Rutkowski and Delandshere *Large-scale Assess Educ (2016) 4:6*
DOI 10.1186/s40536-016-0019-1

Large-scale Assessments in Education
a SpringerOpen Journal

RESEARCH

Open Access

Causal inferences with large scale assessment data: using a validity framework

David Rutkowski[1*] and Ginette Delandshere[2]

*Correspondence:
david.rutkowski@cemo.uio.no
[1] University of Oslo, Oslo, Norway
Full list of author information is available at the end of the article

Abstract

To answer the calls for stronger evidence by the policy community, educational researchers and their associated organizations increasingly demand more studies that can yield causal inferences. International large scale assessments (ILSAs) have been targeted as a rich data sources for causal research. It is in this context that we take up a discussion around causal inferences and ILSAs. Although these rich, carefully developed studies have much to offer in terms of understanding educational systems, we argue that the conditions for making strong causal inferences are rarely met. To develop our argument we first discuss, in general, the nature of causal inferences and then suggest and apply a validity framework to evaluate the tenability of claims made in two well-cited studies. The cited studies exemplify interesting design features and advances in methods of data analysis and certainly contribute to the knowledge base in educational research; however, methodological shortcomings, some of which are unavoidable even in the best of circumstances, urge a more cautious interpretation than that of strict "cause and effect." We then discuss how findings from causal-focused research may not provide answers to the often broad questions posed by the policy community. We conclude with examples of the importance of the validity framework for the ILSA research community and a suggestion of what should be included in studies that wish to employ quasi-experimental methods with ILSA data.

Background

Policy makers often express a need for *scientifically-based evidence* to articulate policy and make funding decisions (e.g., Raudenbush 2008; Stevens 2011; Sutherland et al. 2012). To partially address this need, the United States government, for example, has invested heavily in the What Works Clearinghouse, which attempts to bank educational research findings resulting primarily from randomized controlled trial (RCT) studies so that evidence-based decisions can be made by both policy makers and practitioners. Internationally, the Organisation for Economic Cooperation and Development (Henry et al. 2001) and the World Bank (Jones 2007) as well as conglomerations of many international players (Rutkowski and Sparks 2014) have all placed great focus on attaining scientifically-based evidence so that "evidence-based policy decisions" are possible.

To answer the calls for stronger evidence, educational researchers and their associated organizations increasingly demand more studies that can yield causal inferences. For example, one of the largest educational research organizations in the world, the American Educational Research Association (AERA), commissioned a report on estimating causal effects using observational data (Schneider et al. 2007). In this report, the authors state that "there is a general consensus in the education research community on the need to increase the capacity of researchers to study educational problems scientifically" (Schneider et al. 2007, p. 109). These same authors argue that large cross-sectional educational assessment datasets are an important and often underused resource from which educational researchers and policy makers can draw valid causal inferences. A prime example of the sort of datasets that can be brought to bear in the quest for scientifically-based evidence are international, comparative assessments, such as the Trends in International Mathematics and Science Study, among others.

In spite of a desire on the part of policymakers and researchers to use scientifically-based evidence in policy, research, and practice, there are necessarily limitations to gleaning causal inferences from observational data. It is in this context that we take up a discussion around causal inferences and international large-scale assessments (ILSAs). Although these rich, carefully developed studies have much to offer in terms of understanding educational systems, we argue that the conditions for making strong causal inferences are rarely met. To develop our argument we first discuss, in general, the *nature of causal inferences* and then in the following section titled *limitations of experimental and quasi-experimental design* we suggest and apply a validity framework to evaluate the tenability of claims made in two well-cited studies (Mosteller 1995; Schmidt et al. 2001). The cited studies exemplify interesting design features and advances in methods of data analysis and certainly contribute to the knowledge base in educational research; however, methodological shortcomings, some of which are unavoidable even in the best of circumstances, urge a more cautious interpretation than that of strict "cause and effect." The next section titled *usefulness for policy*, discusses how findings from causal-focused research may not provide answers to the often broad questions posed by the policy community. We conclude with examples of the importance of the validity framework for the ILSA research community and a suggestion of what should be included in studies that wish to employ quasi-experimental methods with ILSA data.

Nature of causal inferences

Causal inferences have primarily relied on so-called "gold standard" experimental designs, especially RCTs (Campbell and Stanley 1963; Cook and Campbell 1979; Shadish et al. 2002; Campbell Collaboration 2015; What Works Clearinghouse 2015; Shavelson and Towne 2002). In studies using experimental designs, cases are randomly assigned to treatment and control groups, with the treatment manipulation under the complete control of the researcher. These types of studies are common in the physical, medical and psychological sciences where environments are controlled in laboratories, allowing scientists to experiment with control and treatment groups to test hypotheses. Even in ideal RCT studies, however, researchers contend with threats such as attrition, experimenter training, and so on. Such experimental conditions are less common in the social sciences because randomization is often difficult (or impossible) and even unethical (e.g., randomly assigning students to low performing vs high performing schools). Given the logistical difficulties and ethical concerns of randomly assigning people to groups and controlling their environments, quasi-experiments are more common in the social sciences. In quasi-experiments random assignment is not possible and statistical control of pre-existing differences between groups has to be carefully exercised. Social science researchers also use data from natural experiments or correlational studies—that is, where observations of naturally occurring phenomena are used—to address causal questions. These natural experiments have been most prevalent in economics and epidemiology but also in political sciences (Dunning 2008). Increasingly, however, quasi-experimental methods and natural experiments are used in connection with ILSA data (Cattaneo and Wolter 2012; Jürges et al. 2005). We discuss this evolution next.

For many years educational researchers have made causal claims based on large-scale observational data (e.g., Coleman Report; National Educational Longitudinal Study 1988–2000; High School and Beyond—HS&B), and there is now a renewed interest in drawing causal inferences from the analysis of large-scale national and international assessment data (e.g., Schneider et al. 2007; West and Woessmann 2010; Woessman 2014). To do so, statisticians, economists and other social scientists have developed methods of analysis (e.g., instrumental variable approach, propensity scores, fixed-effect models) to devise conditions that emulate random assignment by selecting "equivalent" treatment and control groups across a number of non-treatment variables (e.g., matching groups). Such a strategy seeks to equalize differences between groups with the exception of membership to either the "treatment" or "control" group. Each strategy has its own limitations, some of which will be addressed in the remaining articles of this special issue. Here, however, we want to broadly focus on the nature of causal inferences based on large-scale assessment data, which are typically observational and cross-sectional.

