What Works Clearinghouse™

Preview of Regression Discontinuity Design Standards

December 2015

The WWC continues to refine its processes, develop new standards, and create new products. In order to review regression discontinuity designs according to standards that use the most up to date methodology, the WWC has revised the pilot standards for reviewing studies that use RDDs. These revised RDD standards are intended to improve on the pilot regression discontinuity standards in Version 3.0 of the *Procedures and Standards Handbook* and have been proposed by the WWC in collaboration with a team of RDD experts. In addition to removing the "pilot" label from these standards, there are the following substantive changes to the regression discontinuity standards in Version 3.0 of the *Handbook*:

- A new set of procedures for the review of "fuzzy" regression discontinuity designs/impacts,
- Expanded procedures for the review of multi-site and multiple assignment variable regression discontinuity designs, and
- Local bandwidth impact estimation is favored over global regression with flexible functional forms (only the former is eligible to *Meet WWC Group Design Standards without Reservations*).

The revised regression discontinuity design standards will replace the pilot regression discontinuity standards in the next release of the WWC *Handbook*, planned for 2016. These revised standards will <u>not</u> be used by the WWC to review studies until the new version of the WWC *Handbook* is released.

PREVIEW OF REGRESSION DISCONTINUITY DESIGN STANDARDS

Regression discontinuity designs (RDDs) are increasingly used by researchers with the goal of obtaining consistent estimates of the local average impacts of education-related interventions that are made available to individuals or groups on the basis of how they compare to a cutoff value on some known measure. For example, students may be assigned to a summer school program if they score below a cutoff value on a standardized test, or schools may be awarded a grant based on their score on an application.

Under a typical 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" or "running" variable. A regression line (or curve) is estimated for the intervention group and similarly for the comparison group, and the difference in these regression lines at the cutoff value of the forcing variable is the estimate of the effect of the intervention. Stated differently, an effect is said to have occurred 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. RDDs generate asymptotically unbiased estimates of the effect of an intervention if (1) the relationship between the outcome and forcing variable is modeled appropriately (defined in Standard 4 below) and (2) the forcing variable was not manipulated (either behaviorally or mechanically) to influence assignment to the intervention group.

This document presents criteria under which estimates of effects from RDD studies *Meet WWC RDD Standards without Reservations* and the conditions under which they *Meet WWC RDD Standards with Reservations*. These standards apply to both "sharp" and "fuzzy" RDDs (defined below in section C). We provide standards for studies that report a single RDD impact (section C), standards for studies that report multiple impacts (section D), and standards for studies that report pooled or aggregate impacts (section E). As is the case in randomized controlled trials (RCTs), clusters of students (e.g., schools, classrooms, or any other group of multiple units that have the same value of the assignment variable) might be assigned to treatment and comparison groups. These standards apply to both non-clustered and clustered RDDs. While the standards are focused on assessing the causal validity of impact estimates, we also describe two reporting requirements (sections F and G) focused on reporting accurate standard errors.

A. Assessing Whether a Study is Eligible for Review as an RDD

A study is eligible for review as an RDD study if it meets the following criteria:

Treatment assignments are based on a numerical forcing variable; units with numbers at or above a cutoff value (or at or below that value) are assigned to the intervention group whereas 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 RDD 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 RDD if grants are awarded to only those sites with grant application scores that are at least 70. In

some instances, RDDs may use multiple criteria to assign the treatment to study units. For example, a student may be assigned to an afterschool program if the student's reading score is below 30 *or* the student's math score is below 40. Studies that use multiple assignment variables or cutoffs with the same sample are eligible for review under these standards only if they (1) use a method described in the literature (e.g., in Reardon & Robinson, 2012 or Wong, Steiner, & Cook, 2013) to reduce those variables to a single assignment variable or (2) analyze each assignment variable separately. If a study does not do this (e.g., if it uses the response surface method described in Reardon & Robinson, 2012), then it is not currently eligible for review under these standards. As with RCTs, noncompliance with treatment assignment is permitted, but the study must still meet the criteria below to be eligible for a rating of *Meets WWC RDD Standards*.

The forcing variable is ordinal and includes a minimum of four or more unique values below the cutoff and four or more unique values above the cutoff. This condition is required to model the relationship between the outcomes and the forcing variable. The forcing variable must never be based on nonordinal categorical variables (e.g., 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.

