Low-Cost Randomized Controlled Trials in Education

By Nathan Wozny, Cary Balser, and Drew Ives*

Despite the increased frequency of randomized field experiments in the social sciences, they are relatively uncommon in educational research. For instance, Alpert, Couch, and Harmon (2016) conducted a randomized controlled trial (RCT) to estimate the impact of online education in a college-length course and referenced only three previous studies that did the same.

A likely explanation for the lack of RCTs in education is their high cost. A study sponsored by the U.S. Department of Education used an RCT to compare the relative efficacy of four elementary school math curricula, recruiting 110 schools for participation in the study (Agodini et al. 2010). While such large-scale studies may be justified from the high value of robust evidence on established, scalable, and highly standardized educational interventions, education researchers often evaluate techniques that are not well-established or fully reproducible, modifications of existing practices, or practices implemented in different disciplines or settings than have been evaluated previously. Bowen et al. (2013) conduct a large-scale RCT of online learning but also acknowledge that "online learning" refers to a vast array of interventions, requiring any one study to narrow the scope of the programs considered. Indeed, critics of large-scale randomized trials of educational interventions highlight their high costs and limited ability to generalize findings (Thomas 2016).

This paper describes an RCT design appropriate for evaluating a broad class of educational interventions using modest resources. Randomizing a classroom's educational practice within each lesson or block of lessons has the potential to identify causal effects while attaining meaningful statistical power in far smaller-scale trials than are normally required. Randomization ensures that treatment impacts are unbiased, subject to assumptions discussed in more detail in the next section. Varying educational practices across lessons improves statistical precision

* Wozny: Department of Economics and Geosciences, U.S. Air Force Academy, 2354 Fairchild Drive, USAF Academy, CO (email: nathan.wozny@usafa.edu); Balser: Department of Economics, University of Notre Dame, 434 Flanner Hall, Notre dame, IN 46556 (email: cbalser@nd.edu); Ives: Department of Economics and Geosciences, U.S. Air Force Academy, 2354 Fairchild Drive, USAF Academy, CO (email: drew.ives.1@us.af.mil). The authors thank Victoria Bhavsar, Ashley Miller, Amy Munson, Gregor Novak, Sarah Robinson, Lauren Scharff, Kate Silz-Carson, and an anonymous reviewer for comments and suggestions. The views expressed in this article are those of the authors and not necessarily those of the U.S. Air Force Academy, the U.S. Air Force, the Department of Defense, or the U.S. Government.

by enabling fixed effects to capture unobserved individual characteristics and by reducing the effect of clustering on precision. We also offer suggestions for overcoming some of the implementation and analytic challenges of this design.

I. Identification of Treatment Effects

Randomization of treatment is a broadly accepted method to identify program impacts. Levitt and List (2009) explore the growth of field experimentation in the social sciences while also highlighting common threats to identification, such as attrition bias and psychological effects of treatment assignment. We argue that a carefully designed experiment assigning instructional method by lesson can largely avoid these threats, with some advantages over classroom-level assignment. Assigning teaching method by lesson, analogous to crossover designs in clinical trials (Wellek and Blettner 2012), also introduces unique challenges in identifying treatment effects. We describe and address those challenges below.

This paper proposes randomizing classroom-level teaching methods at the section (or classroom) by lesson level. The analysis compares student performance on assessments linked to concepts taught in the treatment condition to the same student's performance on concepts taught in the control condition. The researcher can estimate the following model:

(1) $Y_{isl} = \beta T_{sl} + \alpha_i + \lambda_l + u_{isl}$,

where Y_{isl} is an assessment score (or other relevant outcome) for student i in section s on lesson l, T_{sl} is a binary indicator equal to 1 when lesson l in section s is treated, and α_i and λ_l are student and lesson fixed effects, respectively. β represents the average impact of the treatment on assessment scores compared to the control condition.

A. Modeling Spillover Effects

The key identifying assumption of the proposed design depends on the nature of the experiment. If the teaching method is unconditionally randomly assigned, T_{sl} is uncorrelated with the error u_{isl} by design. However, unconditional random assignment of teaching method may be impractical, for example if fairness considerations or logistical constraints dictate that each section receives equal numbers of lessons with each teaching method. If lessons are randomly assigned conditional on such a constraint, treatment lessons are more likely to be followed by control lessons than by other treatment lessons. If one lesson's teaching method also influences student performance in a subsequent lesson, these spillover effects will likely bias estimated impacts towards zero. For example, if a treatment has a positive impact on outcomes linked to that lesson and to subsequent lessons, then the positive spillover effects will tend to be attributed to subsequent control lessons rather than the latent effect of the treatment.

We propose two methods of minimizing the potential for bias from spillover effects. First, the researcher can modify the experimental design by identifying appropriate units of assignment. If a course contains blocks of closely related lessons, the researcher may vary teaching method across blocks rather than lessons. Second, we propose methods to model spillover effects explicitly after the experiment is implemented. If the specification adequately captures spillover effects, then the treatment status will be uncorrelated with the error even under conditional random assignment of treatment. We view these suggested methods as a starting point, recognizing that requirements for satisfying the identification assumption will vary with the specific situation.

