Rutkowski *Large-scale Assess Educ (2016) 4:8*
DOI 10.1186/s40536-016-0023-5

Large-scale Assessments in Education

EDITORIAL Open Access

Introduction to special issue on quasi-causal methods

Leslie Rutkowski*

*Correspondence:
leslie.rutkowski@cemo.uio.no
University of Oslo, Oslo, Norway

Background

Over the past 50 years, ILSAs have experienced marked development in terms of scope, participants, and sophistication. The first such study, the International Association for the Evaluation of Educational Achievement's Pilot Twelve Country Study was completed in 1961 and measured achievement in math, reading, geography, science, and so-called non-verbal ability (Forshay et al. 1962). Reported achievement was based on fairly simple statistics and item response theory and related methods were relative new-comers and not often used in educational research. This seemingly modest accomplishment of measuring educational achievement in six subjects and twelve countries is really monumental when we consider the state-of-the-science at the time: there was no internet or email and calculating regression parameter estimates could take up to 24 h (Ramcharan 2006). In contrast, current international assessments use highly sophisticated designs, with rotated booklets, complex estimation methods, and computerized platforms for administration; participating systems number in the dozens; and national policy makers wait with bated breath for the results of each study cycle. And commensurate with these developments, a natural interest in understanding variation in achievement has emerged. Perhaps more importantly, researchers and policy makers want to know what, if anything can be done to improve achievement overall and for particular groups of test takers. This, in turn, has motivated interest in making connections between a host of potential causes and achievement; although it is important to note that this interest isn't strictly limited to the achievement domain.

In this special issue of *Large-Scale Assessments in Education*, we offer several papers on the topic of causal inferences with international large-scale assessment (ILSA) data. The papers here are primarily empirical analyses of ILSA data that feature different methods, all with an aim toward estimating the effect of some cause. Each paper also includes a brief introduction to the method at hand along with a discussion of the important statistical assumptions that underpin each method and whether or not the assumptions are plausible in the given circumstance. Although the treatment is selective, the methods featured here are commonly used in practice and serve as a useful introduction to specific methods of data analysis applied in a quasi-experimental context. Important to keep in mind is that the implemented methods are applied to observational data, most of which are also cross-sectional. The final paper (Rutkowski and Delandshere, this issue)

balances the empirical offerings by taking a critical perspective on drawing causal inferences through the lens of a validity framework.

Causality

The concept of causality dates back to ancient Greece and, among the great scholars, Aristotle is known to have written in detail about *forms* of causation (Mulaik 2009). Later, Descartes, Locke, and Hume (among others) also contributed significantly to modern thinking around notions of causality. And although neither this introduction nor this special issue go into meaningful detail around the history or theories of causality, it is often useful to take a moment and consider the current place in time that we find ourselves and the conversations that surround us. As such, it is of utmost importance to appreciate that perspectives on causality differ considerably (Heckman 2005; Holland 1986; Pearl 2009; Rubin 1974) and different traditions dictate what is an admissible cause (Bollen 1989 vs. Holland 1986). As Heckman notes, causality is naturally intuitive but difficult to define. Regardless of the causal camp in which researchers fall, it is generally agreed that there should be *a cause* and *an effect*. At a minimum, causal inferences rely on several conditions; including isolation (that no other variable can reasonably be a cause of another variable), association (that for A to cause B, there should be an association between A and B), and direction of influence (that A causes B and not the other way around; Bollen). Hume (2012), in a similar vein, stated three conditions known as a *regularity theory* of causality: constant conjunction (that if A causes B, we should always experience B when we experience A); temporal priority (that if A causes B, A should precede B); and contiguity (that if A causes B, then A and B should be spatially and temporally adjacent). Each of these ideas can be challenged and, in fact, temporal priority and contiguity are in tension with one another (Mumford and Anjum 2013). Nevertheless, Hume's ideas generally withstand the test of time and are regarded as important contributions to our modern understanding around causality (Holland 1986; Mulaik 2009). In the realm of international assessment, some, if not all of these conditions, can be difficult to establish. This is largely due to the nature of the research design and that the data are not collected with causal questions in mind (see Kaplan, this issue, for a discussion of the later point).

In clinical settings, the gold-standard for making causal claims is the randomized-control trial (RCT; Meldrum 2000), where subjects are randomly assigned to treatment and control groups. Indeed, the U.S. Institute of Education Sciences took the same perspective in 2003, providing guidelines for determining the rigor of evidence, with the RCT taking center-stage as the standard for "scientifically-based research." And in the U.S. the *what works clearing house* privileges well-designed and executed RCTs as the only study design with the possibility of *meeting standards without reservation* (IES 2014). Meeting the *gold-standard* in social sciences and in education, specifically, can be unobtainable for a host of reasons. Randomization into treatment groups might be logistically difficult or even impossible (e.g., adding additional full-time teachers to treatment classes), unethical (e.g., withholding free lunch from needy children), or prohibitively expensive (e.g., large-scale redesign of classrooms).

