Abstract Title Page

Not included in page count.

Title: Examining the foundations of methods that assess treatment effect heterogeneity across intermediate outcomes

Authors and Affiliations:
Avi Feller, Harvard University
Luke Miratrix, Harvard University

Abstract Body

Limit 4 pages single-spaced.

Background / Context:

Description of prior research and its intellectual context.

A large and growing literature addresses the identification and estimation of causal effects for subgroups defined by post-treatment outcomes, known as *principal strata* or *endogenous subgroups*. This is especially true in the ubiquitous setting of randomized evaluations with noncompliance, in which the principal strata are the usual Compliers, Always Takers, and Never Takers. Unfortunately, while these methods are increasingly common, the assumptions behind them are not well understood when the standard exclusion restrictions do not hold. This leads to questions of performance in different situations and sensitive of the resulting inference to violations of assumptions.

In general, covariates play two main roles in estimating principal causal effects (PCEs), i.e., the causal effects for latent subgroups of interest. First, covariates can sharpen inference, e.g., by shortening nonparametric bounds (Flores & Flores-Lagunes, 2013; Grilli & Mealli, 2008; Lee, 2009; Long & Hudgens, 2013), and can potentially make parametric assumptions more plausible, e.g., in model-based principal stratification (Schochet, 2013; Zhang & Rubin, 2003). Second, given appropriate additional assumptions, covariates can be used to directly identify principal causal effects of interest. A broad range of methods fall under this latter umbrella, and despite making similar assumptions, they often seem quite different on the surface.

The goal of this paper is to unify and extend current methods for covariate-based identification and estimation of PCEs. In general, we explore three broad categories of methodology, all gaining increased focus and attention in the education research world. They are:

- **Principal Ignorability**. These approaches generalize the more common propensity score methods that rely on assumptions of *ignorability* or *selection on observables*. Key citations are Hill, Waldfogel, & Brooks-Gunn (2002), Schochet & Burghardt (2007), and Jo & Stuart (2009).

- **Latent Independence**. These approaches generalize standard instrumental variable (IV) methods, positing the existence of an additional (generally binary) covariate that functions like an instrument in standard IV methods. Key citations are Jo (2002), Peck (2003), Ding, Geng, Yan, & Zhou (2011), and Mealli & Pacini (2013).

- **Multi-site, Multi-mediator IV**. These approaches leverage the multi-site design common in large-scale randomized experiments to identify PCEs via site-level regressions. Key citations are Gennetian, Bos, & Morris (2002), Kling, Liebman, & Katz (2007), and Reardon & Raudenbush (2013).

Purpose / Objective / Research Question / Focus of Study:

Description of the focus of the research.

The goal of this study is to better understand how methods for estimating treatment effects of latent groups operate. In particular, we identify where violations of assumptions can lead to biased estimates, and explore how covariates can be critical in the estimation process.

For each set of approaches, we first review the assumptions necessary for identification and discuss practical issues that arise in estimation. We then examine how covariates allow for improved estimation, and determine the conditions necessary for using covariates to identify causal effects in latent groups.

We then compare the different methods using simulation studies built from datasets constructed by imputing missing class membership and potential outcomes from real-world studies. This allows for examining the performance of the different techniques under a variety of plausible circumstances. We finally apply these methods to two common data sets that represent the type of data increasingly available to researchers, the JOBS II study and the Head Start Impact Study (HSIS), and compare the resulting treatment effect estimates to each other and some plausible baseline values.

Setting: N/A.

Population / Participants / Subjects: N/A

Intervention / Program / Practice:
Description of the intervention, program, or practice, including details of administration and duration.
(May not be applicable for Methods submissions)

We use two major datasets as test cases for the different methods we explore. First, we analyze data from the Job Search Intervention Study (JOBS II), a randomized evaluation of an intervention for unemployed workers consisting of a series of training sessions. For example analyses, see Little & Yau (1998), Jo (2002), and Jo & Stuart (2009). In this experiment, only individuals assigned to the treatment group could access the intervention, but only 55 percent of those offered actually participated in the program. This is therefore an excellent example of one-sided noncompliance. While the exclusion restriction for Never Takers seems plausible in this case, the rich set of available covariates makes this a useful test case for assessing different approaches.

