Abstract Title Page

Title: Partial Identification of Treatment Effects: Applications to Generalizability

Authors and Affiliations:
Wendy Chan, Northwestern University

Abstract Body

Background / Context:
Results from large-scale evaluation studies form the foundation of evidence-based policy. The randomized experiment is often considered the gold standard among study designs because the causal impact of a treatment or intervention can be assessed without threats of confounding from external variables. Treatment randomization strengthens the internal validity, or the extent to which causal conclusions can be drawn, of the study results (Campbell, 1957). Policy-makers have become increasingly interested in the external validity of results, or the extent to which the results from a study generalize to a population of interest. Randomization of treatment addresses the question of expected treatment effects within a study, but probability sampling is needed to generalize the results from the sample to the target population of inference. However, recent research revealed that sites in large scale experiments are often not randomly sampled, thereby threatening the external validity of their results (Olsen, Orr, Bell, & Stuart, 2013). Stuart, Cole, Bradshaw, & Leaf (2011) and Olsen, et al. (2013) developed methods to assess the similarity between the sample and population for generalization. Hedges & O'Muircheartaigh (2011) and Tipton (2013) proposed re-weighting methods to adjust for differences between the given sample and target population. These re-weighting methods utilize propensity scores to model selection into the sample given observable covariates.

Purpose / Objective / Research Question / Focus of Study:
The absence of probability sampling causes estimates of population parameters to be biased. Model-based methods used to derive bias-reduced estimators often require multiple assumptions, many of which are untestable with the data. While point estimation is a desirable goal in most research studies, it is important to distinguish between inferences based on the data alone and inferences based on strong assumptions. Interval estimation or partial identification offers an alternative approach to estimating treatment effects without the need for assumptions. In this paper, we derive partial identification bounds on the expected treatment effect for the population under three different frameworks: 1) Worst-case framework with no assumptions on the data; 2) Treatment optimization where subjects select the treatment that optimizes the expected outcome and; 3) Monotone treatment response where the response function is assumed to be weakly increasing in the treatment. Under each framework, we analyze the impact that additional, but weak, assumptions have on the width of the estimation bounds and discuss important differences between the bounds for the expected treatment effect for the sample and those for the treatment effect for the population. We derive the bounds using the original sample and the stratum-specific bounds where the strata are delineated by the propensity scores for generalization.

Significance / Novelty of study:
Current methods to improve generalizability from non-random samples have focused on deriving point estimates for the population. To date, partial identification or interval estimates have not been explored in the field of generalizability.

Statistical, Measurement, or Econometric Model:
Propensity score methods for generalization require several assumptions to be made on the sample and population. The three primary assumptions needed are the stable unit treatment value assumption (SUTVA), strong ignorability of treatment assignment, and strong ignorability of

sample selection (Stuart, et al. (2011), Tipton (2013)). Under SUTVA, the potential response of each individual is independent of both the treatment assignment mechanism and the sample selection mechanism. Furthermore, SUTVA stipulates that the potential responses be independent of the treatment assignment and sample selection assignment of other individuals. Under each ignorability assumption, the potential responses are independent of the treatment and sample assignment mechanism when conditioning on the propensity scores. When these assumptions are met, re-weighting methods such as subclassification and inverse propensity weighting (IPW) are used to derive bias-reduced estimators of the population average treatment effect (PATE).

Partial Identification

Although the three assumptions needed to use propensity scores for generalization allow point estimation of the PATE, they are untestable because they require knowledge about the distribution of potential responses in both the sample and population. The credibility of inferences is often tied to the strength of the assumptions made in deriving the results. The concept of partial identification has been developed extensively through the work of Manski (1999, 2003, 2009) and Horowitz and Manski (2000). Given a dichotomous treatment T, continuous outcome Y, treatment assignment vector Z, and sample selection vector W, the potential outcome under a given treatment is denoted by Y(T). Here, Z=1 (0) indicates whether the individual received treatment (control) and W=1 (0) indicates whether the individual was selected (not selected) into the sample. Each potential outcome can be written as:

$$E(Y(1)) = E(Y(1)|Z=1)\,P(Z=1) + E(Y(1)|Z=0)\,P(Z=0)$$
$$E(Y(0)) = E(Y(0)|Z=1)\,P(Z=1) + E(Y(0)|Z=0)\,P(Z=0)$$

The sample average treatment effect (SATE) is given by the difference between these two potential outcomes. In the context of generalizability, the PATE is estimated using the same difference, but each potential outcome now has two additional terms:

