Abstract Title Page
Not included in page count.

Title: Updated Design Standards and Guidance from the What Works Clearinghouse: Regression Discontinuity Designs and Cluster Designs

Authors and Affiliations:

Russell Cole
Senior Researcher
Mathematica Policy Research

John Deke
Senior Researcher
Mathematica Policy Research

Neil Seftor
WWC Project Director at Mathematica Policy Research

Abstract Body

Limit 4 pages single-spaced.

Background / Context:
The What Works Clearinghouse (WWC) maintains design standards to identify rigorous, internally valid education research. As education researchers advance new methodologies, the WWC must revise its standards to include an assessment of the new designs. Recently, the WWC has revised standards for two emerging study designs: regression discontinuity designs (RDDs), and cluster designs where the clusters (e.g. schools) are of assignment, and data are collected from lower-level unites (e.g. students).

Regression discontinuity designs (RDDs) are considered to be one of the strongest nonexperimental designs available (Shadish, Cook, & Campbell, 2002) for the purpose of identifying the effects of an intervention. These designs are applicable when a continuous "scoring" rule is used to assign the intervention to study units (for example, school districts, schools, or students). Units with scores below a pre-set cutoff value are assigned to the treatment group and units with scores above the cutoff value are assigned to the comparison group, or vice versa. For example, students may be assigned to a summer school program if they score below a preset point on a standardized test, or schools may be awarded a grant based on their score on an application. A consistent estimator of a parameter converges in probability to the true value of the parameter and, thus, is an asymptotically unbiased estimator.

Since Goldberger (1972a and 1972b) showed the theoretical appeal of the approach, numerous researchers have contributed to our understanding of RDD (Cook [2008] reviews this literature). In the past decade, researchers have made a renewed effort to bolster the theoretical underpinnings of RDD and advance the state of the art in estimating impacts and standard errors (Hahn, Todd, & Van der Klauuw, 2001; Imbens & Kalyanaraman, 2009; Lee & Card, 2008).

In cluster designs, researchers may be interested in two different types of impacts: An intent-to-treat (ITT) effect and a place-based (PB) effect (Schochet 2013; Vuchinich et al. 2012). The ITT parameter pertains to students in the study clusters at the time the clusters were randomly assigned to research conditions. This population includes "stayers" who remained in the study clusters during the follow-up period and "leavers" who did not remain in the study clusters. The ITT parameter is the average difference in outcomes between stayers and leavers across research conditions. In contrast, the PB parameter pertains to students who are in the "places" assigned to conditions at the point when outcomes are measured (Boruch and Foley 2000; Bloom 2005). These students include stayers and "joiners" who entered the clusters after random assignment, but exclude leavers. Put differently, the PB effect is the average difference between the outcomes of stayers and joiners in the treatment and control conditions.

Thus, the key difference between the ITT and PB parameters is that the ITT parameter includes leavers and stayers, whereas the PB parameter includes joiners and stayers. Both types of parameters are policy relevant and potentially of interest to WWC consumers, but the approaches to be used to assess the internal validity of studies that estimate ITT or PB parameters are very different. The WWC is revisiting the current guidance in the version 3.0 design standards for the review of cluster designs to address a number of limitations. In particular the revisions are intended to improve the review of cluster RCTs, which have been perceived as being too strict.

Purpose / Objective / Research Question / Focus of Study:
The What Works Clearinghouse (WWC) is a central and trusted source of scientific evidence for what works in education. In addition to identifying all relevant research studies on a particular topic, reviewing those studies against design standards, and synthesizing the findings, the WWC maintains a set of rigorous study design standards. The WWC Design Standards and Guidance are a set of criteria developed by panels of methodological experts to assess research quality. The WWC maintains design standards for randomized controlled trials (RCTs), quasi-experimental designs (QEDs), single-case designs (SCDs), and regression discontinuity designs (RDDs). The design standards focus on the causal validity of the study design and analysis. These design standards allow researchers and research connoisseurs to be confident that any improvement in outcomes is due to the intervention being studied and not some other difference between the treatment and control of districts, schools, teachers, or students. Trained and certified reviewers apply the relevant design standards to each study and assign one of three ratings indicating the degree of causal validity: meets WWC design standards without reservations, meets WWC design standards with reservations, and does not meet WWC design standards.

The presenters will introduce updated standards for RDDs and updated guidance for RCTs and QEDs that employ a cluster design – meaning that the unit of assignment differs from the unit of analysis. Like all WWC design standards, this revised guidance was developed by a panel of methodological experts who carefully considered all aspects of the guidance to ensure that the standards accurately identify studies with strong causal validity whose findings should contribute to WWC products and the general knowledge of what works in education. The presenters will begin by explaining the types of studies to which the given design standards and guidance apply, including specific examples where appropriate. They will then describe the updated WWC design standards and guidance, highlighting differences from earlier versions and the value of the change. Finally, the presenters will reflect on any recent experiences reviewing studies against these design standards or guidance and respond to questions from the audience.

Setting:
Not applicable.

Population / Participants / Subjects:
Not applicable.

Intervention / Program / Practice:
Not applicable.

Significance / Novelty of study:
The revised RDD standards expand and refine the pilot standards for RDD studies (June, 2010). The standards have been expanded to cover "fuzzy" RDDs (studies in which some treatment group members do not receive intervention services or some comparison group members receive services) and to cover RDDs that combine (through aggregation or pooling) multiple impacts (for example, from multiple sites or multiple assignment variables). The RDD standards have also been refined to reflect the evolution of the methodological literature (for example, the

standards now favor studies that estimate impacts within a justified bandwidth around the cutoff value on the assignment variable).