Second, we analyze data from the Head Start Impact Study, a large-scale randomized evaluation of the Head Start program in which children randomized to treatment were offered a seat in a classroom in a Head Start program in fall 2002 for the 2002-2003 school year (Puma, Bell, Cook, Heid, & Shapiro, 2010). This study involved 4,440 children in 351 centers were randomized to treatment or control. The HSIS is an excellent example of two-sided non-compliance, in the sense that there are both Never Takers and Always Takers. One possible complication is that the exclusion restriction for Always Takers might not hold, in the sense that the Always Takers might enroll in, say, lower quality Head Start centers under control than under treatment. Moreover, the multi-site randomization in HSIS suggests that a multi-site IV approach could be fruitful here.

Significance / Novelty of study:
Description of what is missing in previous work and the contribution the study makes.

Our first contribution is to tie together seemingly unrelated methods for estimating the effects of latent groups, which we do not believe currently exists in the literature. We also isolate bias for the different methods under different sets of assumptions, allowing for a direct comparison of these approaches. Furthermore, by specifying what is required to identify causal effects in the latent groups, we obtain a range of estimators of causal effects, some of which are novel in this setting.

Statistical, Measurement, or Econometric Model:
Description of the proposed new methods or novel applications of existing methods.

We describe this problem using the potential outcomes notation (Neyman, 1923 [1990]), as compared to a focus on linear models. The potential outcomes framework allows us to clarify similarities and differences across seemingly disparate methods.

In this setting, we observe N individuals, N_1 of whom randomly receive some encouragement to take up an active intervention (i.e., JOBS II or center-based child care) denoted by $Z_i = 1$, and N_0 of whom are do not receive this encouragement, denoted $Z_i = 0$. For our primary example of compliance, we define two types of potential outcomes. First, let D_i be an indicator for whether individual *i* takes up the treatment, with corresponding potential outcomes $D_i(0)$ and $D_i(1)$. Second, let Y_i denote an observed outcome of interest, which is employment in JOBS II and PPVT score in HSIS, with corresponding potential outcomes, $Y_i(0)$ and $Y_i(1)$. Finally, we assume that we observe a set of pre-treatment covariates, X_i, for each individual. The endogenous groups are then defined as those with specific values of these outcomes. For example, compliers would be those with $D_i(0)=0$ and $D_i(1)=1$. Of course we cannot observe these fully, making those groups latent. Other forms of endogenous subgroups can be similarly defined. The key aspect of these models is that the randomization is solely a function of the assignment of units to treatment; this framework can clarify the estimands of interest in many contexts.

Using these models, we can then categorize methods based on the conditional independence assumptions they depend on. For example, the principal score approach requires *principal ignorability*: conditional on a vector of covariates, the potential outcomes are independent of stratum membership S_i.

$$(Y_i(0), Y_i(1)) \perp S_i \mid \mathbf{X}_i$$

By contrast, the Analysis of Symmetrically Predicted Endogenous Subgroups (ASPES) relies on the assumption that, given stratum membership, a given covariate is conditionally independent of the potential outcomes, also known as a proxy assumption (Bein, 2014).

$$(Y_i(0), Y_i(1)) \perp \widehat{S}_i \mid S_i$$

where $\widehat{S}_i$ is predicted membership as a function of covariates. Related conditional independence assumptions are also necessary for MSMM-IV.

Once different methodologies are so expressed, it becomes easier to assess the plausibility of these approaches in different settings, as we discuss. Furthermore, this framework allows for isolating bias terms in the setting when these assumptions are violated. We can then express bias for different approaches and compare them directly, something not, as far as we know, done before.

Usefulness / Applicability of Method:
Demonstration of the usefulness of the proposed methods using hypothetical or real data.

Applied researchers are increasingly interested in "unpacking the black box" of program evaluation. While useful, the array of methods that currently exist can be somewhat bewildering, leaving practitioners to choose between them without a good understanding of their strengths and weaknesses. Our goal is to detail these strengths and weaknesses and provide guidance to select one method over another given domain knowledge.

Data Collection and Analysis:
Description of the methods for collecting and analyzing data.
(May not be applicable for Methods submissions)

We plan on using data sets that already exist, and that we already have extensive experience using. In particular, our prior work on the Head Start Impact Study gave rise to the methodological questions raised in this abstract, and so we are well situated to complete the circle and use our findings on these data.

Findings / Results: N/A

Conclusions:
Description of conclusions, recommendations, and limitations based on findings.