$$\begin{aligned} E(Y(1)) = {} & E(Y(1)|Z=1, W=1)\,P(Z=1, W=1) + E(Y(1)|Z=0, W=1)\,P(Z=0, W=1) + \\ & E(Y(1)|Z=1, W=0)\,P(Z=1, W=0) + E(Y(1)|Z=0, W=0)\,P(Z=0, W=0) \end{aligned}$$

$$\begin{aligned} E(Y(0)) = {} & E(Y(0)|Z=1, W=1)\,P(Z=1, W=1) + E(Y(0)|Z=0, W=1)\,P(Z=0, W=1) + \\ & E(Y(0)|Z=1, W=0)\,P(Z=1, W=0) + E(Y(0)|Z=0, W=0)\,P(Z=0, W=0) \end{aligned}$$

For each set of potential outcomes, terms such as E(Y(1)|Z=0, W=1) are considered counterfactual outcomes. Since each individual receives at most one treatment, these counterfactuals are unobservable because they represent outcomes under a treatment condition that was not assigned. When the outcome of interest, Y, is bounded, these unobservable counterfactuals are replaced by the bounds of Y in order to derive estimation bounds on the treatment effect. A lower bound for the PATE is given by the difference between the lower bound of E(Y(1)) and the upper bound of E(Y(0)). The upper bound for the PATE is then the difference between the upper bound of E(Y(1)) and the lower bound of E(Y(0)). Tighter bounds are possible with additional assumptions on the response function and treatment selection process.

Research Design:

In generalizability studies, administrative data on the population, such as demographic and socioeconomic information, is used in the propensity score model. In this paper, we derive the estimation bounds for the PATE using the administrative data of the population, combined with the information on the outcome and sample. The potential outcomes used to estimate the PATE are a function of the indicator variables, Z and W. An important difference between the bounds for the SATE and those for the PATE is that the latter contain potential outcomes that are unobservable from not being selected into the experiment. In particular, since treatment is only administered to units selected into the sample, the potential outcomes under each treatment condition for units not selected into the sample are unobservable.

We derive estimation bounds for the PATE with a bounded outcome, Y, under three scenarios. The first, which we refer to as the "full" form, estimates the bounds by treating all of the counterfactuals as unobserved and replacing them with the lower and upper bounds for Y. These are referred to as the "worst-case" bounds. Since the PATE includes terms related to units not selected into the sample, the bounds for the PATE are necessarily wider than those for the SATE. The second framework, referred to as the "reduced counterfactual" scenario, assumes that one of the counterfactuals, $E(Y(0)|Z=0,W=0)$, is identified. This counterfactual outcome represents the expected outcome under the control condition for units that were assigned to control and were not selected into the sample. Under the "reduced counterfactual" case, the distribution of these potential outcomes is assumed to be observable through the use of the administrative data. The rationale for this assumption is that if the control condition represents the "business as usual" condition, then units not selected into the sample realize this outcome by not participating in the experimental study. The last framework, the "reduced" framework, uses the same assumption as the "reduced counterfactual" case, but makes a restriction that units not selected into the sample *do not* receive treatment. Specifically, the counterfactual $E(Y(1)|Z=1,W=0)\ P(Z=1,W=0)$ is treated as zero so that the estimation bounds for the PATE are tighter. This framework would apply to cases where an experimental study was terminated after one iteration so that the probability of subjects participating in future iterations of the study is zero.

The worst case bounds reflect the simplest situation where no assumptions are made on the data. For each of the scenarios described, we considered two additional frameworks for the estimation bounds. The first imposes an additional assumption on treatment selection. In studies where units select the treatment that optimizes their outcome, tighter bounds are possible. An example of this scenario would be job seekers who choose the job that pays the highest wage. Treatment optimization imposes an upper bound on two of the unobservable counterfactual outcomes so that these upper bounds are substituted for the worst case bounds. The second framework makes the assumption that the response function, Y, is weakly increasing (monotone) in the treatment, T. With this assumption, the lower and upper bounds of the PATE are functions of the realized outcomes under each treatment condition.

Usefulness/Applicability of Method:

We provide an example of estimation bounds using data from a cluster randomized controlled trial (Konstantopoulos, Miller & van der Ploeg, 2013). This experiment analyzed the effect of a benchmark assessment system on student achievement in English Language Arts (ELA) and

mathematics. The effect of this system was assessed using the Indiana Statewide Testing for Educational Progress-Plus (ISTEP+) scores. The experimental study consisted of fifty-six K-8 schools who volunteered to implement the system, of which 34 were randomly assigned to the state's benchmark system while 22 served as control schools. In the treatment schools, students were regularly given formative assessments aligned with the state test and their teachers received feedback on their performance as a way to dynamically guide their instruction in the periods leading up to the state exam.

