (Ed.), *Globalisation, ideology and education reforms: Emerging paradigms* (pp. 13–24). Springer.
https://doi.org/10.1007/978-94-024-1743-2_2

Johansson, S., & Strietholt, R. (2019). Globalised student achievement? A longitudinal and cross-country analysis of convergence in mathematics performance. *Comparative Education*, *55*(4), 536–556.
https://doi.org/10.1080/03050068.2019.1657711

Kennedy, A. I., & Strietholt, R. (2023). School closure policies and student reading achievement: Evidence across countries. *Educational Assessment, Evaluation and Accountability, 35.*
https://doi.org/10.1007/s11092-023-09415-4

Leonhardt, D. (2000, July 28). John Tukey, 85, Statistician; Coined the Word 'Software'. *The New York Times*.
https://www.nytimes.com/2000/07/28/us/john-tukey-85-statistician-coined-the-word-software.html

Meyer, H. D., & Benavot, A. O. (Eds.). (2013). *PISA, power, policy. The emergence of global educational governance.* Symposium Books.

Meyer, H. D., Strietholt, R., & Epstein, D. Y. (2017). Three models of global education quality: The emerging democratic deficit in global education governance. In M. Akiba, & G. K. LeTendre (Eds.), *International handbook of teacher quality and policy* (pp. 132–150). Routledge.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5–11.
https://doi.org/10.2307/1175249

Micklewright, J., Schnepf, S. V., & Skinner, C. (2012). Non-response biases in surveys of schoolchildren: The case of the English Programme for International Student Assessment (PISA) samples. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *175*(4), 915–938.
https://www.jstor.org/stable/23355308

Mullis, I. V. S., Martin, M. O., Foy, P., & Drucker, K. T. (2012). *PIRLS 2011 international results in reading.* International Association for the Evaluation of Educational Achievement.

OECD (2019a), *PISA 2018 Results (Volume I): What students know and can do*, OECD Publishing.
https://doi.org/10.1787/5f07c754-en.

OECD (2019b), *PISA 2018 Results (Volume III): What school life means for students' lives*, OECD

Publishing.
https://doi.org/10.1787/acd78851-en

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251).
https://doi.org/10.1126/science.aac4716

Ozga, J. (2012). Governing knowledge: Data, inspection and education policy in Europe. *Globalisation, Societies and Education*, *10*(4), 439-455.
https://doi.org/10.1080/14767724.2012.735148

Jerrim, J. (2021). PISA 2018 in England, Northern Ireland, Scotland and Wales: Is the data really representative of all four corners of the UK?, *Review of Education*, *9*(3).
https://doi.org/10.1002/rev3.3270

Rutkowski, L., & Rutkowski, D. (2010). Getting it 'better': The importance of improving background questionnaires in international large-scale assessment. *Journal of Curriculum Studies*, *42*(3), 411–430.
https://doi.org/10.1080/00220272.2010.487546

Schlotter, M., Schwerdt, G., & Woessmann, L. (2011). Econometric methods for causal evaluation of education policies and practices: A non-technical guide. *Education Economics*, *19*(2), 109–137.
https://doi.org/10.1080/09645292.2010.511821

Steinmann, I., & Olsen, R. V. (2022). Equal opportunities for all? Analyzing within-country variation in school effectiveness. *Large-Scale Assessments in Education*, *10*(1), 1–34.
https://doi.org/10.1186/s40536-022-00120-0

Strello, A., Strietholt, R., Steinmann, I., & Siepmann, C. (2021). Early tracking and different types of inequalities in achievement: Difference-in-differences evidence from 20 years of large-scale assessments. *Educational Assessment, Evaluation and Accountability*, *33*, 139–167.
https://doi.org/10.1007/s11092-020-09346-4

Strietholt, R., Bos, W., Gustafsson, J. E., & Rosén, M. (Eds.). (2014a). *Educational policy evaluation through international comparative assessments*. Waxmann Verlag.

Strietholt, R., Gustafsson, J.-E., Rosén, M., & Bos, W. (2014b). Outcomes and causal inference in international comparative assessments. In R. Strietholt, W. Bos, J.-E. Gustafsson, & M. Rosen (Eds.), *Educational policy evaluation through international comparative assessments* (pp. 9–18). Waxmann Verlag.

Strietholt, R., Fraillon, J., Liaw, Y. L., Meinck, S., & Wild, J. (2021). Changes in digital learning during a pandemic–Findings from the ICILS Teacher Panel. *International Association for the Evaluation of Educational Achievement.*

Strietholt, R., Gustafsson, J.-E., Hogrebe, N., Rolfe, V., Rosén, M., Steinmann, I., & Hansen, K. Y. (2019). The Impact of Education Policies on Socioeconomic Inequality in Student Achievement: A Review of Comparative Studies. In L. Volante, S. V. Schnepf, J. Jerrim, & D. A. Klinger (Eds.), *Socioeconomic Inequality and Student Outcomes Cross-National Trends, Policies, and Practices* (pp.

17–38). Springer Singapore.
https://doi.org/10.1007/978-981-13-9863-6_2

Strietholt, R., Hogrebe, N., & Zachrisson, H. D. (2020). Do increases in national-level preschool enrollment increase student achievement? Evidence from international assessments. *International Journal of Educational Development*, *79*.
https://doi.org/10.1016/j.ijedudev.2020.102287

Strietholt, R., & Scherer, R. (2018). The contribution of international large-scale assessments to educational research: Combining individual and institutional data sources. *Scandinavian Journal of Educational Research*, *62*(3), 368–385.
https://doi.org/10.1080/00313831.2016.1258729

Strietholt, R., & Strello, A. (2022). Socioeconomic inequality in achievement. In T. Nilsen, A. Stancel-Piątak, & J.-E. Gustafsson (Eds.), *International handbook of comparative large-scale studies in education* (pp. 201–220). Springer.
https://doi.org/10.1007/978-3-030-88178-8_11

Wagemaker, H. (2020). *Reliability and validity of international large-scale assessment: Understanding IEA's comparative studies of student achievement*. Springer Nature.
https://doi.org/10.1007/978-3-030-53081-5

Wise, S. L., & DeMars, C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment*, *15*(1), 27–41.
https://doi.org/10.1080/10627191003673216

Recommended Citation

Strietholt, R., & Johannson, S. (2023). Challenges for the design of international assessments: sampling, measurement, and causality. *On Education. Journal for Research and Debate*, *6*(18).
https://doi.org/10.17899/on_ed.2023.18.2

Do you want to comment on this article? Please send your reply to editors@oneducation.net. Replies will be processed like invited contributions. This means they will be assessed according to standard criteria of quality, relevance, and civility. Please make sure to follow editorial policies and formatting guidelines.

Rolf Strietholt

Rolf Strietholt is a researcher in the field of comparative education focusing on educational measurement and educational effectiveness research.

Stefan Johansson

Stefan Johansson is associate professor in education at the University of Gothenburg. His research interests center on the uses and consequences of international large-scale

assessments.