For the first model of spillover effects, we assume that treatment lessons have some effect on a fixed number of subsequent lessons, with the impact depending only on the number of lessons elapsed. Then the researcher can estimate the model:

(2) $Y_{isl} = \beta T_{sl} + \sum_{j=1}^{J} \gamma_j T_{s(l-j)} + \alpha_i + \lambda_l + u_{isl}$,

where $T_{s(l-j)}$ is a binary indicator for treatment j lessons prior.

An alternative is to assume that prior treatment lessons have a cumulative effect, so that the researcher can estimate the model:

(3) $Y_{isl} = \beta T_{sl} + \delta \sum_{j=1}^{l-1} T_{sj} + \alpha_i + \lambda_l + u_{isl}$,

where $\sum_{j=1}^{l-1} T_{sj}$ is the number of prior treated lessons. Equation (2) models latent effects more flexibly at the cost of having to exclude J initial observations to account for the lags. Equation (3) uses the full sample but assumes that spillover effects persist at the same magnitude for all subsequent lessons. A prudent strategy would be to estimate multiple models to account for spillover effects, observing the sensitivity of the $\hat{\beta}$ estimate.[1]

B. Other Threats to Validity

The proposed design furthermore avoids other common threats to experimental validity. Observing all students in both conditions avoids differential attrition, and

[1] Since estimating equation (2) requires restricting the sample, its $\hat{\beta}$ estimate is most appropriately compared to that of equation (1) estimated on the same restricted sample.

therefore attrition bias, by design. While students are likely to be aware of the educational practice being implemented in a given lesson, common testing procedures greatly reduce the likelihood that such awareness affects outcomes. In particular, if graded assessments test concepts taught in treatment lessons and control lessons, students are unlikely to associate each question with the treatment status of the lesson covering the concept. While instructor bias is difficult to avoid in any evaluation of educational practices, each instructor contributes equally to treatment and control conditions under the proposed design. Furthermore, instructors are unlikely to introduce bias when grading assessments since they are also unlikely to be aware of the treatment status of the lesson associated with each question.

II. Statistical Power

A key limitation in precisely estimating effects of a classroom-level teaching method is the clustering adjustment required for correlated effects (Wooldridge 2003). Schochet (2008) calculates that in an experiment where N students are evenly divided among s sections, and half of sections are assigned to the treatment, the variance of the impact estimator is given by $\frac{2(1-\rho)\sigma^2}{N} + \frac{2\rho\sigma^2}{s}$, where ρ is the intra-class correlation and σ^2 is the variance of the outcome's residual.[2] For example, if a researcher wished to have an 80 percent chance of detecting a 0.2 standard deviation impact of a teaching method, she would need 54 sections of 25 students totaling over 1,300 students.[3] An experiment of this magnitude greatly exceeds the resources available for most educational studies.

Varying instructional method by lesson dramatically improves statistical power. In the absence of spillover effects, the proposed design effectively repeats a lesson-level experiment L times, where L is the number of lessons (or blocks) assigned a teaching method. The resulting variance of the impact estimator is $\frac{2(1-\rho)\sigma^2}{NL} + \frac{2\rho\sigma^2}{sL}$. Here, σ^2 represents the variance of the outcome measured across all students and all lessons, while ρ represents the fraction of an outcome's variance that is within a section-lesson combination. While these quantities could in principle be larger or smaller than their analogous definitions for a traditional RCT, we suspect they are in general smaller,

[2] This relationship treats the sample as drawn purposefully, as opposed to a random sample from a broader population to which the study is intended to generalize.

[3] This calculation assumes no control variables but an optimistically low intra-class correlation of 0.1 (Hedges and Hedberg 2007). It also uses a factor (the ratio of the minimum detectable impact to the standard error of the impact estimate) of 2.80, which assumes a large sample.

leading to substantially more precise estimates. This design could detect the same 0.2 standard deviation effect as described above with only 5 sections of 25 students if 11 lessons are assigned to a teaching method.[4]

The design also raises a question of the appropriate level of clustering. The unexplained portion of assessment scores may be correlated both within a section and within a student. Clustered standard errors or multiway clustering (Cameron, Gelbach, and Miller 2011) ensure accurate statistical inference in the presence of such correlations. A prudent strategy would be to estimate models using multiple assumptions about clustering, observing the sensitivity of the $\hat{\beta}$ standard error estimate.

III. Implementation Challenges

Implementing an RCT that varies treatment by lesson introduces some surmountable challenges. The instructor must be well-versed in teaching both methods and must take care not to favor preparation for one method. While this risk of instructor bias must be taken seriously, we argue that such risk is likely greater in a traditional RCT where different instructors implement each teaching method. Furthermore, instructors must communicate clearly with students about course logistics that may vary across sections. Assessments must be designed to test achievement specific to each lesson, with a mechanism for recording disaggregated scores corresponding to each lesson's material. Finally, instructors may wish to take steps to reduce treatment noncompliance, such as implementing access control systems for materials intended for only one condition.

