Abstract Title Page

Title: Estimating long-term program impacts when the control group joins treatment in the short-term: A theoretical and empirical study of the tradeoffs between extra- and quasi-experimental estimates

Author(s):
Andrew Jaciw, Boya Ma, Qingfeng Zhao

Abstract Body

Limit 5 pages single spaced.

Background/context:

Randomized trials of educational interventions often face logistical hurdles (Cook, 2002). For example, sometimes they are prevented from reaching their full-term. This would happen if, as a condition of participation, subjects insist on receiving the treatment program within a certain timeframe. This timeframe may be shorter than the full-term required to get a measure of impact as intended by the program developers. If the controls get exposure to the program before the treatment group gets a full-term exposure, then a valid counterfactual to treatment for measuring the full-term effect is lost, and an alternative has to be established by some other means. In this situation, the researcher has a few options, including: 1) to include an outside comparison group as an alternative control group for when the experimental control group joins treatment (Shadish, Cook and Campbell, 2002); and 2) to estimate impact using an 'extra-experimental' method (Bell and Bradley, 2008) which uses impact estimates from earlier birth cohorts in the same experiment to build up a counterfactual for the treatment group. Benefits of the latter approach are that it does not rely on an outside comparison group, and it rules out selection bias and the effects of secular trends that may affect student performance.

This work builds on the experimental research tradition in education (Bloom, Bos and Lee, 1999; Boruch et al., 2002; Cook, 2002; Raudenbush, 1997). It can be considered is a first step in investigating the merits and limitation of alternative methods of obtaining program impact estimates in situations where a control group joins treatment before a program has had the opportunity to come full-term. We compare the pros and cons of quasi- and extra-experimental impact estimates. The work has both theoretical and empirical components which we outline below.

Purpose / objective / research question / focus of study:

This paper has two components. We describe the objectives in terms of the two parts:

Theoretical component: The objective is to provide a succinct comparison of the quasi- and extra-experimental estimates (we give special focus to the extra-experimental method, which is relatively new.) We derive the standard error for the extra-experimental estimate of a full-term impact and assess both types of estimators in terms of their potential for bias and imprecision.

Empirical component: The objective is to estimate the 2-year impact of a reading intervention used with special education students. Students and teachers originally assigned to the control condition transitioned into treatment after one year. We estimate impact using both the extra-experimental method as well as several alternative quasi-experimental methods. We will compare how closely these estimates correspond to one another in terms of their values and their standard errors.

Setting:

In the empirical part of this work, we analyze the results of an experiment designed to test the effects of an intervention that teaches word recognition skills to students with severe learning disabilities. The experiment is part of a longitudinal study to determine the comparative

effectiveness of the reading program as implemented in several school districts in Florida. We analyze results after two years.

Population / Participants / Subjects:

In the empirical part of this work, we compare estimates of the impact of a reading intervention designed to improve the reading skills of special education students. The program was used in self-contained special education classrooms with students spanning grades three through eight. The students were diagnosed with one or more of the following disabilities: developmental disabilities, autism, and significant learning disabilities. It is important to note that age and grade-level are often weakly correlated with performance for this population (in this experiment the correlation with the pretest was nil.)

Intervention / Program / Practice:

In the empirical part of this work we investigate the effects of a reading program. The intervention is a sight word based program designed to help non-readers become successful readers. It is a mastery-based, individualized program, through which students can learn at their own pace. The program uses various cues and manipulatives to help students learn. The recommended implementation of the program specifies a system of repetition, practice, errorless discrimination, controlled reading, and high-interest activities. Specifically, students learn through a series of steps including learning the word, tracing the word, hands-on practice, independent practice, and repetition of these steps. This is followed by various reviews, assessments, and, finally, practice reading books. The complete program contains word building lessons, supplemental lessons and activities, guided word practice, a trace-and-read workbook, flashcards, and a word viewer. Also embedded in the program are periodic assessments for teachers to administer as part of the learning cycle. Teachers are supplied with a teacher's guide and a checklist for student progress for each level. The program includes reproducible sheets for parents to work on with their children.

Research Design:

In the empirical part of this work we analyze the results of a randomized experiment that also includes a quasi-experimental component. We are interested in the two-year impact of the intervention. There are three groups of interest:

1) Students and teachers randomized to the treatment condition.
2) Students and teachers randomized to the control condition and who are phased into treatment after one year.
3) Students and teachers who were not randomized to conditions and who serve as an outside comparison group.

We will obtain an extra-experimental impact estimate (described below) based on reading outcomes for groups (1) and (2) after one and after two years. We will obtain several quasi-experimental impact estimates by comparing reading outcomes for groups (1) and (3) after two years.

