Large-scale Assessments in Education

RESEARCH Open Access

The limits of inference: reassessing causality in international assessments

David Rutkowski[1*], Leslie Rutkowski[1], Greg Thompson[2] and Yusuf Canbolat[3]

*Correspondence:
David Rutkowski
drutkows@iu.edu
[1]Indiana University, Bloomington, USA
[2]Queensland University of Technology, Brisbane, Australia
[3]Center for Education Policy Research, Harvard University, Cambridge, USA

Abstract

This paper scrutinizes the increasing trend of using international large-scale assessment (ILSA) data for causal inferences in educational research, arguing that such inferences are often tenuous. We explore the complexities of causality within ILSAs, highlighting the methodological constraints that challenge the validity of causal claims derived from these datasets. The analysis begins with an overview of causality in relation to ILSAs, followed by an examination of randomized control trials and quasi-experimental designs. We juxtapose two quasi-experimental studies demonstrating potential against three studies using ILSA data, revealing significant limitations in causal inference. The discussion addresses the ethical and epistemological challenges in applying quasi-experimental designs to ILSAs, emphasizing the difficulty in achieving robust causal inference. The paper concludes by suggesting a framework for critically evaluating quasi-experimental designs using ILSAs, advocating for a cautious approach in employing these data for causal inferences. We call for a reevaluation of methodologies and conceptual frameworks in comparative education, underscoring the need for a multifaceted approach that combines statistical rigor with an understanding of educational contexts and theoretical foundations.

Causality is a complex and multifaceted concept that has been debated by philosophers for centuries. Democritus (460BC-370BC) is credited with saying he would rather discover one true cause than gain the kingdom of Persia. Despite its complexity, causality remains central to how we reason about our world and our ability to predict, control, and respond to events. It is also widely used in scientific research where it functions as a foundation for understanding the interconnectedness of variables and phenomena.

In the field of education, there has been a decades-long focus on evidence-based approaches, often underpinned by a desire to discern the underlying cause to improve outcomes or 'what works' for interventions within educational systems. Partly, this can be explained by the desire of policymakers to understand what they need to do to influence positive, systemic change. For example, Goldacre's (2013) "Building Evidence into Education" policy whitepaper, published by England's Department of Education, exemplifies this desire for more 'scientific' approaches to education research. Similarly, the

establishment of the What Works Clearinghouse in the US in 2002 aimed to gather in a central repository scientific evidence on 'what works' in education programs, products, practices, and policies. Subsequently, there has been an increased interest in making causal inferences using international large-scale assessment (ILSA) data. ILSAs, which are periodic, comparative assessments of educational achievement, have been present in educational research and policymaking since the 1960s and were originally designed to empirically examine and compare educational systems around the world. However, we argue that the nature of ILSA data generally precludes causal inferences no matter the political pressure to make causal claims in the pursuit of data-driven policy. This pressure to make causal claims with ILSAs is deeply rooted in the political and economic realities that shape educational systems worldwide. As policymakers and stakeholders increasingly demand concrete evidence to guide their decisions, the allure of causal inferences derived from ILSA data is becoming more evident in research and policy communities.

In this paper we aim to scrutinize claims of causality—a concept of paramount importance in both the social sciences (Murnane & Willett, 2010; Russo, 2009; Shadish et al., 2002), public policy (Athey & Imbens, 2017; Stone, 1989) and statistics (Holland, 1986)—within the context of ILSAs. Specifically, this paper challenges the use of ILSA data to draw causal inferences, because we contend that it overlooks the fundamental limitations and assumptions inherent in ILSA data. Our critical analysis illuminates the methodological constraints undermining causal claims derived from ILSA datasets including, but not limited to, effects of country heterogeneity and assessment standardization. The paper is organized as follows: First, we provide a concise overview of causality within the context of ILSAs, followed by an overview of randomized control trials and quasi-experimental designs. Subsequently, we analyze two quasi-experimental studies in the field of education, which we believe exemplify the potential of these designs. These studies are then juxtaposed with three quasi-experimental studies that utilize ILSA data that feature substantial limitations to making causal inferences. Finally, we suggest a framework to critically evaluate quasi-experimental designs using ILSAs and advise caution in employing ILSA data under current designs for making causal inferences.

Background: causality in ILSAs

To begin, we first define what we mean by *causality*. In both the social and physical sciences, causality is conceptualized as a relationship between variables wherein a change in one variable (the cause) is understood to induce a change in another variable (the effect). This causal relationship is typically established through empirical evidence and rigorous testing (Pearl, 2009). Furthermore, in most cases, the cause must temporally precede the effect and the occurrence of the cause should reliably lead to the occurrence of the effect (Shadish et al., 2002). Additionally, the identification of a 'cause' must eliminate other plausible alternative explanations for the observed relationship, ensuring that the connection between cause and effect is both direct and unambiguous (Pearl, 2009; Shadish et al., 2002). This rigorous approach to establishing causality is fundamental as it underpins the validity and reliability of scientific findings. We now turn to causality in the ILSA setting.

Well-designed ILSAs have well-developed frameworks, carefully researched and evaluated instruments, and innovative sampling schemes that produce educational data

from representative samples of students and schools in participating countries. Nevertheless, the low-stakes, observational, cross-sectional nature of these studies makes drawing causal inferences a challenge. Proponents of causal inferences using ILSA data advocate for their use in a quasi-experimental design framework, such as Rubin's (1974, 2005) Potential Outcomes Framework (POF) to isolate causal effects. Importantly, as we will discuss, quasi-experimental methods rely on stringent and, in some cases, untestable assumptions and not meeting these assumptions necessarily compromises associated inferences.

A significant obstacle to drawing causal inferences in ILSA studies is the inherent heterogeneity of educational systems, which vary widely in cultural, economic, geographic, and linguistic aspects. This diversity introduces a multitude of confounding factors that can influence relationships within and across these systems, thereby rendering any claims highly context specific. Compounding this issue is the standardization of assessments across diverse systems often strips away the unique contextual elements essential for isolating causal effects. As a result, a causal inference made in one country may not be applicable, *ipso facto*, to another, due to the discordance between the homogenizing nature of standardized assessments and the heterogeneous realities of the educational environments they aim to measure. Although we are not claiming that causal inferences are a totally unreachable goal, as we discuss subsequently, causal inferencing is only plausible in limited circumstances and, even then, the untestability of some assumptions leaves inherent doubt in the validity of any causal claim in most situations.