In order to make causal claims researchers often have to limit the scope of the claim because it must focus on a particular cause and a particular effect in order to establish a relationship between them, holding *everything else equal*. Although a traditional laboratory offers unrivaled control (e.g., a carefully designed and implemented RCT), scholars have questioned the associated causal claims, pointing to the inherent simplicity of the claim (Cronbach 1982). And even proponents of causal inferences in social science research (Cook 2002) have recognized that such studies are best suited for addressing very simple and focused questions. As educational researchers expand their questions to cover more complex topics, these criticisms become all the more relevant, as the risk of violating important assumptions increases. We outline examples of typical violations later in the paper. First, however, we discuss an important distinction between *causal description* and *causal explanation.*

Shadish et al. (2002) explain the distinction between *causal descriptions*—which describe the consequence of varying the cause on the effect or causal relationship—, and *causal explanations*—which provides an account of "the mechanisms through which and the conditions under which that causal relationship holds" (p. 9). Further, Shadish et al. (2002) describes the strength of experiments as having an ability to describe the "consequences attributable to deliberately varying a treatment." The same authors argue that "experiments do less well in clarifying the mechanisms through which the conditions under which that causal relationship holds"—what they define as causal explanations (p. 9). This important distinction allows us to better understand what relationships are being defined by any causal statement. For many policy makers, understanding the mechanism is often of less importance whereas researchers are more likely to value uncovering the specific causal mechanisms or explanations that underlie causal descriptions. For example, if a study finds that spending more time on a subject leads to better performance on a test, such findings are only helpful for researchers if we understand what teachers who spend more time are doing differently than the teachers who spend less time on a topic. In other words, time is not the cause here but how the time is used. Yet in much of social science research this distinction is not clear. Instead, blanket causal statements are made without any further explanation.

Let us now examine the various factors that might affect the validity of causal claims. To do so, we mainly use a validity framework that has been developed and refined over several decades (e.g., Campbell and Stanley 1963; Cook and Campbell 1979; Shadish et al. 2002) in the context of experimental and quasi-experimental designs. We could have also placed our focus on Rubin's Causal Model (RCM) or framework of potential outcomes (Holland 1986; Rubin 1978, 2008) that similarly focuses on the analysis of cause in experiments but also extends to the use of observational studies for approximating randomized experiments. As Rubin (2008) states, however, "many of the appealing features of randomized experiments can and should be duplicated when designing observational comparative studies, that is, nonrandomized studies whose purpose is to obtain, as closely as possible, the same answer that would have been obtained in a randomized experiment comparing the same analogous treatment and control conditions in the same population" (pp. 809–810). RCM outlines conceptual and design considerations that might make it possible to use observational studies to approximate randomized experiments and also requires an explicit consideration of the "assignment mechanism" by which cases were assigned to the treatment and control conditions (Rubin 2008). For example, in a study comparing students attending public and private schools, it is the individual (or his/her parents) who is responsible for deciding on assignment to the treatment (e.g., private school) or to the control (e.g., public school) conditions. The RCM also makes use of a probability model associated with the assignment mechanism and of Bayesian analysis to consider the full set of potential outcomes for each case rather than relying on the observed outcome, which, according to Rubin (2008), is inadequate and "can lead to serious errors" (p. 813). RCM also rests on the important concept of "key covariates," or relevant variables that could explain pre-existing differences between the treatment and control conditions and that can be used to ensure that the distribution of these variables only differ randomly between the two groups—one of the crucial conditions to make causal inferences. In other words, this ensures that the cases in the treatment and control conditions are more or less equivalent on all important variables. Other design considerations relate to clear specification of the hypothesized experiment that is to be approximated (i.e., clear specification of treatment conditions and outcomes), adequacy of sample sizes under all conditions to ensure power, clear understanding of who made the treatment condition assignment and based on what variables, and measurement quality of key covariates (Rubin 2008). As we will see, many of these considerations are similar to and/or compatible with those we consider in the validity framework that we use and describe subsequently.

Drawing from the work of Campbell and Stanley (1963), Cook and Campbell (1979), and Shadish et al. (2002), which focuses on the validity and generalization of causal claims, in the context of randomized experiments, causal claims are affected by a number of factors: (1) the meaning and representation of the constructs related to the claims, and the consequences of using these for a particular purpose—*construct validity*, (2) the study design and proper specification of the model(s) used to test the causal hypothesis and the ruling out of alternative hypotheses—*internal validity*, (3) the sampling of cases, "treatment or cause," outcomes, and settings or contexts—*external validity*, and (4) the use of proper statistical methods to estimate the strength of the relationships between the presumed cause and effect—*statistical conclusion validity*. All these considerations are generally concerned with minimizing various "errors" in making causal claims: error of representation, error in logic of reasoning or in the implied mechanism that underlie the causal relationship, error in estimating differences or relationships, error in extrapolation. All these validity concerns support each other on the one hand, but also compete with one another on the other hand. For example, if the constructs used to represent the cause and effect are problematic in terms of their definitions and measurements, this will have consequences for all other aspects of validity and will seriously affect the validity of the causal claims. On the other hand, narrowing the definition of treatment conditions (or cause), for example, to better fit a particular context may enhance internal validity but may severely limit external validity or extrapolation. We use this validity framework subsequently to evaluate some causal claims.

An additional, yet related issue concerns the comparability of constructs and their measurement across different populations, referred to as construct and measurement equivalence, respectively. This issue is especially important in the international context where complex social phenomena are measured across differing contexts, cultures, and locations. To be fair, in some instances, when important social concepts manifest themselves differently in different countries, for example, (e.g., socio-economic status), accommodations are made regarding the measurement of some of these concepts. For example, adding national options to home background scales. This departure from strictly equivalent measures may enhance the meaningfulness of a concept in a particular context but also makes comparisons across contexts more problematic, since important concepts are measured differently.

