Regression Discontinuity Designs in a Latent Variable Framework

James Soland
University of Virginia & NWEA
(Corresponding Author)

Angela Johnson
NWEA

Eli Talbert
University of Virginia

University of Virginia
School of Education and Human Development
405 Emmet Street
Charlottesville, VA 22904
Phone: 434-924-0742
Email: jgs8e@virginia.edu

Abstract

When randomized control trials are not available, regression discontinuity (RD) designs are a viable quasi-experimental method shown capable of producing causal estimates of how a program or intervention affects an outcome. While the RD design and many related methodological innovations came from the field of psychology, RDs are underutilized among psychologists even though many interventions are assigned on the basis of scores from common psychological measures, a situation tailor-made for RDs. In this tutorial, we present a straightforward way to implement an RD model as a structural equation model (SEM). By using SEM, we both situate RDs within a method commonly used in psychology, as well as show how RDs can be implemented in a way that allows one to account for measurement error and avoid measurement model misspecification, both of which often affect psychological measures. We begin with brief Monte Carlo simulation studies to examine the potential benefits of using a latent variable RD model, then transition to an applied example, replete with code and results. The aim of the study is to introduce RD to a broader audience in psychology, as well as show researchers already familiar with RD how employing an SEM framework can be beneficial.

Keywords: Structural equation modeling (SEM), regression discontinuity, quasi-experimental designs, instrumental variables, measurement error, treatment effects.

Regression Discontinuity Designs in a Latent Variable Framework

Randomized control trials (RCTs) are the gold standard when evaluating interventions and programs in psychology, as well as related disciplines like education and medicine. However, RCTs are often not feasible due to practical, logistical, financial, or other impediments. In such cases, there may be quasi-experimental alternatives that exploit random variation in how study participants are assigned to receive a treatment, participate in a program, or undergo an intervention (Imbens & Rubin, 2017). A particularly rigorous quasi-experimental option is regression discontinuity or RD (Imbens & Lemieux, 2008). RDs are ideal for cases in which program or intervention participation is assigned using a clear threshold on a particular measure, such as when, say, participants receive a treatment if their income is below a certain threshold. If one assumes that participants just on either side of the cut score are identical except for measurement error on the test or other metric upon which the treatment determination is made, then they are as good as randomly assigned to treatment. A key advantage of the RD is its high internal validity from addressing omitted variable bias (Angrist & Pischke, 2009), though inferences about the treatment are limited to participants near the cut score, potentially reducing generalizability.[1]

Yet, even though many methodological developments germane to the RD model have been described and addressed by psychologists, and despite loosely related methods being employed in psychology (e.g., Sequential Multiple Assignment Randomized Trial or SMART [Lei et al., 2012]), the method is underutilized in the field (Cook, 2007). For example, Moscoe et al. (2015) pointed out that RDs are underutilized in public health, and especially in psychiatry,

[1] One should note that internal validity is only strong when the RD assumptions are met, which is not guaranteed. Further, rather than estimating average treatment effects (ATEs) as RCTs do, RDs estimate local average treatment effects (LATEs) using data just on either side of the cut score; this is an important limitation to the RD's external validity.

where drug regimens are often prescribed on the basis of cut scores. Similarly, Cook (2007) argued that RDs suffer from misunderstandings of their limitations and uses that reduce their application in psychology despite how often interventions are based on cut scores developed for clinical purposes, creating ideal circumstances for an RD. There are also likely substantive reasons RDs are not used in psychology. For instance, such designs only allow one to estimate a treatment effect for participants proximal to the cut score, which may not be the estimand of interest in some psychological studies. Further, descriptive studies appear more common in psychology than in fields like economics, due both to the substantive questions of interest and complexities of study designs, which would make RDs less relevant in those cases. Nonetheless, the common use of cut scores for clinical purposes and frequent desire for causal estimates in psychological likely mean RDs are underutilized in the field, with RDs still "waiting for life to arrive in psychology" (Cook, 2007, p. 643).

An additional possible reason for the underutilization of RD in psychology may be that related models are most often presented in an econometric framework. Such econometric models are not typically presented in a way familiar to researchers in psychology, nor are they designed specifically for use when the dependent variable is a short survey scale, which are commonly employed in psychology. Yet, as we show, RDs can be straightforwardly implemented in a structural equation modeling (SEM) framework that is likely more familiar to quantitative psychologists and can include a measurement model for the dependent variable that can help account for the imperfections in the outcome measures used.

Including a measurement model is especially important given growing evidence that using sum scores can produce very different results (e.g., reliability coefficients, rank orderings of respondents) when their assumptions are violated compared to using a factor model making

less stringent assumptions (Kuhfeld & Soland, 2020; McNeish & Wolf, 2020). Yet, evidence suggests sum scores are nonetheless used in the majority of studies within psychology (Flake et al., 2017). Moreover, when meeting or missing the cut score in an RD does not guarantee receiving treatment (e.g., not all patients falling below a cut score on a depression diagnostic receive medication and the decision is made at the doctor's discretion), instrumental variables (IVs) are typically used and can easily be accommodated in an SEM framework. Despite these benefits, RDs are rarely if ever estimated using SEM, though there are several examples that blend other quasi-experimental approaches (especially propensity score methods) with SEM (Leite et al., 2019; Raykov, 2012; Rodríguez De Gil et al., 2015).

In this context of underuse of RDs within psychology, our own study has two broad purposes. First, we explore some of the potential benefits of estimating RDs in an SEM framework when the outcome of interest is captured using multiple indicators measured with error (e.g., survey item responses). We illustrate these benefits using first principles, then conduct brief simulation studies to investigate the magnitude of those potential benefits more concretely. Second, given these benefits, we provide a detailed demonstration of how to implement RDs in an SEM framework. In both simulation studies and the demonstration, we focus on latent variables for the outcome and not the measure used to assign participants to treatment because accounting for measurement error in the latter involves much more complex considerations (as described in the discussion section). Our goal is to make the usefulness of RD methods more apparent to researchers in psychology and the methods more accessible, as well as provide a way to implement an RD that includes a measurement model when survey scores or other like measures are the dependent variable. In our demonstration, code to estimate the RD in an SEM is provided in Mplus and Stata along with the accompanying data.

Background

The Logic of RD Designs

The concept behind RD designs is fairly intuitive. When an RCT is not possible, one can still potentially generate causal claims about the impact of treatment through so-called "natural experiments." In an RD framework, this occurs when assignment to treatment is based on a metric with a clear cut score. As an example, medicine to treat psychological conditions is often prescribed using cut scores from common diagnostic measures (Moscoe et al., 2015), and some interventions are provided partly on the basis of income (Miranda et al., 2002). If one assumes that, in the medication example, patients within a point or two of the cut score are virtually identical given measurement error, then patients very near the cut score are as good as randomly assigned to receive the medication. Thus, estimates of how the patients' conditions improve for those just above or below the cut score can provide a plausibly causal estimate of the medication's effect. Such designs can allow for causal claims, but only for participants near the cut score (a limitation with implications for external validity), and only when the assumptions of the RD are met, which one can examine empirically (as we show in the demonstration).

In terms of the specific RD model, let r_i be the forcing variable for person *i* (also referred to as the "running variable" or the "rating variable", the diagnostic measure in our medication example) in a given sample. This forcing variable must be continuous or semi-continuous with discrete values, such as test scores with integer values (see Lee and Card [2008] and Kolesár and Rothe [2018] for using semi-continuous forcing variables). Let r^* be the cut score for this variable (also called the treatment threshold). Oftentimes, r_i is centered at r^* such that $r^* = 0$, which makes interpretation of regression coefficients more straightforward. A participant i's assigned treatment is represented by z_i such that

$$z_i = \mathrm{I}\{r_i \geq r^*\} = \begin{cases} 1 \; if \; r_i \geq r^* \\ 0 \; if \; r_i < r^* \end{cases} \quad (1)$$

For now, we will assume that z_i is equal to t_i, the treatment the participant *actually* experiences. If our observed (sum) score on the outcome of interest is denoted as o_i, then a basic parametric RD formulation[2] might be:

$$o_i = \beta_0 + \beta_1 r_i + \beta_2 t_i + \beta_3 t_i r_i + \epsilon_i \quad (2)$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. In the above, β_1 represents the unit change in the outcome for every one-unit change in the forcing variable, β_2 represents the mean treatment effect (the parameter of interest, generally), and the interaction term, $\beta_3 t_i r_i$, allows for the slope to differ on either side of the cut score. Further, quadratic terms or higher order polynomials can be added that are the same or different on either side of the cut score. Estimates are only unbiased if the functional form correctly models the relationship between treatment status and the outcome on both sides of the cut score.

The RD design estimates local average treatment effects (LATEs) using data with forcing variable values near the cut score. Results are thus influenced not only by the choice of functional form of the model, but also the bandwidth used to estimate the treatment effect. The bandwidth refers to how wide a range of the forcing variable around the cut score is used to fit the local linear regression. For example, we could use observations with scores on r_i within one standard deviation (SD) of the cut score to estimate treatment effects, assuming the functional form for those in this bandwidth matches the functional form for those very near the cut score.

[2] Given our focus on applied uses of RD designs, we do not present the nonparametric form of the model here. Angrist and Pischke (2009) identified several challenges associated with the nonparametric approach to RD and highlighted that most applied RD work is parametric, and sophisticated nonparametric RD methods have not yet found wide application in empirical practices. However, nonparametric presentations of the model can be vital to understanding the causal assumptions of the model. Interested readers can consult not only Angrist and Pischke (2009), but also Hu and Schennach (2008) and Hahn, Todd, and Van der Klaauw (2001).

The choice of bandwidth involves a tradeoff between bias and precision in the estimation, with implications for internal and external validity. Fortunately, data-driven approaches to calculating optimal bandwidths can be found in Cattaneo et al. (2018), Cattaneo et al. (2019), and Imbens and Kalyanaraman (2012).

Sharp Versus Fuzzy RD

Until now, we have assumed that $z_i = t_i$, which may be feasible in some constrained contexts (e.g., lab studies). That is, if everyone assigned to treatment based on the cut score actually undergoes treatment, then this proposition holds. This scenario is commonly referred to as a "sharp RD". However, sometimes z_i does not equal t_i, which necessitates a "fuzzy RD." For instance, psychologists might combine their own professional judgment with the results from a diagnostic measure. As a result, some patients who would be assigned to treatment based on the cut score do not actually receive treatment and some who would be assigned to control based on the cut score receive treatment anyway. In such a situation, one could simply estimate a treatment effect for those who actually received treatment, (i.e., for whom t_i = 1). However, such an estimate would be biased if an unobserved variable was correlated with whether the study participant complied with assignment to treatment and the outcome of interest, resulting in omitted variable bias.

To avoid such bias, IVs are used to account for a potential correlation between treatment status and the outcome of interest. Here, the assignment, z_i, serves as the instrument that affects the outcome only through its effect on the actual treatment status, t_i. This exclusion of a direct causal link between the instrument and the outcome is called the "exclusion restriction" and helps address concerns about omitted variable bias. That is, we can estimate the indirect effect of z_i through t_i on the outcome, which does not suffer from selection bias, and use that indirect

effect to produce an unbiased estimate of the direct effect of the treatment. In most econometric research, IV treatment effects are estimated through a two-stage least squares (2SLS) approach. In the first stage, the treatment status (i.e., whether treatment was received) is regressed on the whether the participant was assigned to treatment and other exogenous covariates, then the predicted values of t_i are saved. In the second stage, the outcome of interest is regressed on the predicted value of treatment status from the first stage plus other exogenous covariates.[3]

RDs in a Latent Variable Framework

One can also express an RD as an SEM that directly incorporates latent variables. An example for the sharp RD is shown in Figure 1(a). In that figure, y_{i1} through y_{i4} are observed indicators of the construct of interest (e.g., scores from four different clinical measures or survey items) and η_i is a latent variable underlying those observed indicators that measures the outcome of interest. Here, we assume the functional form is the same on either side of the cut score and that there are no polynomials in the model (both done for simplicity).