For these reasons, causal inferences have long been viewed skeptically in the ILSA context. For example, Andres Schleicher, Director for the Directorate of Education and Skills at the OECD and one of the developers of PISA, claimed that PISA "cannot identify clear-cut cause-and-effect relationships" (2009, p. 252). In addition, Singer and Braun (2018) argue that isolating causal effects from ILSAs is frustratingly challenging and ill advised (see also Braun & Singer, 2019). And, in a similar vein, Carnoy (2015) concludes, "cross-sectional surveys such as the TIMSS and PISA are not amenable to estimating the causal effects of school inputs on student achievement gains" (p. 10). These perspectives underscore the inherent complexities in drawing causal conclusions from ILSA data, highlighting the need for caution and critical analysis when interpreting these assessments in the broader context of educational research and policy.

Despite a precedent for avoiding causal claims with ILSA data, in recent years there has been a growing interest in doing so (Cordero et al., 2018; Komatsu & Rappleye, 2021). Recent initiatives reflect this trend; for example, the European Commission funded a project aimed at making causal claims using ILSA data (European Commission, 2018). Furthermore, in a review of academic research that attempted to make causal claims with ILSA data Cordero et al. (2018) found that the "number of [such] studies has increased substantially" from 2004 to 2016 (p. 28). Scholars (See Chmielewski & Dhuey, 2017; Cordero et al., 2018) and organizations that design, administer and collect this data (See Kennedy et al., 2023) are increasingly gravitating towards causal inferencing, reflecting the broader trend in education research toward data-driven decision making. For example, testing organizations like the International Association for the Evaluation of Educational Achievement (IEA) offer quasi-experimental design workshops that encourage the research and policy community to design studies aimed at making causal claims with ILSA data (Kennedy et al., 2023).

Yet, the assumption that causal mechanisms identified through ILSAs are universally applicable is contentious. This is where Hume's Guillotine becomes pertinent: it posits that descriptive statements ("is") cannot directly lead to prescriptive or normative statements ("ought"). For instance, while a causal link might be established, such as private tutoring leading to higher TIMSS scores in Singapore, this doesn't imply a universally applicable policy for different contexts like the United States or South Africa. The challenge lies in ensuring the validity of causal interpretations for each unique policy context (see Shadish et al., 2002). Next, we briefly describe RCTs, the gold standard for establishing causality. We include this discussion as RCTs form the foundation for the quasi-experimental methods which are most often used in attempts to establish causality withing ILSAs.

Randomized control trials

Unlike ILSA study designs, RCTs are purposefully designed to establish causal relationships, thus offering a unique lens for assessing the efficacy of educational interventions and policies. Initially developed in the medical sciences to detect treatment effects through blinding and control groups (Meldrum, 2000), RCTs have been adapted for social science research, particularly in education, to test the impact of various interventions or treatments. The fundamental principle behind RCTs is randomization, which is a process designed to minimize selection bias, enabling researchers to isolate the effect of the intervention from other factors. Randomization ensures that both observed and unobserved characteristics are evenly distributed across groups, making the only systematic difference between them the intervention itself. Key assumptions underpinning the validity of RCTs include the Stable Unit Treatment Value Assumption (SUTVA), which posits that the outcome observed in one individual is not affected by the treatment status of another individual. Ignorability, or the assumption that treatment assignment is independent of potential outcomes, is also crucial. This means the likelihood of receiving treatment is not influenced by the potential outcome it produces. Homogeneity and additivity are additional assumptions. Homogeneity implies that the effect of the treatment is consistent across different subjects, while additivity suggests that the effects of the intervention and other factors influencing the outcome add up linearly. The non-attrition assumption is vital too, implying that participants do not drop out of the study in a way that systematically differs between the treatment and control groups.

These assumptions have been put to work in a variety of ways. Perhaps one of the most ambitious has been the experiments conducted in the English education system. Sims et al. (2023) estimate that the Education Endowment Fund (EEF) in England has commissioned more than 157 randomized experiments since 2010 looking at interventions such as the use of one-to-one numeracy support by teacher assistants for struggling students (Hodgen et al., 2023); the use of financial and experiential incentives to improve Year 11 student motivation in deprived schools (Sibieta et al., 2014) and the impact of dialogic teaching on student achievement (Jay et al., 2017). Ultimately, however, the effects of these RCTs have been underwhelming for both research and policy makers (Lortie-Forgues & Inglis, 2019). In fact, despite the methodological promise of establishing causal relations, in practice confidence has been hard to guarantee and most effect sizes are small. For example, Kraft (2023) found that in randomized control trials focusing on educational interventions with standardized achievement outcomes, 36% of the effect

sizes were smaller than 0.05 standard deviations. With similar findings in the UK, Sims et al. (2023) concluded that "quasi-experimental methods and multi-site trials will often be superior for informing educators' decisions" (n.p.).

One significant challenge in conducting RCTs in educational settings is the variability of individual responses to interventions. This variability can make it hard to detect the average effect of the intervention. Issues of treatment compliance and participant dropout rates further complicate the matter, potentially biasing the results. In response to these limitations, the field of educational research has moved toward quasi-experimental designs as they offer more practical means to establish causal inferences, especially in contexts where conducting RCTs might be impractical or unethical. These designs, while not as rigorous as RCTs, can still provide valuable insights into the effectiveness of educational interventions.

Quasi-experimental designs

Quasi-experimental designs are research methods that aim to evaluate causal relationships using observational data, in scenarios where true experimental designs, like randomized controlled trials, are not feasible. These designs typically leverage natural experiments or existing variations in conditions, such as differences in policies across regions or time, to infer causal effects. Key to their implementation is identifying a credible counterfactual scenario. This allows researchers to approximate the conditions of a controlled experiment by assuming that the observed outcomes would have been the same in the absence of the treatment or intervention, controlling for potential confounding variables.

Like RCTs, quasi-experimental designs rely on stringent theoretical frameworks and assumptions to support causal inference. Notably, Donald Rubin's and Judea Pearl's work (see Pearl, 2009; Rubin, 1974) has provided rich foundations for causal reasoning in the absence of randomization. These frameworks emphasize the need for high quality data and the importance of controlling for confounding variables, acknowledging collider variables (Pearl, 2009), and utilizing statistical techniques such as matching and instrumental variables to mimic the conditions of an RCT (Pearl & Mackenzie, 2018). Yet, the application of these techniques is complex and requires meeting strong assumptions about the underlying data, how the variables map onto theory, and causal relationships. Even small deviations from these assumptions can lead to biased and misleading conclusions (Russo, 2009).

Building on Neyman's (1923) foundational work, Rubin developed and popularized the POF. This framework, when all assumptions are met, allows for causal inferences without an RCT. Within the POF, for each unit under study, there are two potential outcomes: the outcome if the unit is treated (often denoted $Y_{i(1)}$) and the outcomes if the unit is not treated (denoted $Y_{i(0)}$). The causal effect for unit i is then defined as the difference between these two potential outcomes: $(Y_{i(1)} - Y_{i(0)})$. However, a significant challenge arises because, for each unit, only one of these potential outcomes *can* be observed, depending on whether the unit is treated or not. This leads to the 'fundamental problem of causal inference' - the inability to simultaneously observe both potential outcomes for any single unit, resulting in an inherent missing data problem. The unobserved potential outcome must be inferred from the available data. Rubin and others have suggested various methods to address this challenge, all operating under the POF.

