A Correction for Regression Discontinuity Designs with Group-Specific Mismeasurement of the Running Variable*

Otávio Bartalotti, Quentin Brummet, and Steven Dieterle†

October 18, 2018

Abstract

When the running variable in a regression discontinuity (RD) design is measured with error, identification of the local average treatment effect of interest will typically fail. While the form of this measurement error varies across applications, in many cases there is a group structure to the measurement error. We develop a procedure to make use of this group-specific measurement error structure to correct estimates obtained in a regression discontinuity framework using auxiliary data. This procedure extends the prior literature on measurement error on the running variable by leveraging auxiliary information in order to account for more general forms of measurement error. Additionally, we develop adjusted asymptotic variance and standard errors that take in consideration the variability introduced by the nonparametric estimation of nuisance parameters from auxiliary data. Simulations provide evidence that the proposed procedure adequately corrects for measurement error introduced bias and tests using the new adjusted formulas exhibit empirical coverage closer nominal test size than "naive" alternatives. We provide two empirical illustrations to demonstrate that correcting for measurement error can either reinforce the results of a study or provide a new empirical perspective on the data.

Key Words: Nonclassical Measurement Error, Regression Discontinuity, Heterogeneous Measurement Error.

JEL Codes: C21, C14, I12, J65

*We would like to thank Alan Barreca for kindly sharing the data used in one of our applications. We also thank Yang He for his invaluable research assistance, Tim Armstrong, Cristine Pinto, Sergio Firpo, Samuele Centorrino, Jeff Wooldridge and participants at the 2017 Midwest Econometrics Group Meeting and 2017 Brazilian Econometric Society Meeting for valuable comments.

†Bartalotti: Department of Economics, Iowa State University and IZA. 260 Heady Hall, Ames, IA 50011. Email: bartalot@iastate.edu. Brummet: NORC at the University of Chicago, 4350 East-West Highway, 7th Floor, Bethesda Maryland 20814. Email: brummet-quentin@norc.org. Dieterle: School of Economics, University of Edinburgh, 31 Buccleuch Place, Edinburgh, United Kingdom EH8 9JT. Email: steven.dieterle@ed.ac.uk.

1 Introduction

Regression Discontinuity (RD) designs have become a mainstay of policy evaluation in many social science fields. These designs rely on treatment assignment being based on a "running variable" passing a particular cutoff, which is observed by the researcher. In practice, however, there are multiple forms of measurement error in the running variable that, when present, will invalidate this approach.

We consider situations in which a researcher has access to data with the running variable exhibiting group-specific measurement error, where each group faces potentially different measurement error distributions. One prominent example identified by Barreca et al. (2011) (hereafter, BGLW) is the "heaping" of birth weight measures at particular values due to hospitals using scales with different resolutions. In this setting, additional care is given to babies born at a birth weight of strictly below 1500 grams, allowing for an RD analysis of the effect of the additional resources on child outcomes. However, some hospitals record the weight to the nearest gram while others record it at ounce or gram multiples— 5g, 10g, and up to 100g multiples. Therefore, the treated units measured at 1499g are likely to be well measured and accurately reflect the mean outcomes and unobservables at the true weight of 1499g, but the closest untreated units measured at 1500g will reflect the mean outcomes and unobservable factors for babies with a true weight up to 50g away. The problem is further complicated by the nearby ounce multiple measure at 1503g that will have a much different measurement error distribution than babies measured at 1500g. Depending on the gradient of child outcomes with respect to the true birth weight, this could generate a spurious discontinuity at the cutoff of the mismeasured running variable.

Another example comes from geographic regression discontinuity (GeoRD) settings, where the running variable is often measured as the distance from an individual's residence to a border that separates two policy regimes.[1] Ideally, researchers would use a precise distance measure from the residential address to the border. However, due to data limitations it is common to use the distance from the geographic centroid of a larger region to calculate the distance to the border. Due to differences in region size and the population distribution within each region, units in different regions— or groups— will face different measurement error distributions. Importantly, the centroid measure may be closer or farther away from the border than the true distance for many of the units— again creating the possibility of a discontinuous jump in

[1] See Keele and Titiunik (2014) for a general discussion of GeoRD and for examples see Black (1999); Lavy (2006); Bayer, Ferreira, and McMillan (2007); Lalive (2008); Dell (2010); Eugster et al. (2011); Gibbons, Machin, and Silva (2013); Falk, Gold, and Heblich (2014).

unobservable factors at the cutoff in the measured distance.

We propose a measurement error correction procedure that leverages auxiliary information about the measurement error distributions for various groups (see Hausman et al. (1991); Lee and Sepanski (1995); Chen, Hong, and Tamer (2005) and Davezies and Le Barbanchon (2017) for other approaches using auxiliary information to address measurement issues). This information is used to transform the observed data, re-centering the observed running variable around the moments of the underlying latent running variable distribution for each observation or group. Intuitively, this re-centering procedure corrects the distortions caused by the measurement error since, on average, some observations will be closer or farther from the cutoff than the observed running variable would indicate. The re-centering identifies the parameters on the conditional expectation of the outcome with respect to the true (unobserved) running variable rather than the mismeasured one.

The measurement error correction procedure's implementation requires the (potentially non-parametric) estimation of the moments of the multiple measurement error distributions. Hence, we develop procedures for valid inference, developing a novel asymptotic distribution approximation that considers the variability introduced by the multi-sample first stage estimation on the estimates for the ATE at the cutoff.

This procedure can accommodate auxiliary information of many forms, including information that cannot be matched to the primary data set or information that is nested within the primary data. In the birth weight example, we use the non-heaped data to learn about the measurement error properties of the heaped data, while in the GeoRD case the measurement error distributions can be calculated using readily available census population counts at disaggregated levels in the U.S.

Within the RD literature, there is a growing research interest on measurement error issues (Lee and Card (2008); Pei and Shen (2017); Yu (2012); Dong (2015); Davezies and Le Barbanchon (2017); Dong (2017), and Barreca, Lindo, and Waddell (2016)). Importantly, our procedure is applicable to a new class of problems — such as our motivating examples above— not previously covered by the literature. Furthermore, we study both the case in which treatment is determined by the unobserved running variable, which has been the focus of the majority of the existing literature, and the case in which treatment is determined based on the mismeasured running variable, which is common in relevant applications such as the very low birth weight example discussed in Section 4.

While our procedure is motivated by problems not currently covered in the literature characterized by group level mismeasurement, it is flexible enough to handle other measurement error types and provides an easy-to-implement parametric alternative to solve problems covered by the current literature under different assumptions. For instance, this procedure allows non-classical measurement error structures that could depend on the observed running variable and is applicable in both sharp and fuzzy RD designs. It can also be applied to cases where the measurement error can be characterized as discrete measurement of a continuous true measure, as in Lee and Card (2008) and Dong (2015). These previous treatments of discrete running variables focus on more restrictive forms of measurement error as they require the specification error to either be random (Lee and Card, 2008) or identical within each group (Dong, 2015).

Simulation results provide evidence that our procedure performs well and indicate that the importance of measurement error-induced bias is not necessarily mitigated by smaller bandwidths. Additionally, the empirical coverage of the tests implementing the newly proposed standard errors is improved and approaches the test's nominal size, in contrast to naive tests that perform neither the measurement error correction nor the variance adjustment.

In the context of the low birth weight example in Almond et al. (2010) (hereafter, ADKW) and BGLW, we find that correcting for measurement error yields estimates consistent with the original results in ADKW, suggesting a large effect of very low birth weight classification. Further, estimates using our correction are much less sensitive to the exclusion of observations at "heaped" measures near the cutoff than the uncorrected estimates. We also apply our procedure to examine the effect of Unemployment Insurance (UI) benefit extensions during the Great Recession on unemployment studied by Dieterle, Bartalotti, and Brummet (2018). In this paper we focus on the simpler case of the Minnesota-North Dakota border during 2010, where the uncorrected estimates are 18 times larger than the corresponding estimates using the moment-based correction, implying a sizable bias due to the measurement error.

The paper proceeds as follows: Section 2 presents the setup of our paper and derives our measurement error-corrected RD approach; Section 3 presents Monte Carlo evidence about the performance of the method proposed; Section 4 applies our procedure to the very low birth weight example; Section 5 applies our procedure in the GeoRD context; and Section 6 concludes.

2 Running Variable Measurement Error Correction

2.1 Setup and Motivation

Consider a basic RD setup. The interest lies in estimating the average treatment effect of a program or policy in which treatment status ($D = 0, 1$) is determined by a score, usually referred to as "running variable" (X), crossing an arbitrary cutoff (c). Let $Y_1 \equiv y_1(X)$ represent the potential outcome of interest if an observation receives treatment and $Y_0 \equiv y_0(X)$ the potential outcome if it does not. The researcher's intent is usually to estimate $E[Y_1 - Y_0|X = c]$, the local average treatment effect at the threshold. If observable and unobservable factors influencing the outcome evolve continuously at the cutoff then the average treatment effect at the cutoff is identified nonparametrically by comparing the conditional expectation of $Y = DY_1 + (1 - D)Y_0$ on either side of the cutoff:

$$\tau = \lim_{a \downarrow 0} E\left[Y|X = c + a\right] - \lim_{a \uparrow 0} E\left[Y|X = c + a\right]. \tag{2.1}$$

Now, consider the case in which instead of the running variable, X, we observe a mismeasured version, $\tilde{X} = X - e$, where e is the measurement error. This measurement error can be quite general including non-classical forms and can be dependent on either X or $\tilde{X}$.

To gain some intuition about the problems introduced by mismeasurement, consider the special case in which the conditional distribution of the measurement error is continuous. In that case, a researcher that ignores the measurement error and implements standard RD techniques will estimate:

$$\tau^* = \lim_{a \downarrow 0} E\left[Y|\tilde{X} = c + a\right] - \lim_{a \uparrow 0} E\left[Y|\tilde{X} = c + a\right] \tag{2.2}$$

$$= \int_{-\infty}^{\infty} (y_1(c+e)p_1(c+e) + y_0(c+e)p_0(c+e)) f_{e|\tilde{X}}(e|\tilde{X} = c^+) de$$
$$- \int_{-\infty}^{\infty} (y_1(c+e)p_1(c+e) + y_0(c+e)p_0(c+e)) f_{e|\tilde{X}}(e|\tilde{X} = c^-) de \tag{2.3}$$

where $p_0(X)$ and $p_1(X)$ are the probabilities of receiving and not receiving treatment conditional on the unobserved X and $f_{Z|W}(\cdot|W = c^+)$ and $f_{Z|W}(\cdot|W = c^-)$ denote a conditional density of the variable Z evaluated at as W approaches c from above and below, respectively. By looking at the neighborhood of $\tilde{X} = c$ we are effectively analyzing the (weighted) average of the potential outcomes over the values of X for which $\tilde{X} = c$. This quantity in Equation (2.2) will take different forms based on whether treatment is assigned on the observed or unobserved running variable.