Data Collection and Analysis:

The data for this study have been collected and a portion of the planned analyses have been performed. We obtained measures of pre-intervention covariates for students and teachers as

well as measures of performance one and two years after the start of the study. [Details of the data collection procedure, including the full list of covariates, are provided in the full paper.]

The procedure for obtaining the extra-experimental estimate is described in the sections that follow. The quasi-experimental estimates include: (1) a regression-adjusted impact estimate based on a model that includes student- and teacher-level covariates and a random intercept at the teacher-level to control for the effects of clustering; (2) an estimate of the mean difference between the treatment group and matched cases of control students where matching is carried out using propensity scores; (3) same as (1), but where we use matched cases from (2); (4) a combination of (1) and (2), where we use the propensity score and the pretest as covariates in an ANCOVA. [The brevity of this proposal limits the extent to which we can describe the methods used to obtain the quasi-experimental estimates of impact. We provide full descriptions in the paper.]

Findings / Results:

Theoretical results:
We are interested in comparing quasi-experimental (QE) and extra-experimental (EE) estimators. For the theoretical analysis we focus on the situation where we compare impacts over time for a given grade (i.e., as different birth cohorts of students pass through that grade.) We assume that teachers remain constant. The figure in Appendix B depicts changes in average performance of two birth cohorts of students for a particular grade over two years of the experiment. Referring to this figure we note the following quantities:

Treat(X) is the average performance of students, in a given grade level, X years into the experiment, in classes of teachers assigned to treatment at the start of the experiment.
Cont(X) is the average performance of students, in a given grade level, X years into the experiment, in classes of teachers assigned to control at the start of the experiment, but who join treatment after 1 year (i.e., the group joins treatment after cont(1) is measured.)
Comp(X) is the average performance of students, in a given grade level, X years into the experiment in classes of teachers selected to serve as comparison cases. The students and their teachers do not receive treatment

The QE estimates consist of QE(1)=treat(1)-comp(1) and QE(2)=treat(2)-comp(2) after one and two years of the experiment, respectively. The EE estimates consist of the following:

$$EE(1)=treat(1)-cont(1)$$
$$EE(2)= treat(2)-cont(2) + EE(1)= treat(2)-cont(2) + [treat(1)-cont(1)]$$

This approach, which is described fully in Bell and Bradley (2008), iteratively adds unbiased estimates of the difference in performance between the randomized groups in order to infer what the control group performance would be if they were not prematurely taken-up into treatment. Bell and Bradley (2008) point out that provided certain assumptions are met, especially that the treatment effect is constant over time, the EE estimate is unbiased. A potential disadvantage of this approach is that it depends on multiple difference estimates, which can lead to a gradual escalation of the standard error over time.

Derivation of the standard error for two-year EE impact estimates, in a two-level experiment where we are comparing within-grade differences over time.

A benchmark expression against which we compare the standard error of the EE estimates is one for the standard error for the impact estimate in a two-level experiment with randomization at level-2 (adapted from Bloom, 2005):

$$SE = 2\sqrt{\frac{Var\{e_{ij}\}}{nJ'} + \frac{Var\{r_j\}}{J'}}$$

(e_{ij}, and r_j are student- and teacher-level residuals, respectively; *n*, and *J'* are the number of students per teacher, the number of teachers, respectively. The expression assumes a balanced design.) The corresponding expression for the two-year EE impact estimate is[*]:

$$SE = 2.83\sqrt{\frac{Var\{e_{ij}\}}{nJ'} + 2 * \frac{Var\{r_j\}}{J'}}$$

Alternatively, in the paper we show that the minimum detectable effect size (MDES[†]) doubles in going from the experimental to the two-year EE estimate. The inflation is due to the estimate consisting of a sum of average differences as well as the dependencies in observations that exist as a result of different student having the same teachers at the grade level of interest.

Empirical Results[‡]:
We compare the results for the extra-experimental and quasi-experimental estimates[§]:

Extra-experimental impact estimate:
We estimate that students in the treatment group on average recognize 4.68 more words than controls in a 20-word test of word recognition skills (SE=2.68, p=.03). This is the two-year impact estimate. (The total sample of students and teachers used to obtain the EE estimate is 49 and 18, respectively.)

Quasi-experimental impact estimate:
We report only one of the QE estimates[**]. (We are still performing analysis to obtain the others.)

---

[*] Derivations of these expressions are given in the paper.

[†] This is the smallest effect that we can detect at pre-specified rates of type-1 and type-2 error (Bloom, 2005).