These methods aim to estimate the unobserved potential outcomes as accurately as possible, thereby enabling researchers to draw causal inferences from non-experimental data. This approach has become increasingly important in educational research, where the practicalities and ethics of conducting RCTs can often be prohibitive.

Given that quasi-experimental designs strive to emulate the methodological rigidity of RCTs, it follows that the initial assumptions governing RCTs are equally pertinent. Accordingly, within POF it is only when these assumptions are satisfied that quasi-experimental techniques can be used to approximate the conditions of a randomized trial and provide credible causal estimates. In fact, the assumptions must be carefully considered and justified, and sensitivity analyses is often necessary to help gauge the robustness of the results to potential violations. It is this interplay of theoretical robustness and practical complexity that makes POF both a powerful tool and a subject of ongoing investigation and refinement.

Of course, applying the POF to the complexities of international educational systems and ILSAs is far from straightforward; it is an intellectual exercise marked by nuanced complexities and inherent challenges. One significant challenge is the assumption of no hidden variation in treatments, as outlined in the SUTVA. This assumption implies that the treatment is uniform and consistent across all units, yet the heterogeneous nature of international educational systems, each with unique cultural and socio-economic characteristics, complicates this assumption. Consider the example of a policy aimed at reducing classroom sizes, anticipated to yield uniform educational benefits. This policy's effectiveness can vary significantly due to cross-country pedagogical differences linked to class size. For instance, in the United States, smaller classroom sizes might be valued for providing individualized instruction, whereas in Japan, larger classroom sizes are often normalized and not necessarily viewed as detrimental (Ehrenberg et al., 2001). This variation in pedagogical approaches and perceptions of classroom size across countries represents a form of hidden variation in the treatment (classroom size reduction policy) itself. Even within a single country, such a policy could manifest differently in urban and rural settings, further illustrating the challenge of ensuring homogeneity in the application of the treatment. If these variations in how the treatment is implemented or perceived are known but cannot be statistically controlled, it constitutes a violation of the homogeneity assumption of SUTVA. In such cases, it is the responsibility of researchers to demonstrate that this variation does not substantively bias the results of the study.

Defining "treatment" within the context of ILSAs also introduces ambiguity, with the multifaceted nature of interventions aligning imperfectly with POF's binary framework. For instance, even teacher credentials and qualifications – although the same in name (e.g., a bachelor's degree in education) – can vary in kind across country. Whether focusing on a teaching method, curriculum innovation, or policy change, this ambiguity may lead to inconsistencies, weakening the model's findings. Generalizability concerns also arise, where extending findings from one context to another within the diverse landscapes of ILSAs becomes perilous and fraught with quandaries. A literacy program, for example, successful in one cultural setting, might falter in another due to the myriad variables at play such as cultural factors on how parents engage with their children or structural language differences.

The variance in the standardization of educational assessments across different languages and cultures, emerges as a significant hurdle within ILSAs, where translation or

differences in conceptual understanding serve as two examples. This misalignment can subtly introduce bias into causal estimates, undermining POF's foundational assumptions. Ethical considerations should not be overlooked either, as rigid adherence to POF may eclipse the human stories, values, and traditions that infuse the educational landscape, leading to an inadvertent overlook of the complex human and social dimensions of education. In the quest for methodological rigor through standardization, we may compromise the very causal inferences we seek to draw, specifically those grounded in POF. The crux of POF hinges on capturing the heterogeneity within the data, allowing for detailed understandings of causal relationships. However, the act of standardization in ILSAs could inadvertently smooth over this heterogeneity, thereby creating an illusion of uniformity where none exists. This is not a mere academic dilemma but a critical issue. The standardization, although useful for generalizations, masks the heterogeneity of human stories, values, and traditions that define educational experiences across different cultures and social settings. In other words, the standardization efforts in ILSAs—though necessary due to their aim to create universally understandable and applicable questions—can inadvertently introduce bias into causal estimates. This is not only a violation of POF's foundational assumptions but a misstep that can lead to policy recommendations which are empirically flawed as they do not respect human and contextual differences. Thus, ILSAs are stuck in a paradox: the very efforts to standardize and make "objective" comparisons across diverse educational landscapes may be the thing the prevents valid quasi-experimentation.

In the context of assessing the rigor and validity of quasi-experimental designs, particularly in the domain of educational assessment, the principles made clear by Murnane and Willett (2010) offer a foundational framework. These principles encapsulate critical aspects of experimental integrity, starting with the clear definition of the study's participant population, ensuring that the individuals involved are representative of a broader, explicitly defined group. This is complemented by the second principle, which emphasizes the necessity of having explicit and well-defined experimental conditions. This involves a clear delineation of what constitutes the treatment and control scenarios, crucial for maintaining the study's structural clarity. The third principle centers on the equivalence of groups in expectation, approximating the conditions of randomization, thereby addressing potential confounders and biases in group selection. Finally, the fourth principle highlights the importance of measuring outcomes that are directly and sensitively responsive to the treatment, ensuring that the study's findings are both relevant and accurately reflective of the intervention's impact. Together, these principles form a cohesive guide for assessing the quality and credibility of quasi-experimental research, setting a benchmark against which such studies can be rigorously evaluated.

Building upon the foundational principles articulated by Murnane and Willett (2010), we now turn our attention to exemplars within the realm of quasi-experimental research in education. This segment of our exploration is dedicated to showcasing how, despite the inherent complexities and challenges associated with quasi-experimental designs, certain studies have managed to harness the potential of this approach, yielding insightful and robust findings. The selected studies stand out not merely for their methodological rigor but also for their clever utilization of sources of random or exogenous variation, a cornerstone in approximating the conditions of randomized controlled trials (RCTs). By reviewing these studies, we aim to highlight situations in which quasi-experimental

research can transcend its limitations and offer meaningful contributions to our understanding of educational phenomena. Thus, we proceed with a focused analysis of two high-quality papers, each embodying the essence of what can be achieved through careful design and execution.