The question of interest was, if the Indiana benchmark assessment system was implemented in all schools throughout the state, what is the expected treatment effect in terms of students' ELA and mathematics achievement? To answer this question, the original population of 1,587 schools was re-defined to exclude all charter schools, schools with over 95 percent male students, with over 95 percent English Language Learners (ELL), with over 95 percent of students with special education needs and schools with fewer than 100 students. These schools were removed because they were not considered as similar to the experimental sample and subsequent analyses was based on the resulting population of 1,514 schools and sample of 54 schools. A propensity score model was fitted to the data and the population was stratified into three equal-sized strata, seen in **Figure 1**. One of the strata contained only two schools from the sample so that the stratum specific estimate of the PATE would have a larger standard error. Estimates of the PATE using subclassification and IPW, as well as the unweighted estimate (without any re-weighting) are given in **Table 1**. Note that all three estimates are not statistically significant at the $\alpha = 0.05$ level.

To investigate the partial identification bounds of the PATE, we chose 4th grade ELA scores as the outcome of interest and re-defined the population again to only include 4th grade serving schools. For each school, we considered a discrete version of the ELA scores using the ISTEP+ cutoff scores of Pass Plus, Pass, and Not Pass. The outcome, Y, was therefore binary with Y=1 if the school received a "Pass" and Y = 0 if the school received a "Not Pass." The partial identification bounds under the "full," "reduced counterfactual" and "reduced" scenarios for the worst case, treatment optimization, and monotone treatment response frameworks are given in **Table 2**. The stratum-specific bounds for each combination is given in **Table 3**.

Findings / Results:

Compared to the point estimates of the PATE, the partial identification bounds were consistent with the insignificant treatment effect finding. This consistency with the point estimates was also seen in the stratum-specific bounds of the PATE. Furthermore, the partial identification bounds were derived with few to no assumptions on the empirical evidence.

Conclusions:

Partial identification offers an alternative method of deriving information on treatment effects, but with fewer assumptions on the data. The example of this paper demonstrates that the estimation bounds may provide nearly the same information as point estimates in terms of the significance of the treatment effect. Partial identification methods are important when the credibility of inferences rely on strong untestable assumptions.

Appendices

Appendix A. References

Campbell, D.T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54*(4), 297-312.

Hedges, L.V. and O'Muircheartaigh, C.A. (2011). Improving generalizations from designed experiments. Northwestern University. Manuscript submitted for publication.

Horowitz, J. L. and Manski, C.F. (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American Statistical Association, 95,* 77-84.

Konstantopoulos, S., Miller, S.R., and van der Ploeg, A. (2013). The impact of Indiana's system of interim assessments on mathematics and reading achievement. *Educational Evaluation and Policy Analysis, 35,* 481-499.

Manski, C.F. (1999). *Identification problems in the social sciences.* Cambridge, MA: Harvard University Press.

Manski, C.F. (2003). *Partial identification of probability distributions*. New York, NY: Springer Science & Business Media.

Manski, C.F. (2009). *Identification for prediction and decision.* Cambridge, MA: Harvard University Press.

Olsen, R.B., Orr, L.L., Bell, S.H., and Stuart, E.A. (2013). External validity in policy evaluations that choose sites purposively. *Journal of Policy Analysis and Management, 32,* 107-121.

Stuart, E.A., Cole, S.R., Bradshaw, C.P., and Leaf, P.J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 174,* 369-386.

Tipton, E. (2013). Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics, 38,* 239-266.