The cutoff value of the forcing variable must not be used to assign members of the study sample to interventions other than the one being tested. For example, the income cutoff for determining free/reduced-price lunch (FRPL) status cannot be the basis of an RDD because FRPL is used as the eligibility criteria for a wide variety of services that also could affect student achievement. This criterion is necessary to ensure that the study can isolate the causal effects of the tested intervention from the effects of other interventions. A study can examine the combined impact of two or more interventions that all use the same cutoff value; in that case, the study can be eligible for review as an RDD, but the causal statements made must be about the combined impact (because the causal effects of each individual intervention cannot be isolated).

The forcing variable used to calculate impacts is the same as the forcing variable used to assign units to treatment status. The forcing variable used to calculate impacts must be the **actual** forcing variable, not a proxy or estimated forcing variable. A variable is considered to be a proxy if its correlation with the actual forcing variable is less than 1.

If a study claims to be based on an RDD but does not have these properties, the study is not eligible for review as an RDD.

B. Possible Ratings for Studies Using RDDs

Once a study is determined to be an RDD, the study can receive one of three ratings based on the set of criteria described below.

1. *Meets WWC RDD Standards without Reservations.* To qualify, a study must completely satisfy each of the five individual standards listed below.
2. *Meets WWC RDD Standards with Reservations.* To qualify, a study must at least partially satisfy each of the following standards: 1, 4, 5, and either 2 or 3.

3. *Does Not Meet WWC RDD Standards.* A study will receive this rating if it does not at least partially satisfy any of standards 1, 4, or 5, or does not at least partially satisfy both standards 2 and 3.

C. Standards for a Single RDD Impact

The standards presented in this section focus on assessing the causal validity of the impact of a single discontinuity in a single continuous assignment variable on a single outcome. Section D describes how to apply these standards in studies with multiple outcomes or samples. Section E describes how to apply these standards in studies with multiple impacts on the same outcome.

Standard 1: Integrity of the Forcing Variable

A key condition for an RDD to produce consistent estimates of effects of an intervention is that there was no systematic manipulation of the forcing variable. This situation is analogous to the nonrandom manipulation of treatment and comparison group assignments under an RCT. In an RDD, manipulation means that scores for some units were systematically changed from their true obtained values to influence treatment assignments and the true obtained values are unknown. With nonrandom manipulation, the true relationship between the outcome and forcing variable can no longer be identified, which could lead to inconsistent impact estimates.

Manipulation is possible if "scorers" have knowledge of the cutoff value and have incentives and an ability 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* from treatment status noncompliance (which occurs if some intervention 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 less likely to occur 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. Manipulation is also unlikely in cases where the researchers determined the cutoff value using an existing forcing variable (e.g., a score from a test that was administered prior to the implementation of the study).

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

Criterion A. The institutional integrity of the forcing variable must be established by an adequate description of the scoring and treatment assignment process. This description must indicate the forcing variable used; the cutoff value selected; who selected the cutoff (e.g., researchers, school personnel, curriculum developers); who determined values of the forcing variable (e.g., 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" obtained 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 (e.g., 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 must be demonstrated by using statistical tests found in the literature (e.g., McCrary 2008) 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 intervention group (in which case, there may be an unusual mass of units near the cutoff). The statistical test must fail to reject the null hypothesis of continuity in the density of the forcing variable at the 5 percent significance level.

Criterion C. The graphical integrity of the forcing variable must be demonstrated by using a graphical analysis (such as a histogram or other type of density plot) to establish the smoothness of the density of the forcing variable right around the cutoff. There must 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).

A study completely satisfies this standard if criteria A, B, and C are satisfied.

A study partially satisfies this standard if at least two of the three criteria are satisfied.

A study does not satisfy this standard if fewer than two of the three criteria are satisfied.

Standard 2: Attrition

An RDD study must have acceptable levels of overall and differential attrition rates (see the *WWC Procedures and Standards Handbook, Section III.2*). The samples used to calculate attrition must include all sample members who were assigned to the treatment or comparison group using the forcing variable. For example, a study that examines the impact of a prekindergarten program using age as the assignment variable could only have acceptable levels of attrition if it can identify the full set (or exogenous subgroup, meaning a subgroup identified using a variable that is exogenous to intervention participation) of sample members who were assigned to the treatment or comparison groups using that assignment variable (e.g., the analysis could consist of all students born in a specific geographic region). Attrition needs to be assessed separately for each contrast of interest. Another example is that attrition can be calculated within a bandwidth around the cutoff value of the forcing variable.

The way that attrition rates are calculated determines whether an RDD study satisfies this standard completely or partially.

