Abstract Title Page

Not included in page count.

Title of Paper: Design-Based Estimators for Average Treatment Effects for Multi-Armed RCTs

Author: Peter Schochet, Mathematica Policy Research

Abstract Body

Limit 4 pages single-spaced.

Background / Context:

Description of prior research and its intellectual context.

Design-based methods have recently been developed as a way to analyze data from impact evaluations of interventions, programs, and policies (Freedman, 2008; Lin, 2013; Imbens and Rubin, 2015; Schochet, 2013, 2016; Yang and Tsiatis, 2001). The non-parametric estimators are derived using the building blocks of experimental designs with minimal assumptions, and are unbiased and normally distributed in large samples with simple variance estimators. The methods apply to randomized controlled trials (RCTs) and quasi-experimental designs (QEDs) with comparison groups for a wide range of designs used in social policy research. The methods have important advantages over traditional model-based impact estimation methods, such as hierarchical linear model (HLM) and robust cluster standard error (RCSE) methods, and perform well in simulations (Schochet, 2016; Kautz et al, 2017). Design-based estimators are acceptable for What Works Clearinghouse (WWC) evidence reviews (Scher and Cole, 2017).

The literature on design-based methods has focused on RCTs with a *single* treatment and a *single* control group. This theory, however, has not been formally extended to designs with *multiple* research groups. This is an important gap in the literature because multi-armed RCTs can simultaneously examine the effects of multiple interventions in a single study, thereby increasing the amount that researchers and policymakers can learn from impact evaluations. In social policy research, these designs are particularly relevant for interventions that are relatively easy to implement. Multi-armed designs are also useful for rapid-cycle or opportunistic experiments aimed at continuous program improvement, for example, using behavioral-based interventions and encouragement designs.

Multi-armed RCT designs have been used in education research in a variety of contexts. For instance, they have been used to test the effects of different forms of teacher-to-parent communication on student outcomes (Kraft and Rogers, 2014) and the effects of text messaging and peer mentoring on college enrollment rates among high school graduates (Castleman and Page, 2015). Multi-armed RCTs have also been used in larger studies to test the effects of competing math curricula (Agodini et al., 2009) and reading curricula (James-Burdumy et al., 2009). They have also been used internationally, for example, in Honduras to examine the effects of various data-driven assessment tools to improve teaching practices (Toledo et al., 2015).

Purpose / Objective / Research Question / Focus of Study:

Description of the focus of the research.

This session will provide new results on the estimation of average treatment effects (ATEs) for multi-armed designs, building on the design-based literature for the two-group design. The approach is based on the Neyman-Rubin-Holland potential outcomes framework that underlies experiments (Holland, 1986; Neyman, 1923, Rubin, 1974, 1977). The paper will consider both non-clustered and clustered designs as well as designs with blocking and baseline covariates.

The session will discuss how design-based ATE estimators for the two-group design need to be modified for the multi-armed design when comparing pairs of research groups to each other. The session will also present an empirical example using data from a multi-armed RCT testing the effects of various supplemental reading interventions. The empirical analysis shows that these statistical adjustments can matter.

The paper fits with the conference theme by providing new methods to produce rigorous evidence to inform education practice for RCT designs that are becoming increasingly popular in education.

Setting: NA
Description of the research location.
(May not be applicable for Methods submissions)

Population / Participants / Subjects: NA
Description of the participants in the study: who, how many, key features, or characteristics.
(May not be applicable for Methods submissions)

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

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

> This session will present new methods on design-based estimators for RCTs with multiple research groups. The asymptotic properties of the estimators will be presented, along with simple variance estimators, including those for clustered designs, blocked designs, and models with covariates. The literature has not addressed this topic.

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

> The paper considers designs where units are randomly assigned to one of *K* research groups, that could include a control group but does not have to. Under the simplest design with individual-level randomization, design-based theory in the multi-armed setting is based on the following data generating process for the observed outcome for an individual (y_i):

$$(1) \quad y_i = \sum_{k=1}^{K} T_i(k) Y_i(k).$$

In this expression, $Y_i(k)$ is the potential outcome for individual *i* in research condition *k*, and $T_i(k)$ is a research status indicator variable that equals 1 if the person is assigned to research group *k* and 0 otherwise.

The expression in (1) underlies the design-based inference for the multi-armed RCT. In the finite-population (FP) model, the potential outcomes are assumed to be fixed (so that $T_i(k)$ is the only source of randomness), whereas the potential outcomes are considered to be randomly sampled from a broader inference population in the super-population (SP) model framework.

The analysis of data for multi-armed designs typically involves comparing pairs of research groups to each other. Thus, causal inference for the multi-armed estimators can be derived from the design-based estimators for the simple treatment-control group design using the law of iterated expectations and variances by first conditioning on the data for the contrasted pairs and then averaging over possible randomizations to all research groups.