[3] In several cases, 2SLS software generates predicted values of t_i by calculating the probability of a positive outcome, while other software/approaches use the linear prediction. These differences can lead to discrepant scalings of the treatment effect estimate, an issue we address in our demonstration.

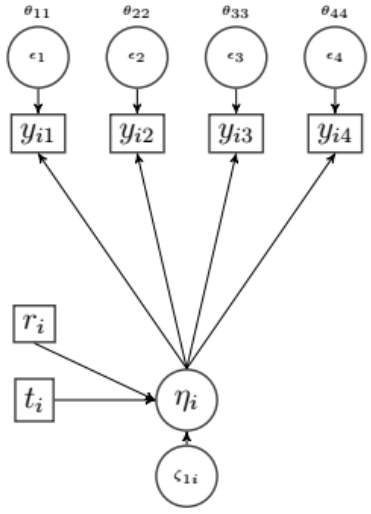

Figure 1(a). *Path diagram for the latent sharp RD model.*

One could express such a path diagrams in equations by having a measurement model for person *i* and indicator *j*:

$$\boldsymbol{y}_i = \boldsymbol{\nu} + \boldsymbol{\lambda}\eta_i + \boldsymbol{\epsilon}_i \quad (3)$$

Where $\boldsymbol{y}_i$ is an $N\ x\ 1$ vector of observed indicator scores/item responses for indicators/items $1\ldots N$, $\boldsymbol{\nu}$ is an $N\ x\ 1$ vector of intercepts, $\boldsymbol{\lambda}$ is an $N\ x\ 1$ vector of loadings, η_i is the single latent variable for person *i*, and $\boldsymbol{\epsilon}_i$ is an $N\ x\ 1$ vector of residuals (each with a mean of zero) where $VAR(\boldsymbol{\epsilon}_i) = \boldsymbol{\theta} = diag(\theta_{11}, \theta_{22} \ldots \theta_{NN})$. Meanwhile, the structural model is

$$\eta_i = \alpha + \gamma_1 t_i + \gamma_2 r_i + \zeta_{1i} \quad (4)$$

with true score variance ϕ.

One could expand this path diagram to include a fuzzy RD design with an IV per Figure 1(b). Unlike in the 2SLS approach, both stages are estimated simultaneously, with treatment status and the outcome of interest included as dependent variables in the same model (Murnane & Willett, 2010). As the figure shows, there is a path from z_i to t_i, but no covariance between z_i

and η_i otherwise—a visual representation of the exclusion restriction. That is, $cov(z_i, \eta_i = 0)$. Further, in this path diagram, the correlation in the error terms between η_i and t_i is expressed directly and estimated with the assumption that $cov(t_i, \eta_i \neq 0)$. As we discuss below, there are several potential benefits to estimating an RD using the below framework.

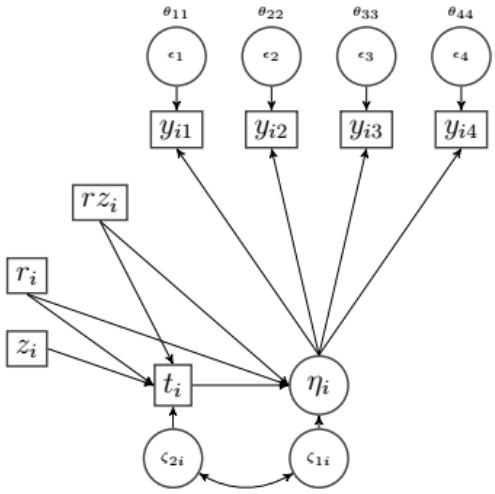

Figure 1(b). *Path diagram for the latent fuzzy RD model.*

Potential Benefits of Using a Latent Variable Model to Estimate an RD

Measurement Model Misspecification When Using Sum Scores. A primary argument in favor of using an SEM to estimate an RD is to avoid using sum scores when their assumptions are violated. McNeish and Wolf (2020) demonstrated that sum score approaches are the equivalent of fitting a highly constrained measurement model that assumes (often wrongly) that indicators should be weighted equally (typically by constraining loadings equal) and that the error terms are equivalent across indicators. As research has shown, such misspecification of the measurement model used to score the dependent variable can lead to wildly different parameter estimates compared to when a measurement model that is not misspecified is used, including

treatment effect estimates (Authors, in press; Bauer & Curran, 2015; Kuhfeld & Soland, 2020; McNeish & Wolf, 2020). To be clear, while estimates based on sum scores versus latent variables can differ due to simple linear or monotonic transformations between the two, that is not always what we describe here. Rather, the scenario we describe involves using a sum score that assumes equal weighting of the indicators (loadings) when the data-generating model involves weights that are not equal. While one could technically use a weighted sum score, where each indicator is weighted in proportion to its loading, and such scores would have the property of being sufficient for the actual latent variable being measured, in practice, sum scores virtually always involve constraining the weights of the indicators equal.

Use of sum scores (or observed measures in general) to produce the dependent variable could lead to different treatment effect estimates in an RD compared to using a less constrained measurement model. Let us return to Equation 4. For now, to focus on recovery of true treatment effects, we will drop $\gamma_2 r_i$ from the model, but what follows would still stand if the term were included (and could even further complicate the differences between sum scores and measurement models that do not constrain the loadings equal). If we assume $\boldsymbol{\nu}$ is a vector of zeros (done for convenience, though not necessary) and α is zero (i.e., the mean η_i for the control group is zero) and substitute the structural equation into the measurement equation, we get

$$\boldsymbol{y}_i = \boldsymbol{\lambda}(\gamma_1 t_i + \zeta_i) + \boldsymbol{\epsilon}_i \quad (5)$$

Given $E(\zeta_i) = 0$ and $E(\boldsymbol{\epsilon}_i) = 0$,

$$E(\boldsymbol{y}_i) = E(\boldsymbol{\lambda}(\gamma_1 t_i + \zeta_i) + \boldsymbol{\epsilon}_i) = \boldsymbol{\lambda}\gamma_1 E(t_i) \quad (6)$$

Since $t_i = 0$ for the control group, the expectation of the vector of observed scores for the control group $E(\boldsymbol{y}_i)$ is simply an $N\ x\ 1$ vector of zeros. Meanwhile, $E(\boldsymbol{y}_i)$ for the treated group is

$$E(\boldsymbol{y}_i\ |t_i = 1)\ = \boldsymbol{\lambda}\gamma_1 E(t_i) = \ \boldsymbol{\lambda}\gamma_1 \qquad (7)$$

Thus, one could express the difference in the expectation of the observed scores between control and treatment groups as

$$E(\boldsymbol{y}_i\ |t_i = 1) - \ E(\boldsymbol{y}_i\ |t_i = 0)\ = E(\boldsymbol{y}_i\ |t_i = 1) - \ 0\ = \boldsymbol{\lambda}\gamma_1 \quad (8)$$

As Equation 8 shows, the difference in the means of the control and treatment observed scores would be the true treatment effect, γ_1, *weighted by the loadings*.

If the loadings are below one, the true control-treatment differences would be larger than the observed (and vice-versa). Thus, when one uses an observed mean difference like with a sum score, those observed scores can misweight the indicators such that treatment effect estimates will differ compared to when the indicator weights are accounted for in the measurement model. Further, if one does not assume $\gamma_2 = 0$, then additional differences could be introduced into the estimate of the coefficient on r_i by failing to account for those weights. Note that differences between the estimates of the treatment effect between sum scores and less constrained measurement models result from the violation of statistical/parametric assumptions in the measurement model when using sum scores. That is, if a researcher assumes the sum score is correct, or if the indicator weights are indeed equal in the data-generating model (such as when using the Rasch Item Response Theory [IRT] model to score achievement tests), then the issue is not consequential. In our analyses, we generally assume the researcher cares about the causal inference with respect to the latent variable, and that the true loadings are not equal.

Measurement Invariance Failures and Response Shifts. Another benefit of the latent variable approach to estimating RDs is addressing measurement invariance failures. This issue is especially important given growing evidence that participating in an intervention can actually induce measurement noninvariance in the dependent variable by changing how treated individuals perceive the construct of interest (Oort, 2005; Oort et al., 2005). For instance, recent research on the effect of an invasive surgery for cancer patients found that accounting for these "response shifts" meaningfully changed estimates of physical health pre- and post-treatment. (Oort, 2005; Oort et al., 2005). There is similar evidence that response shifts have impacted outcomes in depression studies (Fokkema et al., 2013). Such measurement inconsistencies across groups (measurement noninvariance), including control and treatment groups, can be addressed in an SEM framework in a way that largely mitigates any resultant bias in treatment effect estimates (Oort, 2005), but are harder to address using sum scores.[4]

The potential effects of measurement invariance failures between control and treatment can also be expressed mathematically. If we return to Equation 5 above, but no longer assume α and $\boldsymbol{\nu}$ are zero (or a vector of zeros for $\boldsymbol{\nu}$), then we would have

$$\boldsymbol{y}_i = \boldsymbol{\nu} + \boldsymbol{\lambda}(\alpha + \gamma_1 t_i + \zeta_i) + \boldsymbol{\epsilon}_i \quad (9)$$

Given $E(\zeta_i) = 0$ and $E(\boldsymbol{\epsilon}_i) = 0$,

$$E(\boldsymbol{y}_i) = E(\boldsymbol{\nu} + \boldsymbol{\lambda}(\alpha + \gamma_1 t_i + \zeta_i) + \boldsymbol{\epsilon}_i) = \boldsymbol{\nu} + \boldsymbol{\lambda}\alpha + \boldsymbol{\lambda}\gamma_1 E(t_i) \quad (10)$$

If one assumes measurement invariance between control and treatment groups, one could express the vector of differences in mean observed scores as

$$E(\boldsymbol{y}_i \,|t_i = 1) - E(\boldsymbol{y}_i \,|t_i = 0) = \quad (11)$$

[4] While one could perform differential item functioning (DIF) analyses using sum scores and potentially drop items showing DIF, directly accounting for noninvariance is otherwise more complicated with sum scores than when using SEM.

$$(\boldsymbol{v} + \boldsymbol{\lambda}\alpha + \boldsymbol{\lambda}\gamma_1) - (\boldsymbol{v} + \boldsymbol{\lambda}\alpha) = \boldsymbol{\lambda}\gamma_1$$

However, if there is noninvariance in the intercepts or loadings between the groups such that measurement model parameters have a *g* subscript, we would have

$$E(\boldsymbol{y}_i \mid t_i = 1) - E(\boldsymbol{y}_i \mid t_i = 0) = \quad (12)$$
$$(\boldsymbol{v}_g + \boldsymbol{\lambda}_g\alpha + \boldsymbol{\lambda}_g\gamma_1) - (\boldsymbol{v}_g + \boldsymbol{\lambda}_g\alpha)$$

Hypothetically, if the loadings were lower for the control group than the treatment group, then the vector of differences in mean observed scores between control and treatment groups would be larger than the true vector. Further, wrongly assuming the loadings are equal between control and treatment would interact with differences in estimates already introduced by failing to weight the indicators properly when using observed mean differences (Equation 8). Finally, noninvariance in the intercepts between groups would also introduce bias into the vector of observed mean differences. The impact of such measurement invariance failures on recovery of true treatment effects in RCTs has been examined via simulation study and shown to impact estimated treatment effects (Authors, in press).