A study completely satisfies this standard if the reported overall and differential attrition rates are low. This must be done in at least one of the following ways:

(1) Study authors must report the predicted mean attrition rate at the cutoff estimated using data from below the cutoff, and the predicted mean attrition rate at the cutoff estimated using data from above the cutoff. Both numbers must be estimated using a statistical model that controls for the forcing variable using the same approach that was used to estimate the impact on

the outcome. Specifically, the impact on attrition must be estimated either (1) using exactly the same bandwidth and/or functional form as was used to estimate the impact on the outcome or (2) using the same algorithm for selecting the bandwidth and/or functional form as was used to estimate the impact on the outcome. For the purpose of applying this standard, the overall attrition rate will be defined as the average of the predicted mean attrition rates on either side of the cutoff, and the differential attrition rate will be defined as the difference in the predicted mean attrition rates on either side of the cutoff.

(2) Study authors must calculate overall and differential attrition for the sample inside the bandwidth, with or without adjusting for the forcing variable.

A study partially satisfies this standard if the reported overall and differential attrition rates are low when not calculated using one of the two approaches above. Study authors can calculate overall and differential attrition for the entire research sample, with or without adjusting for the forcing variable. If authors calculate overall and differential attrition both ways (i.e., both with and without adjusting for the forcing variable), the WWC will review both and assign the highest possible rating to this part of the study design. Note that approaches should not be mixed: that is, if the rating is based on an overall attrition rate calculated without an adjustment for the forcing variable then the differential should also be unadjusted.

A study does not satisfy this standard if attrition information is not available or if none of the conditions above are met.

Standard 3: Continuity of the Relationship Between the Outcome and the Forcing Variable

To obtain a consistent impact estimate using an RDD, there must be 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 attributed to the intervention.

This smoothness condition cannot be checked directly, although there are two indirect approaches that could 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 must 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 (e.g., pretests).

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 & 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.

Three criteria determine whether a study satisfies this standard.

Criterion A. Baseline equivalence on key covariates (as identified in the review protocol) must be demonstrated at the cutoff value of the forcing variable. This involves calculating an impact at the cutoff on the covariate of interest, and the study must either (1) use exactly the same bandwidth and/or functional form as was used to estimate the impact on the outcome or (2) use the same algorithm for selecting the bandwidth and/or functional form as was used to estimate the impact on the outcome. Authors may exclude sample members from this analysis for reasons that are clearly exogenous to intervention participation (for example, authors may calculate baseline equivalence using only data within the bandwidth that was used to estimate the impact on the outcome). The burden of proof falls on the authors to demonstrate that any sample exclusions were made for exogenous reasons.

The baseline equivalence standards for group designs apply to the results from this analysis (see the *WWC Procedures and Standards Handbook, Section III.3*). Specifically, if the impact for any covariate is greater than 0.25 standard deviations in absolute value (based on the variation of that characteristic in the pooled sample), this criterion is not satisfied. If the impact for a covariate is between 0.05 and 0.25 standard deviations, the statistical model used to estimate the average treatment effect on the outcome must include a statistical adjustment for that covariate to satisfy this criterion. Differences of less than or equal to 0.05 require no statistical adjustment.

For dichotomous covariates, authors must provide the predicted mean covariate value (i.e., predicted probability) at the cutoff estimated using data from below the cutoff and the predicted probability at the cutoff estimated using data from above the cutoff. Both predicted probabilities must be calculated using the same statistical model that is used to estimate the impact on the covariate at the cutoff. These predicted probabilities are needed so that WWC reviewers can transform the impact estimate into standard deviation units.

If the attrition standard is not met, this analysis must be conducted using only sample units with non-missing values of the key outcome variable used in the study.

Criterion B. There must be no evidence, using graphical analyses, of a discontinuity in the outcome-forcing variable relationship at values of the forcing variable other than the cutoff value, unless a satisfactory explanation of such a discontinuity is provided. An example of a "satisfactory explanation" is that the discontinuity corresponds to some other known intervention that was also administered using the same forcing variable but with a different cutoff value. Another example could be a known structural property of the assignment variable; for example, if the assignment variable is a construct involving the aggregation of both continuous and discrete components. The graphical analysis (such as a scatter plot of the outcome and forcing variable using either the raw data or averaged/aggregated data within bins/intervals), must not show a discontinuity at any forcing variable value within the bandwidth (or, for the full sample if no bandwidth is used) that is larger than two times the standard error of the impact estimated at the cutoff value, unless a satisfactory explanation of that discontinuity is provided.

