PISA and mathematics education: What can we learn – and what we cannot

Jenni Ingram

Department of Education, University of Oxford, UK

International large-scale assessments (ILSAs) such as PISA play a powerful role in shaping debates about mathematics education, yet their findings are often used to support claims that extend beyond what their designs can warrant. Responsible use of ILSA data requires alignment between the constructs measured, the study design, the questions asked, and the conclusions drawn. In this paper, I consider three areas in which these issues arise: the policy and public narratives surrounding international comparisons; the methodological and measurement characteristics of ILSAs; and the use of ILSAs to address substantive questions in mathematics education research. The evolving design of ILSAs offers many opportunities to examine important issues within mathematics education, whilst also creating new questions about construct, design and contextual alignment.

Keywords: International large-scale assessments, policy borrowing, comparative education, methodology, validity.

Introduction

International large-scale assessments (ILSAs) such as the OECD's Programme for International Student Assessment (PISA) play a powerful role in shaping debates within mathematics education. However, misunderstandings of the epistemic scope of PISA can distort how the findings are used by politicians, the media and by ourselves as mathematics education researchers. ILSAs have an important role within mathematics education assessment research (Giberti & Maffia, 2020), but we need to be clearer about the methodological affordances and constraints of these large-scale assessments to enable both researchers and other stakeholders to use the findings responsibly. International assessments do not speak for themselves; their meaning is constructed statistically, politically and publicly.

In this paper, I argue that responsible research using ILSA data requires careful alignment between the epistemic affordances of the ILSA design, the questions we ask of it, and the claims being made. These ILSAs still hold a great deal of untapped potential, but they need to be treated responsibly if they are to contribute to our understanding of mathematics education. Responsible use of these data involves considering three dimensions: construct alignment, design alignment, and contextual alignment. That is, does the claim correspond to what was actually measured? Does the study design support the inferences made? And are associations being interpreted at the appropriate level and within their educational context?

To do this, I consider three different but interdependent domains of research that draw on ILSA data. Firstly, research that considers ILSAs in terms of the policy and public narratives that surround them, and the impact on curricula, policy and practice. Secondly, research that considers the more methodological and measurement aspects of ILSAs. Finally, research that uses data from

ILSAs to address wider questions arising from existing research within mathematics education. I will largely focus on PISA, though many of the arguments apply similarly to other ILSAs and international large-scale surveys.

PISA includes an assessment of mathematical literacy using an age-based sample of students aged 15-years-old. It is a cross-sectional study historically conducted every three years (usually). This assessment is accompanied by contextual questionnaires completed by students, schools and, in some countries, parents. These questionnaires cover general topics, including details about students' backgrounds, and also include items specific to students' experiences when learning mathematics. The design and sampling involved in PISA and other ILSAs is complex, with any analysis needing to make use of both sample weights and plausible values. Sample weights account for the differences in students' probabilities of being selected and participating in the study that allow the estimates from the sample to represent the intended population. Plausible values are used to repeatedly estimate students' achievement as each student only completes some of the assessment items; they are designed to support estimates and comparisons at population and group levels (e.g., differences between genders), though they do not enable claims to be made about individual students.

ILSAs as policy and public narratives

PISA in particular has attracted a great deal of public attention, yet policy and public narratives are shaped by political narratives and agendas as much as by the results themselves (Jerrim, 2024). Results that are not reported in the headlines on the day the international report is published disappear from the debate. The rankings that often accompany these studies are a powerful simplification of the results that can be used to celebrate climbing the ranks or be weaponised to justify reform. Well-known examples include the "PISA shock" in Germany, the crisis of "relaxed education" in Japan, the Nordic Awakening, or descriptions such as being the "Runt of the Litter". However, ranks are not necessarily indicators of statistically significant differences – there may be no statistically significant differences between the performance of students in the country ranked 8th for mathematics and the performance of students in the country ranked 14th. Rankings illustrate a simple form of misalignment between the data and the claims made, where crude ordering of average scores in countries is interpreted as evidence of meaningful differences between education systems, even where the assessment does not support that distinction (also an issue for the different league tables in which universities are ranked).