If treatment is sharply assigned based on the true unobserved running variable, standard RD techniques will estimate:

$$\begin{aligned}\tau^* = &\int_{X \geq c} y_1(x)(f_{x|\tilde{X}}(x|\tilde{X} = c^+) - f_{x|\tilde{X}}(x|\tilde{X} = c^-))dx \\ &+ \int_{X < c} y_0(x)(f_{x|\tilde{X}}(x|\tilde{X} = c^+) - f_{x|\tilde{X}}(x|\tilde{X} = c^-))dx \end{aligned} \tag{2.4}$$

which equals zero in the absence of a discontinuity in $f_{x|\tilde{X}}(x|\tilde{X} = c)$ (or equivalently $f_{e|\tilde{X}}(e|\tilde{X} = c)$). This is an example of the loss identification induced by the presence of continuous measurement error in the running variable described in Pei and Shen (2017) and Davezies and Le Barbanchon (2017) in which the measurement error smooths out the conditional expectation of the outcome close to the observed "cutoff" in $\tilde{X}$. Intuitively, this occurs because the measurement error induces the misclassification of treatment to some observations.

If instead treatment is determined by the mismeasured running variable and is therefore observed, a researcher that ignores the measurement error and implements standard RD techniques will estimate:

$$\tau^* = \lim_{a \downarrow 0} E\left[Y|\tilde{X} = c + a\right] - \lim_{a \uparrow 0} E\left[Y|\tilde{X} = c + a\right] \tag{2.5}$$

$$= \lim_{a \downarrow 0} \int_{-\infty}^{\infty} y_1(x) f_{x|\tilde{X}}(x|\tilde{X} = c + a)dx - \lim_{a \uparrow 0} \int_{-\infty}^{\infty} y_0(x) f_{x|\tilde{X}}(x|\tilde{X} = c + a)dx$$

$$= \int_{-\infty}^{\infty} y_1(c + e) f_{e|\tilde{X}}(e|\tilde{X} = c^+)de - \int_{-\infty}^{\infty} y_0(c + e) f_{e|\tilde{X}}(e|\tilde{X} = c^-)de. \tag{2.6}$$

If the measurement error distribution conditional on X is continuous at the policy cutoff, then:

$$\tau^* = \int_{-\infty}^{\infty} (y_1(x) - y_0(x)) \frac{f_{\tilde{X}|X}(c|x)}{f_{\tilde{X}}(c)} dF_X(x) \tag{2.7}$$

Hence, instead of estimating the ATE at the cutoff, the researcher recovers a weighted average treatment effect for the population in the support of X for which $X + e = \tilde{X} = c$. The weights are directly proportional to the *ex ante* likelihood that an individual's value of $\tilde{X}$ will be close to the threshold. This case is similar to the situation described in Lee (2008) and Lee and Lemieux (2010) where individuals can manipulate the running variable with imperfect control. Our approach will recover $E[Y_1 - Y_0|X = c]$ in this setting as well. Note that while this case includes both the geographic RD and birth weight examples discussed later in the paper, our procedure applies to the case where treatment is assigned based on the unobserved running variable provided that the researcher observes true treatment status.

2.1.1 Group-specific Measurement Error Distribution

A central contribution of this paper is to allow for heterogeneity of the measurement error distribution across groups of observations. Let $\tilde{x}_{ig} = x_i - e_{ig}$, where e_{ig} is the measurement error of "type" g. This notation allows each unit to have a measurement error drawn from a separate, individual-specific distribution. It also encompasses the case where individuals in the same group share the same observed value of $\tilde{X}$, such that $\tilde{x}_{ig} = \tilde{x}_g$ for all individuals in group g. This is the situation where, for example, residents in a county have their location reported as the county's centroid or birth weights being rounded to nearest 50 grams or ounce multiples.

Once we allow for different measurement error "types," the identification of the ATE is further complicated by the averaging across groups on both sides of the cutoff. Discontinuous changes in the share of groups at the cutoff introduce bias to estimates of the treatment effect. For example, if all individuals follow the same processes $y_1(x)$ and $y_0(x)$ but suffer from different types of measurement error in the running variable and treatment is assigned based on $\tilde{X}$, then

$$\begin{aligned}\tau^* = \sum_g P(G = g|\tilde{X} = c^+) \int_{-\infty}^{\infty} y_1(x) f_{x|\tilde{X},G}(x|\tilde{X} = c^+, G = g)dx \\ - P(G = g|\tilde{X} = c^-) \int_{-\infty}^{\infty} y_0(x) f_{x|\tilde{X},G}(x|\tilde{X} = c^-, G = g)dx\end{aligned} \tag{2.8}$$

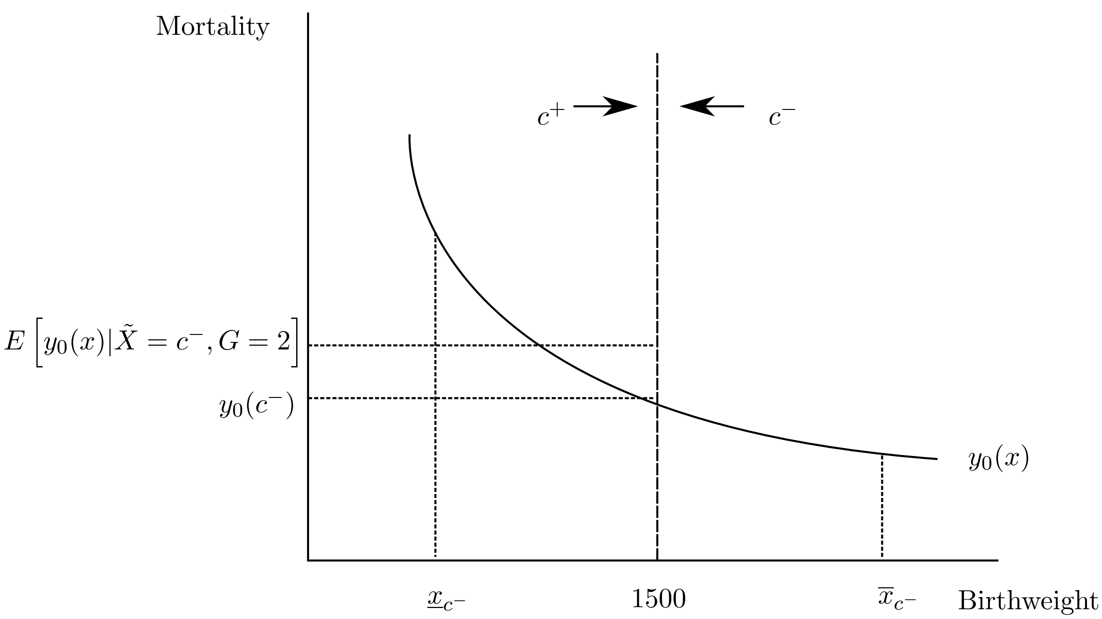
Hence, changes in the share of each group at the cutoff could introduce bias and measurement error correction approaches that ignore the group heterogeneity might fail to identify the intended ATE.

To illustrate the problem, consider the very low birth weight example. For now, assume there are two types of measures— those correctly measured at the individual gram level ($G = 1$) and those rounded to a gram multiple ($G = 2$), where this could be any of the gram-multiple heaps observed in the data (5g, 10g, 25g, 50g, or 100g). In the limit, the conditional expectation for treated units at the cutoff will come from the correctly measured children at 1499g and the distribution of types on the treatment side ($\tilde{X} = c^+$) will be degenerate (i.e. $P(G = 1|\tilde{X} = c^+) = 1$). Meanwhile the untreated units at 1500g will be a mixture of correctly measured and mismeasured units. This implies the following RD estimand:

$$\tau^* = y_1(c^+) - \left[P(G = 1|\tilde{X} = c^-)y_0(c^-) + P(G = 2|\tilde{X} = c^-) \int_{-\infty}^{\infty} y_0(x) f_{x|\tilde{X},G}(x|\tilde{X} = c^-, G = 2)dx\right]$$

If the probability of being a rounded measure is quite high relative to precise measures, the estimate of the conditional expectation for untreated units will be driven by the mismeasured group. In the current example, the observations at 1500g make up 1.75 percent of the overall sample while the adjacent unrounded measure of 1501g only makes up 0.05 percent of the sample, implying $P(G=2|\tilde{X}=c^-) \approx 0.97$. Given evidence of rounding to 100g multiples in some cases, $\int_{-\infty}^{\infty} y_0(x) f_{x|\tilde{X},G}(x|\tilde{X}=c^-, G=2)dx$ may average $y_0(x)$ over a range of true x between 50 grams below to 50 grams above the cutoff. Depending on the shape of $y_0(x)$, this may lead to a very poor estimate of the intended estimand of $\lim_{a\uparrow 0} E\left[Y|X=c+a\right]$. If, for instance, $y_0'(x)<0$ and $y_0''(x)<0$, as is likely the case in the low birth weight example (mortality rate decreases with birth weight, but at a decreasing rate as we approach the natural lower bound of zero), then even if $f_{x|\tilde{X},G}(x|\tilde{X}=c^-, G=2)$ is uniform (i.e. true birth weight is uniformly distributed within the range of true weights associated with a measured weight of 1500g) we will likely overestimate the conditional expectation at the cutoff for the untreated units since $\int_{-\infty}^{\infty} y_0(x) f_{x|\tilde{X},G}(x|\tilde{X}=c^-, G=2)dx > y_0(c^-)$ by Jensen's Inequality. This is depicted in Figure 2.1 where we denote the minimum and maximum values of x associated with $\tilde{X}=c^-$ by $\underline{x}_{c^-}$ and $\overline{x}_{c^-}$, respectively. Our proposed approach will be able to recovers identification of the ATE in these settings.

Figure 2.1



2.2 Assumptions

Assume that the researcher observes the treatment status D, and let $E[Y|x, D = 0] = f_0(x) + R_0, E[Y|x, D = 1] = f_1(x) + R_1,\ E[Y|\tilde{x}, D = 0] = h_0(\tilde{x})$ and $E[Y|\tilde{x}, D = 1] = h_1(\tilde{x})$; where $f_t(x)$ are polynomial approximations to $E[Y|x, D = t]$ of (potentially) unknown order J with approximation error terms given by R_t for $t = 0, 1$. For simplicity, we assume that such polynomial is capable of capturing all relevant features in the pertinent neighborhood of the unobservable X, denoted by S.[2]

We then impose the following assumptions, which follow closely Dong (2015):

A1 $f_1(x)$ and $f_0(x)$ are continuous at $x = c$.

A2 $f_1(x)$ and $f_0(x)$ are polynomials of possibly unknown degree J, and R_1 and R_0 are negligible asymptotically in S.