Additional Benefits and Limitations to Using a Latent Variable Model. While we highlight two of the main reasons using a latent variable framework for RDs is beneficial, there are other possibilities, as well as potential limitations. For example, another potential benefit relates to power. Using a latent variable model could account for measurement error in the dependent variable and thereby plausibly reduce the standard error on a treatment effect, helping avoid Type 2 errors. (One could also use a latent variable for covariates and reduce attenuation of the related coefficient, but that is not our focus in this study.) Using a more efficient estimator is important given RD studies are often underpowered since participants not in the bandwidth

must be removed (Schochet, 2009). However, there are also reasons for which this notion might not hold, including that latent variable models will add uncertainty relative to sum scores given that scores are not treated as known with a simultaneous estimation approach. Further complicating the picture, the first two sum score issues we mentioned (model misspecification and ignoring response shifts) would also likely impact Type 2 error rates, as we show in our simulation studies.

There is also a potential limitation of using a latent variable framework that bears mention. Specifically, the benefits all assume that one has properly specified the measurement model. There is evidence that if one does not properly specify the model, then treatment effects can be biased, so much so that using an observed/sum score might be preferable (Rhemtulla et al., 2020). Thus, fitting RDs as SEMs requires due diligence to safeguard against measurement model misspecification.

In what follows, we conduct two brief simulation studies to examine the specific magnitude of the benefits to using a latent variable model like the one we showed mathematically above, including whether any observed benefits are large enough to be of practical significance to researchers. The first simulation study examines the impact of using a sum score model versus a latent variable RD model for the recovery of true treatment effects under differing violations of assumptions implicit to sum scores. This first study varies measurement model parameters such that using a sum score represents a more or less egregious form of model misspecification dependent on the given condition. The second examines how much accounting for potential measurement invariance failures between control and treatment improves the recovery of true treatment effects compared to using a sum score that ignores noninvariance.

Simulation Study 1. Sum Score Use

In this first of two simulation studies, we simulated RDs in an SEM framework and examined the effects of using a sum score model to estimate treatment effects when the assumption of equal indicator weights is violated. The path diagram in Figure 1(b) served as our baseline generating model. That same figure but with data-generating parameters included is shown in Figure 2. As the figure shows, there is a single latent variable of interest. That variable is measured using four observed indicators, $y_{i1} - y_{i4}$. There is a true treatment effect of .20. The loadings and residuals differ by indicator. While the figure presents a fuzzy RD, the same model is used to create a sharp RD, but with all paths from z_i and rz_i, the correlation between ζ_{1i} and ζ_{2i}, and the path from r_i to t_i constrained to zero. Thus, we explore the effects of using a sum score on both sharp and fuzzy designs when the indicator weights in the data-generating model are not equal.

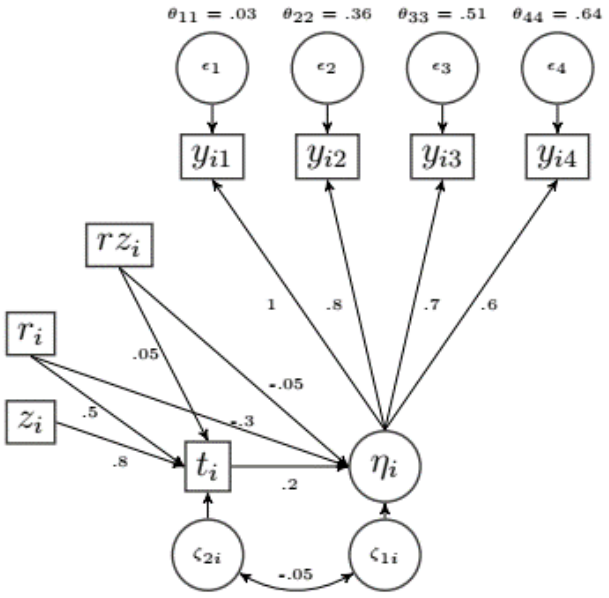

Figure 2. *Path diagram for the data-generating fuzzy RD model.*

In the study, we varied two main variables: the number of participants (simulees) within the bandwidth,[5] and the values of $\boldsymbol{\lambda}$. Sample sizes ranged from 200 to 500 participants within the bandwidth (in increments of 100), and were selected to range from those of small, likely underpowered studies to those of fairly large studies. Loadings were varied to range from cases where the latent variable explains a high proportion of the variance in the observed indicators to cases where that explained variance was low, and to have bigger or smaller gaps in the loadings across measures within a given condition. Specifically, while the first loading was always fixed to one, the other loadings ranged from .6 to .8 in the first condition, .5 to .7 in the second condition, and .4 to .6 in the third condition. Thus, sum scores that wrongly fix the loadings to be equal would likely have differing effects on the estimated treatment effect (as shown in Equation 8).[6]

Once the data were generated, we estimated treatment effects in two ways. In the first, we fit an SEM that matched the generating process to the data. In the second, we fit a similar model, but using sum scores to represent the dependent variable. To avoid scale indeterminacy in the dependent variable that can result by using sum scores, the sum scores were produced by fitting a highly constrained measurement model akin to the one described by McNeish and Wolf (2020). Specifically, we constrained all the loadings equal and set $\theta_{11} = \theta_{22} = \theta_{33} = \theta_{44}$. To ensure this approach matched the use of actual sum scores, we also produced mean scores based on the generated indicator responses, and results matched those produced using our highly constrained SEM.

[5] Here, the bandwidth stays the same (e.g., 1.5 SDs in units of r_i) and we are adding/subtracting the number of simulees with data in the bandwidth.

[6] Using McDonald's ω (McDonald, 2013), reliabilities ranged from a high of .86 to a low of .73. Thus, while we varied the loadings substantively across replications, all of the measures demonstrated plausible reliabilities for short survey scales used in practice.

All conditions were replicated 1,000 times in Mplus version 8.4 (Muthen & Muthen, 2017). Treatment effects based on simulated data were estimated using a Weighted Least Squares Means and Variance (WLSMV) adjusted estimator when the fuzzy RD model was fit, but maximum likelihood when the sharp model was fit. This difference occurred because treatment status (a binary variable) was dependent in the fuzzy specification, but independent in the sharp.[7] In all cases, the observed indicators used to measure the latent variable were continuous, and we assume the latent variable was as well. Across models, the scale of the latent variable was determined by constraining the loading on the first indicator to 1 (matching the parameter).

Finally, estimated treatment effects were examined in several ways. We began by examining the mean and variance of the estimated treatment effects across all replications and scoring approaches (SEM versus sum score). For each set of conditions, we examined parameter bias. Estimated bias was defined as $(\bar{\omega} - \omega)$, where ω is the parameter of interest, $\bar{\omega} = M^{-1} \sum_{m=1}^{M} \hat{\omega}_m$ and M is equal to the number of Monte Carlo replications. Lower bias shows more accurate parameter recovery. Then, we examined the proportion of treatment effect estimates found to be significant at the .05 level by sample size and scoring method.

***Results*.** There were no convergence failures, and parameter recovery when fitting the true model to the generated data was excellent (see Supplemental Materials Table A1). Further, model fit was excellent, with an average RMSEA of ~.035 (other fit statistics like the CFI showed comparable fit). Figure 3 shows box plots of the estimated treatment effects across all

[7] One limitation of this study is that we use two different estimators: WLSMV when treatment status is endogenous, and maximum likelihood when treatment status is considered to be exogenous. We use WLSMV for the former because the model requires a correlation between two residuals, including that of a categorical variable. Such residuals cannot be straightforwardly estimated when using categorical maximum likelihood. Thus, we cannot entirely rule out that some differences exist between results from the two estimators (in fact, results can and do differ because the estimators shift between sharp and fuzzy specifications, which differ in how they are parameterized).

1,000 replications by loading condition and scoring approach for 500 simulees (general trends in the results were similar for smaller sample sizes). As the figure makes clear, estimated treatment effects, on average, matched the true treatment effect when estimated in a latent variable framework. This finding applies both to sharp and fuzzy RD designs. By contrast, estimated treatment effects were much lower when indicators were misweighted using a sum score. For example, under the worst loading condition where λ = [1,.6,.5,.4], the sum score estimate of the treatment effect was ~.13 under the sharp design and ~.12 under the fuzzy. Thus, when the loadings were low, the true treatment effect was understated by roughly 40% using a sum score.

While differences in the estimates using a sum score model versus a less constrained measurement model was the main story of interest, one should also note that using a sum score when its assumptions were violated also impacted power. Using sum scores reduced the proportion of treatment effect estimates found to be significant by approximately ten percentage points, even with a true treatment effect of .20 SDs. Thus, even if one only cared about whether the intervention was found to have a significant impact, misweighting the indicators has a nonnegligible effect.

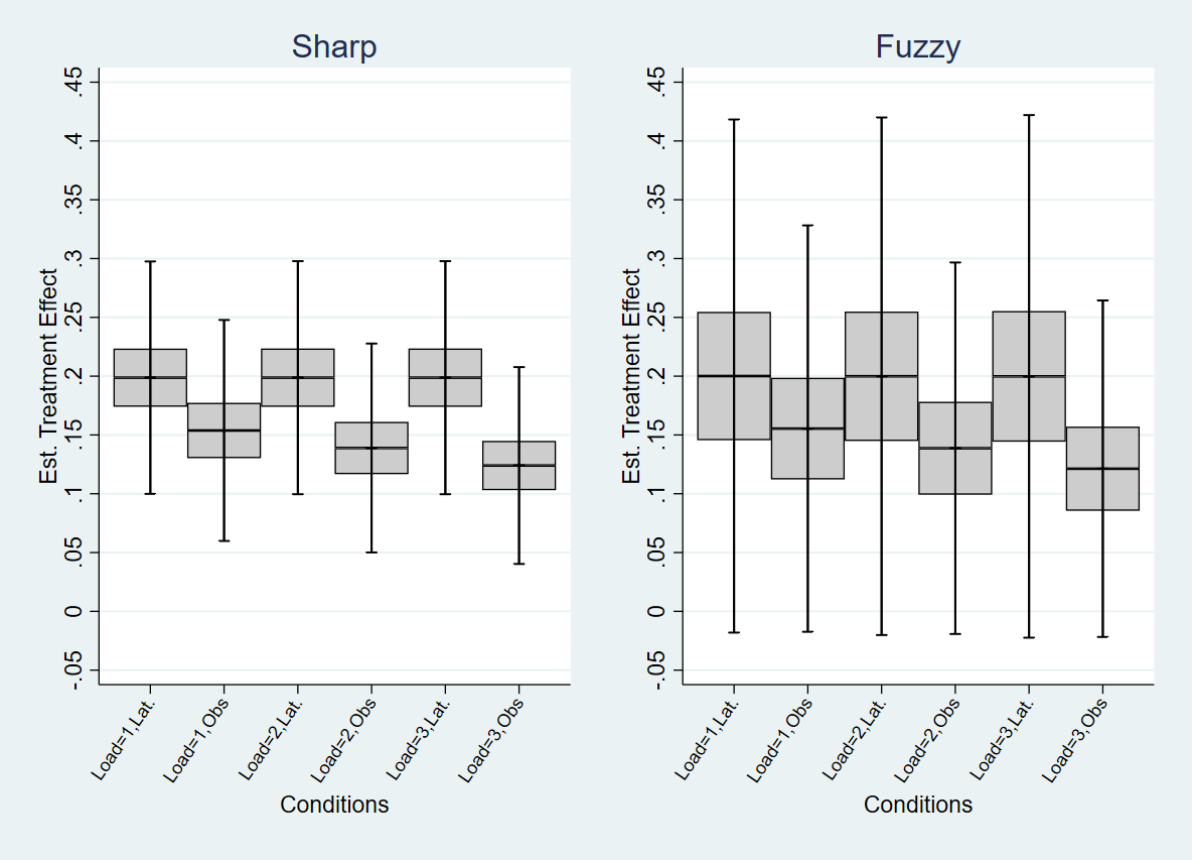

Figure 3. Bar diagrams showing estimated treatment effects across all 1,000 replications by condition for Simulation Study 2 (Sum Score Simulation).