This focus on rankings has also contributed to the emergence of what have been termed "reference societies", such as Finland, Singapore, Shanghai, etc., leading to policy tourism, borrowing, and narratives that justify or support a reform agenda (You & Morris, 2015). Yet policy borrowing and referencing is often selective, with the identified differences in outcomes not necessarily establishing which features produce these differences (Cantley, 2019). Such referencing does not necessarily involve borrowing actual practices but rather borrowing narratives around policies (Jerrim, 2024; Volante & Klinger, 2023), as illustrated in England's focus on Shanghai mathematics, which contributed more to dialogue about mastery than changes in practice (Probert, 2021). The inference that characteristics of practice in higher-performing countries can be used to

explain the high performance of their students and that this practice should therefore be transferred to other contexts raises several questions of alignment. Does the measurement of that practice in these ILSAs accurately reflect the aspects of that practice in classrooms? Can an association between that practice and performance in mathematics in one country be meaningfully transferred to another?

One consequence of these ILSAs has arguably been a homogenisation of mathematics curricula around the world (Johansson & Hansen, 2019), PISA in particular is explicitly normative (Sjøberg & Jenkins, 2022). Kadijevich et al. (2023) illustrate the different influences of PISA and TIMSS, with TIMSS encouraging convergence around content and skills while PISA encourages ideas around competency and mathematical literacy. Construct alignment therefore matters before any statistical analysis can begin. The issue is not simply that these two ILSAs contain different items; their frameworks constrain what can become visible as "successful" mathematics education. Systems may appear relatively stronger or weaker depending on which conception of mathematics is providing the basis for the comparison. These studies are not neutral about what mathematics is, what mathematics education is for, or what should get attention in national and international debates.

These narratives arise because ILSA define what outcomes are valued in ways that mean that education systems can be compared on these outcomes. Higher-performing systems then become treated as having authority and expertise in what high-quality mathematics education is, and become potential sources of policy or practice borrowing. These narratives around which outcomes are valued, combined with this policy borrowing, then lead to greater homogenisation in the drive to improve rankings within these assessments - but is improving rankings the goal or improving mathematics education?

ILSAs as measurement

The political authority of ILSAs depends partly upon their presentation as measurements, yet what is measured is itself the result of substantive decisions informed by existing research, organisational priorities and agendas, as well as current public narratives. These substantive decisions include what mathematics is assessed, what contextual variables are important and how they should be measured, and how achievement should be statistically constructed. In this section I briefly examine each of these to highlight the importance of construct alignment in handling ILSA data responsibly.

There are subtle differences in the nature of mathematics that different ILSAs are assessing. PISA, for example, focuses on mathematical literacy and "preparing students to use mathematics in every aspect of their personal, civic and professional lives, as part of their constructive, engaged, and reflective 21st Century citizenship" (OECD, 2023, p. 52). This emphasis on the utilitarian value of mathematics to society and individuals is consistent with the broader OECD focus on education as a driver of economic growth (Beccuti, 2024). The nature of mathematics is one of utility; it is a tool to be used, which is also reflected in the emphasis on mathematical modelling. In contrast, TIMSS focuses on what students have learned from the school curriculum and mathematics is treated as disciplinary knowledge. Consequently, apparently similar claims about mathematics achievement

do not necessarily refer to equivalent or comparable constructs across ILSAs as they operationalise mathematics differently.

Alongside mathematics, these studies also include items designed to measure latent constructs such as a student's sense of belonging in school or their growth mindset. These measures are often quite limited, with few items used for each construct in order to minimise the time students need to spend completing the questionnaires. They are also restricted to constructs that can be meaningfully measured across a range of country contexts. Cultural and social differences in the interpretation of questions can lead to differences in results that may not reflect differences in the prevalence (or not) of the construct itself. For example, Min et al. (2016) show how culturally patterned responses may have contributed to differences in mathematics self-confidence reported across countries, which complicates the interpretation of comparisons involving this construct. This creates both a construct and contextual alignment issue which requires us to consider the different operationalisations and interpretations before any comparisons are made.