[‡] The empirical results stem from a slightly different design compared to the one discussed in the theoretical section. Rather than examining results for different birth cohorts of students at a given grade-level over time, we follow the same students and teachers over two years. We describe the differences between these variants of the design in greater detail in the paper. The implication is that the design used in the experiment should result in larger standard errors because we are using outcomes for the same students twice (i.e., after each year of the experiment) which implies additional dependencies in the outcomes that are combined to form the EE estimate.

[§] Note that we adjust for the clustering of students in teachers when estimating both the EE and QE impact estimates.

[**] The method involved estimating propensity scores using several student- and teacher-level covariates and then using both the propensity scores and the pretest as covariates in an ANCOVA. That is, we estimate the average difference on the posttest conditioning on the pretest and the propensity to be in the treatment condition. We follow Shadish et al. (2006) in using this approach; they note its effectiveness at reducing selection bias in their study.

We estimate that students in the treatment group on average recognize 3.32 more words than controls in a 20-word test of word recognition skills (SE=1.17, p=.01). (The total sample of students and teachers used to obtain the QE estimate is 82 and 24, respectively.)

We observe that both the EE and QE estimates are positive, and statistically significant. Both effect estimates have the same sign and are similar in magnitude. The standard error for the EE estimate is 2.68 units. The standard error for the QE estimate is 1.17 units. The former is larger, as we might expect, given the dependencies among the observations described in the theoretical results. In this case, it also reflects the larger sample.

Conclusions:

It is worth considering the strengths and weaknesses of both estimation methods and deciding whether we can somehow utilize the benefits of both. The extra-experimental estimate has the benefit of not being affected either by selection bias or secular trends that affect performance. Its drawback is that it consists of a sum of differences which leads to an inflation in the standard error, especially as a result of dependencies among observations that contribute to the two component difference estimates. A quasi-experimental estimate such as the one for which we report results in this work can be expected to have a lower standard error because it is based on a single average score difference. However, it may suffer from selection bias.

In this work we derived the standard error for a two-year extra-experimental impact estimate for a two-level group randomized trial. We also estimated a two-year impact of a reading intervention on word recognition using both extra-experimental and quasi-experimental methods. We observed consistency in the results in the case of this experiment. The fact that two estimates, each with different properties (including different strengths and weaknesses) lead to similar outcomes adds credibility to the overall result.

Can we recommend a general approach for deciding between QE and EE estimates of impact? From the foregoing results we recommend the following approach:

1) Compute the standard error inflation that results from using EE. This can readily be done once the experimental design is known, including the duration of the study.

2) Use the outcomes for the control and comparison group prior to the control group joining treatment to estimate the level of selection bias.

3) Compare the degree of imprecision, such as MDES inflation, (from (1)) and the potential for bias (from (2)), and decide which of the two alternatives is likely to give an estimate that is closer to the parameter of interest. [Another alternative, which requires further study, is to use the QE estimate and subtract off the term for selection bias that is estimated by taking the average difference between the control and comparison groups prior to controls entering treatment.]

Appendices

Not included in page count.

Appendix A. References

Bell, S. H., & Bradley, M. C., (2008, March). *Calculating long-run impacts in RCTs that release the control group into the intervention prior to the end of follow-up.* Paper presented at the Annual Research Conference of the Society for Research on Educational Effectiveness, Crystal City, VA.

Bloom, H. S., (2005). Randomizing groups to evaluate place-based programs. In H. S. Bloom (Ed.), *Learning More From Social Experiments*. New York: Sage.

Bloom, H. S., Bos, J. M., & Lee, S., (1999) Using cluster random assignment to measure program impacts. *Evaluation Review, 23*, 445-469.

Boruch, R., de Moya, D., & Snyder, B., (2002). The importance of randomized field trials in education and related areas. In F. Mosteller & R. Boruch (Eds.), *Evidence matters: Randomized trials in educational research.* Washington, DC: Brookings Institution Press.

Cook, T. D., (2002). Randomized experiments in educational policy research: A critical examination of the reasons the education evaluation community has offered for not doing them, *Educational Evaluation and Policy Analysis, 24,* 175-199.

Raudenbush, S. W., (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods, 2*, 173-185.

Shadish, W. R., Luellen, J. K., & Clark, M. H., (2006) Propensity scores and quasi-experiments: A testimony to the practical side of Lee Sechrest. In R.R. Bootzin & P.E., McKnight (Eds.), Strengthening research methodology: Psychological measurement and evaluation (pp. 143-157). Washington DC: American Psychological Association.

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

Appendix B. Tables and Figures