TESTING STABILITY OF REGRESSION DISCONTINUITY MODELS

Giovanni Cerulli[a], Yingying Dong[b], Arthur Lewbel[c] and Alexander Poulsen[c]

[a]*IRCrES-CNR, National Research Council of Italy, Rome, Italy*
[b]*Department of Economics, UC Irvine, CA, USA*
[c]*Department of Economics, Boston College, Chestnut Hill, MA, USA*

ABSTRACT

Regression discontinuity (RD) models are commonly used to nonparametrically identify and estimate a local average treatment effect. Dong and Lewbel (2015) show how a derivative of this effect, called treatment effect derivative (TED) can be estimated. We argue here that TED should be employed in most RD applications, as a way to assess the stability and hence external validity of RD estimates. Closely related to TED, we define the complier probability derivative (CPD). Just as TED measures stability of the treatment effect, the CPD measures stability of the complier population in fuzzy designs. TED and CPD are numerically trivial to estimate. We provide relevant Stata code, and apply it to some real datasets.

Regression Discontinuity Designs: Theory and Applications
Advances in Econometrics, Volume 38, 317–339

ISSN: 0731-9053/doi:10.1108/S0731-905320170000038013

Keywords: Regression discontinuity; external validity; threshold changes; marginal effects; treatment effect heterogeneity; program evaluation

JEL classifications: C21; C25

1. INTRODUCTION

Consider a standard regression discontinuity (RD) model, where T is a binary treatment indicator, X is a so-called running or forcing variable, c is the threshold for X at which the probability of treatment changes discontinuously, and Y is some observed outcome. The outcome may be affected both by treatment and by X, though the conditional expectation of potential outcomes given X are assumed to be smooth functions of X. The goal in these models is to estimate the effect of treatment T on the outcome Y, and the main result in this literature is that under weak conditions a local average treatment effect (LATE) can be nonparametrically identified and estimated at the point where $X = c$ (see, e.g., Hahn, Todd, & Van der Klaauw, 2001).

The treatment effect identified by RD models only applies to a small subpopulation, namely, people having $X = c$. In fuzzy RD, the relevant group is even more limited, being just people who both have $X = c$ and are compliers. Compliers are defined to be people who have $T = 1$ if $X \geq c$ and have $T = 0$ if $X < c$. Note that since a person can only have a single value of X and of T, one of these defining conditions is a counterfactual statement, and so we can never know exactly who are the compliers.

Given that the estimated RD treatment effect only applies to people having $X = c$, it is important to investigate the stability of RD estimates, that is, to examine whether people with other values of X near c would have expected treatment effects of similar sign and magnitude. If not, that is, if ceteris paribus a small change in X away from c would greatly change the average effect of treatment, then one would have serious doubts about the general usefulness and external validity of the estimates, since other contexts are likely to differ from the given one in even more substantial ways than a marginal change in X.

In this chapter, we argue that an estimator proposed by Dong and Lewbel (2015) called the TED, for treatment effect derivative, can be used to assess the stability of RD LATE estimates. The TED therefore provides a valuable tool for judging potential external validity of the RD LATE estimator. Dong and Lewbel emphasize coupling TED with a local policy invariance assumption, to evaluate how the RD LATE would change if the threshold changed. In contrast, in this chapter we argue that, regardless of whether the local policy invariance assumption holds or not, the TED provides valuable information regarding stability of RD estimates.

TED is basically the derivative of the RD treatment effect with respect to the running variable. A more precise definition is provided below. We argue that a value of TED that is statistically significant and large in magnitude (see Section 5 for guidance on "how large is large") is evidence of instability and hence a potential lack of external validity. In contrast, having TED near-zero provides some evidence supporting stability of RD estimates.

We therefore suggest that one should estimate the TED in virtually all RD applications, and see how far it is from zero as a way to assess the stability and hence external validity of RD estimates. In addition to TED, we define a very closely related concept called the complier probability derivative, or CPD. Just as TED measures stability of the treatment effect, the CPD measures stability of the population of compliers in fuzzy designs.

Both TED and CPD are numerically trivial to estimate. They can be used to investigate external validity of the RD estimates, without requiring any additional covariates (other than the running variable). We provide easy to use Stata code to implement TED and CPD estimation, and apply it to a couple of real datasets.

It is important to note that TED differs substantially from the regression kink design (RKD) estimand of Card, Lee, Pei, and Weber (2015). The two appear superficially similar, because TED equals the difference in derivatives of a function around the threshold, and the estimate of a kink also corresponds to a difference of derivatives of a function around the threshold. However, RKD is the estimate of treatment effect given a continuous treatment with a kink. TED does not involve a continuous treatment in any way. TED applies when the treatment is binary, not continuous, and TED is not the estimate of a treatment effect, rather, TED is a treatment effect derivative.

TED is also not the same as the kink-based treatment effect estimator of Dong (2016a). Dong (2016a) provides an estimate of a binary treatment effect when the probability of a binary treatment (as a function of the running variable) contains a kink instead of a jump at the threshold. In contrast, TED assumes the standard RD jump in the probability of treatment at the threshold, and equals an estimate not of the treatment effect, but a derivative of the treatment effect.

The next section provides a short literature review. This is followed by sections describing TED for sharp designs, and both TED and CPD in fuzzy designs. We then provide some empirical examples, reexamining two published RD studies to see whether their RD estimates are likely to be stable or not.

2. LITERATURE REVIEW

A number of assumptions are required for causal validity of RD treatment effect estimates. Hahn et al. (2001) provide one formal list of assumptions, though some of their assumptions can be relaxed as noted by Lee (2008) and especially Dong (2016b). One such condition is Rubin's (1978, 1980, 1990) "SUTVA," which assumes that treatment of one set of individuals does not affect the potential outcomes of others. Another restriction is that potential outcomes, if they depend directly on X, are continuous functions of X. This relaxes the usual Rubin (1990) unconfoundedness condition, and so is one of the attractions of the RD method. In RD, one instead depends on the "no manipulation" assumption, which is generally investigated using the McCrary (2008) density and covariate smoothness tests.

One of the assumptions in Hahn et al. (2001) is a local independence assumption. This assumption says that treatment effects are independent of X in a neighborhood of the cutoff. Dong (2016b) shows that validity of RD does not actually require this condition, and that it can be replaced by some smoothness assumptions. Dong also shows that the local independence assumption implies that TED equals zero. So one use for TED is to test whether the local independence assumption holds.

