Abstract Title Page
Not included in page count.

Title: The role of pretest and proxy-pretest measures of the outcome for removing selection bias in observational studies.

Author(s): Kelly Hallberg, Peter M. Steiner, & Thomas D. Cook

Abstract Body

Limit 5 pages single spaced.

Background / Context:

Description of prior research and its intellectual context.

Much of educational research is concerned with validly estimating causal effects. Experiments which employ random assignment are the preferred means for estimating causal effects, because, when implemented correctly, they assure that the treatment and control groups are equivalent in expectation on both observed and unobserved characteristics (Rubin, 1974). However, random assignment is not always ethical or feasible. In these cases, researchers must rely on quasi-experimental methods to identify causal effects (Shadish, Cook, & Campbell, 2002). Much debate has centered on whether and under what conditions these methods can produce valid estimates of causal effects (Glazerman, Levy, & Myers, 2003; Bloom, Michalopoulos, & Hill, 2005; Smith & Todd, 2005; Cook, Shadish, & Wong, 2008).

The primary challenge that all quasi-experimental studies face is the fact that without random assignment, the treatment and comparison groups can vary in both observable and unobservable from one another in ways that are independent of receipt of treatment (Rubin, 1974). Many quasi-experimental designs equate the treatment and comparison groups on observable covariates using matching or regression approaches. However, with the exception of regression-discontinuity designs, researchers can never be completely confident that they have adequately accounted for selection because they are not able to rule out the possibility that the groups vary on unobservable characteristics that are related to the outcome of interest (Rosenbaum & Rubin, 1983). The best quasi-experimental researchers can do is select observable covariates to minimize possible bias in effect estimates. For this reason, the selection of covariates in observational studies is of critical importance (Steiner, Cook, Shadish, & Clark, in press).

Two competing approaches to guiding covariate selection appear in the literature on quasi-experimental design. One, most closely associated with Donald B. Rubin (2007) and James Heckman , argues that covariate selection should be primarily concerned with modeling selection. The other, commonly associated with the work of Judea Pearl (2009), is more concerned with covariates that are associated with the outcome.

Within this larger debate sits the more pragmatic question of whether a pretest of the outcome measure plays a special role in reducing bias in observational studies. Some researchers, most notably those associated with the Campbell tradition of causal inference, have placed special emphasis on the pretest, primarily because of its correlation with the outcome (Campbell, 1957; Campbell & Stanley, 1963; Shadish, Cook and Campbell, 2002). While others, including Rubin (2007) and Cronbach (1982), give the pretest no special emphasis, stressing instead compiling a composite of variables that explain selection.

Purpose / Objective / Research Question / Focus of Study:

Description of the focus of the research.

The purpose of this paper is threefold. The first is to test whether the pretest plays a greater role in bias reduction than any other single covariate, which we predict it will. The second is to

examine the marginal improvement in bias reduction offered by having two pretest measurement waves. We predict that there will be some marginal gain in bias reduction as a result of including an additional pretest wave. The third purpose is to examine the extent to which a proxy pretest measure can substitute for a real pretest whose form is invariant between pretest and posttest.

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

Within study comparisons of experimental and quasi-experimental results suggest that the quality of covariates in observational studies matters and that there may reason to think that the pretest plays a unique role in these studies. However, the conclusions are drawn by comparing approaches that are employed across within-study comparisons rather than by examining the importance of certain kinds of covariates within the same study. This paper contributes to the literature by undertaking this task in a case study using a single dataset to examine how true pretest measures of the outcome perform in reducing bias in an observational study compared to other predictors.

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

We conduct a secondary data analysis on the data set used by Hong and Raudenbush (2005) to examine the effect of kindergarten retention on children's cognitive development in reading and mathematics. In this case, there is no randomly assigned experimental benchmark. This is a substantial limitation which means that we remain unable to determine whether any of the methods employed would replicate an experimental benchmark. Hong and Raudenbush (2005) conducted a thorough analysis of the data using a rich set of covariates making the strong ignorability assumption plausible. However, we cannot rule out the possibility that residual statistical bias remains due to the fact that the matched non-retention students come from the lower end of the distribution of non-retained students.

Instead of using an experimental benchmark, this study draws on our understanding of the selection process of kindergarten retention and a comparison of various effect estimates. We know that students are primarily retained "to remedy inadequate academic progress and to aid in the development of students who are judged to be emotionally immature" (Jackson, 1975, p. 614). Alexander, Entwisle, and Dauber (2003) show that prior to retention, future retainees lag behind their non-retained peers on demographic, academic and social predictors of academic success. These predictors of selection into the treatment are likely to be negatively correlated with future academic performance and thus the naïve, unadjusted comparison of retained and non-retained students is likely to be negatively biased. This is supported by the past research on retention which consistently finds a negative bias in the unadjusted mean differences between the two groups (Alexander, R. Entwisle, & Dauber, 2003; Karweit, 1999). For this reason, we view an estimate as more realistic based on how much it reduces this negative bias.