Our headline results show that estimation of latent subgroup effects is difficult without strong predictors of latent class status, even for methods that do not rely on the covariates for identification. This means that, in practice, randomized trials should attempt to collect such covariates by, for example, having expert assessment of likelihood of compliance collected at baseline. If implemented correctly, this could be a major improvement in the designs of future trials. We also show that for identification, many methods require assumptions that are quite strong. We show how, without these assumptions, even if covariates are highly predictive of group membership, they do now allow for point identification of effects of interest. This suggests bounding approaches, an important area of future work.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Bein, E. (2014). Proxy Variable Estimators of Principal Effects. Working Paper.

Ding, P., Geng, Z., Yan, W., & Zhou, X.-H. (2011). Identifiability and Estimation of Causal Effects by Principal Stratification With Outcomes Truncated by Death. *Journal of the American Statistical Association*, *106*(496), 1578–1591.

Flores, C. A., & Flores-Lagunes, A. (2013). Partial Identification of Local Average Treatment Effects with an Invalid Instrument. *Journal of Business & Economic Statistics.*

Gennetian, L. A., Bos, J. M., & Morris, P. A. (2002). Using instrumental variables analysis to learn more from social policy experiments. *New York*.

Grilli, L., & Mealli, F. (2008). Nonparametric Bounds on the Causal Effect of University Studies on Job Opportunities Using Principal Stratification. *Journal of Educational and Behavioral Statistics*, *33*(1), 111–130.

Hill, J., Waldfogel, J., & Brooks-Gunn, J. (2002). Differential effects of high-quality child care. *Journal of Policy Analysis and Management*, *21*(4), 601–627.

Jo, B. (2002). Estimation of Intervention Effects with Noncompliance: Alternative Model Specifications. *Journal of Educational and Behavioral Statistics*, *27*(4), 385–409.

Jo, B., & Stuart, E. A. (2009). On the use of propensity scores in principal causal effect estimation. *Statistics in Medicine*, *28*(23), 2857–2875.

Kling, J. R., Liebman, J. B., & Katz, L. F. (2007). Experimental Analysis of Neighborhood Effects. *Econometrica*, *75*(1), 83–119.

Lee, D. S. (2009). Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects. *The Review of Economic Studies*, *76*(3), 1071–1102. doi:10.2307/40247633?ref=search-gateway:1ee3c513398fb8696187af2f0e46803e

Little, R. J., & Yau, L. H. Y. (1998). Statistical techniques for analyzing data from prevention trials: Treatment of no-shows using Rubin's causal model. *Psychological Methods*, *3*(2), 147–159.

Long, D. M., & Hudgens, M. G. (2013). Sharpening Bounds on Principal Effects with Covariates. *Biometrics*, *69*(4), 812–819.

Mealli, F., & Pacini, B. (2013). Using Secondary Outcomes to Sharpen Inference in Randomized Experiments With Noncompliance. *Journal of the American Statistical Association*, *108*(503), 1120–1131.

Peck, L. R. (2003). Subgroup Analysis in Social Experiments: Measuring Program Impacts Based on Post-Treatment Choice. *American Journal of Evaluation*, *24*(2), 157–187.

Puma, M., Bell, S. H., Cook, R., Heid, C., & Shapiro, G. (2010). Head Start Impact Study. Final Report. *HHS, Administration for Children & Families*.

Reardon, S. F., & Raudenbush, S. W. (2013). Under What Assumptions Do Site-by-Treatment Instruments Identify Average Causal Effects? *Sociological Methods & Research*, *42*(2), 143–163.

Schochet, P. Z. (2013). Student Mobility, Dosage, and Principal Stratification in School-Based RCTs. *Journal of Educational and Behavioral Statistics*, *38*(4), 323–354.

Schochet, P. Z., & Burghardt, J. (2007). Using Propensity Scoring to Estimate Program-Related

Subgroup Impacts in Experimental Program Evaluations. *Evaluation Review*, *31*(2), 95–120.

Splawa-Neyman, J. (1923 [1990]). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, *5*(4), 465–472.

Zhang, J. L., & Rubin, D. B. (2003). Estimation of Causal Effects via Principal Stratification When Some Outcomes are Truncated by "Death." *Journal of Educational and Behavioral Statistics*, *28*(4), 353–368.