Most tests of internal or external validity of treatment effect estimates require covariates with certain properties. For example, one check of validity is the falsification test, which checks whether estimated treatment effects equal zero when the RD estimator is applied after replacing the outcome Y with predetermined covariates. Angrist and Fernandez-Val (2013) assess external validity by investigating how LATE estimates vary across different conditioning sets of covariates. Angrist and Rokkanen (2015) provide conditions that allow RD treatment effects to be applied to individuals away from the cutoff, to expand the population to which RD estimates can be applied, and thereby increase external validity. Angrist and Rokkanen require local independence after conditioning on covariates. Wing and Cook (2013) bring in an additional indicator of being an untreated group, while Bertanha and Imbens (2014) look at conditioning on types. In contrast to all of these, a nice feature of TED is that it does not require any covariates other than X.

More generally, identification and estimation of TED require no additional data or information beyond what is needed for standard RD models. The only additional assumptions required to identify and estimate TED are slightly stronger smoothness conditions than those needed for standard RD, and these required differentiability assumptions are already imposed in practice when one uses standard RD estimators such as local quadratic regression.

TED focuses on changes in slope of the function $E(Y|X)$ around the cutoff $X = c$. Other papers that also examine or exploit slope changes in RD models include Dong (2016a) and Calonico, Cattaneo, and Titiunik (or CCT, 2014).

3. SHARP DESIGN TED

The intuition behind TED is simple. Let

$$Y = g_0(X) + \pi(X)T + e,$$

where $g_0(X)$ is the average untreated outcome, given X, $\pi(x)$ is the treatment effect for compliers who have $X = x$, and e is an error term that embodies all heterogeneity across individuals. Let $\pi'(x) = \partial\pi(x)/\partial x$. The treatment effect estimated by RD designs is $\pi(c)$, and TED is just $\pi'(c)$.

Let $Z = I(X \geq c)$, so Z equals one if the running variable is at or exceeds the cutoff, and is zero otherwise. Sharp RD design has $T = Z$, so $Y = g_0(X) + \pi(X)Z + e$. By just looking at individuals in a small neighborhood of c, we can approximate $g_0(X)$ and $\pi(X)$ with linear functions making

$$Y \approx \beta_1 + Z\beta_2 + (X - c)\beta_3 + (X - c)Z\beta_4 + e. \tag{1}$$

Local linear estimation with a uniform kernel consists precisely of selecting only individuals who have X observations close to (within one bandwidth of) c, and using just those people to obtain estimates $\widehat{\beta}_1$, $\widehat{\beta}_2$, $\widehat{\beta}_3$, and $\widehat{\beta}_4$ in this regression by ordinary least squares (for local quadratic estimation, see below). Under the standard RD and local linear estimation assumptions, shrinking the bandwidth at an appropriate rate as the sample size grows, we get $\widehat{\beta}_2 \rightarrow^p \pi(c)$ and $\widehat{\beta}_4 \rightarrow^p \pi'(c)$ (see Dong & Lewbel, 2015 for details). As a result, $\widehat{\beta}_2$ is the usual estimate of the RD treatment effect, and $\widehat{\beta}_4$ is the estimate of TED.

For any function h and small $\varepsilon > 0$, define the left and right limits of the function h as

$$h_+(x) = \lim_{\varepsilon \rightarrow 0} h(x + \varepsilon) \quad \text{and} \quad h_-(x) = \lim_{\varepsilon \rightarrow 0} h(x - \varepsilon).$$

Similarly, define the left and right derivatives of the function h as

$$h'_+(x) = \lim_{\varepsilon \rightarrow 0} \frac{h(x + \varepsilon) - h(x)}{\varepsilon} \quad \text{and} \quad h'_-(x) = \lim_{\varepsilon \rightarrow 0} \frac{h(x) - h(x - \varepsilon)}{\varepsilon}.$$

Let $g(x) = E(Y|X = x)$. Formally, the sharp RD design treatment effect is defined by $\pi(c) = g_+(c) - g_-(c)$, and Dong and Lewbel (2015) show that the sharp RD design TED, defined by $\pi'(c)$, satisfies the equation $\pi'(c) = g'_+(c) - g'_-(c)$. The above-described local linear estimator is nothing more than a nonparametric regression estimator of $\pi(c)$ and its derivative $\pi'(c)$.

Local quadratic estimation just adds squared terms to Eq. (1). That is, local quadratic regression adds $(X - c)^2\beta_5 + (X - c)^2 Z\beta_6$ to the right side of Eq. (1). But $\widehat{\beta}_2$ will still be a consistent estimate of the treatment effect, and $\widehat{\beta}_4$ will still be a consistent estimate of the TED. Empirical applications also often make use of nonuniform kernels. These correspond exactly to estimating the above regression using weighted least squares instead of ordinary least squares, where the weight of any observation i (given by the choice of kernel) is a function of the distance $|x_i - c|$, with observations closest to c getting the largest weight.

4. FUZZY DESIGN TED AND CPD

For fuzzy RD, where T is the treatment indicator and $Z = I(X \geq c)$ is the instrument, in addition to the outcome Eq. (1) we have the additional linear approximating equation

$$T \approx \alpha_1 + Z\alpha_2 + (X - c)\alpha_3 + (X - c)Z\alpha_4 + u, \tag{2}$$

where u is an error term. Once again, this equation can be estimated by ordinary least squares using only individuals who have X close to c, corresponding to a uniform kernel, or by weighted least squares given a different kernel. Equation (2) is a local linear approximation to $r(x) = E(T|X = x)$. Since T is binary, this equals the probability of treatment for an individual that has $X = x$.

Define a complier as an individual for whom T and Z are the same random variable, so a complier has $T = 1$ if and only if he has $Z = 1$. Equivalently, a complier is treated if and only if his value of X is greater than or equal to c. Let $p(x) = \Pr(T = Z|X = x)$. Note that this is not the probability that $T = Z = 0$, or that $T = Z = 1$, rather, this is the probability that T and Z are the same random variable, conditional on $X = x$. So $p(x)$ is the conditional probability that someone is a complier, conditioning on that person having $X = x$.

