What Works Clearinghouse

STANDARDS FOR REGRESSION DISCONTINUITY DESIGNS

Developed for the What Works Clearinghouse by the following panel:

Schochet, P.
Cook, T.
Deke, J.
Imbens, G.
Lockwood, J.R.
Porter, J.
Smith, J.

June 2010

Recommended citation:
Schochet, P., Cook, T., Deke, J., Imbens, G., Lockwood, J.R., Porter, J., Smith, J. (2010). Standards for Regression Discontinuity Designs. Retrieved from What Works Clearinghouse website: http://ies.ed.gov/ncee/wwc/pdf/wwc_rd.pdf.

STANDARDS FOR REGRESSION DISCONTINUITY DESIGNS

Regression discontinuity (RD) designs are increasingly used by researchers to obtain unbiased estimates of the effects of education-related interventions. 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.

Under an RD design, the effect on an intervention can be estimated as the difference in mean outcomes between treatment and comparison group units, adjusting statistically for the relationship between the outcomes and the variable used to assign units to the intervention, typically 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. Stated differently, an effect occurs if there is a "discontinuity" in the two regression lines at the cutoff. This estimate pertains to average treatment effects for units right at the cutoff. RD designs generate unbiased estimates of the effect of an intervention if (1) the relationship between the outcome and forcing variable can be modeled correctly and (2) the forcing variable was not manipulated to influence treatment assignments.

This document presents criteria under which RD designs *Meet WWC Evidence Standards* and *Meet WWC Evidence Standards with Reservations*.

Assessing Whether a Study Qualifies as an RD Study

A study qualifies as an RD study if it meets *all* of the following criteria:

- ***Treatment assignments are based on a forcing variable; units with scores at or above (or below) a cutoff value are assigned to the treatment group while units with scores on the other side of the cutoff are assigned to the comparison group.*** For example, an evaluation of a tutoring program could be classified as an RD study if students with a reading test score at or below 30 are admitted to the program and students with a reading test score above 30 are not. As another example, a study examining the impacts of grants to improve teacher training in local areas could be considered an RD study if grants are awarded to only those sites with grant application scores that are at least 70. In some instances, RD studies may use multiple criteria to assign the treatment to study units. For example, a student may be assigned to an after-school program if the student's reading score is below 30 *or* math score is below 40.[1] As with RCTs, noncompliance with

[1] For ease of exposition, the remainder of this document will refer to one cutoff.

treatment assignment is permitted, but the study must still meet the criteria below to meet evidence standards. Two additional criteria for the forcing variable are:

- ***The forcing variable must be ordinal with a sufficient number of unique values***. This condition is required to model the relationship between the outcomes and forcing variable. The forcing variable should never be based on cardinal (non-ordinal) categories (like gender or race). The analyzed data must also include at least four unique values of the forcing variable below the cutoff and four unique values above the cutoff.

- ***There must be no factor confounded with the forcing variable***. The cutoff value for the forcing variable must not be used to assign students to interventions other than the one being tested. For example, free/reduced-price lunch (FRPL) status cannot be the basis of an RD design, because FRPL is used as the eligibility criteria for a wide variety of services. This criterion is necessary to ensure that the study can isolate the causal effects of the tested intervention from the effects of other interventions.

If a study claims to be based on an RD design, but does not have these properties, the study does not meet standards as an RD design.

Possible Designations for Studies Using RD Designs

Once a study is determined to be an RD design, the study can receive one of three designations based on the set of criteria described below:

1. ***Meets Evidence Standards.*** To qualify, a study must meet each of the four individual standards listed below without reservations.
2. ***Meets Evidence Standards with Reservations.*** To qualify, a study must meet standards 1, 2, and 4, with or without reservations.
3. ***Does Not Meet Evidence Standards.*** If a study fails to meet standard 1, 2, or 4.

Standard 1: Integrity of the Forcing Variable

A key condition for an RD design to produce unbiased estimates of effects of an intervention is that there was no systematic manipulation of the forcing variable. This situation is analogous to the non-random manipulation of treatment and control group assignments under an RCT. In an RD design, manipulation means that scores for some units were systematically changed from their true values to influence treatment assignments. With nonrandom manipulation, the true relationship between the outcome and forcing variable can no longer be identified, which could lead to biased impact estimates.

Manipulation is possible if "scorers" have knowledge of the cutoff value and have incentives to change unit-level scores to ensure that some units are assigned to a specific research condition. Stated differently, manipulation could occur if the scoring and treatment assignment processes

are not independent. It is important to note that manipulation of the forcing variable is *different* than treatment status noncompliance (which occurs if some treatment group members do not receive intervention services or some comparison group members receive embargoed services).