Study I: private school lottery study

A school choice study found that attending a private school on a public voucher in Louisiana resulted in reduced math, science, and social studies test results (Abdulkadiroğlu et al., 2018). The authors analyzed a program that provided vouchers to low-income children attending underperforming public schools, with eligibility set at family incomes below a given threshold, allowing them to attend private schools of their choice. Exogenous variation came from the fact that from 2012, the program was oversubscribed, and a lottery was used to award vouchers. This study used an instrumental variable approach, where the instrument was whether or not the student was offered a voucher. The authors employed a standard two-stage least squares approach where first, the probability of using a voucher was modeled as a function of a voucher offer and a collection of covariates. Second, the estimated probability of use was used as a predictor in an equation for scores on a standardized test.

In this study, the population was defined as lottery voucher applicants, which were mostly low-income minority applicants. The experimental conditions are whether the student did or did not attend a private school on a voucher. Given a reasonable assumption that the lottery was random and fair, the authors further demonstrate a good balance across covariates, satisfying condition 3. Finally, test scores on the state standardized assessment in math, English language arts, science, and social studies serve as the outcome, the fourth quality condition is met[1]. The authors follow up the main analysis with a series of robustness checks and supplementary analyses to find alternative explanations for the decline in achievement. Using logic supported by empirical evidence, the authors conclude that, due to structural factors, selection of poor-quality private schools into the lottery scheme was a reasonable explanation and that findings were robust.

In this study, the methodological rigor is evident in how it aligns with Murnane and Willett's principles, particularly using a lottery system as a source of exogenous variation, reasonably ensuring randomization and thereby satisfying the condition of creating groups equal in expectation. The clarity in defining the participant population and experimental conditions, along with the measurement of specific, relevant outcomes, further exemplifies the study's adherence to these principles. However, when we pivot to research utilizing ILSA data, which we do subsequently, the landscape shifts significantly. The structured design and conditions evident in the Abdulkadiroğlu et al. (2018) study stand in contrast to the complexities and constraints inherent in ILSA datasets. Unlike the controlled environment of a lottery-based voucher system, ILSA data often encompasses a wide range of uncontrolled variables and diverse educational contexts. This heterogeneity, while offering a rich tapestry of international educational settings, poses significant challenges in defining a consistent participant population, establishing well-defined experimental conditions, and ensuring the equivalence of groups across

[1] Although scores on the test used in this study, the Louisiana iLEAP or LEAP, were not validated for this study, it is reasonable to expect that scores would be sensitive to differences in educational quality.

different national and cultural contexts. Moreover, while the outcome measures in the Abdulkadiroğlu et al. study was directly linked to the intervention, ILSA data often include outcomes that may not be as sensitively attuned to specific treatments or interventions evident in a given context. As we explore studies employing ILSA data, these challenges and their implications for the validity and generalizability of findings will become increasingly apparent, underscoring the nuanced and intricate nature of conducting quasi-experimental research within the realm of international education. Let us turn now to a second high-quality example.

Study II: school consolidation study

In a recent study of school district consolidation, the author found that theory-relevant student outcomes were not impacted by school district consolidation in North Carolina (Chin, 2023). In this study, the author used several data sources at the school-level and county-level. From a national data source, these measures included demographic composition, student enrollment, measures of school segregation, measures of class and school size, and school expenditures, among others. As a measured outcome, the author also drew on the federal data source for longitudinal high school diploma rates at the district level. A second outcome is county-by-birth cohort crime rates.

The author uses a differences-in-differences design and an event study analysis to estimate the effect of consolidation on school attainment, county crime rates, and other outcomes. As a first difference, the author uses outcomes in counties with consolidating districts before and after the *event* (consolidation). As a second difference, the author uses outcomes over time between treatment and control districts. Potential bias in the causal estimates is addressed with a two-stage estimation of the outcome as a function of a fixed effect for county and time. These county- and time-specific estimates are then used to estimate an adjusted coefficient of consolidation on the outcome of interest. This combined approach accounts for time-invariant differences between counties with and without consolidation, statewide shifts in outcomes over time, and bias that could stem from staggered rollout of consolidation. This final issue – although not formally testable – gets at the key parallel trend assumption that causal difference-in-difference estimates rely on.

In this second study, the population was defined as school districts in North Carolina. The experimental conditions are whether a district did or did not experience a consolidation event. As noted above, the analysis strategy accounts for district-level, time-invariant differences and state-level, time varying differences, along with a reasonable investigation of parallel trends, satisfying condition 3. Finally, the outcomes were substantiated with previous literature as theoretically relevant and sufficiently sensitive to the treatment, satisfying the fourth quality condition. Based on a series of robustness checks, the findings were consistent across different model specifications. Notable in this analysis was the merging of multiple datasets at the district level. A better approach would have been to merge data at the student level over time to estimate – at the student level – the causal impact of consolidation; however, as we noted previously, this isn't often possible, due to privacy concerns, limiting the nature of the inferences to the district level. Given what is known about ecological fallacy (Robinson, 1950), the author is cautious to limit their interpretations to the appropriate level and avoid making inferences at the student level.

This study illustrates another high-quality application of quasi-experimental design principles in educational research. By employing a differences-in-differences approach and an event study analysis, the study navigates the complexities of assessing the impact of consolidation on student outcomes and county crime rates. The methodological approach, along with a careful selection of outcomes and the population under study, aligns well with the criteria set forth by Murnane and Willett. The study's rigorous handling of data, including the adjustment for county- and time-specific factors, and the diligent application of robustness checks, further underscore its adherence to these quality conditions. Notably, the study's prudent approach to data integration at the district level, while a necessity due to privacy concerns, highlights a common limitation in educational research—difficulty in obtaining and merging student-level data over time. In contrast to Chin's work, research using ILSA data faces a distinct set of challenges, especially in the context of quasi-experimental designs. While the Chin study could control for a variety of factors at the district level and maintain focus on specific, localized conditions, ILSA data encompasses a vastly broader and more diverse range of educational contexts. This diversity, although valuable for cross-national comparisons, introduces significant complexities in defining a consistent participant population and establishing uniform experimental conditions both across and within countries and cultures. Further, the anonymized nature of the publicly available data in most countries limit data merges to the country level. This limitation then further imposes guardrails on the level of inferences that can be obtained to the country or educational system.

Additionally, while the Chin study could rely on specific, theory-relevant outcomes, ILSA data often involves outcomes that may not be as directly linked to specific educational interventions, thereby complicating the measurement of treatment impact. Moreover, the broader scope of ILSAs can make it challenging to maintain the level of methodological rigor and precision required to establish causality and seen in studies like that of Chin, particularly when it comes to ensuring the equivalence of groups and the sensitivity of outcome measures. As we delve into examples of studies using ILSA data, these challenges will become more evident, underscoring the intricate balancing act of maximizing the rich potential of ILSA data while navigating their inherent limitations.