Let $p'(x) = \partial p(x)/\partial x$. By the same logic as in sharp design estimation in the previous section $p(c) = r_+(c) - r_-(c)$ and $p'(c) = r'_+(c) - r'_-(c)$. Under standard assumptions for fuzzy RD design and local linear estimation, we then have $\widehat{\alpha}_2 \rightarrow^p p(c)$ and $\widehat{\alpha}_4 \rightarrow^p p'(c)$. So $\widehat{\alpha}_4$ is a consistent estimator of what we will call the CPD.

Equation (2) is exactly the same as Eq. (1), replacing the outcome Y with T, replacing $g(x)$ with $r(x)$, and replacing the treatment effect $\pi(c)$ with the complier probability $p(c)$. So the exact same TED machinery as in the sharp design can be applied to Eq. (2), and the CPD is then just the TED when we replace Y with T. Also as before, for local quadratic estimation we just add $(X - c)^2\alpha_5 + (X - c)^2Z\alpha_6$ to Eq. (2), and doing so does not change the consistency of $\widehat{\alpha}_2$ and the CPD $\widehat{\alpha}_4$.

Let $q(x) = E(Y(1)|X = x) - E(Y(0)|X = x)$, so $q(c) = g_+(c) - g_-(c)$. The standard fuzzy design treatment effect is given by $\pi_f(c) = q(c)/p(c)$, and so is consistently estimated by

$$\widehat{\pi}_f(c) = \widehat{\beta}_2/\widehat{\alpha}_2.$$

Applying the formula for the derivative of a ratio,

$$\pi'_f(x) = \frac{\partial \pi_f(x)}{\partial x} = \frac{\partial\big(q(x)/p(x)\big)}{\partial x} = \frac{q'(x)}{p(x)} - \frac{q(x)p'(x)}{p(x)^2} = \frac{q'(x) - \pi_f(x)p'(x)}{p(x)}. \tag{3}$$

So, as shown by Dong and Lewbel (2015), the fuzzy design TED $\pi_f'(c)$ is consistently estimated by

$$\widehat{\pi}_f'(c) = \frac{\widehat{\beta}_4 - \widehat{\pi}_f(c)\widehat{\alpha}_4}{\widehat{\alpha}_2}.$$

5. STABILITY

We can now see how the TED, $\pi'(c)$, measures stability of the RD treatment effect, since $\pi(c+\varepsilon) \approx \pi(c) + \varepsilon\pi'(c)$ for small ε (A related expansion appears in Dinardo and Lee (2011) in a different context, that of extrapolating average treatment effects on the treated). If the TED is zero, then the average treatment effect of an individual with x close to but not equal to c will be $\pi(c+\varepsilon) \approx \pi(c)$, indicating stability of the estimated effect. However, if $\pi'(c)$ is large in magnitude, then people who are almost the same in every way to those at the cutoff, differing only in having a marginally lower or higher value of X, will have dramatically different treatment effects on average. A large value of TED therefore indicates instability.

The exact same logic applies also to fuzzy designs, with $\pi_f(c+\varepsilon) \approx \pi_f(c) + \varepsilon\pi_f'(c)$. However, in fuzzy designs there are two potential sources of instability. As Eq. (3) shows when evaluated at $x = c$, the fuzzy treatment effect could be unstable because $g'(c)$ is far from zero, indicating a true change in the effect on the average complier. Alternatively, the fuzzy treatment effect could be unstable because $p'(c)$ is far from zero. This latter condition is what the CPD tests. Like TED, the CPD is a measure of stability, since having the CPD near zero suggests potential stability of the complier population, whereas a large positive or negative value of the CPD says that the population of compliers changes dramatically with small changes in X. Note that in sharp design $p(c) = 1$ and therefore $p'(c) = 0$. So the CPD is always zero in sharp designs, and therefore only needs to be estimated in fuzzy designs.

Additional support for TED as a stability measure comes indirectly from Gelman and Imbens (2014). They argue that high-order local polynomials should not be used for estimating RD models, because the resulting estimates can be unstable. Unstable estimates from polynomial orders that are too high will typically result in very different slope estimates above and below the threshold, and hence a large estimated TED value.

Either $|\pi(c)/\pi'(c)|$ or $|\pi_f(c)/\pi_f'(c)|$ is approximately how large ε would need to be to change the sign of the estimated treatment effect. The smaller this value is, the more unstable is the RD LATE. Define the relative TED as

$$\text{sharp relative TED} = \left|\frac{\pi(c)}{\pi'(c)b}\right| = \left|\frac{\beta_2}{\beta_4 b}\right|,$$
$$\text{fuzzy relative TED} = \left|\frac{\pi_f(c)}{\pi_f'(c)b}\right| = \left|\frac{\alpha_2\beta_2}{(\alpha_2\beta_4 - \alpha_4\beta_2)b}\right|, \tag{4}$$

where b is the bandwidth used for estimation (meaning that the data used for estimation have values of X in the range from $c - b$ to $c + b$). A relative TED smaller than one implies that the treatment effect would change sign for some subset of people in this range. A simple rule of thumb might be that the RD LATE is unstable if TED is statistically significant and if the relative TED is smaller than about one or two. The CPD cannot change sign, but one might similarly be concerned about stability of the complier population if CPD is significant and the relative CPD, given by $|p(c)/(p'(c)b)|$ and estimated by $|\alpha_2/(\alpha_4 b)|$, were smaller than one.

It is important to note that instability does not mean that the RD estimates are invalid, but rather that they need to be interpreted cautiously. In contrast, a finding of stability (i.e., a small TED) suggests some external validity, since it implies some other people, those away from but near the cutoff, likely have treatment effects of similar magnitudes to those right at the cutoff.

6. EMPIRICAL EXAMPLES

In this section, we provide empirical RD examples that illustrate: (i) estimating LATE, TED and CPD; (ii) testing their significance, and (iii) graphically visualizing the results. We show that the same representation typically used to graph the discontinuity of the outcome (and/or of the probability) at the threshold can be readily extended to include the TED, so one can simultaneously visualize both the magnitude of the treatment effect (LATE) and its stability (TED and CPD). We present two examples drawn from existing RD empirical literature, one using a sharp design, and the other using a fuzzy design.