The revised guidance for reviewing cluster design studies is intended to address four substantive limitations in the version 3.0 standards:

(1) Early joining in cluster RCTs is unlikely to be caused by the intervention, and thus, the current guidance is too strict in how they rate studies with early joiners,
(2) The current guidance penalizes studies for having any joiners, regardless of how many joiners are included in the analytic sample.
(3) The review of studies that make cluster-level inferences does not consider subcluster non-response, but high subcluster response rates are necessary for internally valid cluster-level inferences
(4) Some QEDs making cluster-level inferences examine the effects of multi-year interventions, and the allowance for having an "adjacent cohort" used to demonstrate equivalence for these studies is too restrictive.

Statistical, Measurement, or Econometric Model:
The revised design standards and guidance extended the discussion of best practices for studies that employ a regression discontinuity or cluster design.

Under an RDD, the effect of an intervention is estimated as the difference in mean outcomes between treatment and comparison group units at the cutoff, adjusting statistically for the relationship between the outcomes and the variable used to assign units to the intervention. The variable used to assign units to the intervention is commonly referred to as the "forcing" or "assignment" variable. A regression line (or curve) is estimated for the treatment group and similarly for the comparison group, and the difference in average outcomes between these regression lines at the cutoff value of the forcing variable is the estimate of the effect of the intervention. RDDs generate consistent estimates of the effect of an intervention for units right at the cutoff if (1) the relationship between the outcome and forcing variable is modeled appropriately and (2) the forcing variable was not manipulated to influence assignment to the intervention group.

These revised standards apply to both "sharp" and "fuzzy" RDDs, and to RDDs that report single impacts, multiple impacts, or pooled aggregate impacts. In studies that employ a fuzzy regression discontinuity design, some treatment group members do not receive intervention services or some comparison group members receive embargoed services, but there is still a substantial discontinuity in the probability of receiving services at the cutoff. In these cases, the impact of service receipt is calculated as a ratio of the RDD impact on an outcome of interest to the RDD impact on the probability of receiving services. The WWC identified three conditions that determine the internal validity of a fuzzy RDD developed eight criteria to determine whether a fuzzy RDD satisfies the three conditions.

The revised guidance for the review of cluster designs are in the final stages of development, and will be available by March 2016. The proposed guidance will address the four main limitations described in the previous section.

Usefulness / Applicability of Method:
Studies employing a cluster or regression discontinuity design are becoming increasingly common in educational research. The WWC needs to update design standards and guidance for these designs to enable researchers to understand how to design and execute cluster or RDD studies so that they will yield accurate and valid findings and be eligible to receive the WWC's highest design standards rating, Meets WWC Design Standards Without Reservations. Research connoisseurs can rely on these new standards to distinguish high-quality research.

Research Design:
Not applicable.

Data Collection and Analysis:
Not applicable.

Findings / Results:
Not applicable.

Conclusions:
The WWC serves a critical role in the education research community by establishing and maintaining rigorous study design standards to identify high-quality research from which causal inferences can be made. The WWC design standards are also the basis for systemic reviews and design standards in other subject areas. As the research community increasingly relies on new study designs to discover what works in education, the WWC carefully develops design standards and guidance for these new designs. The WWC also regularly revisits existing design standards and guidance to ensure that they capture best methodological practice. The revised standards for regression discontinuity and guidance for cluster design studies reflect this mission and will be applied to studies using those designs in future WWC reviews.

Appendices

Not included in page count.

Appendix A. References

Bloom, H. S. "Randomizing Groups to Evaluate Place-Based Programs." In *Learning More from Social Experiments: Evolving Analytic Approaches,* edited by H. S. Bloom. New York: Russell Sage Foundation, 2005.

Boruch, R. G., and E. Foley. "The Honestly Experimental Society: Sites and Other Entities as the Units of Allocation and Analysis in Randomized Trials." In *Validity and Social Experimentation: Donald Campbell's legacy* (Vol. 1), edited by L. Bickman. Thousand Oaks, CA: Sage, 2000.

Cook, T. (2008). 'Waiting for life to arrive': A history of the regression-discontinuity design in psychology, statistics and economics. *Journal of Econometrics, 142,* 636-654.

Goldberger, A. S. (1972a). Selection bias in evaluating treatment effects: Some formal illustrations. Madison, WI: Institute for Research on Poverty.

Goldberger, A. S. (1972b). Selection bias in evaluating treatment effects: The case of interaction. Madison, WI: Institute for Research on Poverty.

Hahn, J., Todd, P., & van der Klauuw, W. (2001). Regression discontinuity. *Econometrica, 69*(1), 201-209.

Imbens, G., & Kalyanaraman, K. (2009). Optimal bandwidth choice for the regression discontinuity estimator. NBER Technical Working Paper Series: Technical Working Paper 14726. Cambridge, MA: National Bureau of Economic Research.

Imbens, G., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics 142*(2), 615-635.

Lee, D., & Card, D. (2008). Regression discontinuity inference with specification error. *Journal of Econometrics142*(2), 655-674.

Schochet, P. "Student Mobility, Dosage, and Principal Stratification in School-Based RCTs." *Journal of Educational and Behavioral Statistics,* vol. 38, no. 4, 2012, pp. 323–354.

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

Vuchinich S., B. R. Flay, L. Aber, and L. Bickman. "Person Mobility in the Design and Analysis of Cluster-Randomized Cohort Prevention Trials." Prevention Science, 2012, pp. 1–14.

Appendix B. Tables and Figures

Not included in page count.