Quasi-experiments with ILSAs

Having considered the studies by Abdulkadiroğlu et al. (2018) and Chin (2023), two exemplars of careful, high-quality quasi-experimental designs, we now transition to a different facet of quasi-experimental research in education, one that leverages ILSA data. As previously noted, these assessments, encompassing data from numerous countries and educational systems, present a unique opportunity to explore educational phenomena on a global scale. However, they also introduce a set of complexities and methodological challenges distinct from those encountered in more localized studies and in studies where multiple data sources at varied levels can be combined to create the conditions for causal inference. Here, we critically examine how researchers navigate these challenges, particularly in terms of adhering to the quality conditions outlined by Murnane and Willett. Through this analysis, we aim to further our understanding of the capabilities and constraints of quasi-experimental research within the vast and varied landscape of international education.

ILSA study I: school closure study

In a recent study, Kennedy and Strietholt (2023) examined the relationship between school closure due to the COVID-19 pandemic and reading achievement across 29 countries. Using PIRLS data from two cycles from 2016 to 2021 and school closure data collected by the UNESCO Institute for Statistics (UIS), they applied a fixed effect approach. They exploited the interaction between a continuous school closure variable and a binary variable indicating PIRLS 2021. Thus, the authors compared the change in student achievement between PILRS 2016 and 2021 by the level of school closure across countries. In two separate models, they used five cycles of PIRLS since 2001 and control for baseline achievement levels that vary across countries. They found a negative school closure effect, which is larger for the socioeconomically disadvantaged and for those who do not have computer access. The results were consistent with different samples of countries and alternative measures of school closure (e.g., full or partial closure).

The study offers an important perspective on how trends in achievement across countries have changed before and after the COVID-19 pandemic. That being said, the data and method inhibit its power to prove cause and effect. One of the key limitations of the study to draw causal conclusions is that school closures were not exogenous across countries. Economically less developed countries had longer school closures. The correlation between GDP per capita and the length of school closure is -0.4 across countries. In the absence of school closures, countries with longer school closures might have had a different achievement change because of disparities in economic development and associated educational factors across countries. Though these issues are not addressed in detail in the study, a simple trend analysis we plot in Fig. 1 indicates that countries with higher school closure days had a declining trend compared to countries with shorter school closures even before the school closure. This suggests the pre-trend between PIRLS 2011 and PILRS 2016 confounds the relationship of interest, the trend between PIRLS 2016 and PIRLS 2021 when school closures occurred. The study does not offer insights into the extent to which the method addresses this fundamental issue and whether it recovers causal effects. Therefore, the study fails to provide evidence on condition 3 of Murnane and Willet: equivalence of groups in expectation. Methodologically, if data do not satisfy this assumption, named the parallel trend assumption in difference in differences literature, findings are susceptible to bias and should be treated as correlational, only (Athey & Imbens, 2022). As discussed earlier, for instance, Chin shows that before school district consolidation, control and treatment groups had similar trends in

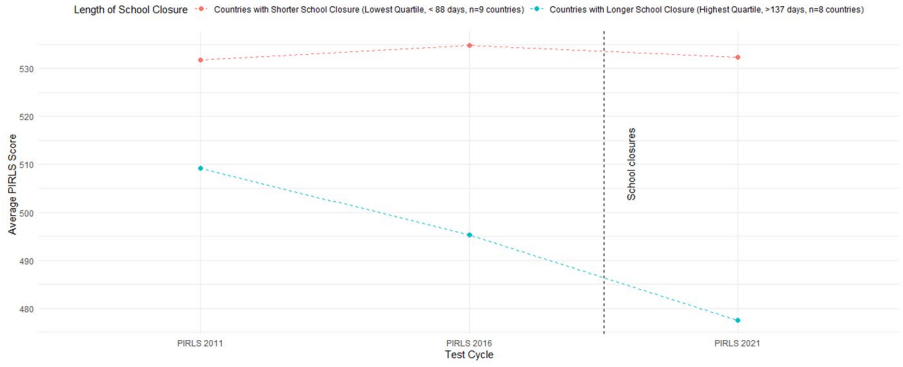

Fig. 1 Trend in PIRLS scores over years and across countries by school closure during COVID-19 pandemic

outcomes enabling the author to attribute the changes in outcome in the post-treatment period to the treatment effect, or lack thereof.

Difficulties in satisfying the equivalence of group in expectations are not unique to Kennedy and Strietholt`s (2023) school closure study but endemic to ILSA data. The underlying, fundamental limitation is that longitudinal studies integrate ILSA data at the country level (or use pooled individual data) since individuals are not followed over time. Those studies assume that countries with varying levels of treatment exposure are comparable in their outcome over time. Typically, they use country-fixed effects to eliminate unobserved time-invariant confounders (Cordero et al., 2018). The unit fixed effect is a useful approach to adjust for those confounders but does not ensure causality because it does not adjust unobserved time-varying confounders. Second, the unit-fixed effect comes at the expense of dynamic causal relationships (Imai & Kim, 2019). Regarding the first one, researchers strictly lean on the assumption that there are no time-varying unobserved confounders, which is not easy to meet in ILSAs since social, economic, and educational characteristics of countries change over time. Given that at least a few cycles of ILSA are needed to construct reasonable panel data that covers at least 6 to 10 years, satisfying no unobserved time-varying confounders is challenging. An additional challenge specific to ILSAs, incorporating time-varying variables relevant to causal relationships of interest is not easy. Country-level rich longitudinal data relevant to both treatment and outcome are not always readily available and not easy to collect. The second assumption of causal unit-fixed effect models requires that past outcomes do not affect current treatment. Because country-level educational outcomes influence policy and intervention decisions, the outcomes in the previous cycles may be linked to the treatment of interest. Thus, researchers end up with multiple strong assumptions that are difficult to meet.

As we briefly touched upon the issues with country-level treatment measures, it is important to highlight that country-level measures omit within-country variation of treatment conditions and outcomes. For instance, Kennedy and Strietholt (2023) rightly acknowledge that country-level school closure data overlooks potential variations within countries. Because inferences at the country level usually do not align with inferences at lower levels, estimates from ILSA studies may not meet the fourth principle that highlights warranting that the findings are accurately reflective of the intervention's impact. Unobserved heterogeneity in the treatment condition and differential effects across students and schools may offset each other leading to ecological fallacy.

Regarding the above issues, whether the study is longitudinal or cross-sectional, leveraging external data is limited to the country level in ILSAs since the individuals and schools are unidentifiable in publicly available data. Therefore, the type of causal questions is bounded by certain, pre-determined variables in background questionnaires. Thus, overreliance on the causal claims with ILSA data may indeed undermine the value and richness of ILSAs. Because causality does not always ensure theoretical and practical relevance, overdependence on causality in ILSA may narrow the scope of potential research questions. Further, the pre-determined nature of ILSA data undercuts the ability to perform robust covariate adjustment techniques to mitigate selection bias when relevant. The rigor of commonly used adjustment approaches such as matching and weighting are bounded by available items in background questionnaires in ILSA that

may not always incorporate the most relevant variables, depending on the causal estimate of interest.