Making causal claims using ILSA data, however, cannot simply be achieved by modifying the measurement of particular concepts or variables for a particular context. The broader issue is whether the causal mechanisms or causal explanations for a particular phenomenon are comparable across contexts (e.g., groups, countries)—an issue that has not received much attention in the literature on causal inference. The primary interest of most international assessment programs is in measuring educational achievement (defined and measured differently, depending on the study) and a pre-defined set of achievement correlates (e.g., the learning environment or the student's home situation).[1] The mechanisms that explain the relationships between these variables and the outcomes measures in the context of each country are rarely problematized or conceptualized and a priori differences in the conceptualization or theorization of these mechanisms across contexts are often not considered. For example, answers to questions regarding the value and purposes of education and schooling in societies as different as Germany, Qatar, and Zimbabwe are not fully articulated prior to designing data collection instruments. Such articulation would likely explain the mechanisms underlying

[1] Such large scale data collection efforts also suffer from a concern for trend analysis (across different waves of data collection) which prevents changes that would compromise comparisons across time.

variability in student achievement or performance in school and their differing associated correlates. In other words, one universal questionnaire administered in all contexts cannot possibly cover all relevant explanatory variables for *all* participating countries. We know, for example, that issues of gender and socio-economic status are understood differently in countries around the world. As such, the conceptualizations of the mechanisms that might explain varied educational achievement would likely yield a number of other variables not currently included in the data collection design.

In the case of ILSA, when a set of variables is imposed across contexts and is examined in terms of the relationships to an outcome variable, some relationships between the variables are bound to be found even if only by chance. The meaning of these relationships, however, is not often clear and plausible alternative causal relationships cannot be examined due to the limited set of variables and the absence of causal mechanisms conceptualized for different contexts. In the field of comparative policy analysis, Falleti and Lynch (2009) have emphasized the importance of causal mechanisms and their interactions with context in order to make credible causal explanations. They argue "that unless causal mechanisms are appropriately contextualized, we run the risk of making faulty causal inferences." They see the importance of context for making causal claims as a "problem of unit homogeneity" (p. 1144), where unit, here, refers to the variables and to the attributes of the units of analysis as well as their meaning and equivalence in the presumed causal mechanisms. So, for example, is a 14-year-old boy from an economically developed country who is picked-up 200 m from his house by the school bus "equivalent" to a 14-year-old boy from an economically developing country who has to walk to school barefoot for several miles every day? Does the same causal mechanism explain variability in achievement test scores? What are the relevant attributes that play out in these different contexts? These are the important and difficult questions that would need to be addressed instead of making comparisons on a universal set of variables, which may not universally apply.

The case of the boys from a developed and developing economy is most likely not a meaningful comparison but it serves here to illustrate the question of meaning and equivalence and the importance of examining these in articulating causal explanations or mechanisms and the setting in which they play out. In addition, the same causal mechanism may have different outcomes in different contexts while different causal mechanisms and multiple causes for the same effect may be at play in the same context. This can be explained by the fact that contexts are multilayered and develop from the interactions of individual characteristics as well as social and institutional norms, values, and functioning. The multiplicity of possible and plausible causal mechanisms defies the usefulness of single simplistic deterministic or probabilistic models as they can only provide a very partial description of "a" possible causal relationship. Most statistical models currently in use cannot readily accommodate the complexity of causal mechanisms, and it is, therefore, necessary to reduce their complexity to test a causal relationship (or hypothesis) between *a* cause and *an* effect. We contend, therefore, that a crucial issue in the nature of causal inference is their need for simplicity and lack of generalizability.

In the next section we outline some limitations to making causal claims in educational research by examining two popular educational studies and submitting them to the validity framework outlined above.

Limitations of experimental and quasi-experimental design

As we argued above, most research studies yielding causal inferences in the social sciences tends to be descriptive rather than explanatory. This is an important distinction as it sets expectations for the claims that can be made when using the results. Causal inferences are inherently linked to their validity and generalization—that is, how true, and how specific or universal are the claims? We illustrate some of these considerations and limitations, first in a randomized experimental study, and then in a study that employs ILSA data where random assignment is not possible. We examine possible reasons why causal claims might be weakened and highlight the importance of acknowledging possible validity threats so that claims can be qualified accordingly. In the end, it may be the case that an experimental or quasi-experimental design is the best choice for the research question at hand; however, the threats to validity may not support causal conclusions.

For our first example we draw from a well-known experimental study in education (Mosteller 1995), and discuss some validity concerns that may affect the study's claims. For the second example (Schmidt et al. 2001), we focus on a study that uses large-scale assessment data to examine the effect of curriculum coverage on achievement gain scores. This example was chosen because it is one of the first instances where researchers explored the Trends in International Mathematics and Science Study (TIMSS) data to design a study that uses matching groups to calculate gain scores and "sophisticated statistical techniques [that presumably] allows them to generate causal hypotheses concerning specific aspects of the curriculum on student learning" (p. 80). Additionally, both of these studies were chosen because they are reasonably straightforward to explain in a limited space, which is not often the case when non-experimental studies are used with the aim of making causal inferences. Although we offer only a brief review of two studies, in this special issue many of the included papers provide resources and examples of how their corresponding topic has been used in research.

Study 1: Tennessee class size study

In the often cited Tennessee Study of Class Size in the Early School Grades (Mosteller 1995), early grades are defined as kindergarten through third grade and the treatment conditions are (1) smaller classes of 13–17 students, (2) larger class size of 22–25 students, and (3) larger class size of 22–25 students with a teacher aid in the classroom. Students and teachers were randomly assigned to classes at least for the first year of implementation. To participate in this study schools had to commit for a period of 4 years, have a minimum of 57 students in each of the targeted grade levels, and guarantee that no new textbooks or curricula would be introduced. Approximately 180 Tennessee schools volunteered to participate, 100 qualified for the study and 79 ultimately participated in the first year of implementation in kindergarten. Achievement in reading and mathematics were measured using the Stanford Achievement Test (SAT) and the Tennessee Basic Skills First (TBSF) which is described as a "curriculum-based measure." The differences between the students in smaller and larger class (with and without aide) are reported as effect sizes (differences in means divided by standard deviation) and range between .13 and .27 in mathematics and between .21 and .23 in reading.