Note. "Lat" means a latent variable model and "Obs" means a sum score model. Load = 1 corresponds to the first loading condition, Load = 2 to the second, and Load = 3 to the third.

Simulation Study 2. Measurement Noninvariance between Control and Treatment

In this simulation study, we examined the effect of measurement invariance failures on treatment effects in an RD framework. Such failures might occur if, say, a growth mindset intervention (with children assigned on the basis of test scores) actually changed how the student perceived growth mindset as a construct, which has been shown to occur in RCTs (e.g., Oort, 2005). Thus, this second simulation study is similar to the first one, but with measurement noninvariance induced between control and treatment groups (see Equation 12).

To simulate noninvariance, we generated data using a multigroup model in Mplus, then estimated treatment effects using a matching latent variable model and a sum score model that ignored group differences (sum score models cannot account for such group differences, at least not directly). To make sure the invariance failures were realistic, we loosely followed measurement model parameters from Fokkema et al. (2013), who examined noninvariance of items from a common depression measure. In line with their findings, we induced noninvariance by using the parameters in Figure 2 for the control group, then reducing the loadings in the treatment group by .20 units for indicators two through four. We also increased the intercept for the first indicator by .10 unit for the treatment group.

***Results*.** As one might expect, bias in the estimated treatment effect was substantial when noninvariance was not accounted for in the model. When estimating the treatment effect using a latent variable model, bias was practically zero (on average) and upwards of 98% of the replications produced results that were significant at the .05 level. By contrast, when using a sum score, the mean estimated treatment effect across the replications was ~.14 for the sharp design and the percent of results found to be significant decreased by more than 10 percentage points. When using a sum score with the fuzzy design, the mean estimated treatment effect was .15 units, also with a roughly 10 percentage point decrease in significant estimates relative to the SEM estimates. Thus, failing to account for measurement noninvariance could meaningfully impact point estimates and Type 2 error rates in an RD design.

Demonstration on How to Fit an RD as a Latent Variable Model

Having illustrated potential benefits of estimating RDs as SEMs mathematically and via Monte Carlo simulation, we now walk through the steps associated with conducting a fuzzy RD as an SEM. To strengthen the connection between the SEM approach and the typical 2SLS

approach used in econometrics, we begin by estimating RDs in a 2SLS framework, then show how to produce similar results in a path diagram framework before extending the model to an SEM that includes latent variables. We present only the fuzzy RD here because the fuzzy design requires all the steps used in a sharp RD plus additional steps that require elaboration.

Note that RDs can support causal claims, but only when its assumptions are met. For example, treatment effect estimates could be biased if there is some form of manipulation of scores near the cutoff, or if there is evidence that there are alternative mechanisms other than treatment assignment that influence outcomes and that treatment assignment is therefore not random. Given the importance of the assumption of random assignment, most RD analyses begin with a series of validity checks to find evidence suggesting that assumptions of the model have not been violated. We do not present those validity checks here, but a description of those preliminary analyses is available in Section B of the Supplemental Materials and the code for the validity checks is in Section C of the Supplemental Materials.

While the data used in this demonstration are simulated, parameters are based roughly on results from an empirical study using real data (Authors, under review). Throughout the demonstration, we will use a hypothetical that loosely follows a study conducted by Kendall et al. (2004). In their study, children were treated early for anxiety, and then the effect of treatment on later substance abuse was examined. In our hypothetical, children are given an anxiety diagnostic measure. Patients with a sufficiently high scores on the anxiety measure (i.e., who are more anxious) receive a cognitive–behavioral treatment for anxiety, though not all of them above the cut score ultimately receive the treatment. Importantly, given our focus on how the dependent variable is measured, the construct of anxiety is not considered as a latent variable in our hypothetical. Rather, the sum or scaled scores for anxiety are used as the "observed" forcing

variable. In the data, variables y_{i1} through y_{i4} are the outcomes of interest (the indicators representing substance abuse, e.g., survey items or scores on separate measures). For simplicity, they are continuous with a mean of zero and variance of one to avoid the complication of interpreting categorical dependent variables, but the same principles apply to categorical data.

When using observed scores in the model, we produce the score by taking the mean of the four indicators. This approach was taken to match what is normally done in practice and to mirror the 2SLS results. However, to ensure differences in treatment effect estimates between mean score and latent variable approaches are due to how the loadings are treated rather than differences in scales, we replicated results by mimicking mean scores using a highly constrained measurement model that scales the mean score and latent variable results comparably (both constrain the first loading to one). All results we report below using actual mean scores held when producing those mean scores using a constrained measurement model.

In the generating model, there is a known treatment effect from being above the cut score of .60 units, measured in units of y_{i1}. Thus, we use data based on known parameters, but the parameters differ from those in the earlier simulation studies to show that results hold even with a different data-generating model. The data also consist of a forcing variable, r_i (here an anxiety diagnostic score centered at a cut score), t_i for a given individual's treatment status (one if the person received treatment, zero otherwise) and an instrumental variable z_i (equal to one when the person was above the cut score on the anxiety measure, zero otherwise).

All 2SLS models are estimated in Stata 16 (StataCorp, 2019) and all SEMs are estimated using Stata's SEM package and Mplus Version 8.4 (Muthen & Muthen, 2018). Code for each model is provided in Section C of the Supplemental Materials. While we use specific code and

statistical programs for the demonstrations, our focus throughout is on conceptual understanding that is not specific to any software.

Estimation Using 2SLS in Stata

After having obtained reassurance about the validity of the RD design (as detailed in the Supplemental Materials), we proceed to estimating the treatment effect through 2SLS. We first produce bin plots (in our example, mean substance abuse scores on the vertical axis estimated separately for set increments of the forcing variable on the horizontal axis). For these analyses, the dependent variable is an observed score obtained by taking the mean of y_{i1} through y_{i4} for each person. In Figure 4 below, the blue markers represent the outcome of interest, binned by values of r_i. A visible discontinuity between the linear fit lines on the two sides of the cut score suggests a treatment effect. We will test this formally using 2SLS.

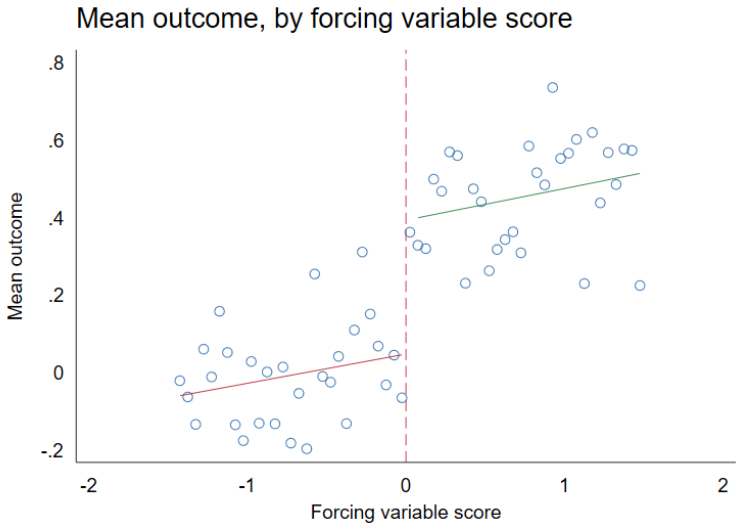

Figure 4. Bin plot of the outcome of interest (self-efficacy) by forcing variable score

We conduct 2SLS estimation in Stata for patients within +/- 1.5 SDs of the cut score. 2SLS estimation produces predicted values of t_i for each study participant as probabilities, leading to an interpretation of the treatment effect as a mean contrast with the control group. In this demonstration, that coefficient on t_i in the second stage is 1.15, indicating that treatment for

anxiety resulted in an improvement of 1.15 units on our observed (mean) substance abuse measure.[8]

Estimation in SEM using Stata/Mplus

In what follows, we briefly compare results from four models to walk readers from the 2SLS estimate to the latent SEM estimate. Those models include (1) mean score, t_i treated as continuous; (2) mean score, t_i treated as binary using a probit link; (3) latent variable estimate, t_i treated as continuous; and (4) latent variable estimate, t_i treated as binary with a probit link. Thus, (1) should parallel the 2SLS estimate (where predicted values of t_i for each study participant are probabilities), (4) is our preferred model, and (2)-(3) show the steps to get from 2SLS to the preferred model.

Mean Score, Continuous Treatment. We begin our presentation of the fuzzy RD approach in an SEM framework by showing the path diagram and estimated coefficients replicating the ivregress results using the path diagramming tool in Stata. We start here to show that, per Figure 5, the ivregress and path diagram estimates match. Here, we treat t_i as continuous not binary.[9] While treating t_i as continuous impacts the scaling of the treatment effect estimate, we can nonetheless produce mean score and latent variable estimates using this scaling (and using identical scales between mean scores and latent variables), then compare them. Unlike when using ivregress, in a path analytic framework the correlation between the residuals $e_1$and e_2 is explicitly estimated. As explained in the background section on the fuzzy RD, this correlation is the mathematical expression of the endogeneity problem: treatment status is

[8] Following our use of ivregress, we also estimate the two-stage equations by hand using the probit link in the first stage.

[9] This approach must also be used when employing the SEM program in Stata because the only estimation option available for binary endogenous variables is maximum likelihood, which does not allow for residual correlations like the one between $e_1$and e_2.

correlated with the error term in the outcome about which we care. Code for this model is in Section C1 of the Supplemental Materials.

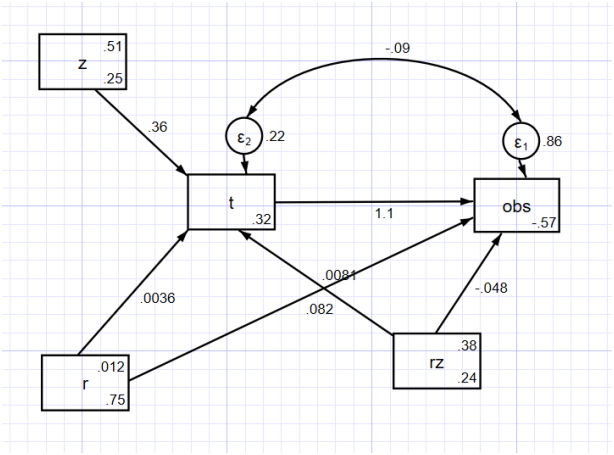

Figure 5. Using observed scores and considering treatment continuous in a path analytic framework

This code can easily be replicated in Mplus as shown in Section C2 of the Supplemental Materials. Table 1 crosswalks results from different model specifications and, as it shows, the treatment effect estimates using Stata versus Mplus match, with coefficients around 1.15. Thus, both sets of results indicate that being in the treatment appears to cause a 1.15-unit increase in the mean score. Again, when treatment status is treated as continuous, the units on the treatment effect estimate will differ compared to the data-generating model and results that use a probit link, but will be comparable between sum score and latent variable models that both treat t_i as continuous.