Study II: private school study

Vandenberghe and Robin (2004) examined student achievement in private schools compared to public schools in 31 countries and concluded that private education does not generate systematic benefits. The authors leveraged the instrumental variable approach by exploiting a binary school location variable (i.e., a big city with a population of more than 100,000 and 0 otherwise) in PISA questionnaires as an instrument. The motivation for using school location as an instrument was students in larger cities are more likely to enroll in private schools. Thus, in theory, the location variable should create randomization and purge out the remaining non-random part of private school enrollment. Their study is one of the early applications of instrumental variables in ILSAs, influencing subsequent studies that used a similar instrumental variable (Perelman & Santin, 2011; Pfeffermann & Landsman, 2011). Although the study offers useful information on whether and the extent to which student achievement differs in public and private schools across countries and incorporates other quasi-experimental approaches to compare results, the results do not warrant causality.

By the exclusion restriction assumption of the instrumental variable approach (Angrist et al., 1996), the relationship between school location and student achievement should be only through private school enrollment. However, there are a host of other factors that influence the relationship between school location and student achievement besides private school enrollment. PISA data shows that schools in urban settings have better educational resources, qualified teachers, and socioeconomically more advantaged students in most participating countries (OECD, 2012). Such evidence indicates that the instrument is endogenous and raises concerns about the validity of the instrument. These issues create major threats to satisfying condition 3 of Murnane and Willett, which requires approximating the conditions of randomization, thereby addressing potential confounders and biases in group selection.

As demonstrated by the reviewed studies, there is a challenge in utilizing randomization and examining causal relationships with ILSA data. This is mainly because, by their design, ILSAs focus more on systemic and structural issues whereas rigorous application of quasi-experimental methods focuses on the impact of specific policies and interventions in a particular institutional context. In most cases, those studies take advantage of institutional, contextual, or policy-specific information to mimic experimental conditions. For instance, school admission lotteries, arbitrary cut-offs in test scores to determine program participation eligibility or quasi-random variation in policy over time and across local or regional education agencies offer opportunities for researchers to utilize quasi-experimental methods. The question of whether ILSAs can incorporate such information in the future deserves a separate discussion but at least in their current form, their limitation to capture policy-specific and institutional information in the subnational units restricts their utility to make causal claims.

Additionally, Vandenberghe and Robin`s (2004) study also represents an example of the paradox of using data from standardized comparisons for valid quasi-experimental studies—an issue that we discussed earlier. Even though private schools are considered a uniform phenomenon across school systems, their organizational structures

considerably differ from country to country. In some school systems, private schools are exempt from most of the state regulations, in others, they are heavily regulated and mirror public schools in their organizational characteristics (OECD, 2012). In addition, there are in-between school types such as charter schools in the US, which are privately managed and publicly funded. Consequently, the treatment (i.e., private school enrollment) and control (i.e., public school enrollment) conditions are not identical across countries. These kinds of cross-country differences even in fundamental phenomenon like private school definitions blurs the structural clarity of experimental designs in ILSA studies violating the second condition. Further, such an issue prevents the research from meeting a key assumption of SUTVA: no hidden variation in treatment. The unobserved variation in the treatment condition creates more than two potential outcomes, the core of POF.

Study III: early tracking study

One of the commonly examined issues with quasi-experimental approaches using ILSA data is early tracking (Cordero et al., 2018). As an example, Lavrijsen and Nicaise (2015) studied the effect of tracking student achievement inequalities by socioeconomic background in 33 countries and found that tracking exacerbates those inequalities. The authors applied a difference-in-differences approach to leverage the fact that students are educated in the same program at elementary school in all countries, but they are tracked at secondary school in some countries. Because PIRLS (4th grade) and PISA (15 years old) sampled students from elementary and mostly from secondary schools respectively, they pooled student-level data from those assessments to exploit variance in tracking over time and across countries. The underlying motivation was those who took PIRLS in 2006 and those who took PISA in 2012 were roughly from the same population. For each country in the sample, they fit regression models to predict variance in reading achievement scores that is explained by socioeconomic background in PIRLS and PISA. Then, those variables are used as control and outcome variable in difference in difference estimation, respectively.

As commonly used by causal ILSA studies, combining achievement and survey results from different assessments to draw causal conclusions is problematic. The purpose and frameworks of those assessments such as PIRLS and PISA differ from each other. While PIRLS assesses the competencies concerning goals and standards for reading education, PISA does not focus on curricula. The estimates may be correlational rather than causal since the country-level change in the outcome from PIRLS to PISA is not necessarily the result of variables of interest but educational characteristics, instructional content, or other policies that change from elementary to secondary schools across countries. Further, the test scores are not on the same scale, making comparison a challenge. Murnane and Willet underline the crucial role of measuring outcomes that are directly and sensitively responsive to the treatment (fifth condition). These issues suggest that researchers need to take extra caution when they make causal claims using data from different tests as they face challenges to ensure that their treatment of interest is aligned with outcomes from different ILSAs simultaneously.

Finally, just because assessments such as PIRLS and PISA sample students from the similar birth cohorts does not mean those students are comparable over time. In fact, unless data are truncated at the fourth grade to ensure a common group of 15-year-olds

several years later (or truncated for 15-year-olds to ensure they are in a common grade), these cohorts are not the same. Further, because ILSAs only sample schooled students but not those who are not enrolled in schools for reasons such as drop-out, the sample in the early grades does not necessarily overlap with later grades. Lavrijsen and Nicaise (2015) do not consider that the characteristics of sampled students in PIRLS 2006 and PISA 2012 systematically differ from each other. For instance, the PIRLS 2006 sample covers most children in Hungary and Denmark, countries with and without early tracking policy, respectively. On the other hand, the PISA 2012 sample covers about 82% and 91% of the 15-year-olds in those countries, respectively (Martin et al., 2007; OECD, 2014). The change in the sample characteristics between PIRLS and PISA may not be uniform across countries as the example indicates. Therefore, using an analytical sample from different assessments does not always ensure that the participants represent the broader group across measurement periods, establishing a threat to condition one of Murnane and Willet. This issue biases the treatment effect since the change in the outcome might be a result of the changing sample characteristics.