From the basic study description, we raise two issues. First, setting cut-offs for determining smaller and larger classrooms is not discussed or justified. Second, the possible causal mechanism of the effect of class size on achievement is limited to the explanation that in smaller classes there are fewer distractions and the teacher has more time to attend to each child than in larger classes. Next, we consider the main conclusion that "[t]he evidence is strong that smaller class size at the beginning of the school experience does improve performance of children on cognitive tests" (p. 123).

We begin by examining the nature of the constructs related to cause (class size) and effect (performance on cognitive tests). Although class size is the variable manipulated in this study, it appears, from the brief explanation provided, that it is a proxy for number of distractions and teacher time spent with each child. Yet there is little information in the study about what teachers were doing in the smaller and larger classes during implementation in terms of working individually with children and minimizing distractions. In addition, although one can easily imagine differences between a class of 13 and a class of 25; it is not entirely clear how a class of 17 and a class of 22 vary. This is an example where the explicit articulation of the causal mechanism(s) that would explain the differences in achievement for students in smaller and larger classes and evidence to support that these mechanisms have actually taken place is missing. For example, it is possible that, since all treatment conditions were implemented in the same school, teachers assigned to larger classes might have felt some resentment or that teachers assigned to smaller class size felt re-energized, which, in both cases, could have affected their teaching performance and, in turn, the performance of their students. These reactions to treatment assignment, in essence, affect the construct of class size being tested in the study. Without an explanation of the causal mechanisms and supporting evidence, attributing differences in achievement to class size is potentially misleading. With regard to the effect of the cause, achievement is measured in mathematics and reading by two different tests—a general standardized test (SAT), and a standardized curriculum-based test (TBSF), which is presumably more sensitive to the Tennessee school curriculum. Given the necessarily limited nature of these measures—in just two content areas—we raise the possibility that other learning constructs should also be measured and compared before any decisions are made on the basis of the evidence from this study.

Next, we highlight issues around the study design and treatment manipulation that relate directly to *internal* and *construct validity* elements that may affect the validity of the study's claim. In particular, the definition of treatment conditions—in this case, class size—is related to treatment manipulation. In this study, in addition to smaller and larger class sizes, the researchers included a treatment condition of larger class size with a teacher aide, presumably to understand whether an additional adult might confer the same advantage as a smaller class. Unfortunately, the role of the aides was not specifically defined; some aides engaged in instructional activities while others did not. The absence of data about what happened in the different treatment conditions during implementation makes it difficult to explain what about class size makes a difference in achievement or to rule out alternative explanations for the differences in achievement between treatment conditions. Taken as a whole, the study provides some evidence of the relationship between class size and student achievement. But threats to both construct and internal validity prevent strong causal inference and an understanding of the actual causal mechanism. The study does appear to establish a causal description that applies to the context of the particular treatment conditions, outcome measures, time, persons, and settings used in this study.

Beyond the identified issues, questions remain about the generalization of the causal relationship—*external validity*—to variations in treatment conditions, test scores, persons and settings. For example, one issue concerns the representativeness of the 79 participating schools out of 180 that initially volunteered. Eighty schools were eliminated because they did not meet the study criteria for participation, including smaller schools with fewer than 57 students per grade. Further, of the 100 schools that qualified, only 79 participated in the first year, raising questions about the degree to which participating schools might have differed from non-participating Tennessee schools. Although Mosteller states that "The study findings apply to poor and well-to-do, farm and city, minority and majority children" (p. 116), he also reports that the effect of class size on achievement was twice as large for minority students. In addition, how generalizable are the findings to other states that might differ in important ways, including education policies, funding structures, and curricula? Finally, the definition of class size is relative to a particular setting, leaving open the possibility of different outcomes if the cut-offs had been defined as *fewer than 20* and *more than 20*, for example. Given what is currently known internationally, findings from this study, assuming they still hold today, should also be reconciled with other reports where some of the highest achieving countries (e.g., Japan, Korea) also have large class sizes. Looking at each of these issues as part of a whole brings us back to the need for clearly articulated causal mechanisms that would explain the presence or absence of a relationship between these two factors and the importance of the interaction with the setting in which this plays out. Although, practically speaking, no study will cover all design and implementation possibilities, we want to highlight the importance of both grounding study decisions in theory to the degree possible while also tempering claims that are associated with potential threats to validity.

In addition to possible problems with construct, internal, and external validity already discussed, another limitation relates to statistical inferences—*statistical conclusion validity*—regarding the co-variation of the cause and the effect. In this study the co-variation between class size and achievement is addressed by testing mean differences between the different treatment conditions. Systematic differences between the smaller and larger class size conditions are implied although Mosteller does not directly report any test of statistical significance, but focuses rather on the magnitude of the effect sizes. Given the large sample sizes we can reasonably assume that differences between means were indeed statistically significant. Differences in effect sizes were observed when different reading and math tests are used; but it is difficult to make sense of these differences without information about homogeneity of variance, psychometric quality of the measures, and confidence intervals about estimated effect sizes.

With regard to the reliability of treatment implementation, Mosteller, reports that after the first year of the study some "incompatible children" were moved from smaller to larger class size, which might have increased the differences between means and effect sizes. The author also reported some "class size drift" with some smaller classes becoming larger and some larger classes becoming smaller than their initial limits, possibly reducing differences between treatment conditions. This study exemplifies how, even when random assignment to treatment conditions is possible, a number of important issues can threaten the validity of researchers' causal claims.

Study 2: TIMSS 1995 curriculum and learning study

As a second example, we turn now to an observational context using ILSA data where, importantly, random assignment to treatment conditions is not possible. The highlighted study (Schmidt et al. 2001, also analyzed in the AERA report Schneider et al. 2007) investigates the relationships between different aspects of curriculum, instruction, and learning using the 1995 TIMSS data. In what follows we consider the degree to which issues related to measurement, sampling, and model choice and specification might raise questions around some of the researchers' claims. To be clear, according to the AERA report, this study does not claim to make causal inferences but rather to "conceptually model and statistically evaluate the potential causal effects of specific aspects of the curriculum on student learning" (p. 84)—a subtle distinction that will likely be overlooked by policy makers who might want to make use of these claims. Further, Schmidt et al. (2001) refer to the model they use as a "causal statistical model" (p. 164).