The proposed design is appropriate only for interventions that can be implemented in a self-contained manner within a lesson or block of lessons. We note that some interventions initially conceptualized as a practice for an entire course may still be quite appropriate for this method. The authors implemented this design to estimate the impact of a flipped classroom where students in the treatment condition for a given section-lesson watched a video lecture before class, enabling instructors to replace in-class lecture with interactive exercises (Wozny, Balser, and Ives, forthcoming). Prunuske et al. (2016) randomly assigned four groups of medical students to sequences of four modules each using one of two online learning methods, although the study's analysis appears to treat students' learning gains as independent despite the clustered design.

[4] To the extent that student fixed effects explain variation of the outcome, precision will further increase as σ^2 represents the unexplained variance in the outcome.

IV. Conclusions

RCTs are understandably uncommon in the evaluation of educational interventions. Large-scale RCTs capable of detecting meaningful impacts are reserved for well-established, highly scalable educational interventions. The remaining majority of educational research plays an essential role in evaluating the extensive diversity of educational techniques in a wide variety of disciplines and settings. This diversity and the likelihood of heterogeneous treatment effects across different settings (Vivalt 2015) limits the external validity of any small-scale study, so that many studies are essential to identify the most effective teaching practices. Although a greater volume of literature can address the limits of a study's external validity, systemic internal validity problems may undermine the conclusions drawn from a body of literature on an educational topic.

This paper proposes an experimental design by which researchers can evaluate educational interventions rigorously while using only modest resources. The more common approach of comparing classrooms with different teaching methods risks confounding the efficacy of the teaching methods with differences in the instructors, student body, or other factors. By contrast, the proposed design ensures that experimental comparisons isolate the effect of the intervention while improving statistical power. Furthermore, such a study could be designed to test specific hypotheses about heterogeneity of treatment effects across lesson types or other subgroups. While the design is not appropriate for all teaching methods, small-scale randomized controlled trials have the potential to bring rigor to evaluating the efficacy of promising teaching practices when large-scale randomized trials are infeasible.

REFERENCES

Agodini, Roberto, Barbara Harris, Melissa Thomas, Robert Murphy, and Lawrence Gallagher. 2010. "Achievement Effects of Four Early Elementary School Math Curricula: Findings for First and Second Graders." Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Alpert, William T., Kenneth A. Couch, and Oskar R. Harmon. 2016. "A Randomized Assessment of Online Learning." *American Economic Review: Papers & Proceedings* 106 (5): 378–382.

Bowen, William G., Matthew M. Chingos, Kelly A. Lack, and Thomas I. Nygren. 2013. "Interactive learning online at public universities: evidence from a six-campus randomized trial." *Journal of Policy Analysis and Management* 33 (1): 94-111.

Cameron, Colin A., Jonah B. Gelbach, and Douglas L. Miller. 2011. "Robust Inference with Multiway Clustering." *Journal of Business and Economic Statistics* 29 (2): 238-249.

Hedges, Larry V. and E. C. Hedberg. 2007. "Intraclass Correlation Values for Planning Group-Randomized Trials in Education." *Educational Evaluation and Policy Analysis* 29 (1): 60-87.

Levitt, Steven D. and John A. List. 2009. "Field Experiments in Economics: The Past, the Present, and the Future." *European Economic Review* 53: 1-18.

Prunuske, Amy J., Lisa Henn, Ann M. Brearley, and Jacob Prunuske. 2016. "A Randomized Crossover Design to Assess Learning Impact and Student Preference for Active and Passive Online Learning Modules." *Medical Science Educator* 26: 135-141.

Schochet, Peter Z. 2008. "Statistical Power for Random Assignment Evaluations of Education Programs." *Journal of Educational and Behavioral Statistics* 33 (1): 62-87.

Thomas, Gary. 2016. "After the gold rush: questioning the 'gold standard' and reappraising the status of experiment and randomized controlled trials in education." *Harvard Educational Review* 86 (3): 390-411.

Vivalt, Eva. 2015. "Heterogeneous Treatment Effects in Impact Evaluation." *American Economic Review: Papers & Proceedings* 105 (5): 467-470.

Wellek, Stefan and Maria Blettner. 2012. "On the Proper Use of the Crossover Design in Clinical Trials." *Deutsches Ärzteblatt International* 109 (15): 276-281.

Wooldridge, Jeffrey M. 2003. "Cluster-Sample Methods in Applied Econometrics." *American Economic Review: Papers & Proceedings* 93 (2): 133-138.

Wozny, Nathan, Cary Balser, and Drew Ives. Forthcoming. "Evaluating the Flipped Classroom: a Randomized Controlled Trial." *Journal of Economic Education*.