Table 1.
Cross-walk of outcomes by model and software

Model No.	Model Description	Stata File	Mplus File	Treat. Effect Stata/Mplus (SEs)
1	Use mean score as the outcome, BW 1.5, *t* treated continuous	Supp. C1	Supp. C2	1.147/1.147 (0.203)
2	Use mean score as the outcome, BW 1.5, *t* treated binary - probit	Supp. C1 (by hand)	Supp. C3	0.441/0.441

				(0.080)
3	Use latent variable as the outcome, BW 1.5, *t* treated continuous	Supp. C1	Supp. C4	1.497/1.497 (0.265)
4	Use latent variable as the outcome, BW 1.5, *t* treated binary	Not possible	Supp. C5	0.579 (0.105)

Mean Score, Binary Treatment. Next, we show results that treat t_i as categorical using the probit link in Mplus, but that continues to use a mean score. Section C3 of the Supplemental Materials provides Mplus code that treats t_i as categorical using WLSMV, the Mplus default estimator, which allows for a correlation between the residuals. As the results in Table 1 show, treating t_i as categorical and using a probit link has led to a rescaling of the coefficient on the path from the treatment to the mean score. Specifically, a one SD increase in t_i*, the latent variable presumed to underlie t_i, appears to cause a .44-unit increase in the mean score. Here, the mean score understates the true score largely because the true loadings on the indicators are quite low (with the exception of the first loading, which equals one in both the true and estimated models), and therefore indicators are misweighted in the observed score model, leading to an understated treatment effect (as described in the background section). As mentioned previously, this result holds when producing mean score results via a highly constrained measurement model that helps avoid scaling differences between mean scores and latent variables.

Latent Variable, Continuous Treatment. Next, one can estimate the fuzzy RD model using a latent variable estimate of the dependent variable in a single SEM framework. Figure 6 below is the path diagram for that model using the path diagramming tool in Stata, replete with estimated coefficients. This model fits the data quite well (e.g., RMSEA<.04), better than the

other competing models in the demonstration. Here, the latent variable is scaled by constraining the loading on the first indicator, y_{i1}, to 1. Since Stata only allows for maximum likelihood estimation in its generalized SEM package, we cannot specify that t_i is a dependent categorical variable *and* include a covariance in the residual terms, e_1 and e_6. Thus, results are comparable to those using a sum score and treating t_i as continuous. Based on the path diagram, receiving the treatment causes a 1.5 unit increase in the outcome measured in units of y_{i1}. This estimated treatment effect is higher than the one produced using a mean score due primarily to the mean score's misweighting of the indicators (constraining all loadings and residuals in the measurement model equal).

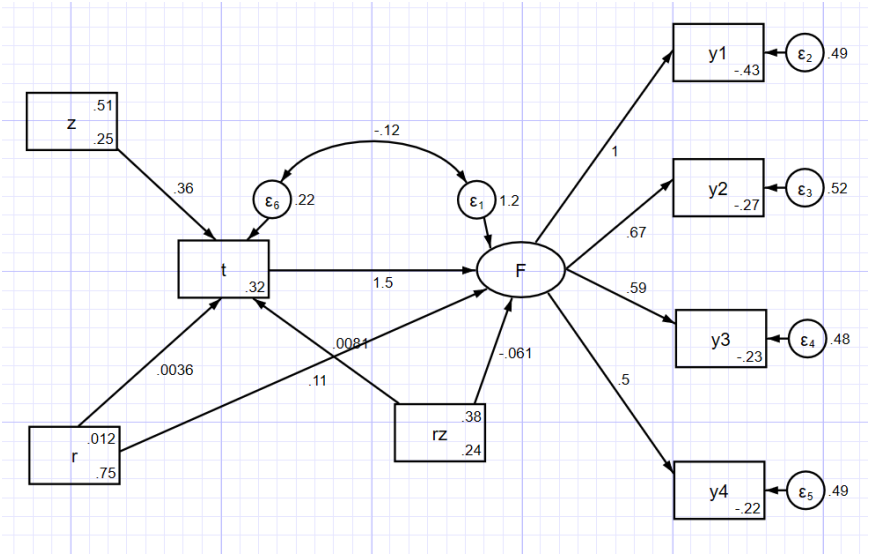

Figure 6. Path diagram for an RD estimate using a latent variable

Equivalent code for this model in Mplus is provided in Section C4 of the Supplemental Materials. Estimates in Stata versus Mplus (shown in Table 1) are identical to the hundredths place.

Latent Variable, Binary Treatment. Last, we fit a latent variable model that treats t_i as binary using a probit link (our preferred model given the data-generating model and results from

the prior simulations). This model can only be estimated in Mplus with an estimator like WLSMV and not in Stata, which only allows maximum likelihood. We estimate a model that treats t_i as binary (per the probit link, the model assumes a normally distributed variable with SD of 1 underlies the observed treatment status) and sets the scale of the latent variable by constraining the loading on the first indicator to 1. The code for this model is not provided separately because it is identical to the Mplus code in Section C4 of the Supplemental Materials, but no longer comments out the line indicating that t_i is categorical. As illustrated in Table 1, the coefficient on treatment status is .58 units, measured in units of y_{i1}. This value is very close to the true value of .60 units, likely differing due to sampling error resulting from using only a single replication of the generated data.

Making Sense of the Demonstration Results

Examining Effects of Using Observed/Mean Scores. Just as in the simulation study, using an observed/mean score results in lower estimated treatment effects than using a latent variable model. For example, when using a probit link for treatment status, the mean score model produced an estimated treatment effect that was .14 SDs lower than when using a latent variable model. This result is not surprising given the loadings tend to be well below one, which would lead to a lower observed treatment effect compared to the latent variable-based estimate (per Equation 8). Thus, even when using a separate data-generating model than in the simulation studies, the results are consistent with misweighting of indicators that can occur when observed/mean scores are used. This result held even when replicating the mean score results using a highly constrained measurement model that placed the mean score and latent variable results on comparable scales.

Examining Bias from Measurement Invariance Failures. Our second simulation study showed that failures of measurement invariance between control and treatment groups can bias treatment effect estimates in both sharp and fuzzy RDs. However, we do not walk through how to examine noninvariance in this demonstration, in part because there are so many existing resources on this process that already exist. For example, there are tutorials on general failures of invariance (Wu et al., 2007), testing for longitudinal invariance (Widaman et al., 2010), and testing for longitudinal invariance with categorical indicators (Liu et al., 2017). More relevant to the RCT and RD contexts, Oort (2005) showed how to test and correct for response shifts that can bias treatment effect estimates in an SEM framework. Such an approach could be extended straightforwardly to the RD-as-SEM approach we describe.

Estimating Measurement Model Parameters Based on All Participants Versus Those Within the Bandwidth Only. A potential wrinkle in fitting RDs as SEMs is that RDs are only estimated using participants within a desired bandwidth. This approach raises a question: Which participants should be used to estimate measurement model parameters? One option might be to estimate the measurement model parameters using only those participants within the bandwidth (i.e., simultaneous with estimation of the structural parameters that constitute the RD model). This is the approach taken in our simulations and demonstrations described above.

However, our simulations and demonstrations assume the optimal bandwidth is quite large (±1.5 SDs from the cut score). Another option might involve estimating the measurement model parameters with the whole sample, then fixing them when estimating the RD parameters using only participants within the bandwidth. The implications of this choice are unknown, especially when the optimal bandwidth is smaller. To help applied researchers understand the implications of this tradeoff, we conducted another simulation study using smaller bandwidths,

then estimating measurement model parameters based on (a) only simulees within the bandwidth (and therefore concurrently with the RD structural parameters) and (b) all simulees then constraining those parameters when estimating the RD parameters limited to simulees within the bandwidth. Those results can be found in Section E of the Supplemental Materials. While treatment effects were generally insensitive to this issue, we did find downward bias in treatment effect estimates when the bandwidth was very small (+/- .25 SDs), likely because the sample size was insufficient to estimate unbiased measurement model parameters. More research is needed on this issue; however, initial results suggest researchers may benefit from first estimating measurement model parameters using the whole sample, then fixing those parameters when estimating the RD parameters only within the bandwidth.

Discussion

Ideally, researchers interested in making a causal claim would implement an RCT. Unfortunately, randomization is often not feasible for a host of reasons. A quasi-experimental alternative is to use an RD design when treatment is assigned on the basis of a cut score. In such cases, if one can assume that study participants immediately on either side of the cut score are identical other than measurement error, then causal claims about treatment can be made for individuals proximal to the cut score. When this assumption is met, the RD design has the advantage of high internal validity (though generalizability may be limited given results only apply to participants within the bandwidth).

To date, most studies that use an RD design do not fit the RD in an SEM framework (in fact, measurement error is often ignored altogether). In this study, we began by showing mathematically and via simulation that relying on sum scores when conducting an RD can result in estimates of the treatment effect that differ substantively compared to using less constrained

measurement models. We investigated two primary reasons for which these differences could occur. First, as McNeish and Wolf (2020) pointed out, using a sum score can be equivalent to fitting a measurement model with extremely restrictive assumptions like equal weighting of the indicators and equal error variances. Wrongly imposing such assumptions can lead to fundamental misinterpretations of test and survey scores, including in experimental settings (Bauer & Curran, 2015; Kang & Hancock, 2017; McNeish & Wolf, 2020). Our results show that this form of model misspecification can severely impact RD-based treatment effect estimates, in some cases resulting in estimated treatment effects that are only 60% of true effects and increasing Type 2 errors.

Second, fitting a measurement model instead of using sum scores allows the researcher to identify and address measurement invariance failures that have been shown to occur in experimental and quasi-experimental studies (e.g., Oort, 2005). Failing to account for response shifts and other forms of noninvariance can lead to bias, possibly compounding the bias already introduced through the model misspecification that sum scores often represent. In our simulation, we showed that control-treatment invariance failures downwardly biased treatment effect estimates when using sum scores by ~.05 SDs (25% of the true treatment effect) and increased Type 2 errors. By contrast, SEM-based estimates of the treatment effect were unbiased and showed practically no Type 2 errors.

Having begun to establish the benefits of fitting RDs as latent variable models, we conducted a demonstration to show how RDs can be fit in an SEM framework using Stata and Mplus. The tutorial is meant to reinforce potential benefits of estimating RDs in a latent variable framework while also giving applied researchers in psychology the tools needed to fit such models. Further, by cross-walking the 2SLS and SEM approaches to RD estimation, we hope

this paper will help econometricians incorporate models more common in psychology into their own work.