In this study the researchers use the TIMSS data from approximately 30 countries to construct a *quasi-longitudinal design* in order to examine the impact of different aspects of curriculum on learning. The conceptual analysis for the study focused on the *intended curriculum* (represented by content standards and textbooks ratings), *implemented curriculum* (represented by textbook ratings and self-reported teacher content coverage), and *attained curriculum* (represented by student achievement gain or increase in percentage of correct items)—all ratings and percentages aggregated at the topic and country levels. The data used were collected at the end of the school year in "two adjacent grades containing the majority of thirteen-year-olds in each country" (p. 5) and consist of school and teacher questionnaires, curriculum documents (e.g., content standards, curriculum guides), textbooks, and achievement test results in mathematics[2] and science, covering a large number of topics and sub-topics. By sampling from the two adjacent grades (e.g., 7th and 8th grades in the US), the researchers constructed presumably equivalent groups to estimate achievement gain scores as the difference in percentage of items correct between the two grades averaged over all items in a topic. Curriculum documents and textbooks were divided in blocks and qualitatively coded to capture the representation and content coverage. These codes were then quantified to characterize the national curriculum for each country relative to twenty mathematics topics included in the quantitative analyses.

Information was collected on how many lessons teachers devoted to specific topics translated into the percentage of teachers in each country addressing topics along with the percentage of instructional time allocated to each topic. At the end, the measures of content standards, curriculum coverage, teacher coverage, instructional time, and achievement gains were aggregated at the topic and country level in order to make the measurement consistent across all variables. This consistency in the level of measurement was perceived as making the impact on achievement gain more sensitive to variability in content and instructional time and coverage. The structural relationships

[2] In this analysis we only focus on mathematics and on the cross-country analyses to illustrate the claims made and the limitations of the study.

between these constructs were then examined. The hypotheses tested in the model were that the "official" curriculum documents (developed at the national, regional or local level) or *content standards* would have an impact on *textbook coverage* used in the classroom, and a direct and indirect relationship to *teacher coverage*, *instructional time* and *achievement gain*. *Textbook coverage* was also hypothesized to have a direct effect on *instructional time* and *teacher coverage* and, through those variables, an indirect effect on *achievement gain,* in addition to a direct effect.

In what follows, we apply the same validity framework to analyze several issues that might justify tempered claims from the Schmidt et al. (2001) study. First, construct measurement is of particular concern. The challenge of retro-fitting the definition and measurement of constructs is inherent to ILSA data that are often collected for different purposes and is not a problem unique to the Schmidt, et al. study. Further, most of the limitations regarding the measurement of the constructs used in this study were appropriately acknowledged by the researchers. For example, the researchers recognized that they are working from a very specific definition of curriculum, that is, one among several possible perspectives. The operationalization of these curriculum definitions and their measurement is another concern.

With regard to coding content standards and textbook ratings, there was considerable variability in document availability across countries—some had multiple documents and textbooks while others had just one. Further, there is only limited information on the rating system and the reliability of the ratings. Coding and rating were initially done at the sub-topic level and then aggregated at the topic and national level. Given that the availability of documents varied across countries, questions naturally arise regarding the meaning and comparability of these concepts and whether they represent well the intended or implemented curriculum. The aggregation procedures (at the national level) also leave open the possibility of committing ecological fallacy—that a relationship that exists at a higher level does not exist or is in a different direction at a lower level. And, as was the case in the class size study, there is no articulation of the causal mechanism that would explain the relationship between content standard national ratings and average gain percent correct response on the items for a particular topic. Documents were also coded relative to the TIMSS frameworks (or *world core curriculum*), which means that topics not included in the frameworks were not coded and therefore not taken into account in the study. Nevertheless, these unaccounted for topics were part of the *intended curriculum* for the different countries but were excluded from the analyses. The study did not describe whether other coding schemes were considered and whether these different approaches would yield different ratings. Such an approach—often referred to as a sensitivity analysis—would have eliminated, to some extent, competing explanations for observed differences.

As in the class-size study, outcomes are limited in scope to measures of achievement in math and sciences. Further, the number of items per topic is a possible area of concern. For example, in mathematics, the study estimates achievement gains for twenty different topics. Due to the nature of the data, 14 of the 20 topics only had 5–10 items, raising concern around construct representativeness. An additional issue worth highlighting surrounds the issue of comparability across countries. For example, researchers extensively describe meaningful variation between countries in the way and the level at which the curriculum is articulated and structured. Further, the researchers found large cross-country differences in the perceived influence that the curricular structure has on the implemented curriculum in schools. Countries varied in terms of topic coverage according to content standards, by textbooks, by teachers, and in instructional time allocated to the different topic. Although it is natural to expect cross-national variation in these relationships, the authors fit an overall structural model to understand the relationship between *content standards, textbook coverage, teacher coverage* (and *instructional time*) for all countries. A key assumption in such an approach is that the constructs are all understood and measured equivalently in each analyzed country (Millsap 2011). The authors do allow for country-by-topic interaction effects, which are further examined by fitting models to each country individually. Nevertheless, these by-country models assume that the same variables and "causal mechanisms" are at play in all countries and that, as we have already mentioned, these variables have the same meaning across different contexts, even though the study found major differences in how the curriculum is structured in different countries.

Causal relationships are based on the principal of *ceteris paribus* or "all else being equal"; however, given evidence to the contrary, the tenability of this assumption is difficult to defend. Therefore, the researchers' general claim that "more curriculum coverage of a topic area—no matter whether manifested as emphasis in content standards, as proportion of textbook space, or as measured by either teacher implementation variable—is related to larger gains in the same topic area" (Schmidt et al. p. 261) can only be descriptive, and is conditional on the definition and measurement of these variables in the particular study. Indeed, the authors recognize the descriptive nature of their claims when they qualify with the following: "the nature of the general relationship is not the same for all countries. This implies that a general relationship between achievement gain and one aspect of curriculum may not even exist at all for some countries." (p. 261). Needless to say, this raises issues regarding the generalizability and the causal nature of their claims.