The likelihood of manipulation will depend on the nature of the forcing variable, the intervention, and the study design. For example, manipulation is likely to be less plausible if the forcing variable is a standardized test score than if it is a student assessment conducted by teachers who also have input into treatment assignment decisions. As another example, manipulation is unlikely if the researchers themselves determined the cutoff value using an existing forcing variable (for example, a score from a test that was administered prior to the implementation of the study).

In all RD studies, the integrity of the forcing variable should be established both institutionally and statistically.

> **Criterion A**. The institutional integrity of the forcing variable should be established by an adequate description of the scoring and treatment assignment process. This description should indicate the forcing variable used, the cutoff value that was selected, who selected the cutoff (for example: researchers, school personnel, curriculum developers), who determined values of the forcing variable (for example, who scored a test), and when the cutoff was selected relative to determining the values of the forcing variable. This description must show that manipulation was unlikely because scorers had little opportunity or little incentive to change "true" scores in order to allow or deny specific individuals access to the intervention. If there is both a clear opportunity to manipulate scores and a clear incentive (for example, in an evaluation of a math curriculum if a placement test is scored by the curriculum developer after the cutoff is known) then the study does not satisfy this standard.
>
> **Criterion B**. The statistical integrity of the forcing variable should be demonstrated by using statistical tests found in the literature or a graphical analysis to establish the smoothness of the density of the forcing variable right around the cutoff. This is important to establish because there may be incentives for scorers to manipulate scores to make units just eligible for the treatment group (in which case, there may be an unusual mass of units near the cutoff). If a statistical test is provided, it should fail to reject the null hypothesis of continuity in the density of the forcing variable. If a graphical analysis is provided (such as a histogram or other type of density plot), there should not be strong evidence of a discontinuity at the cutoff that is obviously larger than discontinuities in the density at other points (some small discontinuities may arise when the forcing variable is discrete). If both are provided then the statistical test will take precedence, unless the statistical test indicates no discontinuity but the graphical analysis provides very strong evidence to the contrary.

To meet this standard without reservations, both criteria must be satisfied.

To meet this standard with reservations, one of the two criteria must be satisfied.

A study fails this standard if neither criterion is satisfied.

Standard 2: Attrition

An RD study must report the number of students (teachers, schools, etc.) who were assigned to the treatment and comparison group samples, and the proportion of students (teachers, schools, etc.) with outcome data who were included in the impact analysis (that is, response rates). Both overall attrition and attrition by treatment status must be reported.

To meet this standard without reservations, an RD study must meet the WWC randomized control trial (RCT) standards for attrition. The study authors can calculate overall and differential attrition either for the entire research sample or for only students near the cutoff value of the forcing variable.

A study fails this standard if attrition information is not available or if the above conditions are not met. A study that fails this standard could potentially be reviewed as a QED if equivalence is established on key baseline covariates (in this case, the forcing variable is not exempt from the equivalence requirement, described below).

Standard 3: Continuity of the Outcome-Forcing Variable Relationship

To obtain a rigorous impact estimate of a key outcome under an RD design, there must be strong evidence that in the absence of the intervention, there would be a smooth relationship between the outcome and the forcing variable at the cutoff score. This condition is needed to ensure that any observed discontinuity in the outcomes of treatment and comparison group units at the cutoff can be attributable to the intervention.

This smoothness condition cannot be checked directly, although there are two indirect approaches that should be used. The first approach is to test whether, conditional on the forcing variable, key *baseline* covariates that are correlated with the outcome variable (as identified in the review protocol for the purpose of establishing equivalence) are continuous at the cutoff. This means that the intervention should have no "impact" on baseline covariates at the cutoff. Particularly important baseline covariates for this analysis are pre-intervention measures of the key outcome variables (for example, pretests). This requirement is waived for any key covariate that is used as the RD forcing variable.

The second approach for assessing the smoothness condition is to use statistical tests or graphical analyses to examine whether there are discontinuities in the outcome-forcing variable relationship at values away from the cutoff. This involves testing for "impacts" at values of the forcing variable where there should be no impacts, such as the medians of points above or below the cutoff value (Imbens and Lemieux 2008). The presence of such discontinuities (impacts) would imply that the relationship between the outcome and the forcing variable at the cutoff may not be truly continuous, suggesting that observed impacts at the cutoff may not be due to the intervention.

Two criteria determine whether a study meets this standard.

Criterion A. Baseline (or pre-baseline) equivalence on key covariates (as identified in the review protocol) should be demonstrated at the cutoff value of the forcing variable. This involves calculating an impact at the cutoff on the covariate of interest. This requirement is waived if the variable on which equivalence must be established is the forcing variable (for example, a baseline test score).

Criterion B. There should be no evidence (using statistical tests or graphical analyses) of an unexplainable discontinuity in the outcome-score relationship at score values other than at the cutoff value. An example of an "explainable" discontinuity is one that corresponds to some other known intervention that was also administered using the same forcing variable but with a different cutoff value.