6.1. Sharp RD Example

The first example we consider is from Haggag and Paci (2014). These authors use a sharp RD design to examine the effect of suggested tip levels offered by credit card machines on consumers' actual tipping behavior, based on data from around 13 million New York City taxi rides. The RD design exploits different tip suggestions offered by the credit card machine depending on whether the fare was above or below \$15. For rides under \$15, tip suggestions are \$2, \$3, and \$4, while for rides above \$15 consumers are presented with 20%, 25%, and 30% tip suggestions. At the \$15 threshold, the shift represents an increase in the suggested tip levels of approximately \$1, \$0.75, and \$0.50, respectively. Haggag and Paci find that the suggested tips have a large local treatment effect. They find that this discontinuous increase in suggested tip amounts yields an increase of \$0.27−0.30 in actual tips, which is more than a 10% increase in the average tip at the \$15 threshold.

Here, we use TED to investigate stability of this \$0.27−0.30 treatment effect. TED provides information on whether and how much this estimated local treatment effect is likely to change for fares that are slightly higher or lower than the \$15 cut off. Treatment effects are estimated in two different ways; either by measuring outcomes (tips) in terms of dollar amounts, or as fractions of the total fare. The estimated RD LATE and TED are presented in Table 1. Columns 1 and 4 (RD1) correspond to the original specifications used in Haggag and Paci (2014, Column 2 of Table 2, Column 1 of Table 3) which is a third-order local polynomial, with a bandwidth that limits fares to be between \$5 and \$25, and controlling for driver fixed effects, pickup day of the week, pickup hour, pickup location, and drop-off location. We then also provide, and focus on, local quadratic regressions (RD2 and RD3), using correspondingly smaller bandwidths (limiting fares to be between \$10 and \$20 or between \$12 and \$18). We use local quadratic regressions because, as noted by Porter (2003), lower than quadratic order local polynomials suffer from boundary bias in RD estimation, while Gelman and Imbens (2014) report that higher than quadratic order local polynomials tend to be less accurate. In this application, the differences across these different polynomials and different bandwidth choices are rather small, probably because the sample size is large.

In Table 1, the estimated RD LATE for the tip amount (tips measured in dollars) ranges from \$0.274−0.287, and the TED estimates are \$0.052−0.061. These TED estimates are relatively large, suggesting that the average impact of the treatment could be five or six cents higher or lower

Table 1. TED and Sharp RD Treatment Effects of Defaults on Tipping.

	Tip Amount			Tip Percent		
	RD1	RD2	RD3	RD1	RD2	RD3
RD LATE	0.276	0.274	0.287	2.025	1.861	1.816
	(0.006)***	(0.008)***	(0.006)***	(0.038)***	(0.038)***	(0.050)***
TED	0.061	0.052	0.056	0.589	0.204	0.231
	(0.006)***	(0.006)***	(0.013)***	(0.038)***	(0.038)***	(0.081)***
Bandwidth	10	5	3	10	5	3
N	6,218,196	2,246,689	1,184,411	6,218,196	2,246,689	1,184,411
Polynomial order	3	2	2	3	2	2

Notes: This table uses the data from Haggag and Paci (2014); The sample is limited to Vendor-equipped cab rides without tolls, taxes, or surcharges; As in Haggag and Paci (2014), all specifications include fixed effects for driver, pickup day of the week, pickup hour, pickup location borough, and drop-off location borough; Columns 1 and 4 (RD1) are the original specifications used in Haggag and Paci (2014). Bandwidth equal to 10 corresponds to $\$5 \leq \text{fare} \leq \$25$ (the original bandwidth used in Haggag & Paci, 2014); Bandwidth equal to 5 corresponds to $\$10 \leq \text{fare} \leq \$20$; Bandwidth equal to 3 corresponds to $\$12 \leq \text{fare} \leq \$18$; Robust standard errors clustered at each fare value ($0.40 intervals).

***Significant at the 1% level.

with just a dollar change in fares. Defining the tip outcome and associated treatment effect in terms of percentage of the fare tells a similar story. The estimated RD treatment effect for tip percentage is 1.816−2.025 percentage points, meaning that the discontinuous jump in suggested tips (around the $15 fare level) increased actual tips by about 2%. The associated TED estimates are 0.204−0.589. This suggests that if the fare were 1% higher, the RD LATE might increase from around 2% to anywhere from 2.204% to 2.589%. Whether the change in the treatment effect would actually be this large if the threshold were actually increased depends on whether the local policy invariance assumption holds in this context (see the next section for details).

These TED estimates are all statistically significant. At the middle bandwidth of 5, the estimated relative TED is 1.05 or 1.82, which is near the borderline suggesting instability of the RD LATE estimates. The magnitude of the TED here means there is a good chance that the magnitude of the treatment effect could be quite different at somewhat lower or higher values of the threshold.

This relatively large TED value (and the associated instability it implies) can be seen in Fig. 1, which uses the tip percentage as the outcome variable. The circles in this figure show cell means, and the curves are the fitted local polynomials. Since this is a sharp design the left dark curve is an estimate of $E(Y(1) \mid X)$ and the right dark curve is an estimate of $E(Y(0) \mid X)$, where $Y(t)$ is the potential outcome given treatment t. As usual, the RD LATE equals the gap between the left and right curves at the cutoff point. The two straight lines in the figure show the estimated slopes of these local polynomials evaluated at the cutoff. Since this is a sharp design, TED just equals the difference in the slopes of these tangent lines. One can see from the figure that the slope of the tangent lines decreases quite a bit from the left to the right side of the threshold. As a result, if one extrapolated the curve on the left a little to the right of the threshold, and evaluated the gap between the curves at this new point (say, at 16 instead of 15), then this new RD LATE would be around 2.3% instead of around 2%. This is the instability that the TED measures.