We also point to a few possible issues related to statistical conclusion validity, particularly as they pertain to cross-country comparability. For each of 29 countries, five regression analyses (one for each pair of coding variables) are fit to the data. Topic ($N = 20$) served as the unit of analysis and results are combined and presented together (see Table 8.1, pp. 274–275). Although the practice of presenting findings for dozens of countries is fairly common in ILSA research, extra care is warranted when the inferential target is causal. These analyses are followed by another 29 regression analyses (one per country) to simultaneously estimate the structural coefficients (direct effects only) between three of the curriculum variables and achievement gains (see Table 8.2, pp. 277–278). Ten countries appear to have a significant ($p \leq .05$) structural coefficient between *textbook coverage* and *achievement gain*; three countries have a significant coefficient between *content standards coverage* and *achievement gain*; and six countries have a significant coefficient between *instructional time* and *achievement*. The reported overall coefficients of determination for these 29 models range from .08 to .70. Given the variability in the findings and questions around comparability, the strength of causal claims arising from this study should likely be revisited.

Usefulness for policy

Policy makers in modern democratic societies often look for causal inferences from the research community to support the perception of "objectivity" in decision making. As Stone (2011) has argued, ideas around objectivity and the subsequent need for "causal theories" is critical to the policy process even if the resulting decisions are not truly based on the causal information provided by the research community. In education, a number of countries, including the US, have invested a great deal of resources in programs like the What Works Clearinghouse and continue to invest in the development and promotion of research that focuses on making causal claims using ILSA data.[3] Although the US government has viewed experiments as the "gold-standard" for social research since the 1950s it is significant that they are now investing resources in promoting research aimed at causal inferences with cross-sectional international assessment data. Much of this is in response to the fact that experimental research is often not feasible in the social sciences; however, policy makers want clear and compelling "evidence" for policy making and distributing resources. As such, in the current manuscript our intention is to acknowledge that experiments and quasi-experiments have an important place in educational research but also to argue that the results from such research are often narrowly focused and rarely succeed in providing answers to larger questions that are most relevant to the policy making process. In what follows we discuss causal inferences in the policy context in light of our previous points.

Policy makers often ask questions in broad terms (e.g., *How can we improve student achievement*? *What are effective instruction, programs, and curricula*? *How do we improve teacher quality*? *How can we close the achievement gaps?*) (Huang et al. 2003) whereas researchers often address narrower questions (e.g., *What is the effect of a particular pedagogical approach on students standardized test scores in mathematics and language arts*? *What is the effect of teacher retention on future student achievement? What is the effect of class size on student achievement*?) due to methodological considerations and data limitations. Under the best conditions, even when using methods that focus on making causal claims, answers to most researchers' questions are qualified and limited in scope. The need to qualify findings and/or limit their scope can be attributed to measurement problems, selection bias, and the lack of ability to control for all relevant variables. When, and if, we are able to attend to the majority of these threats, especially with ILSA data, resulting claims are often limited in scope and not often suitable to address larger policy-focused questions. For example, simply creating a policy that mandates closer alignment to TIMSS will probably not improve test scores if the teachers do not teach the material or if the curriculum becomes too vast to be covered in 1 year. The necessarily narrow focus of most research aimed at causal inferences, especially with ILSA data, unfortunately creates a landscape where learning and achievement are presented as a highly simplified problem. In other words, findings that are overly narrow and do not account for the known complexity of our educational systems can misguide policymakers by ignoring complex interactions just as much as they can inform them about educational systems.

[3] With support from the U.S. National Science Foundation in 2015 AERA held workshops on making causal claims with ILSA data. More information can be found here: http://www.aera.net/ProfessionalOpportunitiesFunding/AERAFundingOpportunities/StatisticalAnalysis-CausalAnalysisUsingInternationalData/tabid/14751/Default.aspx.

In this paper we have also shown some serious and some minor threats to construct, internal, external and statistical conclusion validity, drawing from two well-known, often cited examples in educational research. Each identified problem leads us away from clear explanatory causal claims and can even point to serious concerns about our ability to make descriptive causal claims. The distinction between explanatory and descriptive causal claims is important in part because most of causal-focused research in education emphasizes descriptive claims; however, the same research falls short of articulating valid causal explanations. In the case of the Tennessee study, limitations aside, the findings only provide the causal description that a lower student/teacher ratio leads to higher achievement scores on select assessments for a sample of teachers and students in Tennessee. This is another example of how research aimed at making causal inferences is often focused and narrow in nature. To policy makers, however, findings from a class size study might seem like useful information. Unfortunately, it most likely does not provide the key information policymakers need to create general policies to reduce class size. Few of us in the field of education are naïve enough to believe that simply putting more teachers into classrooms will increase scores. In fact, having poorly qualified teachers in classrooms has been associated with lower scores on standardized assessments (Darling-Hammond 2000). As such, the information that would best inform policy is not simply knowing that we need more teachers but also a clear explanation of what teachers do in small classroom that results in increased student understanding and performance. Hence, we need to know the conditions under which the causal relationship holds.

Another example of how a lack of clear explanation can lead to a range of policy prescriptions can be taken from ILSAs. As ILSAs have grown in both popularity and scope national policy makers have taken a keen interest in identifying policy levers that can improve educational achievement, with ILSA results serving as a frequent pool from which to draw possible solutions. Both class size and curriculum have drawn the attention of policy makers. The recent US discussions around widening income gaps in general and the impact on educational achievement in particular is another example of an important policy issue upon which ILSA data can be brought to bear. An historical and clearly problematic approach to measuring socioeconomic status (SES) in ILSA studies is via the "books in the home" indicator. This single item expresses rough quantities of the number of self-reported books in a child's home. In conjunction with causal models and methods, it is possible to identify what appears to be a "causal effect" of SES (as measured by the number of books in the home) on achievement. Indeed, there is often a fairly strong, positive association between number of books and achievement; however, the explanation for this relationship is unclear and therefore it is important that we also depend on a very clear and well-reasoned argument from the researcher who employed the causal modeling. For example, is the mere presence of books in the home sufficient to stimulate interest in reading, which translates into improved achievement? Or is the number of books serving as a proxy for cultural possessions and indicating a better resourced home environment? In the absence of any clear and well established explanation, the findings are not even causally descriptive, and of limited usefulness for enacting meaningful policy, where possible policies could range from wealth redistribution to providing books for children with fewer resources.