Hong and Raudenbush provided the subset of the ECLS-K used in their initial analysis as well as information about the covariates included in their propensity score models. In their original papers, Hong and Raudenbush examine several different effect estimates to fully understand the various effects of retaining students in kindergarten. For the purpose of this paper, we focus

solely on estimating the effect of retention on retained students. Following Hong and Raudenbush's lead, the ECLS-K data for this analysis were limited to schools in which at least some students are retained in kindergarten. The resulting data set included 10,726 students in 1,080 schools.

Best fitting model using all of the covariates. In any evaluation of a policy where participants are not randomly assigned to treatment or control, the problem of selection is paramount. We know that students who are retained are different from than those that are not retained. These differences can bias estimates of program effectiveness (Shadish, Cook, & Campbell, 2002).

To address this issue, Hong and Raudenbush employed a multi-level propensity score approach. Propensity scores allow for modeling the probability that a given student participates in the program based on available observed characteristics. Because characteristics of the students themselves and their schools are likely to influence whether a student participates in the program, a hierarchical modeling approach was employed to calculate an individual-level propensity score for each student in the data set, denoted as *q*:

$$\hat{q} = P(z_i = 1 | D_j = 1, X_{ij}, W_j, u_j^*)$$

Where *q* is the conditional probability that student *I* in school *j* is retained as a function of his or her individual and school characteristics, X_{ij} *and* W_j respectively, and the residual random effect of school *j*.

We employed a similar approach, starting at first with all 144 pre-treatment covariates as candidates for inclusion in the propensity score model. To select an initial propensity score model, we began by regressing each of the covariates on kindergarten retention. All covariates with a p value of greater than .2 were then included in a forward stepwise regression function to produce an initial propensity score model. Propensity scores and propensity score logits were then estimated using this model. We examined overlap in the treatment and comparison groups and deleted non-overlapping cases. We then looked at balance across the two groups on all 144 covariates. Balance statistics (standardized mean differences and variance ratios) were used to guide model selection.

We balanced pretreatment group differences in observed covariates using a propensity score stratification and marginal mean weighting approach. The propensity score logit was also included in the outcome model to control for within strata differences. Student outcomes were modeled in using two-level hierarchical linear models to account for the nested nature of the data (students within schools).

$$y_{ij} = \gamma_0 + \delta_z Z_{ij} + \gamma_1 (Logit_{\hat{q}}) + \sum_{S=2}^{15} \alpha_s L_{sij} + \gamma_2 (Dur_F)_{ij} + u_{oj} + u_{1j} Z_{ij} + e_{ij}$$

Where y_{ij} is the reading or math score for child *I* in school *j* at the end of first grade, δ_z is the average retention effect on retained students, L_{sij}, s= 2,…,15 are the dummy indicators for the propensity score strata, *Logit_q* is an additional adjustment for the student's propensity of being retained, and *Dur_F* indicates the length of time since the beginning of the treatment year that

passed before the student was assessed. We chose to estimate the average effect of treatment on retained students rather than the average effect of retention on at risk students, as Hong and Raudenbush did, because this is generally the substantive estimate of interest in program evaluations.

Sub Setting the Available Covariates. Some 144 covariates were available in the ECLS-K for possible inclusion in the propensity score. We initially subdivided these covariates into three primary group: pretests (pre-intervention data on the same measure as the outcome), proxy pretests (teachers ratings of students' academic performance), and all other covariates. The "all other covariates " group was later subdivided into eight additional groups: child demographics, child social skills, classroom demographic composition, classroom learning environment, home environment, school structures and supports, school demographic composition, and teacher demographics.

Each group of covariates was then included separately in the propensity score model. The initial propensity score model for each set of covariates was created using stepwise regression as described above. The model was then finalized by examining balance statistics to determine whether additional covariates (from that group of covariates), higher order terms, and interactions should be included. Once the propensity score model was finalized, the predicted values were used to model the effect of kindergarten retention on mathematics and reading achievement using the procedure described above. Effect estimates were then compared with one another. Boot strap standard errors were calculated to allow us to determine whether the estimates varied significantly.

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

This paper will provide valuable insight to applied researchers grappling with covariate selection in observational studies or trying to determine whether propensity score matching is a valid strategy given the available covariates. It will also provide guidance to consumers of research in their assessment of whether the appropriate variables were included in observational studies.

Findings / Results:
Description of the main findings with specific details.
(May not be applicable for Methods submissions)

Table 1 shows the effect estimates for each set of covariates. The results support the notion that the true pretest of the outcome plays a special role in bias reduction compared to other covariates as well as the hypothesis that two pretest waves is preferable to one. However, the propensity score model with two proxy pretest waves performed almost as well.

The propensity score that included all covariates except the pretests and proxy pretests did not do as well as the pretest and proxy pretest models. However, this model was able to reduce a fair amount of the bias. The researchers hypothesized that this is a result of the particularly rich set of other covariates available in this data set which spanned multiple domains of child development. When the group of all non-pretest and proxy pretest covariates was further subdivided into

domain specific subgroups, no single sub groups performed as well as all of the other covariates combined (see Table 2).

