Causal Inference With Large-Scale Assessments: Challenges and Opportunities

David Kaplan

Department of Educational Psychology

THE UNIVERSITY of WISCONSIN MADISON

IEA IRC, Cape Town, 2015

- The outline of the talk is as follows:
  1. The potential outcomes framework
  2. Possible approaches to causal inference with large-scale assessments
  3. Implications for large-scale assessment design.
  4. Challenges and opportunities

Potential Outcomes Framework

- The dominant framework of causal inference in the education sciences is the potential outcomes framework of Rubin (1974) based on earlier work of Neyman (1923).
- Start by defining a selection variable S that assigns a unit i (e.g. a student) who is a member of population to either a treatment condition, $T = 1$ or a control condition, $T = 0$.
- In randomized experiments, S is created by the experimenter, but in observational studies such as LSAs, assignment to a treatment condition often occurs naturally.
- In the RCM, the critical characteristic is that the value S_i for each individual could potentially be different.

- The role of the outcome variable Y in the RCM is also crucial to the framework.
- First, for the variable Y to measure the effect of the cause, Y must be measured (or presume to occur) post-exposure - that is after exposure to the treatment.
- The value of the post-exposure outcome variable must be a result of either the cause t or the cause c defined on a particular student.
- Therefore, the RCM conceives of the same student providing an outcome after being exposed to the treatment, Y_{1i} or after being exposed to the control Y_{0i}.

- The causal effect defined within the RCM framework is then the difference between Y_1 and Y_0 for student i.
- That is for individual i, the goal, ideally, would be to observe the individual under receipt of the treatment and under non-receipt of the treatment.
- This defines the *potential outcomes* framework for causal inference and can be expressed formally as

$$Y_i = T_i Y_{1i} + (1 - T_i) Y_{0i}, \tag{1}$$

- Holland (1986) points out that the potential outcomes framework has a serious problem – namely, it is rarely possible to observe the values of Y_0 and Y_1 on the same individual i, and therefore rarely possible to observe the effects of $T = 1$ and $T = 0$.

- Holland refers to this as

> The Fundamental Problem of Causal Inference

- A statistical solution to the Fundamental Problem offered by Holland (1986) is to make use of the population of individuals.

- In this case, the *average causal effect*, can be defined (relative to the control group) as the expected value of the difference between Y_1 and Y_0 over the units in the population – viz.

$$T = E(Y_1) - E(Y_0). \tag{2}$$

- To quote Holland (1986),

> "The important point is that the statistical solution replaces the impossible-to-observe causal effect of T on a specific unit with the possible-to-estimate average causal effect of T over a population of units".

- More from Holland (1986)

> “Put as bluntly and as contentiously as possible... I take the position that causes are only those things that could, in principle, be treatments in experiments. The qualification, “in principle” is important because practical, ethical, and other considerations might make some experiments infeasible, that is, limit us to contemplating *hypothetical experiments*”.

and

> “No causation without manipulation”

- LSAs do not involve random assignment of participants to conditions
- What methods can we use to warrant causal claims in the context of LSAs?
- We consider *propensity score analysis*

- An implication of the RCM is that because we are unable to observe the outcomes of an individual under both treatment and control we need to find individuals in both groups that serve as each others counterfactuals.
- Thus, in order to warrant causal inferences in the setting of LSAs, individuals in treatment conditions should be matched as closely as possible to those in the control condition on observed pre-treatment assignment variables.

- Consider the effect of pre-primary education attendance on reading achievement using data from PIRLS 2011.
- To warrant the claim that pre-primary education attendance increases reading achievement, a researcher would need to find children who attended pre-primary education who are as similar as possible to those children who did not attend pre-primary education on characteristics that might lead to selection into pre-primary education or not.
- These characteristics should have been measured (or hypothetically present) before the child's selection into pre-primary education (e.g. parental socio-economic status).

- Rosenbaum and Rubin (1983) proposed propensity score analysis as a practical tool for reducing selection bias through balancing treatment and control groups on measured covariates.

- Consider the potential outcomes model in (1). Under this model, the probability that individual i receives the treatment can be expressed as

$$e_i = p(T = 1|Y_{1i}, Y_{0i}, Z_i, U_i), \tag{3}$$

where U_i contain unobserved covariates. Notice that in an LSA, (Y_{0i}, Y_{1i}, U_i) are not observed. Thus, it is not possible to obtain the true propensity score. Instead, we estimate the propensity score based on covariates z. Specifically,

$$\hat{e}(Z) = p(T = 1|Z), \tag{4}$$

which is referred to as the *estimated propensity score*.

- The estimated propensity score $\hat{e}(Z)$ has many important properties. Perhaps the most important property is the *balancing* property, which states that those in $T = 1$ and $T = 0$ with the same $\hat{e}(Z)$ will have the same distribution on the covariates Z.

- Formally, the balancing property can be expressed as

$$p\{Z|T = 1, \hat{e}(Z)\} = p\{Z|T = 0, \hat{e}(Z)\}, \quad (5)$$

or equivalently as

$$T \perp Z|\hat{e}(Z). \quad (6)$$

Implementation of the Propensity Score

- There are four approaches that are commonly used in implementing the propensity score:
  1. Stratification on $\hat{e}(Z)$,
  2. Propensity score weighting,
  3. Optimal full matching, and
  4. Propensity score regression.
- Bayesian approaches that account for uncertainty in the propensity score equation are also now available (Kaplan & Chen, 2013).

Implications for the Design of Large-Scale Assessments

- We did not exhaust the range of possibilities for causal inference in large-scale assessments, e.g.
- Other possible methods include
  1. Causal mediation analysis
  2. Instrumental variable estimation
- However, these rest on very strong assumptions also.
- Our approach is pragmatic. There many different types of causal questions and there is no “one size fits all methodology (see.e.g. Cartwright’s *Hunting Causes and Using Them*).

- But with all of this potential confounding, how do we warrant causal claims?
- We argue that statistical warrants for causal claims about a real or hypothetical treatment are always made within the context of a specific set of observed and unobserved explanatory variables measured before the onset of the treatment.
- Mackie suggests that the problem in distinguishing between conditions and causes is addressed by considering that causes take place in a context, or what Mackie (1974, pg. 35) refers to as a *causal field*.

What is said to be caused, then, is not just an event, but an event-in-a-certain-field, and some conditions can be set aside as not causing this-event-in-this-field simply because they are not part of the chosen field, though if a different field were chosen, in other words if a different causal question were being asked, one of those conditions might well be said to cause this-event-in-that-other-field.

- What advice can be given for the design of context questionnaires to support causal inference?

  1. The causal variable should reflect an actual event in the life of the respondent (e.g. pre-primary education attendance) that is relevant to the policy purposes of the survey.
  2. The causal variable should encode a counterfactual statement – a hypothetical manipulation that could have occurred in a real-life experiment.
  3. Reliable reporting of exposure to conditions. We can't follow the bodies.
  4. Additional variables should be obtained that represent confounders within the relevant causal field.
  5. Sensitivity to assumptions should be obtained whenever possible.

- This paper argues for a theoretical framework for causal inference in LSAs. We recognize that additional support for basic research on causal inference with large-scale assessments is needed.

  1. It is important to study precisely how causal variables can be reliably measured and used in statistical models such as those described in this paper.
     - The field-trial stage of an LSA operation can play a role here.
  2. Alternative frameworks and methods for causal inference should be studied in terms of their value in the context of large-scale assessments.

- The hope is that this paper stimulates a broader discussion of the challenges and opportunities of causal inference with large-scale assessments.

THANK YOU