As formalized mathematically in this article, the paper finds that key components of the design-based theory for the two-group design apply also to multi-armed RCTs. However, two modifications are required:

1. Under the FP model, ATE estimators for each pairwise contrast pertain to the *entire* randomized sample, not just to the two groups being compared. Thus, variance estimators for the FP model for the two-group design need to be adjusted slightly to reflect the broader inference population.
2. For similar reasons, analysis weights for each pairwise comparison need to be scaled to reflect the size of the full randomized sample for each block and subgroup.

The paper shows the simple adjustments that are required and proves that the simple differences-in-means and OLS estimators with covariates are asymptotically normal, which is critical for hypothesis testing. The paper concludes with an empirical example that shows that these adjustments can matter.

The session will also mention that the free *RCT-YES* software has been updated to allow for multi-armed trials.

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

> We believe that the new design-based methods will be useful to education researchers conducting multi-armed trials. Analysts typically ignore the required statistical adjustments in the multi-armed context. Thus, the new methods can improve the statistical rigor of causal inference for these designs.

In addition, as mentioned, the free *RCT-YES* software has been updated to accommodate multi-armed trials, which will facilitate the applicability and accessibility of the new methods.

Research Design: NA
Description of the research design.
(May not be applicable for Methods submissions)

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

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

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

> This session will present new research on design-based estimators for multi-armed impact evaluations for a wide range of designs used in education research. Because the analysis in the multi-armed setting typically involves pairwise contrasts across the research groups, the key methodological question addressed in the article is: How do the estimators for the two-group design need to be adjusted for multi-armed trials? The critical insight is that in multi-armed trials, the samples for each pairwise contrast are representative of the full set of randomized units, not just of themselves. The implications are that the design-based estimators for the treatment-control design need to be adjusted to reflect this generalization. The key lesson is that researchers should be wary about using the two-armed estimators in the multi-armed context.

References

Agodini, R., B. Harris, S. Atkins-Burnett, S. Heaviside, T. Novak, R. Murphy (2009). *Achievement effects of four early elementary school math curricula: Findings from first graders in 39 schools*. Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Castleman, B.L. and L.C. Page (2015). Summer nudging: Can personalized text messages and peer mentor outreach increase college going among low-income high school graduates? *Journal of Economic Behavior & Organization*, 115, 144-160.Cheng, J. & D. Small (2006). Bounds on causal effects in three-arm trials with non-compliance. *Journal of the Royal Statistical Society, Series B,* 68(5), 815-837.

Freedman, D. (2008). On regression adjustments to experimental data. *Advances in Applied Mathematics* 40, 180-193.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association,* 81(396), 945–960.

Imbens, G. & D. Rubin (2015). *Causal inference for statistics, social, and biomedical sciences: an introduction.* Cambridge UK: Cambridge University Press.

James-Burdumy, S. et al. (2009). *Effectiveness of selected supplemental reading comprehension interventions*. Washington, DC: U.S. Department of Education Institute of Education Sciences.Kramer, C.Y. (1956). Extension of the multiple range test to group means with unequal numbers of replications. *Biometrics,* 12, 307-310.

Kautz, T., Schochet, P. Z. & Tilley, C. (2017). *Comparing impact findings from design-based and model-based methods: An empirical investigation.* Washington, DC: Analytic Technical Assistance and Development, National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Kraft, M.A. (2014). *The underutilized potential of teacher-to-parent communication: Evidence from a field experiment*. Cambridge MA: Harvard Kennedy School Working Paper RWP14-049

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *Annals of Applied Statistics* 7, 295-318.

Neyman, J. (1923). On the application of probability theory to agricultural experiments: Essay on principles. Section 9, Translated in *Statistical Science*, 1990: 5(4).

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Education Psychology,* 66, 688–701.

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Education Statistics, 2*(1), 1–26.

Scher, L. & Cole, R. (2017). *Evidence review standards considerations when using RCT-YES.* Washington, DC: Analytic Technical Assistance and Development, National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Accessed at www.rct-yes.com.

Schochet, P. Z. (2010). Is regression adjustment supported by the Neyman model for causal inference? *Journal of Statistical Planning and Inference, 140*, 246–259.

Schochet, P. Z. (2013). Estimators for clustered education RCTs using the Neyman model for causal inference. *Journal of Educational and Behavioral Statistics, 38*(3), 219–238.

Schochet, P. Z. (2016 Second Edition; 2015 First Edition). *Statistical theory for the RCT-YES software: Design-based causal inference for RCTs* (NCEE 2015–4011). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. https://ies.id.gov/ncee/pubs/20154011/pdf/20154011.pdf

Toledo, C., Humpage-Liuzzi, S., Murray, N., & Glazerman S. (2015). *Data-driven instruction in Honduras: An impact evaluation of the EducAccion Promising Reading intervention*. AEA RCT Registry: https://www.socialscienceregistry.org/trials/780

Yang, L. & Tsiatis, A. (2001). Efficiency study of estimators for a treatment effect in a pretest-posttest trial, *American Statistician* 55(4), 314-321.