A3 Polynomial approximations of order J for $h_1(\tilde{x})$ and $h_0(\tilde{x})$ are identified above and below the cutoff.

A4 The treatment status is observed by the researcher.

A5 (a) For all integers $k \leq J$, $E(e^k|\tilde{x}, G = g) = \mu_g^{(k)}(\tilde{x})$, these moments exist and are identified in the support of $\tilde{X}$. (b) The conditional distribution of the measurement error for each group in the primary and auxiliary samples is the same, i.e., $f_a(e|\tilde{X}, G) = f_p(e|\tilde{X}, G)$. (c) The known group affiliation, denoted by G is redundant if there is no measurement error, that is, $E[Y|x, D = t, G] = E[Y|x, D = t]$.

A6 $\tilde{x}$ is assumed to be redundant, i.e., $E[Y|x, \tilde{x}, D = t] = E[Y|x, D = t] = f_t(x)$ for $t = 0, 1$.

The first assumption is the usual RD identifying assumption that the potential outcomes are continuous at the threshold, so that the observed "jump" at the threshold can be associated with the causal effect of the treatment. $A2$ is a parametric functional form approximation, since local methods to eliminate the approximation error will no longer be appropriate due to the measurement error in the running variable. This assumption is flexible in the sense that it allows for a variety of approaches to approximate the conditional expectation of the outcome. For simplicity one could simply assume that $f_t(x)$ are correctly specified, implying $R_t = 0$ (Dong, 2015), or that $E[R_t|x, D = t] = 0$ (Lee and Card, 2008).

[2] More precisely, let $\tilde{X} = g(X) = X - e$, then note that for a given measurement error distribution we can map any value of $\tilde{X}$ into the support of X. Let that set be $G^{-1}(\tilde{X}) \equiv \{X : X + e = \tilde{X}\text{with probability greater than zero}\}$. Specifically, let an arbitrary neighborhood around $\tilde{X} = c$ be given by $B = [c - h, c + h]$. Then, let $A = \bigcup_{\tilde{X} \in B} G^{-1}(\tilde{x})$ be the relevant support of X and define $S = [\inf A, \sup A]$.

Alternatively, if there is concerned about misspecification bias arising from the use of a polynomial of order J to approximate $f_t(x)$ for $t = \{0, 1\}$, we can obtain "honest confidence intervals" that cover the true parameter at the nominal level uniformly over the parameter $\mathcal{F}$ for $f_t(x)$ by taking the approach in Armstrong and Kolesár (2018); Armstrong and Kolesár (2018b). We constrain the parameter space by bounding the derivatives on $f(\cdot)$, hence bounding the potential bias due to misspecification. Here we consider $\mathbb{F} = \mathcal{F}_J(M)$ where M indexes the smoothness of the class of functions considered. As in Armstrong and Kolesár (2018); Armstrong and Kolesár (2018b) we restrict or attention to functions such that the approximation error from a $J-th$ order Taylor expansion about $x = 0$ is bounded, uniformly over the support of X. For example, consider that around $X = 0$, R_t is bounded by $M|x|^{J+1}$ for a chosen constant M. We discuss this approach in section 2.5.[3]

$A3$ states that we can identify a polynomial of order J that describes the mean outcome as a function of the mismeasured $\tilde{X}$. This requires $\tilde{X}$ to have sufficient variation to identify $h_t(\tilde{X})$. As described below, our approach will exploit the mapping between $h_t(\cdot)$ and $f_t(\cdot)$ implied by separability and additivity of the measurement error to recover the treatment effect parameters. This aspect of the procedure is closely related to the approach proposed by Hausman et al. (1991) and Dong (2015), while Assumption $A3$ serves a similar purpose to the completeness condition required by Davezies and Le Barbanchon (2017). This assumption is more likely to hold in practice when the mismeasured running variable is continuous, and in the discrete case requires that the researcher has access to several points in the support of $\tilde{X}$ that have positive density.

Assumption $A4$ implies different data requirements in fuzzy as opposed to sharp RD designs. In a sharp design, if $\tilde{X}$ is always on the same side of the cutoff as X or if treatment is determined by $\tilde{X}$, this is not restrictive at all, as discussed in the geographic RD case analyzed in Section 5. However, if the measurement error causes the observed running variable to cross the threshold, Assumption $A4$ requires the researcher to observe the true D. In other words, the measurement error must not prevent us from observing the treatment status. While restrictive, this could be considered a relatively mild requirement on the data and certainly is the case in many interesting applications. Finally, in the fuzzy RDD context covered in Davezies and Le Barbanchon (2017) and Pei and Shen (2017), estimating the treatment effect using $\tilde{X}$ ignoring the measurement error will lead to a "smoothing out" of the discontinuity, resulting in estimates that could be severely biased. In that case, Assumption $A4$ needs to be strengthened so that the researcher

[3]We thank Tim Armstrong for suggesting this point.

observes not only the true treatment status but also $\mathbf{1}[x > c]$, which indicates which side of the threshold each observation lies based on the unobserved X. In the absence of the information required in Assumption $A4$ the procedure proposed here will not be feasible, but the approaches of Davezies and Le Barbanchon (2017) and Pei and Shen (2017) still provide potential solutions to the measurement error problem under alternative assumptions.

Assumption A5 is central to our approach, and requires that the $k \leq J$ uncentered moments of the measurement error distribution conditional on the observed mismeasured running variable are identified based on the information in the auxiliary data for each group. This assumption allows these moments to depend on the observed running variable and to differ for each group. Hence, it complements the existing literature on measurement error by permitting dependence in the true and mismeasured running variables, and measurement error that is not identically distributed across groups. This encompasses a large number of empirical applications, as exemplified in Sections 4 and 5. In the very low birth weight example, different hospitals have measurements of varying precision, while in the geographic case regions may be of different size and have different population densities relative to the border. Furthermore, Assumption $A5$ could be strengthened if the measurement error is assumed to be independent of $\tilde{X}$, then $E(e^k|\tilde{x}, G = g) = \mu_g^{(k)}(\tilde{x}) = \mu_g^{(k)}$.

As mentioned previously, this setup easily encompasses the situation where treatment is determined by the observed running variable, $\tilde{X}$. In settings where treatment is determined by the true running variable, X, if the researcher observes treatment status coupled with the mismeasured running variable— perhaps in survey data where participants are asked about participation in a means tested program determined by true income falling below a certain threshold, but income (X) is only reported in discrete bins in the survey— then our approach provides an easy to implement alternative to the previously mentioned approaches proposed by Pei and Shen (2017) and Davezies and Le Barbanchon (2017).

2.3 Identification and Estimation

Researchers are faced with two potential identification problems when implementing RD designs in the presence of measurement error. The first source, common to all RD designs, is the potential for local misspecification of the conditional mean function for the outcome. The second is the measurement error itself, which can distort the estimates of the jump at the cutoff. Crucially, the introduction of measurement error renders infeasible the usual local nonparametric approach to deal with the original misspecification problem.

With measurement error a local approach based on shrinking bandwidths ($h \to 0$) around the cutoff is ineffective since the researcher can only observe $\tilde{x}$ and would not be able to guarantee that the true value of the running variable falls within a neighborhood of the cutoff. In other words, even observations that seem close enough to the threshold for treatment might in reality be far away and be a poor comparison to observations just on the other side of the cutoff.

Identification of τ can be obtained through the combination of a local polynomial approximation of the conditional mean of the outcome and information about the group specific measurement error distributions.

Theorem 2.1. *Let assumptions A1-A5 hold. Then τ can be identified even if x is not observed.*

See Appendix A for the proof of Theorem 2.1.

To illustrate the problem and the proposed correction, consider the simple case where $\mu_g^{(k)}(\tilde{x}) = \mu_g^{(k)}$ and the (local) quadratic approximations for $f_t(x_{ig})$ for $t = 0, 1$ with parameters $b_{p,t}$ are used:

$$\begin{aligned} f_t(x_{ig}) &= b_{0,t} + b_{1,t}(x_{ig}) + b_{2,t}x_{ig}^2 \\ &= b_{0,t} + b_{1,t}(\tilde{x}_{ig} + e_{ig}) + b_{2,t}(\tilde{x}_{ig} + e_{ig})^2 \\ &= b_{0,t} + b_{1,t}\left(\tilde{x}_{ig} + e_{ig}\right) + b_{2,t}\left[\tilde{x}_{ig}^2 + 2e_{ig}\tilde{x}_{ig} + e_{ig}^2\right] \end{aligned}$$

Note that

$$\begin{aligned} E[Y|\tilde{x}, D = t, G] &= E\left[f_t(x_{ig})|\tilde{x}, G\right] \\ &= \left(b_{0,t} + b_{1,t}\mu_g^{(1)} + b_{2,t}\mu_g^{(2)}\right) + \left(b_{1,t} + 2\mu_g^{(1)}\right)\tilde{x}_{ig} + b_{2,t}\tilde{x}_{ig}^2 && (2.9) \\ &= b_{0,t} + b_{1,t}\left(\tilde{x}_{ig} + \mu_g^{(1)}\right) + b_{2,t}\left[\tilde{x}_{ig}^2 + 2\mu_g^{(1)}\tilde{x}_{ig} + \mu_g^{(2)}\right] && (2.10) \end{aligned}$$

Expression (2.10) is helpful for motivating our correction procedure, while (2.9) is useful to highlight the problems with using the mismeasured running variable. Starting with the latter, note that using the mismeasured running variable will yield a biased estimate of τ:

$$\widetilde{b}_{0,1} - \widetilde{b}_{0,0} = \tau + b_{1,0}(\overline{\mu_1^{(1)}} - \overline{\mu_0^{(1)}}) + \overline{\mu_1^{(1)}}(b_{1,1} - b_{1,0}) + b_{2,0}(\overline{\mu_1^{(2)}} - \overline{\mu_0^{(2)}}) + \overline{\mu_1^{(2)}}(b_{2,1} - b_{2,0}),$$

where $\overline{\mu_t^{(k)}} = E[\mu_{g,t}^{(k)}]$ are the expected values of measurement error moments over the support of $\tilde{X}$ used in estimation. In particular, the bias becomes more serious when the measurement error distributions differ across the cutoff. In the birth weight example, this arises due to heaping near the cutoffs. In the GeoRD case, this refers to the population distributions within geographic areas and is made worse when the population mass is farther from the border, the slope is steeper, or there is higher curvature near the boundary.