Working with these large-scale studies as researchers is challenging. They are designed to support inferences about populations from samples, and it is therefore essential that we consider what population that particular sample is representative of (Atasever et al., 2025; Micklewright et al., 2012). This is key to design alignment. This sampling means that weights are used, yet there is debate about the extent to which these weights matter – ignoring them often produces biased estimates but does not often change conclusions (Jerrim et al., 2017). Historically there have been many published articles that have essentially ignored the sample design, yet whether this necessarily means that their conclusions should be ignored is debated (Atasever et al., 2025; Jerrim et al., 2017). Design alignment is not simply a matter of using established statistical methodologies that include these weights; it requires us to be clear about the population to which we are intending to generalise and whether the sampling design can support this.

Achievement is estimated using plausible values to generate population-level scores because students answer only a fraction of the items available in the assessment, and this estimation also considers background characteristics such as gender and socioeconomic background. Achievement in mathematics as measured in these assessments is therefore a population-level latent estimate, not observed individual scores on the assessments. Responsible analysis requires us to match the levels of our claims to the levels at which the assessment was designed to provide reliable data.

Over time, these assessments have also changed the methodology of the assessments themselves, with updated frameworks and more complex assessment designs. This includes shifts from paper-based assessments to computer-based assessments, and to the use of an adaptive design where students are given different items depending upon their performance in previous items. Key constructs are also not fixed over time; items change, frameworks evolve, domains enter and leave, and latent constructs are redefined. Many of these evolutions challenge what can and what should be compared over time.

Finally, the sheer number of variables and constructs that these ILSAs contain also pose a risk of data mining, particularly in an age of machine learning and AI tools which can analyse these large

datasets. The sophistication of these methods is increasing faster than our discussions about what should and can be researched.

Large-scale datasets require specialised methodological care. The closer we move from system-level patterns about which countries perform highly, where inequalities exist and what has changed over time, towards claims about teaching and learning, and what teachers and schools should do, the more careful we need to be. ILSA designs are less able to support the inferences made as we move from system-level foci to teacher- or student-level foci. This goes beyond issues of attempting to make explicitly causal statistical claims. Similar issues arise when descriptive international differences are translated into explanations of effective teaching. For example, several policies have focused on advocating for specific pedagogical practices borrowed from high-performing countries, yet Bokhove (2022) showed that there was no consistent pattern of instructional practices both within East Asian countries and between East Asian and Western countries using TIMSS 2019 data. ILSAs can be used to identify systematic relationships between the measured features of instructional quality and mathematics achievement (e.g., Bellens et al., 2019; Liu et al., 2024; Nilsen et al., 2026), but the meaning, stability and nature of these relationships require considerable care. The international nature of these assessments enables us to ask questions about whether constructs matter similarly across countries and student profiles, yet the measurements included may not be invariant across countries, time, modes or even groups of students. Issues of construct and design alignment can be intertwined where we need to establish what is being measured over time given changes to the study designs.

ILSAs can also tell us only so much. One key area that ILSAs have particularly highlighted is educational inequalities, particularly those related to students' gender, socioeconomic or migration background. These studies make educational inequalities visible, but do not naturally answer the questions about how to reduce them. They can often allow us to answer questions about what constructs can explain some of the variances associated with these demographic variables. For example, the construct of Opportunity to Learn (OTL) has been widely researched since its original identification as a potential explanation for differences in performance associated with socioeconomic background (Schmidt & McKnight, 1995) and, more recently, race (Schmidt et al., 2024). OTL provides one possible mechanism connecting socioeconomic background to performance in mathematics, supporting arguments that students' opportunities to learn are more limited if they have lower socioeconomic status, which consequently impacts their mathematics attainment, but also that their mathematics attainment can also impact their subsequent opportunities to learn (Ingram, 2026). Yet these associations are context-specific, existing in some educational contexts (predominantly Western contexts) but not others. A cross-sectional association between OTL and attainment in mathematics can be used to support several substantially different arguments (Hansen & Strietholt, 2018).

ILSAs as a research opportunity

In this final section, I consider the potential questions that ILSAs are particularly suited to answering, with a focus on emerging opportunities from the more recent methodological innovations and refinements.

The first opportunity is a focus on population distributions. Although attention is frequently on means and average performance, the large samples enable us to also look at distributions. In an age of evidence-based policies and practices, these variations are increasingly important to understand – what works for the 'average' student, or for students in one country context where a study took place, may not work or may even further disadvantage other students. ILSAs can enable us to examine the heterogeneity across education systems and challenge some of these universal 'what works' claims.