Criterion C. There must be no evidence, using statistical tests, of a discontinuity in the outcome-forcing variable relationship at values of the forcing variable other than the

cutoff value, unless a satisfactory explanation of such a discontinuity is provided. The statistical tests must (1) use the same algorithm for selecting the bandwidth and/or functional form as was used to estimate the impact on the outcome, and (2) be conducted for at least four values of the forcing variable below the cutoff and four values above the cutoff (these values can be either within or outside the bandwidth). At least 95 percent of the estimated impacts on the outcome at other values of the forcing variable must be statistically insignificant at the 5 percent significance level. For example, if impacts are estimated for 20 values of the forcing variable, at least 19 of them must be statistically insignificant.[1]

A study completely satisfies this standard if criteria A, B, and C are satisfied.

A study partially satisfies this standard if criterion A, and either criteria B or C, are satisfied.

A study does not satisfy this standard if criterion A is not satisfied, or both criteria B and C are not satisfied.

Standard 4: Functional Form and Bandwidth

Unlike with RCTs, statistical modeling plays a central role in estimating impacts in an RDD study. The most critical aspects of the statistical modeling are (1) the functional form specification of the relationship between the outcome variable and the forcing variable and (2) the appropriate range of forcing variable values used to select the analysis sample (i.e., the *bandwidth* around the cutoff value). Five criteria determine whether a study satisfies this standard.

Criterion A. The local average treatment effect for an outcome must be estimated using a statistical model that controls for the forcing variable. 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. Ideally the study should use a local regression (either linear or quadratic) or related nonparametric approach in which impacts are estimated within a justified bandwidth, meaning a bandwidth selected using a systematic procedure that is described and supported in the methodological literature, such as cross-validation. For example, a bandwidth selection procedure described in an article published in a peer reviewed methodological journal that describes the procedure and demonstrates its effectiveness would be a justified bandwidth. An article published in an applied journal where the procedure happens to be used does not count as justification. (A study that does not use a justified bandwidth does not completely satisfy this standard, but could partially satisfy this standard if criterion C is satisfied).

[1] If impacts are estimated for fewer than 20 values of the forcing variable, all of them must be statistically insignificant at the 5 percent significance level.

Criterion C. If the study does not use a local regression or related nonparametric approach or uses such an approach but not within a justified bandwidth, then it may estimate impacts using a "best fit" regression (using either the full sample or the sample within a bandwidth; the bandwidth does not need to be justified). For an impact estimate to meet this criterion, the functional form of the relationship between the outcome and forcing variable must be shown to be a better fit to the data than at least 2 other functional forms. Any measure of goodness of fit from the methodological literature can be used, such as the Akaike Information Criterion (AIC) or adjusted R-squared.

Criterion D. The study needs to provide evidence that the findings are robust to varying bandwidth or functional form choices. At least one of five types of evidence is sufficient to meet this criterion[2]:

(1) In the case that criterion B applies, the sign and significance of impact estimates must be the same for a total of at least two different justified bandwidths. For example, this criterion would be satisfied if the sign and significance of an impact is the same using a bandwidth selected by cross-validation[3] and a bandwidth selected by the method described in Imbens and Kalyanaraman (2012). Two impact estimates are considered to have the same significance if they are both statistically significant at the 5 percent significance level, or if neither of them are statistically significant at the 5 percent significance level. Two impact estimates are considered to have the same sign if they are both positive, both negative, or if one is positive and one is negative but neither are statistically significant at the 5 percent significance level.

(2) In the case that criterion B applies, the sign and significance of impact estimates must be the same for at least one justified bandwidth and at least two additional bandwidths (that are not justified).

(3) In the case that criterion C applies, the sign and significance of impact estimates must be the same using a total of at least two different goodness-of-fit measures to select functional form. For example, this criterion would be satisfied if the impact corresponding to the functional form selected using the AIC is the same sign and significance as an impact corresponding to the functional form selected using the regression R-squared (note, both measures may select the same functional form).

(4) In the case that criterion C applies, the sign and significance of impact estimates must be the same for at least three different functional forms (including the "best fit" regression).

[2] If a study presents more than one type of evidence, and one type shows findings are robust while another type does not, then this criterion is still satisfied. That is, studies are not penalized for conducting more sensitivity analyses.

[3] An implementation of cross-validation for RDD analysis is described in Imbens and Lemieux (2008).

(5) If the study meets both criteria B and C, then the sign and significance of impact estimates must be the same for the impact estimated within a justified bandwidth and the impact estimated using a "best fit" regression.

Criterion E. The report must include a graphical analysis displaying the relationship between the outcome and forcing variable, including a scatter plot (using either the raw data or averaged/aggregated data within bins/intervals) and a fitted curve. The display cannot be obviously inconsistent with the choice of bandwidth and the functional form specification for the analysis. Specifically, (a) if the study uses a particular functional form for the outcome-forcing variable relationship, then the study must show graphically that this functional form fits the scatter plot reasonably well, and (b) if the study uses a local linear regression, then the scatter plot must show that the outcome-forcing variable relationship is indeed reasonably linear within the chosen bandwidth.