As psychologists hopefully begin to use the models like the ones we describe more, a few key differences between RCTs and RDs should be kept in mind, especially limitations of the latter. In RCTs, the treatment and control data overlap along all values of the assignment variables. In RDs, the treatment and control data do not overlap at all, and parametric or nonparametric regression methods are used to separately *predict* treatment and control group values of the outcome at the cutoff. Such prediction depends heavily on appropriate choice of functional form and bandwidth, making the RD prone to bias (Chaplin et al., 2018). Another key limitation to RDs is their external validity: RDs estimate LATEs, and the results are not generalizable to participants whose running variable values are far away from the cut score. These limitations should be kept in mind when considering whether to use RD.

Limitations and Future Directions

A few limitations of the study bear mention. First, we did not examine the effect of including a measurement model for covariates used in the RD. Such an approach could change results, and possibly expand the benefits of fitting RDs in a latent variable framework. For example, RDs often include participant background characteristics as covariates even though they should not impact the treatment effect assuming random assignment within the bandwidth. Such covariates are included because they can reduce the standard errors on the treatment effect. If the covariates had their own measurement model and measurement error was accounted for, then controlling for the (more precisely represented) covariates could benefit the RD by further improving precision. Examining the effect of accounting for measurement error in covariates in an RD using SEM is worthy of future exploration.

Relatedly, we did not consider the use of latent variables for the forcing variable, in part because RD *assumes* the forcing variable is measured with error, and that participants just on either side of the cut score are ignorably assigned. While there are some potential benefits to using latent variables to better understand the forcing variable (e.g., Rokkanen, 2015), we did not investigate that issue given the complexities of such a decision, which are beyond the scope of our study. On one hand, removing *all* measurement error from the forcing variable would be problematic because then there would be substantive differences on that variable for participants just on either side of the cut score, though one could debate just how substantive the differences are. On the other, latent variable approaches can be used with the forcing variable. For example, many RDs in education employ achievement tests scored using IRT models that could help account for basic model misspecification that occurs using sum scores, but do not purge scores of all measurement error. Additional detail on this issue is provided in other studies, including Rokkanen (2015) and Angrist and Rokkanen (2018).

In fact, one could even imagine scenarios where failing to account for measurement model misspecification in the forcing variable might threaten the fundamental assumption of exchangeable study participants proximal to the cut score. For example, if the forcing variable was based on a depression survey scale, then a sum score would weight items about trouble sleeping and suicidality equally, but a latent variable model likely would not. Hypothetically, participants just on either side of a sum score-based cutoff might be very different in terms of their true depression if a sleep item moved one respondent above the threshold and the suicidality item moved another respondent below. Whether mistakenly weighting such items equally occurs in a systematic way around the cut score or simply washes out is difficult to say, but is certainly worthy of additional investigation. Regardless, this form of measurement model

misspecification is very different than the random measurement error in the forcing variable that is beneficial to the RD design.

Related to the potential benefits of our approach, our simulations were meant primarily as brief demonstrations and were not meant to be exhaustive. For example, the simulations could have examined a host of different measures and measurement models, such as using different numbers of items, testlets, or multiple scales with item cross-loadings. Understanding how results differ dependent on such decisions is important for future research. Further, our brief simulation examining how the decision about estimating measurement model parameters with the full sample or only within the bandwidth was very limited given it was meant to supplement the main study. The impact of the decision about which participants to use when estimating measurement model parameters should be investigated more in future research, including in cases where the data generation assumes measurement model parameters differ for participants within the bandwidth relative to those outside.

In practical terms, many study designs that employ RDs assign individuals to treatment on the basis of some cluster (for example, patients are clustered within doctors). Failing to account for such clustering could affect results in at least two ways. First, treatment effect estimates could be biased, though this problem would likely be less worrisome if, for example, all clusters were the same size. Second, power is already an issue in RD designs because only units near the cut score are used (Schochet, 2009), and that issue could be exacerbated in a cluster design. For instance, if there is a non-trivial intraclass correlation coefficient (ICC), the standard errors would not be consistent, which could in turn impact Type 1 error rates. Given our own study is meant as a tutorial on how to implement RDs in an SEM framework, fully exploring this issue is beyond the scope of the study, including issues of power more generally.

However, the model we propose could be straightforwardly adapted to have a multilevel structure (Morell et al., 2020). Fully enumerating that model is worthy of additional research.

Conclusion and Additional Resources

For those who want to learn more, there are a variety of resources. We highly recommend Sande and Ghosh (2018) for a gentle introduction to the logic of IVs, including in a path analytic framework. There are also several articles that examine how to approach quasi-experimental methods other than RD using SEM (e.g., Leite et al., 2019; Raykov, 2012; Rodríguez De Gil et al., 2015). For a broader view of the econometric and quasi-experimental literature outside of SEM including RDs, Angrist and Pischke (2008) provide a thorough and highly readable introduction. Researchers coming from a quasi-experimental background but who know less about SEM might examine the Stata tour of models in the user manual, which is quite thorough, or the examples in the Mplus user's manual, especially in conjunction with a classic text like Bollen (1989).

We hope that our results make the potential benefits of fitting RDs in a latent variable framework more apparent, and that we have given applied researchers in psychology the background and tools needed to estimate such models. Our results indicate that estimating an RD with sum scores as the dependent variable when the indicator weights are not uniform in the data-generating model can impact treatment effect estimates in ways that are nonnegligible, including increasing Type 2 error rates. We also show that fitting RDs as SEMs largely mitigates those sources of bias when model assumptions are met.

References

Angrist, J. D., & Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Angrist, J. D., & Pischke, J.-S. (2009). Mastering 'metrics: The path from cause to effect. Princeton, New Jersey: Princeton University Press.

Authors (In Press).

Bauer, D., & Curran, P. (2015). The discrepancy between measurement and modeling in longitudinal data analysis. *Advances in Multilevel Modeling for Educational Research: Addressing Practical Issues Found in Real-World Applications*, 3–38.

Bollen, K. A. (1989). Structural equations with latent variables Wiley. *New York*.

Cattaneo, M. D., Jansson, M., & Ma, X. (2018). Manipulation Testing Based on Density Discontinuity. The Stata Journal, 18(1), 234–261. https://doi.org/10.1177/1536867X1801800115

Cattaneo, M. D., Titiunik, R., & Vazquez-Bare, G. (2019). The Regression Discontinuity Design. arXiv preprint arXiv:1906.04242.

Chaplin, D.D., Cook, T.D., Zurovac, J., Coopersmith, J.S., Finucane, M.M., Vollmer, L.N. and Morris, R.E. (2018). THE internal and external validity of the regression discontinuity design: A meta-analysis of 15 within-study comparisons. *Journal of Policy Analysis Management, 37*(1), 403-429.

Dimitrov, D. M. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development*, *43*(2), 121–149.

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

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

Imbens, G. W., & Rubin, D. B. (2017). Rubin causal model. *The New Palgrave Dictionary of Economics*, 1–10.

Jacob, B. A., & Lefgren, L. (2004). Remedial education and student achievement: A regression-discontinuity analysis. *Review of Economics and Statistics*, *86*(1), 226–244.

Jacob, B. A., & Lefgren, L. (2009). The effect of grade retention on high school completion. *American Economic Journal: Applied Economics*, 33–58.

Kang, Y., & Hancock, G. R. (2017). The effect of scale referent on tests of mean structure parameters. *The Journal of Experimental Education*, *85*(3), 376–388.

Kendall, P. C., Safford, S., Flannery-Schroeder, E., & Webb, A. (2004). Child anxiety treatment: outcomes in adolescence and impact on substance use and depression at 7.4-year follow-up. *Journal of Consulting and Clinical Psychology, 72*(2), 276.

Kolesár, M., & Rothe, C. (2018). Inference in regression discontinuity designs with a discrete running variable. American Economic Review, 108(8), 2277-2304.

Kuhfeld, M., & Soland, J. (2020). Avoiding bias from sum scores in growth estimates: An examination of IRT-based approaches to scoring longitudinal survey responses. *Psychological Methods*.

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

Lee, D. S. & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, *48*(2), 281-355.

Leite, W. L., Stapleton, L. M., & Bettini, E. F. (2019). Propensity score analysis of complex survey data with structural equation modeling: A tutorial with Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(3), 448–469.

Lei, H., Nahum-Shani, I., Lynch, K., Oslin, D., & Murphy, S. A. (2012). A" SMART" design for building individualized treatment sequences. *Annual Review of Clinical Psychology, 8*, 21-48.

Liu, Y., Millsap, R. E., West, S. G., Tein, J. Y., Tanaka, R., & Grimm, K. J. (2017). Testing measurement invariance in longitudinal data with ordered-categorical measures. *Psychological Methods*, *22*(3), 486.

McDonald, R. P. (2013). Test theory: A unified treatment. New York, NY: Psychology Press.

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*, 1–19.

Morell, M., Yang, J. S., & Liu, Y. (2020). Latent variable regression discontinuity design with cluster level treatment assignment. *Multivariate Behavioral Research, 55*(1), 146-146.

Muthén, L.K. and Muthén, B.O. (1998-2012). Mplus User's Guide. Seventh Edition. Los Angeles, CA: Muthén & Muthén

Oort, F. J. (2005). Using structural equation modeling to detect response shifts and true change. *Quality of Life Research*, *14*(3), 587-598.

Oort, F. J., Visser, M. R., & Sprangers, M. A. (2005). An application of structural equation modeling to detect response shifts and true change in quality of life data from cancer patients undergoing invasive surgery. *Quality of Life Research*, *14*(3), 599-609.

Raykov, T. (2012). Propensity score analysis with fallible covariates: A note on a latent variable modeling approach. *Educational and Psychological Measurement*, *72*(5), 715–733.

Rodríguez De Gil, P., Bellara, A. P., Lanehart, R. E., Lee, R. S., Kim, E. S., & Kromrey, J. D. (2015). How do propensity score methods measure up in the presence of measurement error? A Monte Carlo study. *Multivariate Behavioral Research*, *50*(5), 520–532.

Rhemtulla, M., van Bork, R., & Borsboom, D. (2020). Worse than measurement error: Consequences of inappropriate latent variable measurement models. *Psychological Methods*, *25*(1), 30.

Rokkanen, M. (2015), "Exam Schools, Ability, and the Effects of Affirmative Action: Latent Factor Extrapolation in the Regression Discontinuity Design," Discussion Paper 1415-03, Columbia University, Department of Economics, New York, NY.

Sande, J. B., & Ghosh, M. (2018). Endogeneity in survey research. *International Journal of Research in Marketing*, *35*(2), 185–204.

Schochet, P. Z. (2009). Statistical power for regression discontinuity designs in education evaluations. *Journal of Educational and Behavioral Statistics*, *34*(2), 238–266.

Soland, J., Kuhfeld, M., Wolk, E., & Bi, S. (2019). Examining the State-Trait Composition of Social-Emotional Learning Constructs: Implications for Practice, Policy, and Evaluation. *Journal of Research on Educational Effectiveness*, *12*(3), 550–577.

StataCorp. (2019). *Stata Statistical Software: Release 16*. College Station, TX: StataCorp LLC.

Stefanski, L. A. (1985). The effects of measurement error on parameter estimation. *Biometrika*, *72*(3), 583–592.

Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives*, *4*(1), 10-18.

Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research, and Evaluation, 12*(1), 3.

Section A. Additional Simulation Study Details

Table A1.