Over time, these studies have shown how education systems are remarkably stable in terms of average performance despite a context of curriculum reform, policy change, and shifting government priorities. There is substantial persistence in system-level patterns in relation to the relative performance of students in different education systems, inequalities, and relationships between performance and some contextual variables. These assessments reveal how difficult education change actually is. Furthermore, curriculum reform and policy change do not happen in isolation, so attributing change to particular reforms is often challenging (Nilsen et al., 2026). Existing research has highlighted the inconsistencies across different predictors over time and cultural contexts (Wang et al., 2023), yet this tells us something about the contextual dependence of these relationships. Instead of questions about what predicts mathematics achievement, questions need to be asked around the educational contexts in which associations appear, disappear or even reverse. Contextual alignment therefore includes considering the scope of a generalisation itself as an empirical question. In which education systems and for which groups of students under which educational conditions are these contextual constructs associated with mathematics achievement?

The most recent PISA cycle also offers new potential research directions, as the assessment measures not only outcomes but also behaviour during the assessment. This includes response times, item navigation, how often students revisit items, what they skip or omit, particular sequences of actions and potentially problem-solving pathways. The improvements in technologies and methodologies allow us as researchers to identify patterns in millions of interactions, perhaps enabling us to go beyond students' responses to items to aspects of their reasoning processes. However, more data does not necessarily mean better explanations. It will be interesting to see whether this richer process data expands the epistemic scope of ILSAs, or simply allows us to identify more complex associations. Questions that can potentially be addressed relate to all three dimensions: construct alignment – what does response time actually measure? Design alignment – does observing a sequence of behaviours within or across items allow us to infer a reasoning process? And contextual alignment – are comparable digital behaviours comparable across students and cultures?

Care is also needed around the questions we ask of ILSAs that they were never designed to answer. These studies can help us to identify patterns, not causes (Rutkowski et al., 2024). Yet research responsibilities go beyond avoiding causal claims; we also need to ensure that different, sometimes competing, interpretations or explanations of associations are visible. Responsible use is not about a more cautious or conservative use of ILSAs. It is about aligning the questions we ask, the constructs we use, the ILSA design with the claims and conclusions we make.

References

Atasever, U., Huang, F. L., & Rutkowski, L. (2025). Reassessing weights in large-scale assessments and multilevel models. *Large-scale Assessments in Education*, *13*(1), Article 9. https://doi.org/10.1186/s40536-025-00245-y

Beccuti, F. (2024). Meaning and subjectivity in the PISA mathematics frameworks: a sociological approach. *Educational Studies in Mathematics*, *116*(1), 49–65. https://doi.org/10.1007/s10649-023-10296-z

Bellens, K., Van Damme, J., Van Den Noortgate, W., Wendt, H., & Nilsen, T. (2019). Instructional quality: catalyst or pitfall in educational systems' aim for high achievement and equity? An answer based on multilevel SEM analyses of TIMSS 2015 data in Flanders (Belgium), Germany, and Norway. *Large-scale Assessments in Education*, *7*(1), Article 1. https://doi.org/10.1186/s40536-019-0069-2

Bokhove, C. (2022). Are instructional practices different between East and West? An analysis of Grade 8 TIMSS 2019 data. *Asian Journal for Mathematics Education*, *1*(2), 221–241. https://doi.org/10.1177/27527263221109752

Cantley, I. (2019). PISA and policy-borrowing: A philosophical perspective on their interplay in mathematics education. *Educational Philosophy and Theory*, *51*(12), 1200–1215. https://doi.org/10.1080/00131857.2018.1523005

Giberti, C., & Maffia, A. (2020). Mathematics educators are speaking about PISA, aren't they? *Teaching Mathematics and its Applications*, *39*(4), 266–280. https://doi.org/10.1093/teamat/hrz018

Hansen, K. Y., & Strietholt, R. (2018). Does schooling actually perpetuate educational inequality in mathematics performance? A validity question on the measures of opportunity to learn in PISA. *ZDM – Mathematics Education*, *50*(4), 643–658. https://doi.org/10.1007/s11858-018-0935-3

Ingram, J. (2026). Relationships between Opportunity to Learn and class prior attainment in mathematics. *International Journal of Science and Mathematics Education*, *24*(5), Article 63. https://doi.org/10.1007/s10763-026-10687-7

Jerrim, J. (2024). Has peak PISA passed? An investigation of interest in International Large-Scale Assessments across countries and over time. *European Educational Research Journal*, *23*(3), 450–476. https://doi.org/10.1177/14749041231151793

Jerrim, J., Lopez-Agudo, L. A., Marcenaro-Gutierrez, O. D., & Shure, N. (2017). To weight or not to weight?: the case of PISA data. *Investigaciones de Economía de la Educación*, *12*, 285–302.