To derive the RD LATE, Hahn et al. (2001) invoke a local independence assumption. This assumption says that treatment effects are independent of X in a neighborhood of the cutoff. As noted earlier, Dong (2016b) shows that validity of RD does not actually require this condition, that it can be replaced by some smoothness assumptions, and that the local independence

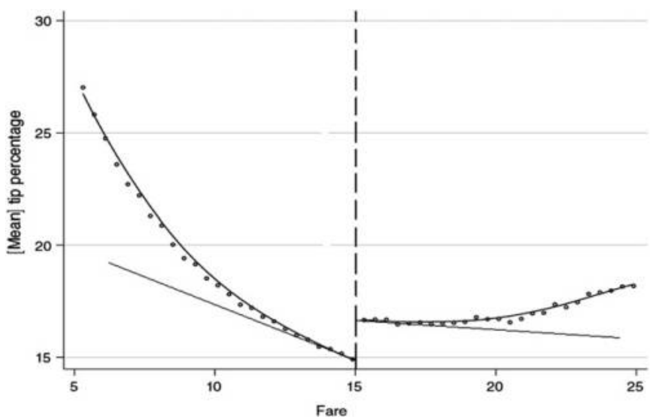

Fig. 1. Sharp RD Discontinuity in the Outcome Variable (Percentage Tips) and Tangents Lines at Threshold. *Source*: Dataset from Haggag and Paci (2014). *Note*: Each dot is the average within a discrete fare amount.

assumption implies that TED equals zero. So in the present application, our finding of a significantly nonzero TED means that this assumption is violated.

All the TED estimates in this application are significantly different from zero at the 1% significance level. Note that the sample size is very large in this application. With very large samples, even a behaviorally tiny estimate of TED could be statistically significant. This shows why it is important to consider the magnitude of the estimated TED, and not just its statistical significance, in judging stability.

6.2. Fuzzy RD Example

We now consider the fuzzy design RD model in Clark and Martorell (2010, 2014), which evaluates the signaling value of a high school diploma. In about half of US states, high school students are required to pass an exit exam to obtain a diploma. Clark and Martorell assume the random chance that leads to students falling on either side of the passing score threshold generates a credible RD design. They use this exit exam rule to evaluate the impact on earnings of having a high school diploma, since the difference in average actual learning between students with or without the diploma should be negligible, when only considering student who had grades very close to the passing grade cutoff. In this application a fuzzy RD design is appropriate, because students need to fulfill other requirements in addition to passing the exit exam in order to obtain a diploma, and some eligible students can be exempted from taking the exit exam. These other requirements include, for example, maintaining a 2.0 GPA and earning a required number of course credits. See Clark and Martorell (2010) for more details on these requirements and exemptions.

Using Texas and Florida school administrative data combined with the earnings information from the Unemployment Insurance (UI) records, Clark and Martorell find that having a high school diploma per se has little impact on earnings. This is an important finding for comparing human capital theory to signaling theory as possible explanations of the returns to education. We therefore want to investigate whether their estimates appear stable near the RD cutoff. Here, the outcome Y is the present discounted value of earnings 7 years after one takes the last round of exit exams. The treatment T is whether a student receives a high school diploma or not. The running variable X is the exit exam score (centered at the threshold passing score). Following Clark and Martorell (2010, 2014), we focus on the last chance sample, that is, those who take the last round of exit exams in high school. In our sample, X ranges from -100 to 50. About 46.7% of these students receive a high school diploma, and their average earnings are \$25,721. Detailed information on the construction of the sample can be found in Clark and Martorell (2010).

Figs. 2 and 3 show, respectively, the probability of receiving a high school diploma and earnings as a function of the exit exam score. As we can see from Fig. 2, the probability of receiving a high school diploma changes from about 40% to about 90% at the threshold passing score. This figure shows a modest change in slope at the threshold, from slightly increasing to slightly decreasing, indicating a small negative CPD. Fig. 3 shows very little if any discontinuity in outcomes around the threshold, which is the basis of Clark and Martorell's finding that having a high school diploma *per se* has little impact on earnings. The tangent lines shown in Fig. 3 are close to parallel, indicating that TED (which depends on the difference in the slopes of these lines) is also close to zero. This suggests that Clark and Martorell's results are stable, and not just a quirk of

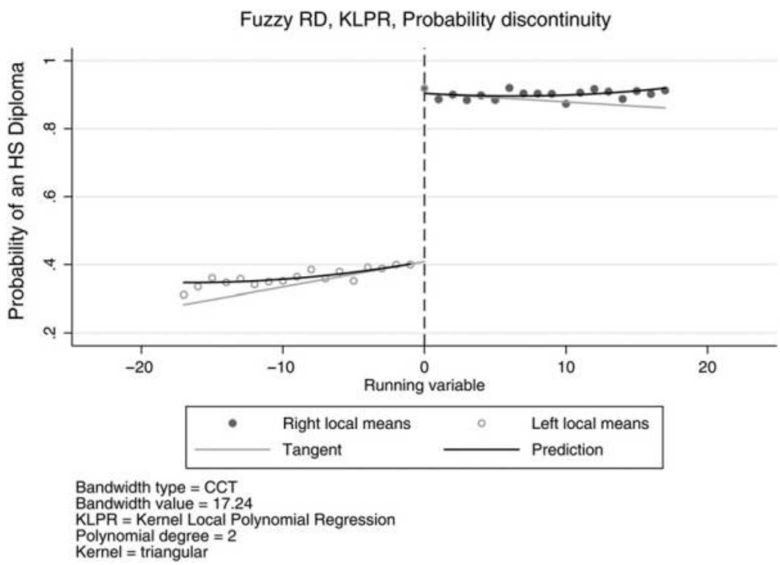

Fig. 2. Fuzzy RD Discontinuity in the Probability and Tangents Lines at Threshold. *Source*: Dataset from Clark and Martorell (2010).

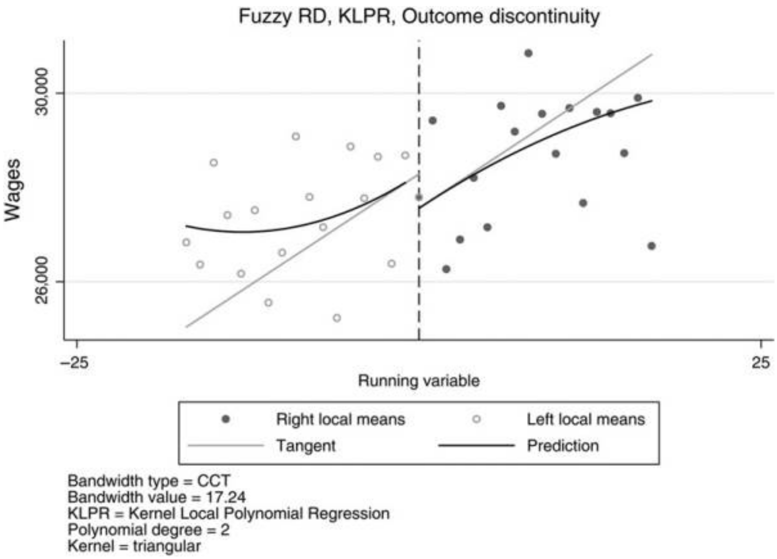

Fig. 3. Fuzzy RD Discontinuity in the Outcome and Tangents Lines at Threshold.
Source: Dataset from Clark and Martorell (2010).

where the threshold is located. This small TED suggests that if the threshold had been somewhat lower or higher, the estimated LATE would likely still have been close to zero.