Criterion F. The relationship between the forcing variable and the outcome must not be constrained to be the same on both sides of the cutoff.

A study completely satisfies this standard if criteria A, B, D, E, and F are satisfied.

A study partially satisfies this standard if criteria A, B or C, and E are satisfied.

A study does not satisfy this standard if either criterion A or criterion E is not satisfied or if both criteria B and C are not satisfied.

Standard 5: Fuzzy Regression Discontinuity Design (FRDD)

In a sharp RDD, all intervention group members receive intervention services and no comparison group members receive services. In a FRDD some intervention 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 an FRDD analysis the impact of service receipt is calculated as a ratio. The numerator of the ratio is the RDD impact on an outcome of interest. The denominator is the RDD impact on the probability of receiving services. This analysis is typically conducted either using two stage least squares (TSLS) or a Wald estimator. FRDD analysis is analogous to a complier average causal effect (CACE) or local average treatment effect (LATE) analysis—consequently many aspects of this standard are analogous to the WWC standards for CACE analysis in the context of RCTs.

The internal validity of an FRDD estimate depends primarily on three conditions. The first condition, known as the exclusion restriction, requires that the only channel through which assignment to the treatment or comparison groups can influence outcomes is by affecting take-up of the intervention being studied (Angrist, Imbens, & Rubin, 1996). When this condition does not hold, group differences in outcomes would be attributed to the effects of taking up the intervention when they may be attributable to other factors differing between the treatment and comparison groups. The exclusion restriction cannot be completely verified, as it is impossible to determine whether the effects of assignment on outcomes are mediated through unobserved channels. However, it is possible to identify clear violations of the exclusion restriction—in particular, situations in which groups face different circumstances beyond their differing take-up of the intervention of interest.

The second condition is that the discontinuity in the probability of receiving services at the cutoff is large enough to limit the influence of finite sample bias. The FRDD scenario can be interpreted as an instrumental variables (IV) model in which falling above or below the cutoff is an instrument for receiving intervention services (the participation indicator). IV estimators will be subject to finite sample bias if there is not a substantial difference in service receipt on either side of the cutoff; that is, if the instrument is "weak" (Stock & Yogo, 2005). FRDD impacts need not be estimated using TSLS methods (e.g., they can be estimated using Wald estimators), but authors must run the first stage regression (of the participation indicator on the indicator for being above or below the cutoff and the forcing variable) and provide either the F-statistic or the t-statistic from this regression.

The third condition is that two relationships are modeled appropriately: (1) the relationship between the forcing variable and the outcome of interest (Standard 4) and (2) the relationship between the forcing variable and receipt of services. Ideally the FRDD impact would be estimated using a justified bandwidth and functional form, where justification is focused on the overall FRDD impact, not just the numerator or denominator separately. There are methods in the literature for selecting a justified bandwidth that targets the ratio (e.g., Imbens & Kalyanaraman, 2012; Calonico, Cattaneo, & Titiunik 2014). However, in practice authors often use the bandwidth for the numerator of the FRDD, which is consistent with advice from Imbens and Kalyanaraman (2012)[4].

Eight criteria determine whether a study satisfies this standard. All eight criteria are waived for impact estimates calculated using a reduced form model (in which the outcome is modeled as a function of the forcing variable, an indicator for being above or below the cutoff, and possibly other covariates, but the *participation* indicator is not included in the model). This type of model is analogous to an "intent-to-treat" (ITT) analysis in the context of RCTs.[5]

Criterion A. The participation indicator must be a binary indicator for taking up at least a portion of the intervention (e.g., it could be a binary indicator for receiving any positive dosage of the intervention).

Criterion B. The estimation model must have exactly one participation indicator.

Criterion C. The indicator for being above or below the cutoff must be a binary indicator for the groups (treatment and comparison groups) to which study participants are assigned.

[4] On page 14 the authors write "In practice, this often leads to bandwidth choices similar to those based on the optimal bandwidth for estimation of only the numerator of the RD estimate. One may therefore simply wish to use the basic algorithm ignoring the fact that the regression discontinuity design is fuzzy."

[5] An important consideration when interpreting and applying these standards is that they are focused on the causal validity of impact estimates, not on appropriate interpretation of impact estimates. While the reduced form impact estimate may be a valid estimate of the effect of being below (or above) the RDD cutoff, interpreting that impact can be challenging in some contexts. In particular, while the reduced form RDD impact is methodologically analogous to the ITT impact from an RCT, the substantive interpretation can be entirely different. Addressing these interpretive issues is beyond the scope of these standards but we urge users of these standards to think carefully about interpretation.