Discussion

Using quasi-experimental design with ILSAs, as currently designed, presents a complex web of methodological, ethical, and epistemological challenges. These challenges, which range from the difficulty of applying standardized measures across heterogeneous populations to the complexities inherent in the structure of ILSAs, defy straightforward resolution. In response to these challenges, if ILSA programs aspire to establish credible causal inferences, a more structured and collaborative strategy is required. This strategy should involve defining a fixed set of causal questions rooted in educational theories and policy issues, ensuring their relevance and feasibility for quasi-experimental designs. Additionally, it necessitates concerted contributions from participating countries in selecting and agreeing upon the necessary variables for a robust quasi-experimental framework. Implementing such designs, while continuously evaluating and adapting to the dynamic nature of educational contexts, is crucial. However, this focused pursuit of causal inference will lead ILSAs away from their original mission. Traditionally, ILSAs have functioned as a global thermometer, measuring educational progress and challenges across diverse contexts. By shifting towards a design primarily aimed at causal analysis, ILSAs risk narrowing their scope and impact, potentially transforming from a comprehensive global educational barometer into a tool more specialized in addressing specific causal questions in education.

Moreover, the very idea of seeking causality at the international level warrants critical examination. While quasi-experimental designs are adept at tracing causal relationships when stringent assumptions are met, they often fall short in elucidating the underlying 'why' and 'how' of these relationships. This limitation presents a significant epistemological challenge, particularly in translating controlled research conditions into the complex, dynamic realm of real-world educational settings, especially cross-nationally. When analysis occurs at the country level, the notion of using one nation as a counterfactual for another becomes particularly problematic. Given the unique cultural, policy, and educational systems of each country, comparing them in a binary fashion oversimplifies the intricate reality of global education. This perspective heightens the complexity of applying the POF to ILSAs and raises fundamental questions about the methodology

and conceptual framework of comparative education. These considerations underscore the need for careful reflection on the direction and methodologies of ILSAs, particularly in their endeavor to contribute meaningfully to the field of educational research and policy.

Pearl's (2000) insight on causal analysis provides a critical perspective that resonates deeply with the challenges faced in integrating the POF with ILSAs. Pearl cautions against overreliance on statistical methods for causal inference: "But why would anyone play down the cautionary note of Rosenbaum and Rubin when doing so would violate the golden rule of causal analysis: No causal claim can be established by a purely statistical method, be it propensity scores, regression, stratification, or any other distribution-based design" (p. 350). This statement underscores the inherent limitations of relying solely on statistical approaches in the complex landscape of ILSAs. The cautionary note Pearl strikes aligns with the challenges in applying quasi-experimental designs to ILSAs, where the diverse educational contexts and inherent variability defy straightforward statistical solutions. It reinforces the necessity of a multifaceted approach that combines statistical rigor with a deep understanding of the educational contexts and theoretical underpinnings, ensuring that causal claims are not only statistically sound but also contextually relevant and theoretically grounded.

Considering the framework proposed by Murnane and Willett, the complexities in adapting their principles to ILSAs become evident. Applying these conditions to ILSAs highlights significant hurdles. First, defining a 'study population' in the context of diverse and varied international educational systems is fraught with challenges. The heterogeneity of these populations often defies the standardization required for a clear-cut definition, essential in quasi-experimental design. Second, establishing 'well-defined experimental conditions' in ILSAs is problematic. The varied educational policies, cultural contexts, and implementation strategies across countries make it difficult to create uniform conditions that could be compared in a meaningful way. This issue is further complicated by the difficulty in ensuring that the groups in different experimental conditions are 'equal in expectation', a critical requirement for mimicking randomization. The inherent diversity in educational systems and policies across countries makes this equivalence a challenging, if not impossible, task. Last, the measurement of 'appropriate outcomes' sensitive to the impact of treatment in ILSAs also poses a challenge. The variation in educational outcomes and how they are valued and measured across different cultures and education systems complicates the identification of universally applicable and sensitive outcomes to particular interventions.

We want to conclude by saying that, although we have very serious reservations about the ability to make causal inferences with ILSA data in most circumstances, we are not absolutists. First, a careful submission of a causal question to Willett and Murnane's evaluative framework is a pre-condition for conducting any such study. Then, as we described above, examples of high-quality quasi-experimental studies using educational data are possible; however, many use enriched data that include census or register data with more comprehensive information. We can imagine, then, that with this additional data, it becomes plausible to merge data from individual countries or small sets of highly homogeneous countries in a way that supports causal inference. A second possibility is in a context where a natural experiment occurred, such as a lottery or similar. Although, as noted in our critique of the school closure study, even natural experiments are not

silver bullets and, to the extent possible, a careful analysis of assumptions should be undertaken. A third avenue for improved inferences could be in extending ILSAs into repeated measures or longitudinal studies, to track the same students over time.

Finally, we are not arguing against *using* quasi-experimental designs with ILSA data. Indeed, careful application of these methods can ensure that some alternative explanations are excluded, or that possible confounding variables are included. In this challenging endeavor, rigorous quasi-experimental designs using ILSA data, when applied meticulously, can help to mitigate biases, and consider potential confounders, contributing towards more reliable conclusions. However, achieving unassailable causal inference remains a formidable task, underscoring the adage that in the realm of rigorous scientific inquiry, particularly in causal inference, easy answers are rare, and the most valuable insights often require the most arduous journeys.

Acknowledgements
We would like to express our gratitude to the journal editors and the anonymous reviewers for their insightful and constructive feedback. Additionally, our thanks go to Professor Christian Kjeldsen and Professor David Kaplan for their early reviews of our manuscript.

Author contributions
David Rutkowski paper conception and writing; Leslie Rutkowski: paper conception and writing; Greg Thompson: paper conception and writing; Yusuf Canbolat: writing.

Funding
Not Applicable.

Data availability
Not Applicable.

Declarations

Ethics approval and consent to participate
Not Applicable.

Consent for publication
Authors give consent for the paper to be submitted, reviewed and published. The paper is not under review with any other journal.

Competing interests
The authors have no competing interests.

Received: 8 February 2024 / Accepted: 7 March 2024
Published online: 01 April 2024

References

Abdulkadiroğlu, A., Pathak, P. A., & Walters, C. R. (2018). Free to choose: Can school choice reduce student achievement? *American Economic Journal: Applied Economics*, *10*(1), 175–206. https://doi.org/10.1257/app.20160634

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*(434), 444–455.