An important barrier to supporting causal explanations in the ILSA context also includes the design of the studies (e.g., they are cross-sectional and observational). *Prima facie* these design features do not lend themselves to causal explanations. And although quasi-experimental methods *can* overcome this barrier, a host of validity assumptions *must* be tested before causal explanations are supported. For example, except in limited cases such as the Schmidt et al. (2001) study, it is very difficult to know if the cause precedes the effect. Not being able to provide such information to policy makers greatly reduces the usefulness of any causal claims being made. Although the findings might be able to suggest to a policy maker that a causal relationship exists, the claim does not provide policy makers with the breadth of information that they need to enact change.

As we have argued, even though many policy makers emphasize the need for causal inferences to support "objective" policy decisions, the reality of the policy process is much more complex and influenced by a host of social and political values and interests. That said, as a research community we often embrace uncertainty and operate with caution as we move forward with research, especially when using ILSA data. Although policy makers' and researchers' goals are not mutually exclusive, since they both aim to improve education, the two groups approach problems differently and, as a result, often have different objectives for the findings. Understanding and communicating these differences will be an important step as we move forward with more causal modeling of ILSA data.

Conclusion

Experimental and quasi-experimental designs play a key role in developing an understanding of important phenomena in educational research. In fact, we contend that many such studies assist us in better understanding our educational system and also allow both policy makers and researchers to explore different types of questions. For example, each author in this special issue provides interesting examples of how a given quasi-experimental design or method can be used in educational research to explore important topics and, to a certain degree, eliminate alternative explanations for identified relationships. Nevertheless, even in an ideal setting, where subjects are randomly assigned to treatment and control groups, threats to the validity of inferences are persistent and should be recognized when interpreting findings. In quasi-experimental research, these issues are even more prevalent given the fact that no random assignment has occurred and only an approximation of this process is possible. Finally, regardless of whether subjects are randomized, it is important to recognize the narrow focus of most research studies that aim at making causal inferences as well as the critical difference between causal descriptions and explanations. With this in mind, we offer a few recommendations to researchers who are using and interpreting the results of research that attempts to approximate randomized experiment using ILSA data.

Paying attention to both *conceptual* and *design* considerations is key to approximating randomized experiments with ILSA data. A clear articulation of the causal mechanism(s) investigated would greatly enhance the design of a study by highlighting the important elements that should be included in the design as well as those that would allow for testing alternative explanations across different contexts. We agree with Rubin (2008) and contend that ILSA researchers should make use of a probability model associated with the assignment mechanism as well a Bayesian analysis to consider the full set of potential outcomes for each case. This process would strengthen resulting inferences that can be made from the research. Researchers should also justify their selection of what Rubin termed "key covariates" and ensure that they only differ randomly between control and treatment groups. Again, "key covariates" can only be identified if researchers have carefully articulated the possible mechanisms that could explain the association between the constructs and events being investigated. Further, all research should clearly articulate: specification of the treatment and outcome condition; sample sizes that ensure acceptable power; who made the treatment condition assignment and based on what variables; and providing evidence of the measurement quality of the key covariates. Within ILSA research each of these poses its own set of challenges. For example, missing rates and disagreement between students and parents on identical, policy relevant variables has been shown to be high (see Rutkowski and Rutkowski 2010). Similarly, scale reliabilities can vary widely between countries, from high to unacceptably low (see Rutkowski and Rutkowski 2013).

The validity framework offers important considerations that the research community can use to further minimize errors when quasi-experimental designs are used with ILSA data. Reminding ourselves that error exists throughout the entire research process and reasoning helps clarify that statistical techniques alone do not establish causal claims. In other words, sound statistical conclusion validity does not lead to an acceptable causal claim unless it is supported by a compelling causal mechanism that has been clearly articulated and taken into account in the design of the study. When designing a study that approximates a randomized experiment using ILSA data issues of construct, internal, and external validity are of critical importance, as we have illustrated in the context of the two studies we discussed earlier. As we have noted, there are a number of ways to examine the validity of findings in relation to experimental and quasi-experimental studies. In this paper we depended largely on the framework that was first developed by Campbell and Stanley (1963) and later refined by Cook and Campbell (1979) and Shadish et al. (2002). This framework allowed us to productively examine the validity of claims made by two well cited studies in education. As such, we recommend that all quasi-experimental studies that use ILSA data: (1) choose an established validity framework to work from; and (2) clearly explain threats to the validity of their claims. For example, if the validity framework outlined in this paper is chosen the study should include a discussion of: *construct validity*, *internal validity*, *external validity*, and *statistical conclusion validity*. Shadish et al. (2002) provide a detailed description of possible threats to validity and are a useful resource for both researchers and the reviewers of these studies.

Finally, we would like to point out some issues that are especially important given the design of ILSA data collection. The following issues do not constitute an exhaustive list but are simply examples where the validity of inferences may be threatened. With respect to construct validity, we are always at the mercy of the available data. That is, the validity of the claims made about a self-efficacy construct, for example, rests on the availability of sufficient variables or items to meaningfully represent this construct. Defending this proposition is the responsibility of the researcher, using self-efficacy literature as well as a thorough psychometric investigation of the data and providing supporting evidence for the validity of the claims made. Similarly, a primary issue with internal validity relates to model specification and ensuring that all relevant variables have been included in the model to support a thorough investigation of the hypothesized causal mechanism(s) and to make possible ruling out alternative hypotheses for an estimated effect. Again, based on a thorough examination of the substantive literature, the researcher is responsible for evaluating whether possible (reasonable) alternative explanations can be tested and whether relevant variables are included in the model. Given that ILSA data only provides a fixed set of variables researchers need to be transparent about the variables that were not included and have an open discussion about how that weakens their conclusions. Regarding external validity, a key issue in ILSA data is the operationalization of the outcome variable(s) which relies on appropriate and relevant but necessarily limited measurements. That is, one cannot reasonably use ILSA data to estimate the causal effect of some variable on "schooling" or "education" writ large, since most ILSAs are limited to only a few schooling outcomes (e.g., math, science, reading, and affective variables) and a highly selective sample of students (i.e., 8th graders or 15 year olds).