Criterion D. The same covariates (one of which must be the forcing variable) must be included in (1) the analysis that estimates the impact on participation and (2) the analysis that estimates the impact on outcomes. In the case of TSLS estimation, this means that the same covariates must be used in the first and second stages.

Criterion E. To satisfy this criterion, an FRDD estimate must have no clear violations of the exclusion restriction. Defining participation inconsistently between the assigned treatment and assigned comparison groups would constitute a clear violation of the exclusion restriction. Therefore, the study must report a definition of take-up that is the same across assigned groups. Another violation of the exclusion restriction is the scenario in which assignment to the intervention group changes the behavior of study participants even if they do not take up the intervention itself. In this case, the treatment assignment might have effects on outcomes though channels other than the take-up rate. There must be no clear evidence that assignment to the intervention influenced the outcomes of study participants through channels other than take-up of the intervention.

Criterion F. The study must provide evidence that the forcing variable is a strong predictor of participation in the intervention. In a regression of program participation on a treatment indicator and other covariates, the coefficient on the treatment indicator must report a minimum F-statistic of 16 or a minimum t-statistic of 4[6]. For FRDD studies with more than one indicator for being above or below the cutoff, see the WWC Group Design Standards for RCTs that report CACE estimates for the minimum required first-stage F-statistic.

Criterion G. The study must use a local regression or related nonparametric approach in which FRDD impacts are estimated within a justified bandwidth, meaning a bandwidth selected using a systematic procedure that is described and supported in the methodological literature. Ideally this method would be justified for the FRDD impact estimate, not just the numerator of the FRDD estimate. However, two other approaches are acceptable. First, it is acceptable to use separate bandwidths for the numerator and denominator, if both are selected using a justified approach (for example, the IK algorithm applied separately to the numerator and denominator). Second, it is acceptable to use the bandwidth selected for the numerator if that bandwidth is smaller than (or equal to) a justified bandwidth selected for the denominator.

Criterion H. If criterion G is not met, the study can still partially satisfy the standard by satisfying this criterion. This criterion is satisfied if the FRDD impact is estimated using a

[6] Stock and Yogo (2005). The F-statistic must be for the instrument only—*not* the F-statistic for the entire first stage regression. If the unit of assignment does not equal the unit of analysis, the F-statistic (or t-statistic) must account for clustering using an appropriate method (such as boot-strapping, hierarchical linear modeling (HLM), or the method proposed in Lee & Card, 2008). Also, a working paper by Marmer, Fier, & Lemieux (2014) suggests that in the FRDD context, the minimum first-stage F-statistic that ensures asymptotic validity of a 5 percent two-sided test is much higher than would be required in a simple IV setting (specifically, they suggest 135). Until a published paper provides an F-statistic cutoff that is appropriate for FRDD studies that use a justified bandwidth, the F-statistic of 16 will be used as the interim criterion for assessing instrument strength.

bandwidth that is only justified for the numerator (even if it is larger than a bandwidth justified for the denominator). This criterion is also satisfied if the denominator is estimated using a "best fit" functional form. That is, the functional form of the relationship between program receipt and the forcing variable must be shown to be a better fit to the data than at least 2 other functional forms. Any measure of goodness of fit from the methodological literature can be used, such as the AIC or adjusted R-squared.

A study completely satisfies this standard if criteria A-G are satisfied.

A study partially satisfies this standard if criteria A-F and criterion H are satisfied.

A study does not satisfy this standard if any of the following criteria are not satisfied: A-F, or H.

D. Applying Standards to Studies that Report Multiple Impact Estimates

Some RDD studies report multiple separate impacts, for example impacts for different outcomes or subgroups of interest. Each of the standards described above will be applied to each outcome-subgroup combination, resulting in a separate rating for each combination. The overall rating for the study will be the highest rating attained by any outcome-subgroup combination, and will apply to only the combination(s) with that rating. In section E we address the special case of impacts that are pooled or aggregated across multiple combinations of forcing variables, cutoffs, and samples.

E. Applying Standards to Studies That Involve Aggregate or Pooled Impacts

Some RDD studies may report pooled or aggregate impacts for some combinations of forcing variables, cutoffs, and samples. By "pooled impact" we mean that data from each combination of forcing variable, cutoff, and sample are standardized and grouped into a single data set for which a single impact is calculated. By "aggregate impact" we mean a weighted average of impacts that are calculated separately for every combination of forcing variable, cutoff, and sample.