Our numerical estimates confirm what is seen in these figures. Table 2 presents both the estimated first-stage discontinuity and the RD LATE, while Table 3 provides the CPD and the TED. We compare estimates based on three popular bandwidth selectors: the CCT (Calonico et al., 2014), the IK (Imbens & Kalyanaraman, 2012), and the CV (Ludwig & Miller, 2007) bandwidths. The table also considers two different kernel functions, the triangular kernel, which is shown to be optimal for estimating the conditional mean at a boundary point (Fan & Gijbels, 1996) and the uniform kernel, which is commonly used for its convenience. As noted earlier, we use local quadratic regressions based on Porter (2003) and Gelman and Imbens (2014). As discussed in Dong and Lewbel (2015), for fuzzy designs one can estimate TED either by a local two-stage least squares (2SLS), using Z, ZX, and ZX^2 as excluded IVs for T and TX in the outcome model to get point estimates, or one can estimate local quadratic regressions separately for the reduced-form outcome and treatment equations, and then construct TED

Table 2. Fuzzy RD Estimates of the Impacts of HS Diploma on Wages.

	CCT1	IK1	CV1	CCT2	IK2	CV2
First-stage discontinuity	0.497	0.499	0.516	0.497	0.502	0.523
	(0.016)***	(0.012)***	(0.008)***	(0.017)***	(0.012)***	(0.007)***
RD LATE	−1556.1	−1659.0	−113.8	−2742.3	−1067.5	−45.4
	(2640.7)	(2418.0)	(1602.7)	(2728.9)	(2384.9)	(1592.3)
Bandwidth	11.63	19.58	50.00	8.29	15.39	42.50
N	13,364	21,694	40,795	10,051	17,744	37,715

Notes: This table uses the Florida data from Clark and Martorell (2014); All RD LATE estimates are based on the bias-corrected robust inference proposed by Calonico et al. (2014) using local linear regressions; CCT refers to the optimal bandwidth by CCT; IK refers to the optimal bandwidth proposed by Imbens and Kalyanaraman (2012); CV refers to the cross validation optimal bandwidth proposed by Ludwig and Miller (2007); 1 uses a triangular kernel, and 2 uses a uniform kernel; Standard errors are in parentheses.
*Significant at the 10% level, **significant at the 5% level, ***significant at the 1% level.

Table 3. TED and CPD of Fuzzy RD Treatment Effects of HS Diploma on Wages.

	CCT1	IK1	CV1	CCT2	IK2	CV2
CPD	−0.006	−0.006	−0.005	−0.010	−0.006	−0.004
	(0.003)*	(0.003)*	(0.001)***	(0.006)*	(0.003)*	(0.001)***
TED	287.9	296.0	44.3	509.3	360.3	−23.0
	(529.1)	(499.2)	(243.7)	(464.2)	(702.3)	(194.2)
Bandwidth	24.41	25.16	50.00	23.46	21.04	50.00
N	21,694	26,846	40,795	17,744	23,460	41,220

Notes: This table uses the Florida data from Clark and Martorell (2014); All estimates are based on local quadratic regressions; CCT refers to the optimal bandwidth proposed by Calonico et al. (2014); IK refers to the optimal bandwidth proposed by Imbens and Kalyanaraman (2012); CV refers to the cross validation optimal bandwidth proposed by Ludwig and Miller (2007); 1 uses a triangular kernel, and 2 uses a uniform kernel; Bandwidth and sample size *N* refer to those of the outcome equation; Bootstrapped Standard errors based on 500 simulations are in parentheses.
*Significant at the 10% level, **significant at the 5% level, ***significant at the 1% level.

from the estimated intercepts and slopes in the two equations as described in Section 4. For convenience, we chose the latter method, using the bootstrap to calculate standard errors.

Consistent with findings in Clark and Martorell (2010), Table 2 shows that the probability of receiving a high school diploma increases by about 50% at the threshold, which is statistically significant at the 1% level and is largely insensitive to different bandwidth and kernel choices. In contrast, all of the RD LATE estimates in Table 2 are numerically small and statistically not significant.

The estimates of CPD in Table 3 range from −0.004 to −0.010, which are all statistically significant. The normalized exam score ranges from −100 to 50. These estimates suggest that, given a 10-point decrease in the exit exam score, the percent of students who are compliers would increase from about 50% to somewhere between 54% and 60%. The relative CPD implied by the estimates in Tables 2 and 3 is around five. So the set of compliers looks unstable. However, as one would guess from Fig. 3, the estimates of TED in Table 3 are rather small and not statistically

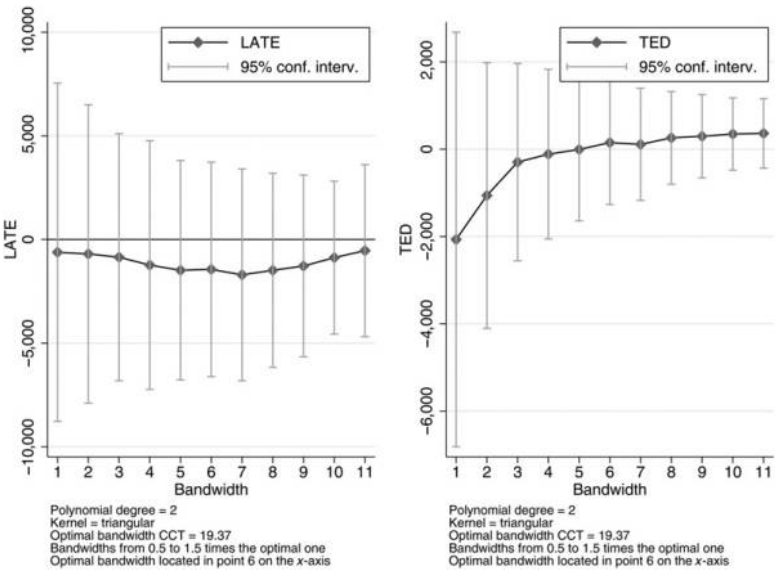

Fig. 4. Fuzzy RD LATE and TED Point Estimations and Confidence Intervals over a Range of Bandwidths. Bootstrap Confidence Intervals Based on 200 Replications. *Source*: Dataset from Clark and Martorell (2010).

significant. The implied relative TED estimates are all well below one. Together these results indicate that although the set of compliers is not stable, Clark and Martorell's conclusion (that among students with test scores near the cutoff, there is little effect of having diploma or not) does appear stable.