Johansson, S., & Hansen, K. Y. (2019). Are mathematics curricula harmonizing globally over time? Evidence from TIMSS national research coordinator data. *EURASIA Journal of Mathematics, Science and Technology Education*, *15*(2). https://doi.org/10.29333/ejmste/99516

Kadijevich, D. M., Stephens, M., Solares-Rojas, A., & Guberman, R. (2023). Impacts of TIMSS and PISA on mathematics curriculum reforms. In Y. Shimizu & R. Vithal (Eds.), *Mathematics curriculum reforms around the world: The 24th ICMI Study* (pp. 359–376). Springer.

Liu, X., Hansen, K. Y., Valcke, M., & De Neve, J. (2024). A decade of PISA: student perceived instructional quality and mathematics achievement across European countries. *ZDM – Mathematics Education*, *56*(5), 859–891. https://doi.org/10.1007/s11858-024-01630-7

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://doi.org/10.1111/j.1467-985X.2012.01036.x

Min, I., Cortina, K. S., & Miller, K. F. (2016). Modesty bias and the attitude-achievement paradox across nations: A reanalysis of TIMSS. *Learning and Individual Differences*, *51*, 359–366. https://doi.org/10.1016/j.lindif.2016.09.008

Nilsen, T., Strietholt, R., Johansson, S., Kennedy, A., Eryilmaz, N., & Atasever, U. (2026). Teaching quality: Its influence on student achievement and equity in mathematics. Evidence from TIMSS longitudinal. *Studies in Educational Evaluation*, *90*. https://doi.org/10.1016/j.stueduc.2026.101637

OECD. (2023). *PISA 2022 assessment and analytical framework*. OECD Publishing. https://doi.org/10.1787/dfe0bf9c-en

Probert, S. (2021). Policy transfer and isomorphism: A case study of the England-China maths teacher exchange. *British Journal of Educational Studies*, *70*(3), 305–321. https://doi.org/10.1080/00071005.2021.1926915

Rutkowski, D., Rutkowski, L., Thompson, G., & Canbolat, Y. (2024). The limits of inference: reassessing causality in international assessments. *Large-scale Assessments in Education*, *12*(1). https://doi.org/10.1186/s40536-024-00197-9

Schmidt, W. H., Guo, S., & Sullivan, W. F. (2024). Inequality in USA mathematics education: The roles race and socio-economic status play. *ZDM – Mathematics Education*, *56*(3), 497–509. https://doi.org/10.1007/s11858-024-01593-9

Schmidt, W. H., & McKnight, C. C. (1995). Surveying educational opportunity in mathematics and science: An international perspective. *Educational Evaluation and Policy Analysis*, *17*(3), 337–353.

Sjøberg, S., & Jenkins, E. (2022). PISA: a political project and a research agenda. *Studies in Science Education*, *58*(1), 1–14. https://doi.org/10.1080/03057267.2020.1824473

Volante, L., & Klinger, D. A. (2023). PISA, global reference societies, and policy borrowing: The promises and pitfalls of ‘academic resilience’. *Policy Futures in Education*, *21*(7), 755–764. https://doi.org/10.1177/14782103211069002

Wang, X. S., Perry, L. B., Malpique, A., & Ide, T. (2023). Factors predicting mathematics achievement in PISA: a systematic review. *Large-scale Assessments in Education*, *11*(1). https://doi.org/10.1186/s40536-023-00174-8

You, Y., & Morris, P. (2015). Imagining school autonomy in high-performing education systems: East Asia as a source of policy referencing in England. *Compare: A Journal of Comparative and International Education*, *46*(6), 882–905. https://doi.org/10.1080/03057925.2015.1080115