The overall rating for the study will be the highest rated impact (including pooled and aggregate impacts) presented. Authors may improve the rating of a pooled or aggregate impact by excluding combinations of forcing variables, cutoffs, and samples that do not meet WWC RDD standards for reasons that are clearly exogenous to intervention participation. For example, in a multi-site study, a site that fails the institutional check for manipulation could be excluded from the aggregate impact, resulting in a higher rating for the aggregate impact. However, potentially endogenous exclusions will not improve the rating of an aggregate impact because standards will be applied as if those exclusions were not made. For example, excluding sites that have a high differential attrition rate from an aggregate impact will not improve the rating of that impact because for the purpose of applying the attrition standard, we will include those sites. The burden of proof falls on the authors to demonstrate that any exclusions from the aggregate impact were made for exogenous reasons.

For each impact that is based on a single forcing variable, cutoff, and sample the standards can be directly applied as stated in section C.

For pooled or aggregate impacts that are based on multiple forcing variables, cutoffs, or samples, additional guidance for applying the standards is provided here.

Standard 1: Integrity of the Forcing Variable

Criterion A. If the institutional integrity of the forcing variable is not satisfied for any combination of forcing variable, cutoff, and sample that are included in a pooled or aggregate impact, then this criterion is not satisfied for that pooled or aggregate impact. However, it is permissible to exclude from a pooled or aggregate impact cases that do not satisfy this criterion. For example, if a pooled or aggregate impact is estimated using data from 5 sites, and the institutional integrity of the forcing variable is not satisfied in one of those 5 sites, then the pooled or aggregate impact does not satisfy this criterion. However, a pooled or aggregate impact estimated using data from only the 4 sites for which the institutional integrity of the forcing variable is satisfied would satisfy this criterion.

Criterion B. For an aggregate or a pooled impact this criterion is satisfied if it is satisfied for every unique combination of forcing variable, cutoff, and sample that contributes to the pooled or aggregate impact. In the case of a pooled impact, applying an appropriate statistical test to the pooled data can also satisfy this criterion. It is permissible to exclude from a pooled or aggregate impact cases that do not satisfy this criterion.

Criterion C. For an aggregate or a pooled impact this criterion is satisfied if it is satisfied for every unique combination of forcing variable, cutoff, and sample that contributes to the pooled or aggregate impact. In the case of a pooled impact, providing a single figure based on the pooled data can also satisfy this criterion. It is permissible to exclude from a pooled or aggregate impact cases that do not satisfy this criterion.

Standard 2: Attrition

In the case of a pooled impact, the attrition standard described in section C can be applied directly if the authors calculate and report overall and differential attrition using the pooled sample. Any sample excluded for endogenous reasons from calculating the pooled or aggregate impact cannot be excluded from the attrition calculation.

In the case of an aggregate impact, the WWC attrition standard can be applied to the overall and differential attrition rates calculated as weighted averages of the overall and differential rates calculated for each unique combination of forcing variable, cutoff, and sample that contribute to the aggregate impact. Authors must calculate overall and differential attrition for each of those unique combinations in a way that is consistent with the standard described in section C, and the weights used in aggregation must be the same weights used to calculate the weighted impact being reviewed. The attrition standard for the aggregate impact can only completely satisfy the standard if all individual overall and differential attrition rates are low based on acceptable calculations, as described earlier. The attrition standard for the aggregate impact can only partially satisfy the standard if all individual overall and differential attrition rates at least partially satisfy the standard.

Standard 3: Continuity of the Outcome-Forcing Variable Relationship

Criterion A. In the case of a pooled impact this criterion can be applied as described in section C without modification. In the case of an aggregate impact baseline equivalence can be established by applying the same aggregation approach to the impacts on baseline covariates as is used to aggregate impacts on outcomes.

Criterion B. In the case of a pooled impact this criterion can be applied as described in section C without modification. In the case of an aggregate impact, the requirements for this criterion must be applied cumulatively across all combinations of forcing variables, cutoffs, and samples. Specifically, there must not be evidence of a discontinuity larger than twice the standard error of the impact at any non-cutoff value within the bandwidth of any forcing variable for any sample (this means that a graphical analysis must be presented for every combination of forcing variable, cutoff, and sample). In cases where impacts from disjoint (non-overlapping) samples are being aggregated, it is acceptable to exclude from the aggregate impact any impacts from samples that do not satisfy this criterion (such an exclusion is considered exogenous).