Quasi-experimental strategies

Empirical quantitative analyses in education often aim to reveal a *cause-and-effect* relationship, even if explicitly the methods are only correlational. Consider a relatively recent example from my own work (Rutkowski et al. 2013), where my colleagues and I used multilevel modeling to understand whether an association exists internationally between bullying and educational achievement. Although we managed to show a consistent negative relationship between these two constructs across many countries, we did not provide evidence that reducing bullying victimization would *cause* better mathematics achievement. But, certainly, we would have liked to make this claim. A number of features of the research design and analyses imposed challenges to making causal claims about this relationship, including the fact that the data (2007 Trends in International Mathematics and Science Study) were cross-sectional and observational. And using our chosen methods, we had established an association; however, we could not be certain of the direction of the relationship. We were also left wondering if there could be another common cause (something that causes both bullying victimization and achievement) that was omitted from our model. And all of these doubts arose in spite of the fact that we were careful in reviewing the literature to select and include covariates that could play such a role. Indeed, no amount of literature review could ensure that we were safe to assume a causal relationship. The moral of this story: working with extant data necessarily brings with it natural limitations. Importantly (but not exhaustively), subsequent analysts have no control over the available variables, leaving the possibility that an important confounder or common cause is not available for inclusion. Given the research design(s) typically used in ILSAs, it is indeed a real challenge to definitively establish the directionality of a relationship. Isn't it equally plausible that (higher) math achievement might *cause* (less) bullying victimization? So, what is the typical social sciences researcher to do with such data?

Thanks to work on potential outcomes and the counterfactual theory of causality—stemming from the early ideas of Hume (e.g., Rubin 1974)—a number of quasi-experimental strategies have emerged that (under a set of fairly stringent assumptions) offer the possibility of estimating a causal effect from data that arises from non-experimental research designs. Kaplan (this issue) provides an overview of the potential outcomes framework (also referred to as the Rubin model or the Neyman–Rubin model, in honor of the Polish statistician, Jerzy Neyman, who is first attributed with formalizing the idea of potential outcomes; Splawa-Neyman 1923). Here, this causal model emphasizes the *what if* aspect of a sequence of events. What would we have observed if we could see the outcomes for one subject that had received *both* the treatment and the control? Of course, this is impossible in practice and is referred to as the *fundamental problem of causal inference* (Holland 1986 p. 947). Rubin's causal model places emphasis of the *effects of the cause* and permits an estimate of the *average* causal effect of the treatment over a population of subjects. Importantly, observable information from different subjects can be used to inform us about the causal effect of the treatment.

Using Rubin's causal model as a foundation, some approaches rely on natural experiments. Take, for example, a 2007 retrospective study that used variation in adoption of lead reduction policies across U.S. states to establish a causal connection between childhood lead exposure and subsequent state crime levels (Wolpaw Reyes 2007). Clearly, this

study capitalized and relied on variations in policy implementation that were outside of the researcher's control. Further, the case for a causal effect of lead reduction on crime relies on the assumptions that there is no other explanation for declining crime rates than reduced lead exposure. Ruling out other nuisance variables can be exceedingly difficult to establish in practice, particularly when the researcher is relying on extant observational data.

With regard to international assessments and surveys, natural experiments also occur, at least at the system level (e.g., nation or other sub-national participants). For example, take two relatively similar countries (Norway and Sweden) and consider a situation where Sweden chooses to privatize their educational system while Norway chooses not to follow suit. Under certain assumptions, we can treat this situation as a natural experiment and estimate the difference in some outcome between the two countries. As one example, Rosén and Gustafsson (this issue) use country-level trend data to pursue a kind of *differences-in-differences* approach, whereby they estimate the effect of home computer use on achievement. An advantage of this method is that there is some control over omitted variables at the country-level that did not change over time (e.g., compulsory education age, teacher training requirements). Importantly, however, a weakness of this method is the parallel trend assumption—that nothing changed in one country that did not also change in another country. In general, this is a fairly challenging assumption to test or support with confidence. Nevertheless, the findings are encouragingly in line with other studies and the article is a useful introduction to this particular differences-in-differences approach.

From another perspective, Hogrebe and Streitholt (this issue) use propensity score matching—a classic instantiation of the Rubin causal model—to investigate the effect of not attending preschool on subsequent reading literacy. Matching methods appeal to the desire for *unit homogeneity* (Holland 1986) or that the characteristics of two groups differ only with respect to belonging to the control or treatment group. If subject groups can reasonably be regarded as homogeneous on relevant covariates, average differences on the outcomes could be attributed to the treatment. Propensity score matching represents the counterfactual theory of causality in that the sample of subjects is comparable on all observed covariates. A weakness of this approach is that treatment assignment can be explained by unobserved (or unobservable) covariates, producing biased estimates of the average treatment effect.