Finally, in Fig. 4, we check how sensitive these estimates are to bandwidth choice (a comparable exercise was not needed in our previous empirical application because the extremely large sample size there resulted in very little dependence on bandwidth). Consistent with Tables 2 and 3, Fig. 4 shows that the point estimates of both RD LATE and TED at varying bandwidths are almost all near zero. The only exception is that the TED estimate moves away from zero at the lowest bandwidth, however, the confidence band around the estimate also widens considerably at that bandwidth, so in all cases both the RD LATE and the TED are statistically insignificant.

7. COVARIATES

As noted earlier, one advantage of TED over other tools for evaluating RD estimates is that TED does not require covariates. Nevertheless, one often does have covariates, and one may readily estimate TED (and CPD) conditioning on covariate values. For example, given a binary covariate W, one may replace each β_j in Eq. (1) with $\beta_{j0}(1-W)+\beta_{j1}W$. Then in sharp designs β_{20} and β_{21} are the average treatment effects conditioning on $W=0$ and $W=1$ respectively, and similarly β_{40} and β_{41} are conditional TED values, conditioning on $W=0$ and $W=1$ respectively. For fuzzy design, one would also replace each α_j in Eq. (2) with $\alpha_{j0}(1-W)+\alpha_{j1}W$. Then $\pi_{fw}(c)=\beta_{2w}/\alpha_{2w}$ is fuzzy conditional average treatment effect, conditioning on $W=w$ (for w equal to zero or one), and it follows immediately that the corresponding fuzzy design conditional TED is $\left(\beta_{4w}-\pi_{fw}(c)\alpha_{4w}\right)/\alpha_{2w}$. The extension to more covariate values is straightforward.

One possible use for conditional TED values is that RD estimates could turn out to be relatively stable for some sets of covariate values but not others. One might then have more confidence in making policy recommendations in other contexts based on the more stable RD estimates. Alternatively, one would have still more confidence in RD estimates if they appear stable not just unconditionally, but also conditionally on each covariate value.

8. TED AND MTTE

We have stressed the use of TED as a stability measure, but as derived in Dong and Lewbel (2015) (see their paper for more details) under a local policy invariance assumption TED equals the MTTE, that is the marginal threshold treatment effect. Here, we provide a little more insight into what TED means, by comparing the difference between TED and MTTE.

For sharp design RD, define $S(x,c) = E[Y(1) - Y(0)|X = x, C = c]$, that is, $S(x,c)$ is the conditional expectation of the treatment effect $Y(1) - Y(0)$, conditioning on having the running variable X equal the value x, and conditioning on having the threshold equal the value c. If the design is fuzzy, then assume $S(x,c)$ is defined as conditioning on $X = x$, $C = c$, and on being a complier. It follows from this definition that for sharp designs $\pi(x) = S(x,c)$, and for fuzzy designs $\pi_f(x) = S(x,c)$. So $S(c,c)$ is the LATE that is identified by standard RD estimation.

The level of the cutoff c is a policy parameter. This notation makes explicit that the RD LATE depends on this policy. When $x \neq c$, the function $S(x,c)$ is a counterfactual. It defines what the expected treatment effect would be for a complier who is not actually at the cutoff c, but instead has his running variable equal to x, despite having the policy for everyone be that the cutoff equals c.

Define the function $s(x)$ by $s(x) = S(x,x)$. What TED equals is given by

$$\text{TED} = \pi'(c) = \frac{\partial S(x,c)}{\partial x}\Big|_{x=c}.$$

In contrast, the MTTE is defined by

$$\text{MTTE} = \frac{ds(c)}{dc} = \frac{\text{dS}(c,c)}{dc} = \frac{\partial S(x,c)}{\partial x}\Big|_{x=c} + \frac{\partial S(x,c)}{\partial c}\Big|_{x=c} = \pi'(c) + \frac{\partial S(x,c)}{\partial c}\Big|_{x=c}.$$

As the above equations show, TED equals MTTE if and only if $\left[\partial S(x,c)/\partial c\right]\big|_{x=c} = 0$, that is, if one's expected average treatment effect at a given value x of the running variable would not change if c marginally changed. This is the local policy invariance assumption defined by Dong and Lewbel (2015). This assumption is similar to, but weaker than, the general policy invariance condition defined by Abbring and Heckman (2007).

If we knew the MTTE, then we could evaluate how the treatment effect would change if the cutoff marginally changed from c to $c + \varepsilon$, using

$$s(c+\varepsilon) \approx s(c) + \frac{ds(c)}{dc}\varepsilon.$$

To illustrate these concepts, consider the empirical applications analyzed in the previous section. In the Haggag and Paci (2014) application, local policy invariance implies that, holding my taxi fare fixed at $c = \$15$, the difference in the amount I would tip depending on which set of tip suggestions I saw would not change if, for everybody else, the threshold for switching between the two sets of tip suggestions changed from c to $c+\varepsilon$ for a small ε. In this application, local policy invariance is quite plausible, since it is unlikely that people even notice the discontinuity at all. If local policy invariance does hold here, then we can use our estimate of TED to estimate how the tip suggestion LATE would change if the cutoff at which the change in tip suggestions occurred were raised or lowered a little.