Equation (2.10) suggests a simple solution. If μ_g^j are known (or estimable) for every $g = 1, ..., G$ and $j = 0, ..., J$, we can obtain the intercepts above and below the cutoff, $b_{0,0}$ and $b_{0,1}$ by appropriately adjusting the data and recover $b_{0,1} - b_{0,0}$. For example, for the case of $J = 2$, we transform the data and use $x^*_{1,ig} = \tilde{x}_{ig} + \mu_g^1$ and $x^*_{2,ig} = \tilde{x}^2_{ig} + 2\mu_g^1 \tilde{x}_{ig} + \mu_g^2$ to replace the mismeasured running variables. In general, the vector of the mismeasured running variable $\tilde{X}' = [1, \tilde{x}, \tilde{x}^2, \ldots, \tilde{x}^J]$, is replaced by the vector of the "corrected" running variable of same dimensions $X^{*\prime} = [1, x^*_1, x^*_2, \ldots, x^*_J]$, where $x^*_j = \sum_{k=0}^{j} \binom{j}{k} \mu_g^{(j-k)} \tilde{x}^k$. The uncentered moments for the measurement error can be replaced by consistent estimates.

Most relevant, information about μ_g^k does not need to come from the same data set as the individual level outcomes. That is, we can use auxiliary data to estimate the measurement error distributions. For example, in the GeoRD problem in Section 5 we only need precise location information to estimate the population distribution relative to the border in each county in order to correct our estimates of the treatment effect. This approach can be applied in any case where auxiliary information on the distribution of the measurement error is available.

More generally, our strategy estimates a regression of order J for treated and untreated observations as described below:

$$\hat{\tau} = \hat{\beta}_+ - \hat{\beta}_-$$

$$(\hat{\beta}_+, \hat{\beta}_+^{(1)}, \ldots, \hat{\beta}_+^{(J)})' = argmin_{b_0, b_1, \ldots, b_J} \sum_{i=1}^{N_p} (Y_i - b_0 - b_1 \hat{x}^*_{1,i} - \cdots - b_J \hat{x}^*_{J,i})^2$$

$$(\hat{\beta}_-, \hat{\beta}_-^{(1)}, \ldots, \hat{\beta}_-^{(J)})' = argmin_{b_0, b_1, \ldots, b_J} \sum_{i=1}^{N_p} (Y_i - b_0 - b_1 \hat{x}^*_{1,i} - \cdots - b_J \hat{x}^*_{J,i})^2.$$

Where, N_p is the size of our primary sample on which we observe the outcome and mismeasured running variable. Also, let $\hat{x}^*_j = \sum_{k=0}^{j} \binom{j}{k} \hat{\mu}_g^{(j-k)}(\tilde{x}) \tilde{x}^k$ and $\hat{\mu}_g^{(j)}(\tilde{x})$ be a consistent estimator of $\mu_g^{(j)}(\tilde{x})$.

Since $E(e^k | \tilde{x}, G = g) = \mu_g^{(k)}(\tilde{x})$, it is natural to use a local kernel estimator:

$$\hat{\mu}_g^{(j)}(\tilde{x}) = N_{a,g}^{-1} \sum_{i=1}^{N_{a,g}} K_h(\tilde{x}) e^j_{i,g} \tag{2.11}$$

Where the $N_{a,g}$ is the size of the auxiliary sample for group g on which we observe the measurement error, h_g the group-specific smoothing parameter, and $K_h(\tilde{x}) = K\left(\frac{\tilde{x}_i - \tilde{x}}{h_g}\right)$ is a bounded kernel with usual properties (Fan and Gijbels, 1996).

From the applied researcher perspective, it is useful to keep in mind that what is needed is information about the measurement error distribution moments for each group. As described in assumption A5, if this information is being acquired from auxiliary data this requires it to represent the same population as the main data, or at least have the same measurement error distribution. Note that it is not required that the auxiliary data match specific observations, nor does it need to be nested within the main sample. Furthermore, the auxiliary data is required to contain information about X and $\tilde{X}$ for each group but could omit the outcome Y. These requirements are similar to assumption 3 in Davezies and Le Barbanchon (2017), with the extra requirement that we have information available to each group for both treated and untreated observations. The auxiliary datasets could be different for each group as mentioned above. The availability of such auxiliary information will certainly vary on a case-by-case basis and the willingness of the researcher to assume how the measurement error takes place. For example in the VLBW application below we perform the measurement correction under two different explicit assumptions on how the measurement error is generated, that is (a) by rounding errors that lead to accumulation at specific points of the observed weight distribution, and (b) by a similar process, but allowing the distribution of the error terms to vary with mother's education by effectively changing the definition of the group structure. While these are certainly important assumptions, this highlights the flexibility of the proposed procedure to accommodate different structures of measurement error by the researcher. This permits easy robustness checks on the importance of assumptions such as independence of measurement error and $\tilde{X}$, or that the measurement error distributions does not depend on some covariate observed on both the main and auxiliary data, like mother's education. This is more appealing and less restrictive than assuming that the measurement error distribution is known (Dong, 2015), follows the CEV assumption (Pei and Shen, 2017), or is homogeneous across groups (Davezies and Le Barbanchon, 2017).

2.4 Large Sample Properties

To perform inference about the parameters of interest, we propose a novel studentized test which incorporates the uncertainty associated with estimation for each group of the several moments of the measurement error conditional distributions, $\mu_g^{(j)}(\tilde{x})$, using auxiliary information. We assume that the auxiliary data are independent from each other and the primary data. Define

N_p and $N_{a,g}$ to be the sample sizes for the primary dataset and auxiliary dataset for group g, respectively. Additionally, let $\lambda_g = lim_{N_p \to \infty} \frac{N_p}{N_{a,g} h_g}$ for all g, which essentially requires both N_p and $N_{a,g} h_g$ to go to infinity at the same rate so that the approximation captures the effect of estimating $\hat{\mu}_g$. This is similar to the asymptotic conditions in Davezies and Le Barbanchon (2017), however requiring it at the "group level," and would not be needed if identification was not based in auxiliary data (Pei and Shen, 2017; Dong, 2015). Also define e' as the conformable row vector of zeros except for the first entry equal to one.

Theorem 2.2. *Let assumptions A1-A5 hold, and λ_g be a finite constant for every g. Define, $X_i^{*\prime} = [1, x_{i,1}^*, \ldots, x_{i,J}^*]$, $\mu_g'(\tilde{x}) = [1, \mu_g^{(1)}(\tilde{x}), \ldots, \mu_g^{(J)}(\tilde{x})]$, $B_+' = [\beta_+, \beta_+^{(1)}, \ldots, \beta_+^{(J)}]$ and equivalently for B_-, and $\Gamma_i = L_{J+1} \circ Q$ is the Hadamard product of the lower diagonal Pascal matrix, L_{J+1}, and the matrix Q_i, where $Q_{(b,c)} = \tilde{x}_i^{c-b}$. Also, $\varepsilon = Y - E[Y|\tilde{x}, D, G]$. Finally, assume that $h_g \to 0$, $N_{a,g} h_g \to \infty$ as the sample sizes increase for all g, and that a CLT holds for the vector of measurement error moment estimators using the auxiliary data such that $(N_{a,g} h_g)^{\frac{1}{2}} (\hat{\mu}_g(\tilde{x}) - \mu_g(\tilde{x})) \to N(0, V_g(\tilde{x}))$ for all relevant points of the support of $\tilde{X}$. As $N_p \to \infty$, then*

$$\sqrt{N_p}(\hat{\tau} - \tau) \to^d N(0, \Omega) \tag{2.12}$$

where,

$$\Omega = e' \left[\Omega_+ + \Omega_-\right] e \tag{2.13}$$

$$\Omega_+ = A^{-1} E\left[X^{*\prime} \varepsilon_i \varepsilon_i' X^*\right] A^{-1} + A^{-1} \left[\sum_{g=1}^{G} \lambda_g F_{+,g}' V_g(\tilde{x}) F_{+,g}\right] A^{-1} \tag{2.14}$$

$$F_{+,g} = E\left[\left(x_{lg}^{*\prime} \otimes B_+' \Gamma_{lg}\right)\right] \tag{2.15}$$

$$A = E\left[X^{*\prime} X^*\right] \tag{2.16}$$

and similarly for Ω_- on the other side of the cutoff.

See Appendix A for the proof of Theorem 2.2.

Theorem 2.2 provides the asymptotic approximation to the distribution of $\hat{\tau}$, which we can use for hypothesis testing and creating valid confidence intervals.

The asymptotic variance approximation, Ω, incorporates the uncertainty introduced due to the estimation of the parameters μ_g using auxiliary datasets/information for each group-specific measurement error. The results are related to the auxiliary data/two-sample results in the literature (Chen, Hong, and Tamer, 2005; Lee and Sepanski, 1995) and extend their conclusions to the case in which heterogeneous measurement is present for groups. Our results are more specialized in the sense that we focus on RD designs and assume a local polynomial approximation for the conditional mean of the outcome. Nevertheless, the flexibility allowed by the proposed approach in both the definition of groups and the degree of the polynomial function used, and nonparametric estimation of the conditional moments, $\mu_g(\cdot)$, greatly expands the nature of the measurement error problems that can be addressed.

Our approach is able to correct for multiple sources of measurement error using distinct auxiliary datasets. This allows us to tackle measurement error problems in a wider range of contexts by using of several different sources of auxiliary data that are increasingly available to researchers. The explicit adjustment in the variance formula takes into account the amount of information available for estimation of μ_g in each auxiliary dataset in order to obtain confidence intervals with correct coverage when testing hypotheses. When these moments are not known but estimated from an auxiliary sample, extra variability will be introduced to the estimates. Ignoring such variability could lead to confidence intervals with inadequate coverage. This point is similar to the one made by Calonico, Cattaneo, and Titiunik (2014) for the usual RD case when the first-order bias correction term is estimated.