To meet this standard without reservations, both criteria must be satisfied. If criterion A is waived (see above), it can be regarded as satisfied.

A study fails this standard if either criterion is not satisfied. If criterion A is waived (see above), it can be regarded as satisfied.

Standard 4: Functional Form and Bandwidth

Unlike with RCTs, statistical modeling plays a central role in estimating impacts in an RD study. The most critical aspects of the statistical modeling are (1) the functional form specification of the relationship between the outcome and the forcing variable, and (2) the appropriate range of forcing variable values for selecting the sample (that is, the *bandwidth* around the cutoff value). Five criteria determine whether a study meets this standard.

Criterion A. The average treatment effect for an outcome must be estimated using a statistical model that controls for the forcing variable. Other baseline covariates may also be included in the statistical models, though they are not required. For both bias and variance considerations, it is never acceptable to estimate an impact by comparing the mean outcomes of treatment and comparison group members without adjusting for the forcing variable (even if there is a weak relationship between the outcome and forcing variable).

Criterion B. A graphical analysis displaying the relationship between the outcome and forcing variable—including a scatter plot and a fitted curve—must be included in the report. The display must be consistent with the choice of bandwidth and the functional form specification for the analysis. For example, if the graphical analysis shows a nonlinear relationship between the outcome and the forcing variable, then the functional form of the impact regression should also be nonlinear, or the bandwidth should be restricted to the range of data that is approximately linear on either side of the cutoff. One way to assess whether the bandwidth or functional form was appropriately chosen is to measure the sensitivity of impacts to the inclusion of observations in the tails of the forcing variable distribution.

Criterion C. Evidence must be provided that an appropriate parametric, semi-parametric, or nonparametric model was fit to the data. For a parametric approach, the adopted functional form (for example, a polynomial specification) must be shown to be the best fit to the data using statistical significance of higher order terms or a recognized "best fit" criterion (for example, the polynomial degree could be chosen to minimize the Akaike Information Criteria). Alternatively, a local regression or related nonparametric approach can be used, where the chosen bandwidth is justified using an approach such as cross-validation (or other similar approaches found in the literature). In the event that competing models are plausible, evidence of the robustness of impact findings to alternative model specifications should be provided.

Criterion D. If the estimate of the relationship between the outcome and the forcing variable is constrained to be the same on both sides of the cutoff (for example, a line that is constrained to have the same slope on both sides of the cutoff), then empirical support (either a statistical test or graphical evidence) for that constraint must be provided.

Criterion E. If the reported impact is an average of impacts across multiple sites (where, for example, a different cutoff or forcing variable is used in each site), each site impact should be estimated separately. The model used in each site should be justified using the criteria discussed above.

To meet this standard without reservations, all five of the criteria must be satisfied.

To meet this standard with reservations, Criteria A and D must be satisfied. In addition either B or C must also be satisfied.

A study fails this standard if Criterion A is not satisfied, or criterion D is not satisfied, or if both criteria B and C are not satisfied.

Reporting Requirement

Truly continuous forcing variables are likely to be rare in education studies. For example, test scores are not truly continuous – they often have a finite number of unique values because every test has a finite number of questions. If a forcing variable has a very small number of unique values (for example, a letter grade on an A-F scale) then it is not possible to estimate the relationship between the outcome and the forcing variable. Thus, we require at least 4 categories above and below the cutoff for a study to be eligible for review as an RD design. However, even in cases with a larger (but still discrete) number of unique values of the forcing variable standard errors must be estimated appropriately to account for the clustering of students at unique values of the forcing variable (see Lee and Card 2008).

As is the case in RCT designs, clustering of students should not cause biased estimates of the impact of the intervention, so if study authors do not appropriately account for the clustering of students, a study can still meet WWC standards if it meets the standards described above.

However, since the statistical significance of findings is used for the rating of the effectiveness of an intervention, study authors must account for clustering using an appropriate method (for example, the method proposed in Lee and Card 2008) in order for findings reported by the author to be included in the rating of effectiveness. If the authors do not account for clustering, then the WWC will not rely on the statistical significance of the findings from the study. However, the findings can still be included as "substantively important" if the effect size is 0.25 standard deviation or greater.

Study authors may also demonstrate that clustering of students into unique test score values does not require adjustments in the calculation of standards errors. This can be done by showing that the forcing variable is continuous around the cutoff and there is no clustering of observation around specific scores.

References

Imbens, G. and T. Lemieux (2008). Regression Discontinuity Designs: A Guide to Practice. Journal of Econometrics 142 (2), 615-635.

Lee, David and David Card (2008). Regression Discontinuity Inference With Specification Error. Journal of Econometrics142 (2), 655-674.