Parameter Recovery

Parameter	Gen.	Est.	Difference
η *by* y_1 (λ_1)	1.00	1.000	0.000
η *by* y_2 (λ_2)	0.80	0.797	0.003
η *by* y_3 (λ_3)	0.70	0.701	-0.001
η *by* y_4 (λ_4)	0.60	0.600	0.000
y_1 *residual* (θ_1)	0.05	0.050	0.000
y_2 *residual* (θ_2)	0.36	0.356	0.004
y_3 *residual* (θ_3)	0.51	0.510	0.001
y_4 *residual* (θ_4)	0.64	0.638	0.002
t on z	0.80	0.808	-0.008
t on r	0.50	0.501	-0.001
t on rz	0.05	0.055	-0.005
η *on t*	0.20	0.204	-0.004
η *on r*	-0.30	-0.301	0.001
η *on rz*	-0.05	-0.048	-0.002
η *with t*	-0.05	-0.057	0.007
threshold t	0.00	0.002	-0.002
residual variance η	0.85	0.859	-0.009

Note. Estimates represent averages across 1,000 replications with a sample size of 1,000

Section B. RD Validity Checks

Preliminary Analyses and Estimation in Stata

In this section we demonstrate how to implement the fuzzy RD in Stata. We start with tests for validity of our RD design, proceed to estimation, and finally check the robustness of our estimates. The corresponding Stata do-file is provided in Section C1 and the data file is included in the supplemental materials. The validity of the RD design (both sharp and fuzzy) hinges on two main assumptions (Lee & Lemieux, 2010). Therefore, the plausibility of these assumptions is assessed prior to estimation of the treatment effect. First, we assume the forcing variable is not precisely manipulatable. If, in our hypothetical, patients can precisely manipulate the test scores used to determine participation in treatment, pushing some patients over/under the cut score to gain/lose access to treatment, treatment assignment would no longer be as good as randomly assigned. If there is evidence of such "heaping", or significantly higher density, on either side of the cut score, we might suspect units above and below the cut score differ in certain ways, which would make treatment assignment not as good as random. We therefore perform visual and statistical tests on density of r_i just above and below the cut score.[10]

Second, we assume that all patients are fully exchangeable around the cut score, differing primarily due to measurement error on their test score. Thus, all potential outcomes other than our dependent variable should be continuous at the cut score. Since all potential outcomes are not observed, we check to see if variables related to the outcomes are continuous. This check is to ensure that units (in our case patients) assigned to the treatment at the cut score are comparable

[10] One should note that manipulation could still occur without visual evidence of heaping if an equal number of students were moved up and moved down in accordance with some unobservable reason behind the teacher's decision-making that is correlated with potential outcomes

to units that are assigned to control. To further examine whether units around the cut score are comparable, we test the balance of pretreatment covariates (x_i in this example) at the cut score.

Finally, fuzzy RD requires one additional validity check beyond the first two, which are the same for sharp RDs. The key difference between the sharp and the fuzzy RD designs is that treatment take-up is not 100% in the latter. As with all 2SLS estimation, a strong first stage, signifying that assignment strongly predicts actual treatment, is vital to avoid bias and imprecision in the treatment effect estimate (having an instrument that strongly predicts treatment is also desirable in SEM). We therefore perform an additional test for the strength of the instrument prior to implementing the RD model.

Validity Test 1. Density Near the Cut Score. To test this aspect of the RD design, we examine the density of the forcing variable around the cut score visually and statistically. First, a histogram of r_i is generated (Stata code C1, line 37), shown in Figure B1. The density immediately to the right of the cut score (red vertical line at 0) looks similar to the density to the left of the cut score, suggesting no manipulation.

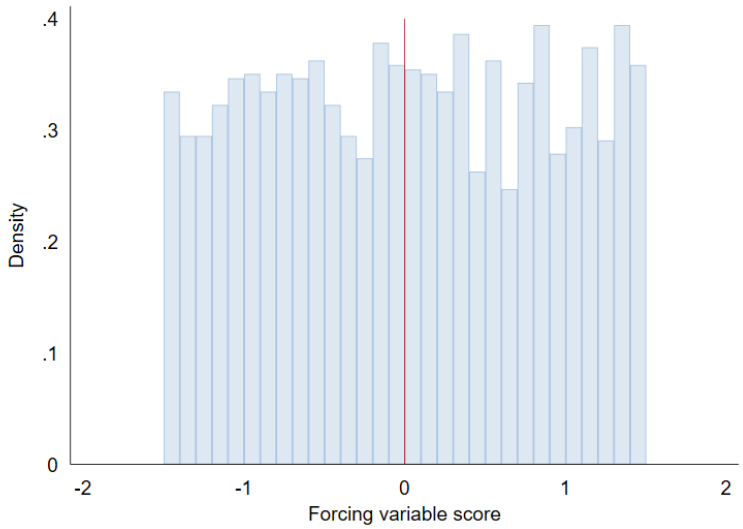

Figure B1. Density plot of the forcing variable

To further interrogate density balance, McCrary (2008) and Cattaneo et al. (2018) density tests are implemented (C1, lines 42-49). For parsimony, we will not provide details on those tests here (the citations should suffice for interested readers). Figure B2 shows the results of the McCrary (2008) test. Unlike the histogram, the density of r_i above the cut score appears higher than below the cut score; however, the difference is not statistically significant, and the confidence intervals in the figure are overlapping. Thus, we gain some reassurance that scores on the forcing variable are not being manipulated.

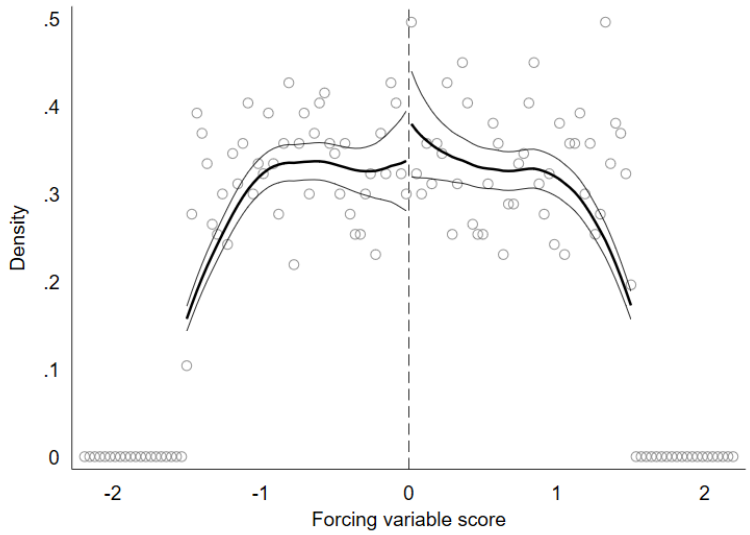

Figure B2. Results from the McCrary (2008) test

Results from the Cattaneo et al. (2018) test, presented in Figure B3 below, are very similar (p=.157) to those from the McCrary (2008) test, providing further evidence that any heaping is not statistically significant.

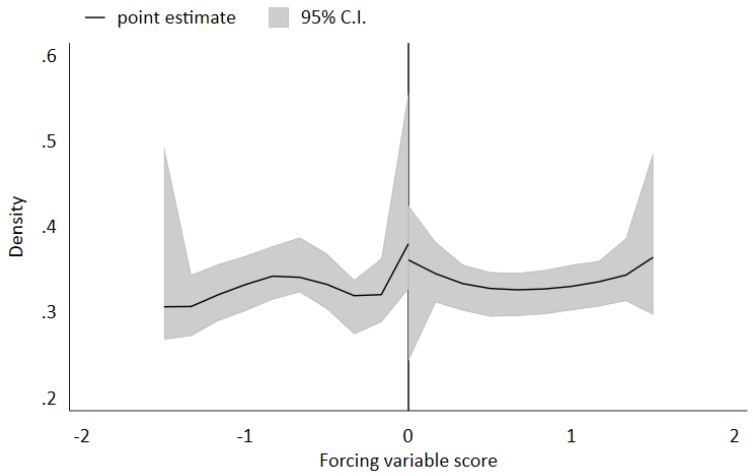

Figure B3. Results from the Cattaneo et al. (2018) test

Altogether, while some graphs suggest there may be slight difference in the density of r_i on either side of the cut score, those differences are not statistically significant. Therefore, there does not appear to be evidence of density-related manipulation at the cut score. In field studies, researchers also interrogate potential for manipulation in practice in addition to conducting density tests. For instance, one might ask if the tests were scored by doctors and if thresholds were known to them in advance, a scenario that could invite manipulation.

Validity Test 2. Covariate Balance. Next, we test if units assigned to treatment and control close to the cut score are comparable based on observed baseline characteristics (as we would hope). We look for evidence that pretreatment covariate x_i is continuous at the cut score. For both this balance test and for estimating the treatment effect, we retain only data within a selected bandwidth. For illustrative purposes, we use -1.5 < r_i < 1.5. Covariate x_i is regressed on z_i, r_i, and the interaction term rz_i (C1, line 54). The estimate for z_i is the coefficient of interest, with statistical significance indicating the units in the treatment and control conditions may not be comparable and the RD design potentially invalid. For example, if one were to think of x_i as a measure of socioeconomic status (SES), the logic of RD designs assumes patients on either

side of the cut score are similar in terms of SES, but a significant coefficient on z_i in this case would suggest they are not. In our data, the non-significant t-test suggests the baseline covariate x_i is continuous at the cut score.

Beyond this specific example, one would perform this test using several covariates. If multiple covariates show significant discontinuities, then the unfortunate reality is that the project may not be feasible given one would worry that results on the main outcome are also spurious. However, if only a single covariate among many produces significant results, then one option is to control for that covariate in the RD model. In general, quasi-experimental studies often involve complicated decisions around how to proceed if a single validity check fails.

Validity Test 3. Instrument Strength. As explained earlier, the 2SLS framework requires that the instrument z_i (being above the achievement test cut score) strongly predicts actual treatment t_i (taking advanced courses). We begin testing this assumption by visually examining treatment take-up for a subsample of units within a selected bandwidth of the cut score, -1.5 < r_i < 1.5 (C1, lines 68-81). For ease of visualization, values of r_i are binned in increments of 0.05 (C1, line 69), and the variable of interest (probability of treatment, here attending advanced courses) are averaged for each bin (C1, line 72).

As shown in Figure B4 below, units (students) below the cut score have a low (but not zero) probability of receiving the treatment, while units above the cut score have a high (but not one) probability of receiving treatment. This reflects imperfect compliance with the treatment assignment: some units assigned to treatment did not receive it, while others not assigned to treatment did. However, there is a visible jump at the cut score, suggesting that assignment strongly predicts actual treatment. We will return to test this discontinuity statistically after looking at the mean outcome visually.

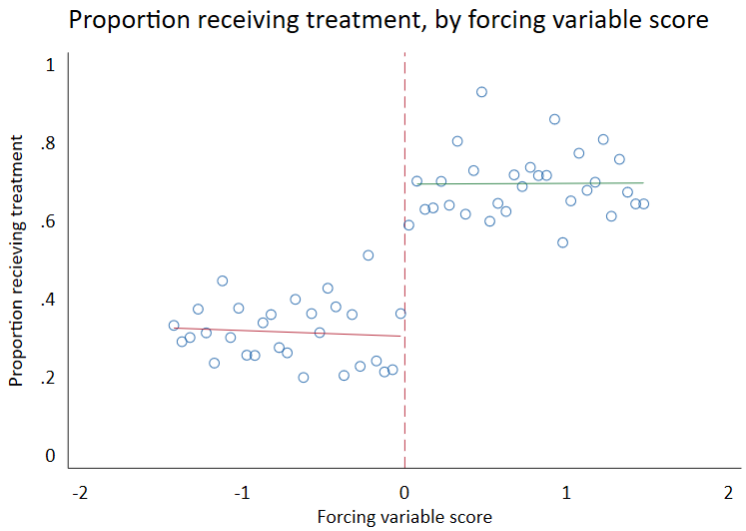

Figure B4. Bin plot of treatment status by forcing variable score

We test the strength of the instrument more formally using data within the selected bandwidth (-1.5 < r < 1.5) and a first-stage model with linear splines, by regressing t_i on z_i, r_i, and the interaction term $r_i z_i$ (C1, line 90). Specifically, we conduct an F-test for the estimate of the regression of treatment status on the instrument to determine if z_i is a strong predictor of t_i (C1, line 91). The commonly used threshold for a strong instrument is an F-statistic of 10 (or ideally much higher). In our data, the instrument z_i produced an F-statistic of 96.28, suggesting it is a very strong instrument and providing support for the 2SLS approach.