Appendix B. Tables and Figures

Figure 1: Distribution of Propensity Score Logits

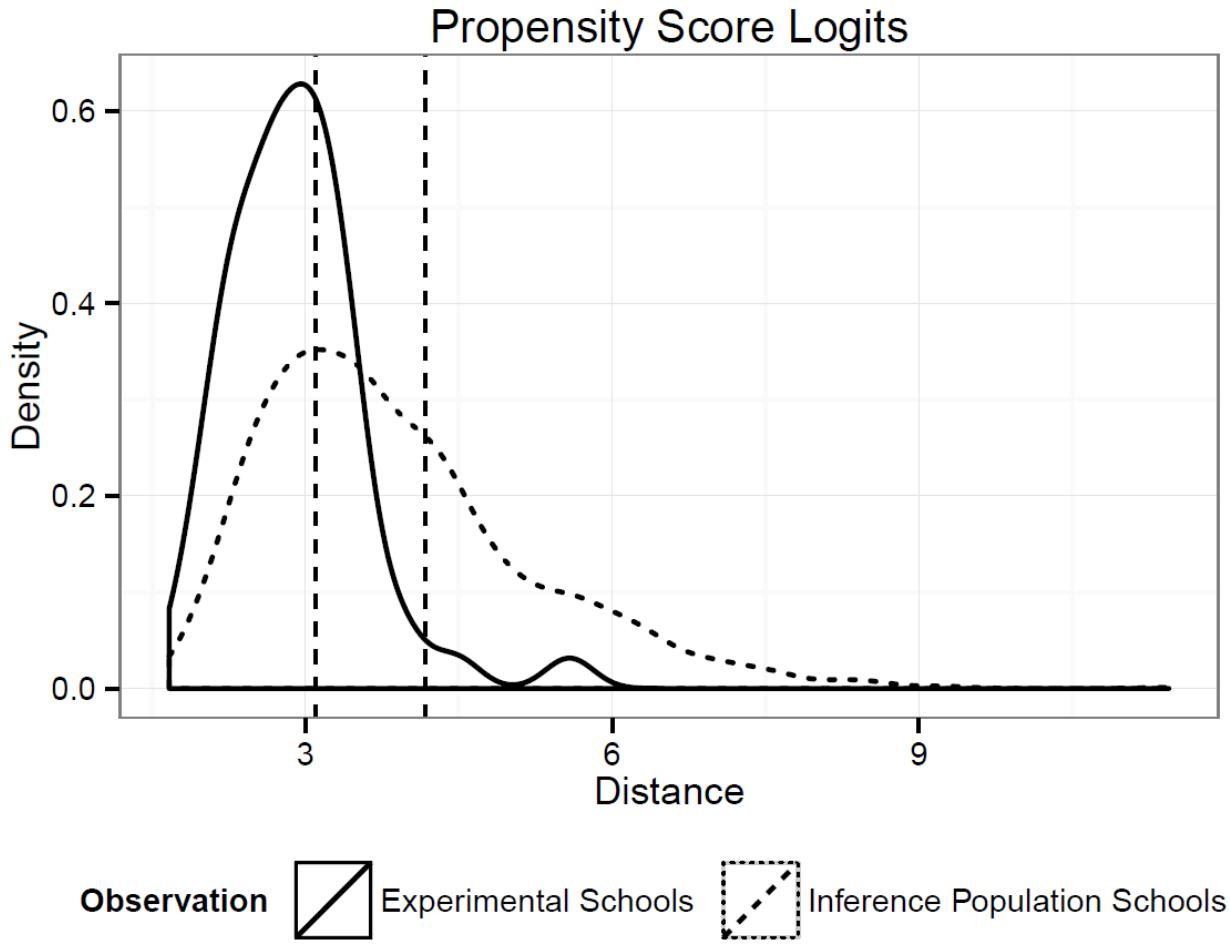

Table 1: Estimates of the PATE

	Unweighted	**IPW**	**Subclassification**
Estimate	0.20	0.19	0.26
Standard Error	0.12	0.12	0.15

Table 2: Bounds on the Expected Treatment Effect

	Worst Case			**Outcome Opt**			**Monotone Trt**	
	Full	RC	Reduced	Full	RC	Reduced	Full	Reduced
LB	-0.97	-1.88	-0.93	-0.95	-1.82	-0.91	0.00	0.00
UB	0.98	1.02	0.07	0.97	1.01	0.06	0.03	0.07

Note: "RC" refers to the reduced counterfactual case. "LB" and "UB" refer to the lower bound and upper bound, respectively.

Table 3: Bounds on the Expected Treatment Effect by Stratum

		Worst Case			**Outcome Opt**			**Monotone Trt**	
		Full	RC	Reduced	Full	RC	Reduced	Full	Reduced
Stratum 1	LB	-0.97	-1.9	-0.95	-0.95	-1.86	-0.93	0.00	0.00
	UB	0.98	1.00	0.05	0.97	0.99	0.04	0.03	0.05
Stratum 2	LB	-0.97	-1.91	-0.96	-0.96	-1.88	-0.95	0.00	0.00
	UB	0.97	0.99	0.04	0.96	0.97	0.03	0.03	0.04
Stratum 3	LB	-0.96	-1.81	-0.86	-0.93	-1.68	-0.83	0.00	0.00
	UB	0.99	1.09	0.14	0.97	1.08	0.13	0.05	0.15