Athey, S., & Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, *31*(2), 3–32. https://doi.org/10.1257/jep.31.2.3

Athey, S., & Imbens, G. W. (2022). Design-based analysis in difference-in-differences settings with staggered adoption. *Journal of Econometrics*, *226*(1), 62–79. https://doi.org/10.1016/j.jeconom.2020.10.012

Braun, H. I., & Singer, J. D. (2019). Assessment for monitoring of education systems: International comparisons. *The ANNALS of the American Academy of Political and Social Science*, *683*(1), 75–92. https://doi.org/10.1177/0002716219843804

Carnoy, M. (2015). International test score comparisons and educational policy: A review of the critiques. *National Education Policy Center*. https://eric.ed.gov/?id=ED574696

Chin, M. J. (2023). School district consolidation in North Carolina: Impacts on school composition and finance, crime outcomes, and educational attainment. *Economics of Education Review*, *95*, 102432. https://doi.org/10.1016/j.econedurev.2023.102432

Chmielewski, A. K., & Dhuey, E. (2017). *The analysis of international large-scale assessments to address causal questions in education policy*. National Academy of Education. http://naeducation.org/wp-content/uploads/2017/06/ChmielewskiDhuey_Revision_04_06_2017_akc_web-version-1.pdf

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

Ehrenberg, R. G., Brewer, D. J., Gamoran, A., & Willms, J. D. (2001). Does class size matter? *Scientific American*, *285*(5), 78–85.

European Commission. (2018). *Outcomes and causal inference in international comparative assessments (OCCAM)*. European Commision. https://cordis.europa.eu/project/id/765400
Goldacre, B. (2013). *Building evidence into education*. *Department for Education London*. https://core.ac.uk/download/pdf/9983746.pdf
Hodgen, J., Adkins, M., & Ainsworth, S. E. (2023). Can teaching assistants improve attainment and attitudes of low performing pupils in numeracy? Evidence from a large-scale randomised controlled trial. *Cambridge Journal of Education*, *53*(2), 215–235. https://doi.org/10.1080/0305764X.2022.2093838
Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945. https://doi.org/10.2307/2289064
Imai, K., & Kim, I. S. (2019). When should we use unit fixed effects regression models for causal inference with longitudinal data? *American Journal of Political Science*, *63*(2), 467–490. https://doi.org/10.1111/ajps.12417
Jay, T., Willis, B., Thomas, P., Taylor, R., Moore, N., Burnett, C., Merchant, G., & Stevens, A. (2017). *Dialogic teaching: Evaluation report and executive summary*. https://shura.shu.ac.uk/17014/
Kennedy, A. I., & Strietholt, R. (2023). School closure policies and student reading achievement: Evidence across countries. *Educational Assessment Evaluation and Accountability*, *35*(4), 475–501. https://doi.org/10.1007/s11092-023-09415-4
Kennedy, A., Strello, A., & Strietholt, R. (2023). Methods for causal inference with observational data from international assessments. *Pre-Confrence Workshop*. https://www.iea.nl/news-events/news/irc-2023-pre-conference-workshops-announced
Komatsu, H., & Rappleye, J. (2021). Rearticulating PISA. *Globalisation Societies and Education*, *19*(2), 245–258. https://doi.org/10.1080/14767724.2021.1878014
Kraft, M. A. (2023). The effect-size benchmark that matters most: Education interventions often fail. *Educational Researcher*, *52*(3), 183–187. https://doi.org/10.3102/0013189X231155154
Lavrijsen, J., & Nicaise, I. (2015). New empirical evidence on the effect of educational tracking on social inequalities in reading achievement. *European Educational Research Journal*, *14*(3–4), 206–221. https://doi.org/10.1177/1474904115589039
Lortie-Forgues, H., & Inglis, M. (2019). Rigorous large-scale educational RCTs are often uninformative: Should we be concerned? *Educational Researcher*, *48*(3), 158–166. https://doi.org/10.3102/0013189X19832850
Martin, M. O., Mullis, I. V. S., & Kennedy, A. M. (2007). *Progress in international reading literacy study (PIRLS): PIRLS 2006 technical report*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.
Meldrum, M. L. (2000). A brief history of the randomized controlled trial: From oranges and lemons to the gold standard. *Hematology/Oncology Clinics of North America, 14*(4), 745–760. https://doi.org/10.1016/S0889-8588(05)70309-9
Murnane, R. J., & Willett, J. B. (Eds.). (2010). *Methods matter: Improving causal inference in educational and social science research*. Oxford University Press. https://books.google.com/books?hl=en&lr=&id=lA0qSsQk_AgC&oi=fnd&pg=PR5&dq=methods+matter&ots=mwfccCvBIf&sig=3FlzUWDg24iP2rAJ7iLZd6QuNj4
OECD (2012). *Public and Private Schools: How Management and Funding Relate to their Socio-economic Profile*. OECD. https://doi.org/10.1787/9789264175006-en
OECD (2014). *PISA 2012 Technical Report*. OECD Publishing.
Pearl, J. (2000). Models, reasoning and inference. *Cambridge UK: CambridgeUniversityPress*, *19*(2), 3.
Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys*, *3*, 96–146.
Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic books.
Perelman, S., & Santin, D. (2011). Vandenberghe. *Education Economics*, *19*(1), 29–49. https://doi.org/10.1080/09645290802470475
Pfeffermann, D., & Landsman, V. (2011). Are private schools better than public schools? Appraisal for Ireland by methods for observational studies. *The Annals of Applied Statistics*, *5*(3), 1726–1751. https://doi.org/10.1214/11-AOAS456
Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, *15*(3), 351–357. https://doi.org/10.2307/2087176
Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. https://doi.org/10.1037/h0037350
Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, *100*(469), 322–331. https://doi.org/10.1198/016214504000001880
Russo, F. (2009). Causality and causal modelling in the social sciences. *Springer Netherlands*. https://doi.org/10.1007/978-1-4020-8817-9
Schleicher, A. (2009). Securing quality and equity in education: Lessons from PISA. *PROSPECTS*, *39*(3), 251–263. https://doi.org/10.1007/s11125-009-9126-x
Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Company.
Sibieta, L., Greaves, E., & Sianesi, B. (2014). Increasing pupil motivation: Evaluation report and executive Summary. *Education Endowment Foundation*. Education Endowment Foundation. https://eric.ed.gov/?id=ED581249
Sims, S., Anders, J., Inglis, M., & Lortie-Forgues, H. (2023). Quantifying promising trials bias in randomized controlled trials in education. *Journal of Research on Educational Effectiveness*, *16*(4), 663–680. https://doi.org/10.1080/19345747.2022.2090470
Singer, J. D., & Braun, H. I. (2018). Testing international education assessments. *Science*, *360*(6384), 38–40. https://doi.org/10.1126/science.aar4952
Splawa-Neyman, J. On the application of probability theory to agricultural experiments. Essay on Principles. (, Dabrowska, D., & Speed, T. (1923). Trans.). *Statistical Science*, *1990*(5), 465–472.
Stone, D. A. (1989). Causal stories and the formation of policy agendas. *Political Science Quarterly*, *104*(2), 281–300. https://doi.org/10.2307/2151585
Vandenberghe, V., & Robin, S. (2004). Evaluating the effectiveness of private education across countries: A comparison of methods. *Labour Economics*, *11*(4), 487–506.

Publisher's Note