Criterion C. In the case of a pooled impact this criterion can be applied as described in section C without modification. In the case of an aggregate impact, the requirements for this criterion must be applied cumulatively across all combinations of forcing variables, cutoffs, and samples. That is, at least 95 percent of estimated impacts at values of the forcing variables other than the cutoffs (across all samples) must be statistically insignificant. In cases where impacts from disjoint (non-overlapping) samples are being aggregated, it is acceptable to exclude from the aggregate impact any impacts from samples that do not satisfy this criterion (such an exclusion is considered exogenous).

Standard 4: Functional Form and Bandwidth

In the case of a pooled impact this standard can be applied as described in section C without modification.

In the case of an aggregate impact, criteria A-C, E, and F of this standard must be applied to every impact included in the aggregate. Any impacts excluded from the aggregate because they do not satisfy one of those criteria will be treated as attrition. The aggregate impact will receive the lowest rating from among all of these impacts.

Criterion D can just be applied to the aggregate impact. That is, it is sufficient to demonstrate robustness of the aggregate impact—it is not necessary to show robustness of every impact included in the aggregate (although showing robustness for every individual impact is also acceptable).

Standard 5: FRDD

In the case of a pooled impact this standard can be applied as described in section C without modification.

In the case of an aggregate impact this standard must be applied to every impact included in the aggregate. Any impacts excluded from the aggregate will be treated as attrition, with two

exceptions—impacts may be excluded if they don't meet Criterion E or F. The aggregate impact will receive the lowest rating from among all of these impacts.

F. Reporting Requirement for Studies with Clustered Sample

As is the case in RCTs, clusters of students (e.g., schools, classrooms, or any other group of multiple units that have the same value of the assignment variable) might be assigned to treatment and comparison groups. Clustering affects standard errors but does not lead to biased impact estimates, so if study authors do not appropriately account for the clustering of students, a study can still meet WWC RDD standards if it satisfies the standards described above. However, because the statistical significance of findings is used for the rating of the effectiveness of an intervention, when observations are clustered into groups and the unit of assignment (the cluster) differs from the unit of analysis (the individual), study authors must account for clustering using an appropriate method (such as boot-strapping, HLM, or the method proposed in Lee & 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 deviations or greater.

G. Reporting Requirement for Dichotomous Outcomes

For dichotomous outcomes, study authors must provide the predicted mean outcome (i.e., predicted probability) at the cutoff estimated using data from below the cutoff and the predicted probability at the cutoff estimated using data from above the cutoff. Both predicted probabilities must be calculated using the same statistical model that is used to estimate the impact on the outcome at the cutoff. These predicted probabilities are needed in order for findings reported by the author for those outcomes to be included in the rating of effectiveness.

References

Angrist, J., Imbens, G., & Rubin, D. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*, 444–472.

Calonico, S., Cattaneo, M., & Titiunik, R. (2014). Robust nonparametric confidence intervals for regression discontinuity designs. *Econometrica*, *82*(6), 2295–2326.

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

Imbens, G. W., & Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *Review of Economic Studies, 79*(3), 933–959.

Judd, C. M., & Kenny, D. (1981). Estimating the effects of social interventions. New York: Cambridge University Press.

Lee, D., & Card, D. (2008). Regression discontinuity inference with specification error. *Journal of Econometrics, 142*(2), 655–674.

Marmer, V., Fier, D. & Lemieux, T. (in press). Weak identification in fuzzy regression discontinuity designs. *Journal of Business & Economic Statistics*.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics, 142*(2), 698–714.

Reardon, S., & Robinson, J. P. (2012). Regression discontinuity designs with multiple rating-score variables. *Journal of Research on Educational Effectiveness*, *5*(1), 83–104.

Stock J., & Yogo, M. (2005). Testing for weak instruments in linear IV regression. In D. W. K. Andrews (Ed.), *Identification and inference for econometric models* (pp. 80–108). New York: Cambridge University Press.

Wong, V., Steiner, P., & Cook, T. (2013). Analyzing regression-discontinuity designs with multiple assignment variables: A comparative study of four estimation methods. *Journal of Educational and Behavioral Statistics, 38*(2), 107–141.

WWC RDD Standards Panels

Panelists - 2014 Revisions

Dr. Tom Cook
Dr. John Deke*
Dr. Lisa Dragoset
Dr. Sean Reardon
Dr. Rocio Titiunik
Dr. Petra Todd
Dr. Wilbert van der Klaauw
Dr. Glenn Wadell

Panelist - Original 2010 Standards

Dr. Tom Cook
Dr. John Deke
Dr. Guido Imbens
Dr. J.R. Lockwood
Dr. Jack Porter
Dr. Jeffrey Smith
Dr. Peter Schochet*

* - Denotes Panel Chair