In contrast, consider the Clark and Martorell (2010, 2014) application. There, local policy invariance roughly means that, holding my grade level fixed at $x = c$, my expected earnings difference between having a diploma or not would not change if, for all other compliers, the grade cutoff changed marginally from c to $c+\varepsilon$ for a tiny ε. In this application, local policy invariance might not hold due to general equilibrium effects, for example, a change in c might change employer's perception of the value of a diploma. More generally, local policy invariance might not hold for the same reasons that Rubin's (1978) stable unit treatment value assumption (SUTVA) could be violated, that is, if treatment of one individual might affect the outcomes of others. However, suppose local policy invariance does hold for this application (or that $\left[\partial S(x,c)/\partial c\right]\big|_{x=c}$ is close to zero, as it would be if these general equilibrium effects are small). Then the TED we reported would not just be a stability measure, but it would also tell us how much the RD treatment effect of getting a diploma would change if the cutoff grade were raised or lowered by a small amount.

These cutoff change calculations are relevant because many policy debates center on whether to change thresholds. Examples of such policy thresholds include minimum wage levels, the legal age for drinking, smoking, voting, medicare, or pension eligibility, grade levels for promotions, graduation, or scholarships, and permitted levels of food additives or of environmental pollutants. In addition to measuring stability, TED at minimum provides information that is relevant for these debates, since even if local policy invariance does not hold, TED at least comprises one component of the MTTE.

9. CONCLUSIONS

We reconsider the TED estimator defined in Dong and Lewbel (2015), and we define a related CPD estimator for fuzzy RD designs. We note how both of these are nonparametrically identified and can be easily estimated using only the same information that is needed to estimate standard RD models. No covariates or other outside information is needed to calculate the TED or CPD.

Dong and Lewbel (2015) focus on using the TED to evaluate the impact of a hypothetical change in threshold given a local policy invariance assumption. In contrast, we show here that both the TED and CPD provide valuable information, regardless of whether local policy invariance holds or not. In particular, we claim that the TED should be examined in nearly all RD applications, as a way of assessing the stability and hence the potential external validity of RD estimates. Additionally, in fuzzy RD applications, one should examine the CPD (in addition to the TED), since the CPD can be used to evaluate the stability of the complier population. We provide a simple rule of thumb for determining if TED or CPD are large enough to indicate instability.

We illustrate these claims using two different empirical applications, one sharp design and one fuzzy design. We find that the sharp RD treatment effects of taxi tip suggestions reported in Haggag and Paci (2014) are not stable, indicating that the magnitudes of their estimated treatment effects might change significantly at slightly lower or higher tip levels. The second application we consider is the fuzzy RD model of Clark and Martorell (2014). The CPD of their model suggests that the set of compliers is not stable, but the TED is numerically small and statistically insignificant. This near-zero TED suggests that their finding of almost no effect of receiving a diploma on wages (conditional on holding the level of education, as indicated by test scores, fixed) is stable, and so would likely remain valid at somewhat lower or higher test score levels.

ACKNOWLEDGEMENT

We would like to thank Damon Clark for sharing the Florida data.

REFERENCES

Abbring, J. H., & Heckman, J. J. (2007). Econometric evaluation of social programs, part III: Distributional treatment effects, dynamic treatment effects, dynamic discrete choice, and general equilibrium policy evaluation. In J. J. Heckman, & E. E. Leamer (Eds.), *Handbook of Econometrics* (Vol. 6). Berlin; Heidelberg, Germany: Elsevier.

Angrist, J. D., & Fernandez-Val, I. (2013). ExtrapoLATE-ing: External validity and overidentification in the LATE framework. In *Advances in economics and econometrics: Theory and applications, tenth world congress, volume III: Econometrics*. Econometric Society Monographs.

Angrist, J. D., & Rokkanen, M. (2015). Wanna get away? Regression discontinuity estimation of exam school effects away from the cutoff. *Journal of the American Statistical Association*, *110*(512), 1331–1344.

Bertanha, M., & Imbens, G. W. (2014). *External validity in fuzzy regression discontinuity designs*. No. w20773. National Bureau of Economic Research.

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

Card, D., Lee, D. S., Pei, Z., & Weber, A. (2015). Inference on causal effects in a generalized regression kink design. *Econometrica*, *83*(6), 2453–2483.

Clark, D., & Martorell, P. (2010). *The signaling value of a high school diploma*. Princeton IRS Working Paper No. #557.

Clark, D., & Martorell, P. (2014). The signaling value of a high school diploma. *Journal of Political Economy*, *122*(2), 282–318.

Dinardo, J., & Lee, D. S. (2011). Program evaluation and research designs. In O. Ashenfelter, & A. Card (Eds.), *Handbook of labor economics* (Vol. 4a, pp. 463–536). Amsterdam: Elsevier.

Dong, Y. (2016a). *Jump or kink? Regression probability jump and kink design for treatment effect evaluation*. Unpublished Manuscript.

Dong, Y. (2016b). *An alternative assumption to identify LATE in regression discontinuity designs*. Unpublished Manuscript.

Dong, Y., & Arthur Lewbel. (2015). Identifying the effect of changing the policy threshold in regression discontinuity models. *Review of Economics and Statistics*, *97*(5), 1081–1092.

Fan, J., & Gijbels, I. (1996). *Local polynomial modelling and its applications*. London: Chapman and Hall.

Gelman, A., & Imbens, G. (2014). *Why high-order polynomials should not be used in regression discontinuity designs*. NBER Working Paper No. 20405.

Haggag, K., & Paci, G. (2014). Default tips. *American Economic Journal: Applied Economics*, *6*(3), 1–19.

Hahn, J., Todd, P. E., & Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica, 69*, 201–209.

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

Lee, D. S. (2008). Randomized experiments from non-random selection in US House elections. *Journal of Econometrics*, *142*(2), 675–697.

Ludwig, J., & Miller, D. L. (2007). Does head start improve children's life chances? Evidence from a regression discontinuity design. *The Quarterly Journal of Economics*, *122*, 159–208.

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

Porter, J. R. (2003). *Estimation in the regression discontinuity model*. Unpublished Manuscript.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics* *6*, 34–58.

Rubin, D. B. (1980). Using empirical Bayes techniques in the law school validity studies. *Journal of the American Statistical Association*, *75*(372), 801–816.

Rubin, D. B. (1990). Formal mode of statistical inference for causal effects. *Journal of Statistical Planning and Inference*, *25*(3), 279–292.

Wing, C., & Cook, T. D. (2013). Strengthening the regression discontinuity design using additional design elements: A within-study comparison, *Journal of Policy Analysis and Management*, *32*, 853–877.