2.5 Honest Confidence Intervals

If the researcher is concerned that the polynomial order chosen,"J," does not fully capture the relevant features of the conditional mean of the outcome, $f_t(\cdot)$, leading to some potential misspecification bias we can resort to honest confidence intervals (Armstrong and Kolesár, 2018; Armstrong and Kolesár, 2018b). Honest CIs cover the true parameter at the nominal level uniformly over the possible parameter space for $\mathcal{F}(M)$ for $f_t(\cdot)$. Here we focus on the class of functions that place bounds on the derivatives of f_t, with M denoting the smoothness of the functions being considered, which is chosen by the researcher. These honest CIs are built by considering the non-random worst-case bias of the estimator $\hat{\tau}$ for functions in $\mathcal{F}(M)$. For more details, see (Armstrong and Kolesár, 2018; Armstrong and Kolesár, 2018b). By applying the insights about identification and inference in the previous sections we can extend the honest CIs approach in Armstrong and Kolesár (2018) to the measurement-error-corrected RD setting. Intuitively we will approximate $f_t(\cdot)$ by a polynomial of order J which can be recovered from the observed data following the procedures described in section 2.3 and obtain an approximation to the worst-case bias that can be used to generate the honest CIs. In this setting,

$$f_t(x) = \sum_{j=0}^{J} x_j\beta_j + R_t(x), |R_t(x)| \leq \bar{R}_t(x) f_t(x) = \sum_{j=0}^{J} x_j^*\beta_j + R_t^*(x) \tag{2.17}$$

where

$$R_t^*(x) = \sum_{j=0}^{J} \left[\sum_{k=0}^{j} \binom{j}{k} \left(e^{(j-k)} - \mu_g^{(j-k)} \right) \tilde{x}^k \right] \beta_j + R_t(x) \tag{2.18}$$

Note that we can sum across the individuals on each group,

$$N_g^{-1} \sum_{i=1}^{N_g} f(x_i) = N_g^{-1} \sum_{i=1}^{N_g} \sum_{j=0}^{J} x_{j,i}^*\beta_j + N_g^{-1} \sum_{i=1}^{N_g} R_t^*(x_i) \tag{2.19}$$

$$N_g^{-1} \sum_{i=1}^{N_g} R_t^*(x_i) = N_g^{-1} \sum_{i=1}^{N_g} R_t(x_i) + o_p(1) \tag{2.20}$$

Hence we can obtain the worst-case bias by focusing on $N_g^{-1} \sum_{i=1}^{N_g} R_t(x_i) = R_{t,g}$. The bounds for $|R_{t,g}|$ will be directly obtained from the conditions and class function adopted for the original $f_t(x)$. It is worth noting that these are defined in terms of the original conditional means of the outcome, considering the model with no measurement error on the running variable. That is positive since the properties and constraints the researcher imposes on the DGP are more intuitive and natural in that setting. As in (Armstrong and Kolesár, 2018; Armstrong and Kolesár, 2018b) we focus on the Taylor class of functions such that

$$f_t \in \mathcal{F}_J(M) = \left\{ f : |f(x) - \sum_{j=0}^{J} f^{(j)}(0) \frac{x_j}{j!}| \leq \frac{M}{p!} |x^{J+1}|, \text{for all } x \in \mathcal{X} \right\} \tag{2.21}$$

Then, by similar arguments we can rewrite the bounds on misspecification in terms of the observed transformed data.

$$|R_t(x)| \leq \frac{M}{p!} |x^{J+1}| \tag{2.22}$$

$$|R_{t,g}| \leq \frac{M}{p!} N_g^{-1} \sum_{i=1}^{N_g} |x_i^{J+1}| \leq \frac{M}{p!} N_g^{-1} \sum_{i=1}^{N_g} |x_{J+1,i}^*| + o_p(1) \tag{2.23}$$

Hence, we can rely on the results in Armstrong and Kolesár (2018); Armstrong and Kolesár (2018b) to obtain "honest" CIs in the presence of measurement error as described above.

In particular, the estimator proposed in section 2.3 can be rewritten to match more closely the notation in those papers

$$\hat{\tau} = \sum_{i=1}^{n} w^n(x_i^*) y_i, \tag{2.24}$$

$$w^n(x^*) = w_+^n(x^*) - w_-^n(x^*) \tag{2.25}$$

$$w_+^n(x^*) = e_1' Q_{n,+}^{-1} X^* D \tag{2.26}$$

$$Q_{n,+} = \sum_{i=1}^{n} D_i X_i^{*\prime} X_i^* \tag{2.27}$$

And similarly for $Q_{n,-}$ and weights $w_-^n(x^*)$. Note that when we replace $w^n(x^*)$ by $\hat{w}^n(\hat{x}^*)$ an additional term will be added to the estimator's residuals. This analysis fits within the framework developed by Armstrong and Kolesár (2018) Theorem F.1. with the following adjustments in notation:

$$\hat{L} = \hat{\tau} = \sum_{i=1}^{n} \hat{w}^n(\hat{x}_i^*) y_i \tag{2.28}$$

$$u_i = \hat{x}_i^{*\prime} \left[\left(x_i^* - \hat{x}_i^* \right) B_+ + \varepsilon_i \right] \tag{2.29}$$

with $s_{n,\mathcal{Q}} = \Omega^{\frac{1}{2}}$, and $\hat{se}_n = \hat{\Omega}^{\frac{1}{2}}$ where the variance matrix Ω and its estimators are those described in theorem 2.2. Then, coupling the bias bounds derived above with the results in Armstrong and Kolesár (2018), the largest possible bias of the estimator over the parameter space $\mathcal{F}_J(M)$ is asymptotically given by

$$\bar{bias}_{\mathcal{F}_J(M)}(\hat{L}) = \frac{M}{p!} \sum_{i=1}^{n} |w_+^n(x_i^*) + w_-^n(x_i^*)||x_{J+1,i}^*|. \tag{2.30}$$

and the "honest" Cis can be obtained as

$$\hat{L} \pm cv_\alpha \left(\frac{\bar{bias}_{\mathcal{F}_J(M)}(\hat{L})}{\hat{se}_n} \right) \cdot \hat{se}_n. \tag{2.31}$$

where $cv_\alpha(t)$ is the $1-\alpha$ quantile of the absolute value of a $N(t,1)$ distribution.

3 Simulation Evidence

This section presents simulation evidence that the RD measurement error correction and tests based on the adjusted standard error proposed in the previous section provide reasonable estimates of the treatment effect at the threshold as well as significant improvements in the test's empirical coverage, substantially reducing the empirical bias and size distortions due to the measurement error.

3.1 Data Generating Process

For ease of comparison with the previous literature, we focus on a DGP similar to Calonico, Cattaneo, and Titiunik (2014) except that the running variable may be measured with error. Throughout we still define X to be the true running variable and $\tilde{X} = X - \epsilon$ to be the mismeasured running variable observed by the researcher. The simulated data are generated as follows:

$$\tilde{X}_i \sim U(-1, 1),$$
$$X_i = \tilde{X}_i + \epsilon_i,$$
$$Y_i = m_j(X_i) + v_i,$$

where $v_i \sim i.i.d.\ N(0, 0.1295^2)$. The expectation of the outcome conditional to X, is given by:

$$m(x) = \begin{cases} 1.27x + 7.18x^2 + 20.21x^3 + 21.54x^4 + 7.33x^5 & \text{if } D = 1 \\ 0.84x - 3.00x^2 + 7.99x^3 - 9.01x^4 + 3.56x^5 & \text{otherwise,} \end{cases}$$

The model is based on a fifth-order polynomial fitted to the data in Lee (2008) in his analysis of incumbency effects on electoral races to the U.S. House of Representatives. As discussed in Section 2, we assume that the misspecification in the model is asymptotically negligible. Since the correction proposed becomes more demanding on the auxiliary data for higher order polynomials, and to focus on the importance of measurement error we present results for DGPs based on polynomials of order 1 through 5 using only the terms up to order j of the equation $m(x)$. Hence, for every simulation result below we adopt a correct specification.

We present the empirical bias and coverage rates based on a 5% nominal size test for the null hypothesis that τ equals its true value. As discussed in Section 2, uncentered moments of the measurement errors in the running variable can be used to recover the coefficients of the conditional mean of interest, correcting for multiple group-specific types of mismeasurement of the running variable.

In the following tables and figures, the results are designated as follows:

1. "No Error" - No measurement error in the running variable.

2. "Naive" - Mismeasured running variable without a measurement error correction.

3. "Corrected" - Mismeasured running variable with a measurement error correction using estimated error moments.

4. "Adjusted S.E." - Mismeasured running variable with a measurement error correction using estimated error moments and adjusted standard errors.

We expect that tests from cases 2 and 3 will have empirical size deviating from the test's nominal level.

3.2 Treatment Based on True Running Variable

We first present the case where the true running variable is continuous and treatment is determined based on X. In addition, the measurement error is such that the resulting observed running variable, $\tilde{X}$, is still continuous at the cutoff, and we allow for seven different types of measurement error, e_i:

1. $U(0, a)$: "rounding down" with uniform distribution,

2. $U(-a, 0)$: "rounding up" with uniform distribution,

3. $U(-a, a)$: "rounding to midpoint" with uniform distribution,

4. $N(\mu, \sigma^2)$ truncated by $(0, a)$: "rounding down" with truncated normal distribution,

5. $N(\mu, \sigma^2)$ truncated by $(-a, 0)$: "rounding up" with truncated normal distribution,

6. $N(\mu, \sigma^2)$ truncated by $(-a, a)$: "rounding to midpoint" with truncated normal distribution,

7. No measurement error.

where $a = 0.1$, $\mu = 0.05$, $\sigma = 0.05$.

Each observation i is randomly assigned to one of these seven groups from which the measurement error will be drawn, and these "group" assignments are observed by the researcher.

Each primary sample has 200 observations and 4000 replications are used to calculate the empirical bias and coverage of the estimated treatment effect. The auxiliary samples have 2000 observations.

Table 3.1 provides evidence that, as expected, the measurement error correction is successful at recentering the estimator around the true parameter, even though the magnitude of improvement is dependent of the specific DGP used to generate the data. Also, ignoring the variability introduced by the estimation of the moments of the measurement error distribution for each

Table 3.1: Simulation 1

Wald Statistic	Level	Empirical Bias and Rejection rates				
Polynomial Order		1	2	3	4	5
Bias	No Error	-0.0001	-0.0018	-0.0020	-0.0015	-0.0265
	Naive	0.0871	0.2379	0.1485	0.1307	0.1158
	Corrected	0.0080	-0.0007	0.0598	0.0648	0.0178
Empirical Coverage	No Error	95.2	94.0	92.2	91.9	90.3
	Naive	69.1	53.7	81.8	85.8	77.6
	Corrected	94.9	93.57	93.2	92.5	90.1
	Adjusted S.E	95.0	93.85	94.1	93.7	91.4

Simulation results based on 4000 replications. $N_p = 200$, $N_{aux} = 2000$.

group can distort the test's coverage, but that is mitigated by the adjustment to the standard errors based on Theorem 2.2.

3.3 Treatment Based on Observed Running Variable

We now examine a case in which treatment is determined based on the observed running variable, $\tilde{X}$. In particular, the measurement error induces heaping on the data, mimicking the features of the birth weight data analyzed in Section 4 and is related to simulations performed by Barreca, Lindo, and Waddell (2016), who point out that this measurement error-induced heaping will cause the expected value of the estimates to shift as heaps are included or excluded on each side of the cutoff, making inference unreliable.