Table 3 shows the correlation of the propensity scores created with each sub set of covariates with both selection into treatment and the outcome. By examining these relationships we are better able to understand why some groups of covariates perform better or worse than others.

Conclusions:

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

This case study provides a unique opportunity to examine the relationship between covariate selection and bias reduction in observational research. Past research has suggested the importance of covariate selection and the importance of certain types of covariates, such as the true pretest of the outcome. However, this study is somewhat unique in its comparison of the performance of different sets of covariates in the same study. From this study, we can glean initial insights into how much bias certain types of covariates can reduce in observational studies. However, as a case study, the issue of generalizability remains at the forefront. It is unclear the extent to which the findings from this study generalize more broadly.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Alexander, K. L., R. Entwisle, D., & Dauber, S. L. (2003). *On the success of failure,: A reassessment of the effects of retention in the primary school grades.* New York: Cambridge University Press.

Allen, C. S., Chen, Q., Wilson, V. L., & Hughes, J. N. (2009). Quality of research design moderates effects of grade retention on achievement: A meta-analytic, multilevel analysis. *Educational Evaluation and Policy Analysis , 31* (4), 480-499.

Bloom, H., Michalopoulos, C., & Hill, C. (2005). Using experiments to assess nonexperimental comparison-group methods for measuring program effects. In H. Bloom, *Learning more from social experiments.* New York: Russell Sage Foundation.

Cook, T., Shadish, W., & Wong, V. (2008). Three conditions under which observational studies produce the same results as experiments. *Journal of Policy Analysis and Management , 274*, 724-50.

Glazerman, S., Levy, D., & Myers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. *The Annals of the American Academy , 589*, 63-91.

Hambleton, R., Swaminathan, H., & Rogers, H. (1991). *Fundamentals of Item Response Theory.* Newbury Park, CA: Sage.

Hong, G., & Raudenbush, S. (2005). Effects of kindergarten retention policy on children's cognitive growth in reading and mathematics. *Education Evaluation and Policy Analysis , 27* (3), 205-224.

Jackson, G. B. (1975). The research evidence on the effects of grade retention. *Review of Educational Research , 45* (4), 613-635.

Karweit, N. (1999). *Grade retention: Prevalence, timing, and effects.* Baltimore, MD: The John Hopkins University, Center for Social Organization of Schools.

Rosenbaum, P., & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika , 70* (1), 41-55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology ,* 688-701.

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

Smith, J., & Todd, P. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics* , 305-353.

Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. (in press). The importance of covariate selection in controlling for selction bias in observational studies. *Psychological Methods* .

Westat, University of Michigan, & Educational Testing Service. *Early Childhood Longitudinal Study: Kindergarten Class of 1998-1999.* Washington, DC: The National Center for Education Statistics, U.S. Department of Education.

Appendix B. Tables and Figures

Not included in page count.

Table 1. Effect Estimates from Covariate Subsets

	Unadjusted	All Covariates	One Pretest	Two Pretests	First Pretest and Slopes	One Proxy Pretest	Two Proxy Pretests	All Other Covariates
Reading	-19.80	-9.06	-11.57	-8.69	-9.28	-14.22	-9.29	-12.45
	(0.38)	(0.43)	(0.43)	(0.38)	(0.40)	(0.48)	(0.39)	(0.46)
Math	-11.86	-5.21	-6.23	-4.92	-5.03	-8.37	-5.02	-6.91
	(0.17)	(0.32)	(0.30)	(0.29)	(0.28)	(0.34)	(0.29)	(0.32)

Table 2. Effect Estimates with Domain-Specific Sets of Covariates Included

	Reading	Math
All covariates	-9.06	-5.21
	(0.43)	(0.32)
All other covariates	-12.45	-6.91
	(0.46)	(0.32)
Child demographics	-17.19	-10.39
	(0.51)	(0.35)
Child social skills	-13.87	-7.69
	(0.49)	(0.34)
Classroom demographic composition	-19.50	-11.87
	(0.55)	(0.36)
Classroom Learning Environment	-18.18	-10.93
	(0.56)	(0.36)
Home environment	-17.87	-10.88
	(0.52)	(0.36)
School structures and supports	-20.38	-12.16
	(0.55)	(0.36)
School demographic composition	-20.14	-12.03
	(0.55)	(0.36)
Teacher demographics	-20.14	-12.17
	(0.55)	(0.36)

Table 3. Propensity Score Correlation with Selection and Outcomes

Propensity score with…	Correlation with Selection	Correlation with the Reading Outcome (Control Group Only)	Correlation with Math Outcome (Control Group Only)
All pretests	0.59	-0.30	-0.29
All proxy pretests	0.61	-0.27	-0.26
All covariates except pretests and proxy pretests	0.59	-0.20	-0.21