Of course, even RCTs can fail to meet ideal conditions (e.g., the Tennessee study). In ILSA research, there will always be further threats and more justifications will be necessary to allay concerns surrounding the validity of causal claims. Clearly articulated and reasonable research questions, well specified research design consistent with hypothesized causal mechanism(s), relevant and quality data, and well-specified models continue to be critical to support research claims made in this context. It is equally important to be thorough and transparent in acknowledging weaknesses in the causal chain of inferences as well as other limitations. As such, we urge everyone who works with ILSA data, and especially in applying "causal models" to ILSA data, to openly engage in a thorough and self-critical process that utilizes a well-recognized validity framework such as the one we discussed in the current paper. Through this process, we can have honest conversations about what the data and models can reasonably tell us about educational inputs, processes, and outcomes. We can also better engage with policy makers about the usefulness and limitations of research claims to inform policy.

Authors' contributions
The work represents extensive collaboration and discussion between DR and GD. Both authors participated in the development of the rationale for the study as well as the writing. Both authors read and approved the final manuscript.

Authors' information
David Rutkowski is a professor of educational assessment at the Center for Educational Measurement (CEMO) at the University of Oslo, Norway. David's research is focused in the area of educational policy and technical topics within international large-scale assessment and program evaluation. His interests include how large scale assessments are used within policy debates, the impact of background questionnaire quality on achievement results, and topics concerning immigrant students at the international level.
Ginette Delandshere is a professor of Inquiry Methodology and Chair of the Counseling and Educational Psychology Department in the School of Education at Indiana University, Bloomington. Her research interests are measurement and assessment and the associated validity of inferences and research claims as well as the study of the socio-political practice of assessment and its purpose and meaning in the context of teaching and learning.

Author details
[1] University of Oslo, Oslo, Norway. [2] Indiana University, Bloomington, IN, USA.

Competing interests
The authors declare that they have no competing interests.

Received: 12 January 2016 Accepted: 20 January 2016
Published online: 26 February 2016

References

Campbell Collaboration. (2015). *The Campbell collaboration: What helps? What harms? Based on what evidence?* Retrieved 21 July 2015, from http://www.campbellcollaboration.org/.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand McNally.

Cattaneo, A., & Wolter, S. C. (2012). *Migration policy san boost PISA results: Findings from a natural experiment* (SSRN Scholarly Paper No. ID 1999328). Rochester: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=1999328.

Cook, T. (2002). Randomized experiments in educational policy research: a critical examination of the reasons the educational evaluation community has offered for not doing them. *Educational Evaluation and Policy Analysis*, *24*(3), 175–199.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis for field settings*. Chicago: Rand McNally.

Cronbach, L. J. (1982). Prudent aspirations for social inquiry. *The social sciences: Their nature and uses, 61*, 81.

Darling-Hammond, L. (2000). Teacher quality and student achievement. *Education Policy Analysis Archives, 8*, 1.

Dunning, T. (2008). Improving causal inference: Strengths and limitations of natural experiments. *Political Research Quarterly, 61*(2), 282–293.

Falleti, T. G., & Lynch, J. F. (2009). Context and causal mechanisms in political analysis. *Comparative Political Studies, 42*(9), 1143–1166.

Henry, M., Lingard, B., Rizvi, F., & Taylor, S. (2001). *The OECD, globalisation and education policy*. Published for IAU Press, Pergamon. Retrieved from http://www.lavoisier.fr/livre/notice.asp?id=OOSWOKA2KK6OWG.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistics Association, 81*, 945–970.

Huang, G., Reiser, M., Parker, A., Muniec, J., & Salvucci, S. (2003). Institute of education science findings from interviews with education policymakers. Institute of Education Sciences. Retrieved from http://eric.ed.gov/?id=ED480144.

Jones, P. W. (2007). *World Bank financing of education: Lending, learning and development*. New York: Routledge.

Jürges, H., Schneider, K., & Büchel, F. (2005). The effect of central exit examinations on student achievement: Quasi-experimental evidence from TIMSS Germany. *Journal of the European Economic Association*, *3*(5), 1134–1155. http://doi.org/10.1162/1542476054729400.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York: Routledge.

Mosteller, F. (1995). The Tennessee study of class size in the early school grades. *The Future of Children, 5*(2), 113–127.

Raudenbush, S. W. (2008). Advancing educational policy by advancing research on instruction. *American Educational Research Journal, 45*(1), 206–230.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics, 6*(1), 34–58.

Rubin, D. B. (2008). Objective causal inference, design trumps analysis. *The Annals of Applied Statistics, 2*(3), 808–840.

Rutkowski, L., & Rutkowski, D. (2010). Getting it better: The importance of improving background questionnaires in International Large-Scale Assessment. *Journal of Curriculum Studies*, *42*(3), 411–430. http://doi.org/10.1080/00220272.2010.487546.

Rutkowski, D., & Rutkowski, L. (2013). Measuring socioeconomic background in PISA: One size might not fit all. *Research in Comparative and International Education, 8*(3), 259–278.

Rutkowski, D., & Sparks, J. (2014). The new scalar politics of evaluation: An emerging governance role for. *Evaluation*, *20*(4), 492–508. http://doi.org/10.1177/1356389014550561.

Schmidt, W. H., McKnight, C. C., Houang, R. T., Wang, H., Wiley, D. E., Cogan, L. S., et al. (2001). *Why schools matter: A cross-national comparison of curriculum and learning*. New York: Wiley.

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, R. J. (2007). *Estimating causal effects using experimental and observational designs*. Washington: American Educational Research Association.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin Company.

Shavelson, R. J., & Towne, L. (2002). *Scientific research in education*. Washington: National Academy Press.

Stevens, A. (2011). Telling policy stories: An ethnographic study of the use of evidence in policy-making in the UK. *Journal of Social Policy, 40*(2), 237–255.

Stone, D. (2011). *Policy paradox: The art of political decision making* (3rd ed.). New York: W. W. Norton & Company.

Sutherland, W. J., Bellingan, L., Bellingham, J. R., et al. (2012). A collaboratively-derived science-policy research agenda. *PLoS One,*. doi:10.1371/journal.pone.0031824.

West, M. R., & Woessmann, L. (2010). Every catholic child in a catholic school: Historical resistance to state schooling, contemporary private competition and student achievement across countries. *The Economic Journal, 120*(546), F229–F255.

What Works Clearinghouse. (n.d.). Homepage. Retrieved July 21, 2015, from http://ies.ed.gov/ncee/wwc/.

Woessman, L. (2014). *The economic case for education* (No. 20). European Expert Network on Econmics of Education.