In particular, we generate a primary dataset based on the very low birth weight example. The true running variable X is generated for the 1400-1600 grams interval. To obtain the observed $\tilde{X}$, we impose that 40% of the data is rounded to nearest ounce, 40% is rounded to nearest 50 gram multiple, and 20% of the data is correctly measured. The functional form of the conditional expectation of the outcome still follows m_j as described above.[4]

The measurement error correction recovers correctly centered estimates for the treatment effect and adapts to the inclusion/exclusion of heaps reflected on the different choices of bandwidth around the threshold on Table 3.2. Empirical coverage improves markedly in most cases, becoming closer to the test's nominal size. Both empirical bias and coverage vary significantly with the polynomial order and bandwidth choice, which drive the amount of bias introduced by the heaping in the data, in line with the conclusions in Barreca, Lindo, and Waddell (2016). Even though the empirical coverage obtained by tests implementing the measurement error cor-

[4]Since the functional form of $m(x)$ was designed for a running variable in the support $[-1, 1]$ we re-scale the interval 1400-1600 before generating the outcome data to fit that support. This does not affect the results presented.

rection and adjusted standard errors still diverges from 95% in several cases, the coverage is much more stable with regards to bandwidth choice when compared to the naive procedure's performance.

Table 3.2: Simulation 2

Wald Statistic	Bandwidth		Empirical Bias and Rejection rates				
Polynomial Order			1	2	3	4	5
Bias	20	Naive	0.0000	-0.1753	-0.1696	-0.1813	-0.1713
		Corrected	0.0000	-0.0003	0.0032	0.0002	0.0135
	40	Naive	-0.0013	-0.1653	-0.1557	-0.1880	-0.1809
		Corrected	-0.0016	-0.0023	0.0010	-0.0032	-0.0156
	60	Naive	0.0000	-0.1480	-0.0949	-0.1664	-0.2047
		Corrected	0.0000	-0.0014	-0.0174	0.0023	-0.0391
	80	Naive	-0.0005	-0.1336	-0.1359	-0.2212	-0.1515
		Corrected	-0.0003	0.0000	-0.0057	0.0000	-0.0024
Empirical Coverage	20	Naive	94.50	32.62	72.40	70.30	63.87
		Corrected	94.07	91.92	88.02	82.57	77.32
		Adjusted S.E	94.10	91.95	88.02	82.57	77.32
	40	Naive	93.60	25.67	59.37	70.90	75.97
		Corrected	93.00	92.90	91.00	86.42	80.17
		Adjusted S.E	93.00	92.92	91.00	86.42	80.17
	60	Naive	94.32	41.17	82.22	66.30	73.20
		Corrected	93.52	92.77	94.05	91.77	85.72
		Adjusted S.E	93.52	92.77	94.05	91.87	85.77
	80	Naive	94.05	56.57	69.02	49.62	75.82
		Corrected	93.07	92.40	93.65	91.30	84.60
		Adjusted S.E	93.12	92.45	93.70	91.32	84.65

Simulation results based on 4000 replications. $N_p = 200$, $N_{aux} = 800$.

4 Application I: Very Low Birth Weight

Here we apply our approach to the case studied by ADKW looking at the effect on infant mortality of additional care received by newborns classified as Very Low Birth Weight (VLBW). They take advantage of the fact that VLBW is classified based on having a measured birth weight of strictly less than 1500 grams. This setup lends itself to estimating the effect of these additional resources and services by RD where the measured birth weight is the running variable and treatment is switched on when passing 1500g from above.

ADKW focus on a window of measured birth weights from 1415g-1585g and estimate the treatment effect controlling for a linear function in birth weight on each side of the cutoff. Doing so, they estimate fairly large effects of additional care. Their baseline estimates with no controls suggest a 0.95 percentage point decline in the one-year mortality rate from crossing the 1500g threshold and receiving additional care, a fairly large effect given a mean mortality rate of 5.53 percent for the untreated just above the cutoff.[5]

BGLW suggest caution in interpreting these results by noting that the observed distribution of birth weights shows large "heaps" at ounce multiples and at multiples of 100g as well as smaller heaps at other points (multiples of 50g, 25g, etc.). BGLW focus on the fact that some of the heaped measures tend to have higher mortality rates than neighboring unheaped measures. In particular, they emphasize that the observations measured at 1500g have "substantially higher mortality rates than surrounding observations on *either* side of the VLBW threshold." BGLW view this as evidence of potential non-random sorting into a 1500g birth weight measure and propose a simple sensitivity check called the Donut RD. They test the robustness of the estimated treatment effect to dropping the heaped observations very near the cutoff, creating a "donut hole" with no data around the cutoff. They start by dropping the 1500g observations and then progressively increase the size of the donut hole until they are excluding observations with measured birth weights between 1497g-1503g. Importantly, 1503g corresponds to one of the large heaps at 53 ounces. BGLW find that the estimated treatment effect falls substantially when omitting observations near the cutoff.

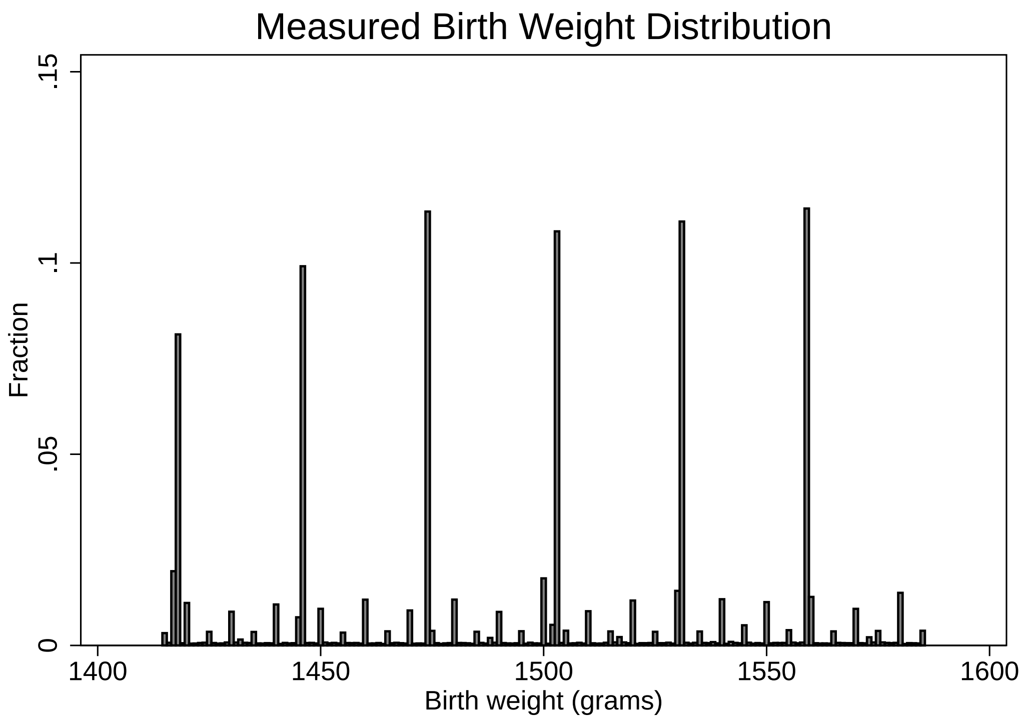
4.1 Data

The main data on birth weight and infant mortality are drawn from the National Center for Health Statistics linked birth/infant death files.[6] The data are discussed in detail in ADKW and BGLW. Briefly, the data include information from birth certificates for all births in the US between 1983-1991 and 1995-2002 and are linked to death certificates for infants up to one year after birth. For the main analysis sample used here this yields 202,078 separate births with an overall one year mortality rate of 5.8 percent. The histogram in Figure 4.1 shows the distribution of measured birth weights for the main sample used with the largest heaps occurring at the six ounce multiples within the 1415g-1585g window used by ADKW and BGLW.

[5] ADKW's main results include additional control variables, but here we focus on the RD results without controls. See Frölich and Huber (2018) for a discussion of RD with and without covariates.

[6] Raw data files available at `https://www.cdc.gov/nchs/data_access/vitalstatsonline.htm`. We thank Alan Barreca for providing the data files from BGLW.

Figure 4.1



Source: National Center For Health Statistics Linked Birth-Infant Death Files. N=202,078.

4.2 Measurement Error Correction

The analysis in BGLW highlights the potential importance of heaping in running variables. Here, we hope to extend their analysis by addressing the underlying measurement problem that leads to heaping.[7] Heaping at ounce and gram multiples is likely due to rounding errors. Specifically, BGLW note that the scales used by hospitals to weigh newborns differ in their precision and there may be a human tendency to round numbers when recording the birth weight. Here we explore the importance of the differential rounding error by using our measurement error correction, and assume that measured birth weights at ounce, 100g, 50g, 25g, 10g, or 5g multiples reflect true birth weights that were rounded to that nearest multiple. All other observed measures— those not at one of the multiples— are assumed to be correctly measured. For instance, it is assumed that the true birth weight for those measured at 1500g will range from 1450g to 1549g and were simply rounded to the nearest 100g multiple. Similarly, those measured at 1503g (53oz) had true birth weights between half an ounce above and below (from 1489g-1517g).

This sort of differential rounding leads to an interesting pattern of potential measurement

[7]Note that while our correction accounts for potential discontinuities in measurement error near the cutoff, it does not address potential endogeneity in hospital measuring systems near the cutoff, similarly to the previous literature.

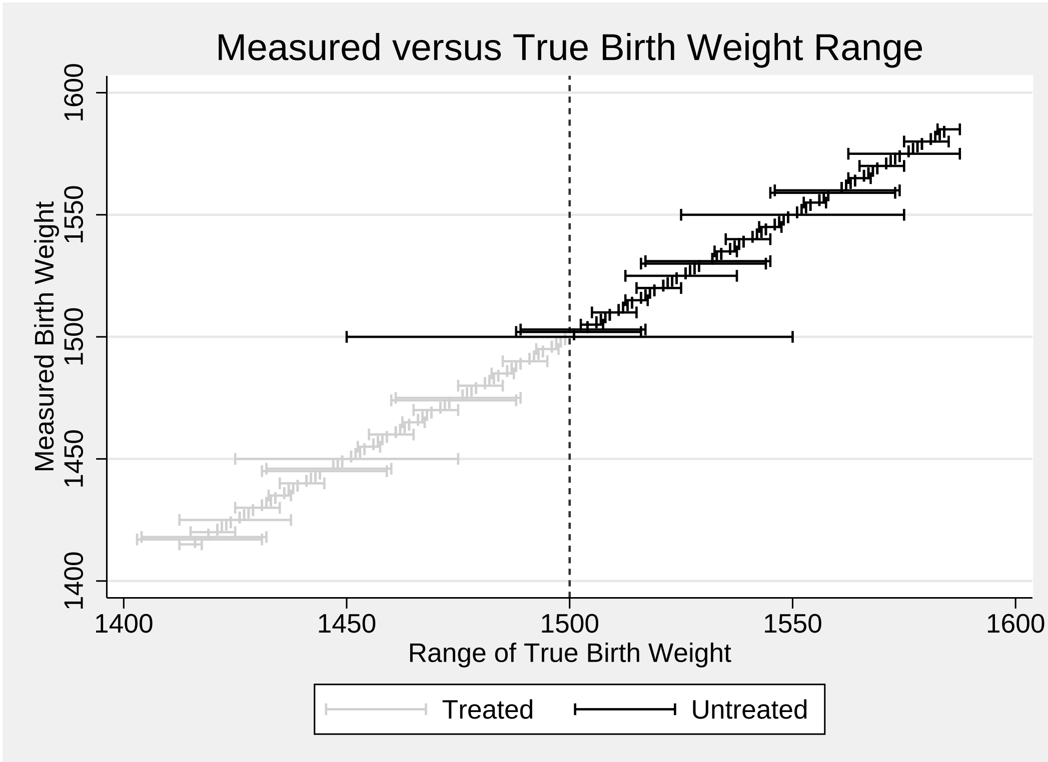
errors. In Figure 4.2, we plot the observed birth weight measure on the vertical axis and the range of potential true birth weights on the horizontal distinguishing between observed measures that receive treatment (observed measure less than 1500g) and those that do not. First note that among those with a measured birth weight just to the right of the cutoff, many may have true birth weights well below the cutoff. This provides a potential explanation for why the mortality rate at 1500g is noticeably higher— namely, these children do not receive the additional care, but many will have similar birth weights and associated unobservable factors to children at much lower birth weights who do receive additional treatment. Also note that this "misclassification" only occurs for untreated units as none of the true weight ranges for treated units cross the threshold. Finally, note that the 1500g measure exhibits the largest potential measurement error range in the described rounding error scenario.