RD Estimation. Having obtained reassurance for the validity of the RD design and a strong first stage, we proceed to estimating the treatment effect through 2SLS. As with examining the strength of the instrument, we first produce bin plots, in this case with the outcome on the vertical axis and the forcing variable on the horizontal (C1, lines 106-124). For these analyses, the dependent variable is an observed score obtained by taking the mean of y_1 through y_4 for each person (C1, line 106). In Figure B5 below, the blue markers represent the

outcome of interest, binned by values of r_i. A visible discontinuity between the linear fit lines on the two sides of the cut score suggests a treatment effect. We will test this formally using 2SLS.

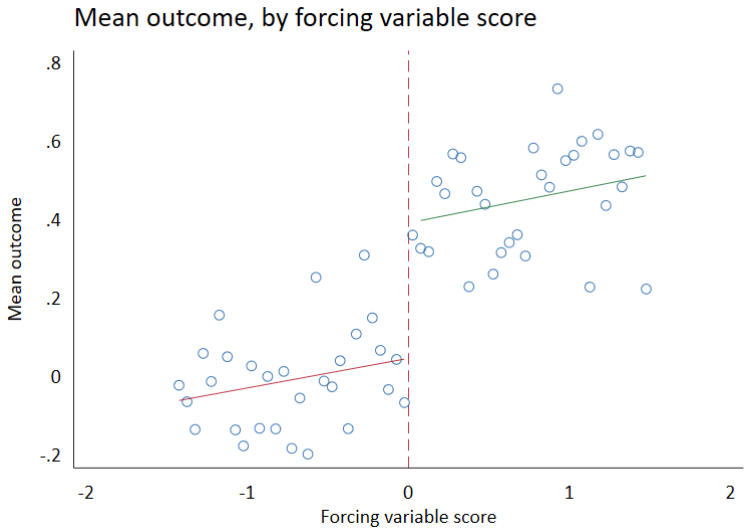

Figure B5. Bin plot of the outcome of interest (self-efficacy) by forcing variable score

Stata has several commands for 2SLS estimation, but the most relevant is likely ivregress (in Stata versions 15+, one could also use eregress[11]). In the ivregress code (C1, line 139), we specify that z_i is the instrument for t_i. After the "if" statement, we are (a) limiting to a bandwidth of our choice (+/- 1.5 units of r_i) and (b) asking for output from the first stage of the two-stage regression (using the "first" option below).

As mentioned previously, 2SLS estimation will produce predicted values of t_i for each study participant as probabilities, leading to an interpretation of the treatment effect as a mean contrast with the control group. In this demonstration, that coefficient on t_i in the second stage is 1.15, indicating that treatment (attending advanced courses) resulted in an increase of 1.15 units on our observed (mean) survey score. This is essentially the discontinuity in Figure B5 divided

[11] The main difference between ivregress and ergress is that the latter allows one to fit the first stage using either a probit or linear probability model. However, ivregress only allows one to fit a linear probability model that assumes the treatment indicator is continuous. Despite this limitation of ivregress, it is often used for fuzzy RDs, even those with a binary treatment status.

by the discontinuity in Figure B4, or, in the IV framework, rescaling the intent-to-treat effect by the first-stage estimate (Equation 7). Following our use of ivregress, we also estimate the two-stage equations by hand using the probit link in the first stage. In one version, we use predicted probabilities for $\hat{t}_i$, which matches the ivregress estimate (though note that the standard errors are slightly different between the two, and incorrect for the by-hand version). In the other version, we use a linear prediction for $\hat{t}_i$. As you will see in an SEM framework, this point estimate will match the one that uses a probit link for treatment status in Mplus following a t^* interpretation.

***Generalizability of the Fuzzy RD Results*.** The RD design generates credibly causal estimates of treatment effects, but with some key limitations to generalizability. First, the estimates are local average treatment effects. That is, results may not be generalizable to units (patients) whose forcing variable values are far from the cut score. Second, in the presence of treatment effect heterogeneity, the estimates would only be defined for compliers. In other words, the results may not be generalizable to always-takers (patients who would always be treated regardless of their score) and never-takers (patients who would never end up getting treatment). For more details, see Bertanha and Imbens (2014).

Section C. Stata and Mplus Code

Syntax files and data are available separately as part of the supplemental materials.

Section D. Robustness Checks for the RD Design in Stata

Since our treatment effect estimate may be sensitive to our choice of bandwidth and the functional form of the RD model, we test for robustness by varying the bandwidth and by adding quadratic terms to our equation (Stata code C1, lines 150-156). Although RD studies often test relative model fit (mostly between the linear and quadratic models, as the use of higher-order polynomials is not recommended) and report results for a preferred specification, model fit is not the focus. Thus, the preferred econometric approach often involves testing different functional forms and their fit, but ultimately estimating the model with different functional forms and bandwidths to ensure results are insensitive to modeling decisions. This approach is quite different than how someone more familiar with SEM might approach the problem, namely testing model fit, selecting a preferred model, and reporting only relevant results.

Table D1 shows the results from using bandwidth $-1 < r_i < 1$. The treatment effect estimate (1.31) has a confidence interval that overlaps our original estimate using a bandwidth of ($-1.5 < r_i < 1.5$).

Table D1. 2SLS results from using bandwidth=1

First-stage regressions

Number of obs	3,365
F(3, 3361)	510.40
Prob > F	0.0000
R-squared	0.3130
Adj R-squared	0.3124
Root MSE	0.4142

t	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
r	.0107889	.0356951	0.30	0.762	-.0591974	.0807751
rz	.0555745	.0496341	1.12	0.263	-.0417415	.1528905
z	.5204753	.0283547	18.36	0.000	.4648811	.5760695
_cons	.2438448	.0204995	11.90	0.000	.203652	.2840375

Instrumental variables (2SLS) regression

Number of obs	3,365
Wald chi2(3)	450.33
Prob > chi2	0.0000
R-squared	0.1208
Root MSE	.90802

obs	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
t	1.310602	.1194214	10.97	0.000	1.07654	1.544664
r	-.0303099	.0790608	-0.38	0.701	-.1852663	.1246465
rz	-.0764714	.1087516	-0.70	0.482	-.2896205	.1366778
_cons	-.6875922	.0689883	-9.97	0.000	-.8228067	-.5523777

Instrumented: t
Instruments: r rz z

As shown in Table D2, adding quadratic terms r_i^2 and $r_i^2 z_i$ produces an estimate of 1.41, with a confidence interval that also overlaps the estimate from the linear model. Though not reported, using other bandwidths produced similar results. Thus, we can feel fairly certain that varying the bandwidth and functional form does not substantively change our results.

Table D2. 2SLS results from using quadratic model

First-stage regressions

Number of obs	5,057
F(5, 5051)	467.17
Prob > F	0.0000
R-squared	0.3162
Adj R-squared	0.3155
Root MSE	0.4133

t	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
r	.0766803	.0767755	1.00	0.318	-.073833	.2271936
rz	.0002828	.1069011	0.00	0.998	-.2092896	.2098553
r2	.0504111	.0491962	1.02	0.306	-.0460348	.1468571
r2z	-.08387	.0689296	-1.22	0.224	-.2190018	.0512618
z	.5095976	.0345631	14.74	0.000	.4418389	.5773563
_cons	.257436	.0250703	10.27	0.000	.2082874	.3065847

Instrumental variables (2SLS) regression

Number of obs	=	5,057
Wald chi2(5)	=	725.57
Prob > chi2	=	0.0000
R-squared	=	0.1149
Root MSE	=	.9226

obs	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
t	1.413054	.1513952	9.33	0.000	1.116325	1.709783
r	-.1231107	.17892	-0.69	0.491	-.4737874	.227566
rz	-.2101865	.2386192	-0.88	0.378	-.6778716	.2574986
r2	-.0839994	.1141488	-0.74	0.462	-.3077269	.1397281
r2z	.3256773	.1635386	1.99	0.046	.0051477	.646207
_cons	-.7305156	.0884009	-8.26	0.000	-.9037782	-.557253

Instrumented: t
Instruments: r rz r2 r2z z

Section E. Simulation Investigating Impact of Estimating Measurement Model Parameter with All Participants Versus Those Within the Bandwidth

A potential wrinkle in fitting RDs as SEMs is that RDs are only estimated using participants within a desired bandwidth. One option might be to estimate the measurement model parameters using only those participants within the bandwidth (i.e., simultaneous with estimation of the structural parameters that constitute the RD model). Another option might involve estimating the measurement model parameters with the whole sample, then fixing them when estimating the RD parameters using only participants within the bandwidth. The implications of this choice are unknown.

To help investigate this issue, we redid the simulation studies with a range of bandwidths and slightly larger sample sizes (N = 5,000 and 10,000). Two approaches were used to estimate the parameters in the measurement submodel of the SEM. First, the data were subset to include only those simulees within the selected bandwidth and measurement and structural parameters were estimated together, including the treatment effect. Second, measurement model parameters were estimated using the full sample, then those parameters were fixed when estimating the structural portion of the model based only on simulees within the bandwidth. These two approaches were taken because RD analyses are, in practice, typically conducted using only the data within the optimal bandwidth (Imbens & Kalyanaraman, 2011; Robinson, 2011).

Table E1 includes treatment effect estimates and RMSE for a model that estimates measurement model parameters based only on simulees near the discontinuity (i.e., within the desired bandwidth) and based on the whole sample. In general, differences between models that estimate measurement model parameters within the bandwidth versus using the full sample differ little. However, there is an exception to this finding. When the sample size is 5,000 and the

bandwidth is .25, calibrating based only on simulees within the bandwidth produced a downwardly biased treatment effect estimate.

Table E1

Simulation Results for Measurement Model Parameters Based on All Simulees Versus those within the Bandwidth

N	Bandwidth	Latent (Full Sample)	s.e.	RMSE	Latent (Bandwidth)	s.e.	RMSE	Sum	s.e.	RMSE
5000	0.25	0.206	0.010	0.006	0.121	0.118	0.149	0.180	0.009	0.020
5000	0.5	0.199	0.006	0.001	0.197	0.006	0.003	0.173	0.006	0.027
5000	1	0.199	0.004	0.001	0.198	0.004	0.002	0.173	0.004	0.027
10000	0.25	0.204	0.006	0.004	0.187	0.006	0.013	0.177	0.005	0.023
10000	0.5	0.199	0.004	0.001	0.199	0.004	0.001	0.174	0.004	0.026
10000	1	0.197	0.003	0.003	0.197	0.003	0.003	0.172	0.003	0.028