Where matching might overlook some unobserved confounder, Pokropek (this issue) implements an instrumental variables approach for dealing with potentially endogenous treatment variables in an analysis of the effect of having a neighborhood friend in the class on achievement. Endogeneity refers to the situation mentioned above—that treatment assignment and the outcome are simultaneously correlated with a variable not included in the model. Take, for example, an analysis that seeks to estimate the effect of attending private school, making *private school attendance* the treatment. But it is reasonable that socioeconomic status is correlated with both achievement and whether or not a student attends private school, thus creating endogeneity in the treatment (and a biased estimate of the average treatment effect). Although a good instrument—a variable that is correlated with the outcome *only through the treatment variable*—goes a long

way to correcting for endogeneity, identifying a strong, plausible instrument can be a real challenge.

Kaplan (this issue) takes an altogether different approach and considers the problem of making causal inferences in international assessment through a Bayesian lens. In doing so, the author aligns with the three previous papers in that he opts for the Rubin causal model. A particularly interesting aspect of this contribution is that Kaplan argues that organizations charged with conducting international assessments must first prioritize causal questions in the development phase of the study. Although this is a lofty aim, indeed, an effective program of querying causal questions rests on a study design that places this sort of inquiry alongside other goals. And, in fact, I would argue, this sort of prioritization will certainly help fulfill the other policy and research aims of international studies (i.e., understanding the context and correlates of achievement).

The final paper in this set (Rutkowski and Delandshere, this issue) provides a critical and complementary perspective on causal claims. The authors acknowledge a growing demand for causal inferences in the ILSA context; however, they also advocate a healthy dose of caution in interpreting effects as *causal*. This is accomplished by applying a validity framework to the matter at hand—that of determining the tenability of causal claims. This perspective is, to my mind, welcome in the current atmosphere, where the risk of making inappropriate causal inferences carries potentially high costs to children, their teachers, and their schools.

In sum, the following is a non-comprehensive presentation of several common methods used to draw causal inferences from non-experimental data that are, under several non-trivial conditions, suitable for analysis with ILSA data. Importantly, the degree to which causal conclusions are valid and plausible rests entirely on the degree to which these conditions are met. Although all inferences are subject to the same scrutiny, it is especially important in the case of causal inferences, given that they are more likely to be used as the basis for action or to inform policy. Now, without further ado, I sincerely thank the authors for their valuable contributions, the reviewers for their constructive feedback, and Justin Wild for his organizational support early in the development of this issue. And I hope that you enjoy reading this special issue as much as I enjoyed editing it.

Authors' information

Leslie Rutkowski is a professor of Educational Measurement in the Centre for Educational Measurement at the University of Oslo; e-mail: leslie.rutkowski@cemo.uio.no. Her research is focused in the area of international large-scale assessment from both methodological and applied perspectives. Her interests include latent variable models for international assessments, heterogeneous measurement models, and the impact of background questionnaires on achievement results.

Received: 9 February 2016 Accepted: 22 April 2016
Published online: 06 May 2016

References

Bollen, K. (1989). *Structural equations with latent variables*. New York: Wiley.

Forshay, A. W., Thorndike, R. L., Hotyat, F., Pidgeon, D. A., & Walker, D. A. (1962). *Educational achievements of thirteen-year-olds in twelve countries: Results of an international research project, 1959–1961*. Hamburg: UNESCO Institute for Education. Retrieved from http://iea.nl/pilot_twelve-country_study.html.

Heckman, J. J. (2005). The scientific model of causality. *Sociological Methodology, 35*(1), 1–97. doi:10.1111/j.0081-1750.2006.00164.x.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*(396), 945. doi:10.2307/2289064.

Hume, D. (2012). *A treatise of human nature*. Oxford: Oxford University Press (reprinted from original, 1739). Retrieved from http://www.gutenberg.org/files/4705/4705-h/4705-h.htm.
IES. (2014). *What works clearinghouse procedures and standards handbook* (3.0 ed.). IES National Center for Education Statistics. Retrieved from https://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v3_0_standards_handbook.pdf.
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. doi:10.1016/S0889-8588(05)70309-9.
Mulaik, S. (2009). Causation. *Linear causal modeling with structural equations* (pp. 63–110). Boca Raton: Chapman & Hall/CRC Press.
Mumford, S., & Anjum, R. L. (2013). *Causation: a very short introduction*. Oxford: Oxford University Press.
Pearl, J. (2009). Causal inference in statistics: an overview. *Statistics Surveys, 3*, 96–146.
Ramcharan, R. (2006). Regressions: why are economists obsessed with them? *Finance and Development*, 43(1). Retrieved from http://www.imf.org/external/pubs/ft/fandd/2006/03/basics.htm.
Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), 688–701. doi:10.1037/h0037350.
Rutkowski, L., Rutkowski, D., & Engel, L. (2013). Sharp contrasts at the boundaries: school violence and educational outcomes internationally. *Comparative Education Review, 57*(2), 232–259. doi:10.1086/669120.
Splawa-Neyman, J. (1923). On the application of probability theory to agricultural experiments. *Essay on Principles. Statistical Science, 1990*(5), 4.
Wolpaw Reyes, J. (2007). Environmental policy as social policy? The impact of childhood lead exposure on crime. *The Journal of Economic Analysis Policy, 7*(1), 1–41. doi:10.2202/1935-1682.1796.