Figure 4.2 also suggests that this setting fits well with our correction procedure: there are groups of observations that face different measurement error distributions and these groups are identified by their measured birth weight. To apply our procedure we need to approximate the true birth weight distributions within each measurement group. For our baseline estimates, we use all births with observed birth weights between 1000g-2000g that are not at one of the heaped values and use a kernel density estimate of the distribution for the unheaped observations.[8] The estimated density is then used to calculate the moments of the birth weight distribution in each measurement group, where the measurement groups are defined by the observed measure ($\widetilde{X}$). Since the distributions were not importantly skewed, the first moments do not differ much from the observed measure since the mid-point (observed measure) is quite close to the mean (first corrected moment). This implies that when fitting a linear function of the running variable, the measurement error will play a minor role. However, when higher order polynomials are used to reduce the approximation error in the conditional mean of the outcome, measurement error could substantially distort estimates.

In Table 4.1, we present the corrected and uncorrected estimates for different samples. We use a cross validation procedure to pick the order of the polynomial as suggested by Lee and Lemieux (2010). We use the Akaike Information Criteria (AIC) to choose among all combinations of polynomials of length one to five on either side of the VLBW cutoff. For the uncorrected, the AIC was minimized with a fifth order polynomial on the right side (untreated) and a linear function on the left (treated), while for the corrected estimates it was with a fourth order

[8] Nearly identical results were obtained including all births— both heaped and unheaped— when estimating the density while using a wide bandwidth in order to smooth out the heaps. This suggests that in terms of estimating the true birth weight distribution our choice to focus on unheaped measures only does not lead to a problematic selection issue.

Figure 4.2



Ranges refer to the range of potential true birth weights for a given measured birth weight.

polynomial and linear function, respectively. Starting in the first row, using the same sample as BGLW, we see a large difference between the uncorrected and corrected estimates in Columns (a) and (b), respectively. The uncorrected estimates suggest a 3.1 percentage point drop in the mortality rate when receiving additional care, while the corrected estimate is only 0.67 percentage points. Note that in Column (b), adjusting the standard errors for the first-step estimation of the measurement error moments has little effect on on the estimated standard errors. This is due to two factors: (1) many of the observations (those not at an ounce or 5g multiple) are assumed to be correctly measured and require no adjustment when calculating the standard errors and (2) the auxiliary samples in this case are quite large, implying fairly precise first-step estimates.

To provide some intuition for the difference in the two estimates, Figure 4.3 depicts the estimated functions on either side of the cutoff along with mean mortality rates within five gram bins of the observed birth weight measure. First, we see that the two approaches yield similar estimates of the conditional mean function to the left of the cutoff. However, we see very

Table 4.1

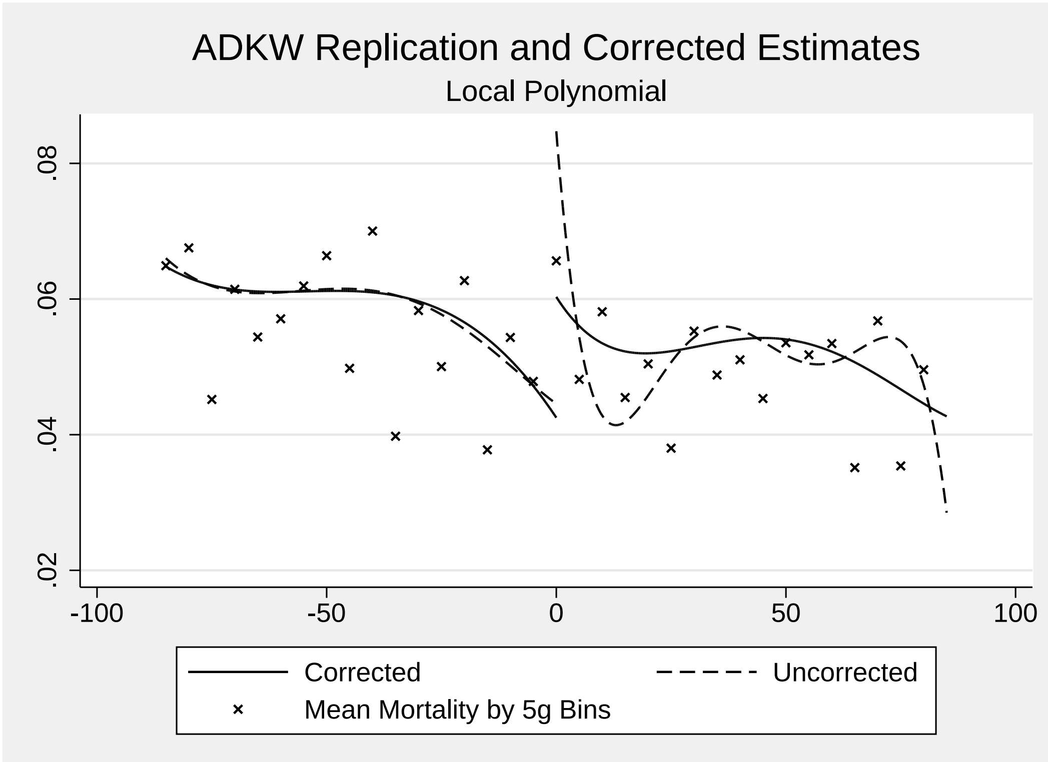
RD VLBW Estimates: Corrected and Uncorrected			
Estimator:	*Uncorrected*	*Corrected*	*Corrected*
Measurement Error Groups:		$\widetilde{X}$	$\widetilde{X}$ *by Education*
	(a)	*(b)*	*(c)*
(1) All BGLW Sample	-0.0311	-0.0067	-0.0067
$N_p = 202,078$	(0.0037)	(0.0025)	(0.0025)
		[0.0025]	[0.0031]
(2) Omitting 1500g	-0.0139	-0.0067	-0.0067
$N_p = 198,530$	(0.0043)	(0.0025)	(0.0025)
		[0.0025]	[0.0032]
(3) Omitting 1499-1501g	-0.0149	-0.0069	-0.0069
$N_p = 198,334$	(0.0043)	(0.0025)	(0.0025)
		[0.0025]	[0.0031]
(4) Omitting 1498-1502g	-0.0156	-0.0069	-0.0069
$N_p = 197,135$	(0.0045)	(0.0025)	(0.0025)
		[0.0025]	[0.0031]
(5) Omitting 1497-1503g	-0.0038	0.0102	0.0010
$N_p = 175,108$	(0.0147)	(0.0105)	(0.0105)
		[0.0110]	[0.0585]

Source: National Center For Health Statistics Linked Birth-Infant Death Files. $\widetilde{X}$ Groups: $N_a = 27,846; \min(N_{a,g}) = 10,145; \max(N_{a,g}) = 12,279$. $\widetilde{X}$ by Mother's Education Groups: $N_a = 27,846; \min(N_{a,g}) = 1,033; \max(N_{a,g}) = 5,067$. N_p is the primary sample size, N_a is the total auxiliary sample size, and $N_{a,g}$ is the auxiliary sample size for group g. Unadjusted standard errors in parentheses and adjusted standard errors in square brackets.

different fitted functions to the right of the cutoff. Intuitively, the uncorrected function gets very steep near the cutoff because it treats the observations at a measured weight of 1500g that have a higher mean mortality rate as being precisely measured at 1500g and tries to fit that point. In contrast, our correction recognizes that many of those with a measured birth weight of 1500g may have a true birth weight away from 1500g and the regression function above the cutoff is not influenced as much by these observations.

We also revisit the donut RD from BGLW here, holding the order of the polynomial fixed as in BGLW. In Row (2) of Table 4.1, we see that the corrected estimate is unaffected by dropping the 1500g heap, while the uncorrected falls by over half as in BGLW. The fact that the corrected estimate is robust to dropping the 1500g heap is encouraging that our correction is helping to control for the underlying measurement problem that led to the heap. Intuitively, the uncorrected estimator treats every observation measured at 1500g as precisely measured. As discussed before, the group measured at 1500g has a relatively high mortality rate, so omitting these observations removes a large mass with a high mortality rate from a single point right at

Figure 4.3



Source: National Center For Health Statistics Linked Birth-Infant Death Files. $N_p = 202,078; N_a = 27,846; \min(N_{a,g}) = 10,145; \max(N_{a,g}) = 12,279$. N_p is the primary sample size, N_a is the total auxiliary sample size, and $N_{a,g}$ is the auxiliary sample size for group g. Lines represent local polynomial regressions with order chosen using AICc.

the cutoff. Instead, the corrected accounts for the fact that most observations with a measured weight of 1500g actually have true birth weights above or below 1500g. Therefore, it is as if we are removing observations from the range of true birth weights associated with a measured weight of 1500g. Additionally, the treatment cutoff of 1500g still falls in the support of the true birth weights for the 1503g measure, as shown in Figure 4.2. Together, this suggests that it is similar to randomly dropping some observations from a range of true X while keeping the cutoff in the support of the data in the perfectly measured case, which we would not expect to affect the estimate of the conditional mean drastically.

While dropping observations with measured birth weights up to 3g away from the cutoff alters the corrected estimate in Row (5), the corrected estimate otherwise appears remarkably stable across the different size donut holes. In contrast, uncorrected estimates are quite sensitive to the different size donut holes used. Importantly, 1503g corresponds to 53oz and ounce measures seem to be the most common type in the data. As ADKW note in their reply to BGLW, this actually removes about 20 percent of the data to the right of the cutoff while barely dropping any to the left (Almond et al., 2011). This is because the closest ounce measure from below is at 1474g. In justifying their approach, BGLW note that dropping those within 3g and at the cutoff represents an incremental difference in birth weights since the implied gap in birth weights between the observations to the left and right of the cutoff is roughly equivalent in weight to seven paper clips (7g). However, when viewed from the perspective that those measured at ounce multiples are rounded to the nearest ounce, this implies that the gap between most true birth weights when dropping 1503g is actually between 29g-85g since the largest ounce measure below the cutoff is 1474g with a true range from 1460-1488g and the first ounce measure above 1503g is at 1531g with a true range of 1517-1545g. In particular, now most the data on the untreated side are for babies with much higher birth weights who have much lower mortality rates regardless of treatment. This suggests some caution considering the basic RD identification argument when dropping the 1503g heap as the babies on either side of the threshold may no longer be comparable along unobservables.

An additional issue raised by BGLW is that the measurement technology available may differ by hospitals that serve women with different backgrounds. In particular, hospitals in higher poverty areas may have less precise scales (more likely to have a rounded birth weight). If the true birth weight distribution differs across different maternal backgrounds— for example, more mass at lower birth weights for disadvantaged mothers— this could lead to differences in the measurement error distributions for babies at the same observed measure. To address this possibility, we allow the measurement error to differ by mother's education level (less than high school, high school, some college, college and above, and missing education data). Specifically, we simply redefine our measurement error groups to be based on the observed measure and mother's education. We then re-estimate the birth weight density for each education level to generate a new set of corrected moments. Column (c) of Table 4.1 displays the results using the mother's education specific measurement groups. The results are very similar to those in Column (b) of Table 4.1 with an estimated treatment effect of 0.0067 that is robust to dropping observations within 2g of the cutoff. The corrected standard errors are now noticeably larger than the uncorrected due to dividing the sample into more measurement error groups with smaller auxiliary samples.

5 Application II: UI Benefit Effects using Geographic RD

In this section, we apply our correction procedure to the problem of estimating the effect of Unemployment Insurance (UI) extensions on unemployment during the Great Recession using a

GeoRD. During the Great Recession, the duration of UI benefits was extended from 26 weeks to as many as 99 weeks. The realized benefit duration varied at the state level and was determined by state-level labor market aggregates passing pre-specified trigger levels.[9] In theory, such extensions may lead to increased unemployment through reduced job search effort by workers and a contraction of vacancies by firms. The main econometric challenge in estimating the effect on unemployment is to isolate the differences due to the policy from the differences due to the factors driving adoption of the policy.

Hagedorn et al. (2015) and Dieterle, Bartalotti, and Brummet (2018) both study this case in detail, attempting to exploit differences in UI extensions at state boundaries in estimation.[10] Here, the goal is to compare the preferred RD estimates using the measurement error correction proposed in Section 2 to those using a mismeasured, centroid-based, distance to the state borders. During the recession, there were many instances in which neighboring states faced different UI regimes due to the fact that the extensions were triggered by state level aggregate unemployment. To focus our discussion on the correction procedure, we will consider one such case: the Minnesota- North Dakota boundary in the second quarter of 2010. The average available UI benefit duration over the entire quarter in Minnesota was 62 weeks while it was only 43 weeks in North Dakota.

5.1 Data

We use county-level data on the unemployment rate from the Bureau of Labor Statistics' (BLS) Local Area Unemployment Statistics (LAUS), and the duration of UI benefits provided by US Department of Labor.[11] Our sample includes all counties located in either state for which the MN-ND boundary is the closest state boundary.

5.2 Measurement Error

The main issue with implementing the RD strategy in this case is that geographic location is reported at the county level, but the underlying running variable is a continuous measure of distance to the border. Because of this, researchers often calculate the distance to the border based on the geographic center of the county. This geographic centroid based distance measure is the

[9] See Hagedorn et al. (2015) and Rothstein (2011) for a more detailed discussion of the institutional details of Unemployment Insurance benefit extensions.

[10] Dieterle, Bartalotti, and Brummet (2018) implements the measurement error correction procedure as proposed in this paper for the whole U.S.

[11] See `http://ows.doleta.gov/unemploy/trigger/` and `http://ows.doleta.gov/unemploy/euc_trigger/`. Here, we use the county-level unemployment rate as given. See Dieterle, Bartalotti, and Brummet (2018) for a discussion of potential issues with using an aggregate outcome measure in this setting.

mismeasured running variable in this context. To implement the measurement error correction in this GeoRD example, we require information on the geographic location of counties and the within county population distribution relative to a state boundary to calculate the moments of the measurement error present on the data. We use the TIGER geographic shapefiles that contain population counts by census block from the 2010 Census. The geographic information gives precise location of census block, county, and state borders. For the centroid-based distance we can therefore calculate the distance from the geographic center of a county to the state border. We can also calculate the distance from the center of each census block to the state boundary. Since census blocks are typically very small, we can use this to approximate a continuous measure of the population weighted distance to the border needed for our measurement error correction.

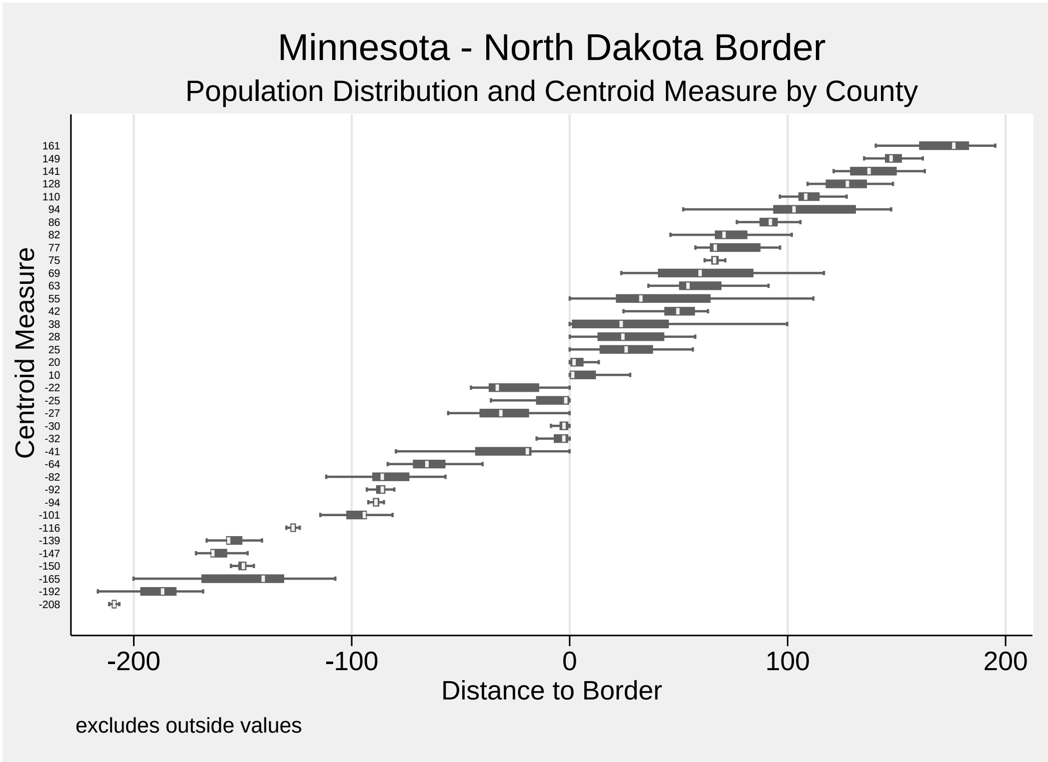
To calculate the population moments and the centroid-based distance from the TIGER shapefiles, we use the nearstat package in Stata (Jeanty, 2010). Since particular areas within a county may have a different nearest neighbor, we determine the modal nearest state boundary among the census blocks in a county and then calculate distances and moments based on the modal neighbor.

Figure 5.1 depicts box plots of the population distribution within each county in our sample. Here, the population distributions are directly linked to the measurement error distributions. We have ordered the plots by the centroid measure, starting at the county farthest from the border in North Dakota 208 km away up to the farthest in Minnesota at 161 km away. Several features of the measurement error are worth highlighting. First, the measurement error distributions do not cross the cutoff, so the treatment is identically defined by the true and mismeasured running variables. In many cases, when comparing two counties the one that is measured to be closer by the centroid based measure actually has most of the population mass farther away. For many of the counties the population distributions are far from symmetric and, importantly, they vary substantially across each group (county). Together this suggests that the group-specific measurement error correction may be particularly important in this setting.

5.3 Results

We estimate the effect of the difference in available UI duration at the Minnesota-North Dakota border in the second quarter of 2010 on log unemployment by GeoRD using both the uncorrected centroid based measure and our moments based correction. Again we follow Lee and Lemieux (2010), and use a cross-validation procedure to choose the order of the polynomial opting for

Figure 5.1



Source: US Census TIGER Geographic Shapefiles. N=93,530 census blocks

the small sample version of the Akaike Information Criteria (AICc) due to the relatively small sample size on both sides of the border. The AICc chosen polynomial in each case is a quadratic.

Table 5.1 presents the uncorrected and corrected estimates for the ATE at the boundary. The uncorrected estimates is large and negative, but imprecise. The point estimate for the uncorrected case would suggest a 25 percent reduction in unemployment from the 19 extra weeks of UI available in Minnesota. The corrected estimate is much smaller in magnitude— nearly zero— and more precisely estimated. The lack of an estimated effect when using our correction is consistent with the evidence of UI policy spillovers discussed in Dieterle, Bartalotti, and Brummet (2018).

To provide more intuition for the correction procedure, Figure 5.2 depicts the corrected and uncorrected estimated polynomials along with the centroid measures. We also overlay the range, twenty-fifth to seventy-fifth percentile range, and the median of the population distribution to provide some sense of how the population distribution differs from the centroid measure. On the North Dakota side of the border (distance less than zero), we see the uncorrected polynomial is influenced by a few counties that have a centroid distance roughly 10-40km away from the border, but have households living right up to the border. The corrected estimates take into

Table 5.1

Geographic RD: Corrected and Uncorrected Estimates	
(1)	(2)
Uncorrected	*Corrected*
-0.2855	-0.0155
(0.1943)	(0.1436)
	[0.1548]

Source: LAUS and TIGER Geographic Shapefiles. $N_p = 38; N_a = 93,530; \min(N_{a,g}) = 845; \max(N_{a,g}) = 7,921$. N_p is the primary sample size, N_a is the total auxiliary sample size, and $N_{a,g}$ is the auxiliary sample size for group g. Unadjusted standard errors in parentheses and adjusted standard errors in square brackets.

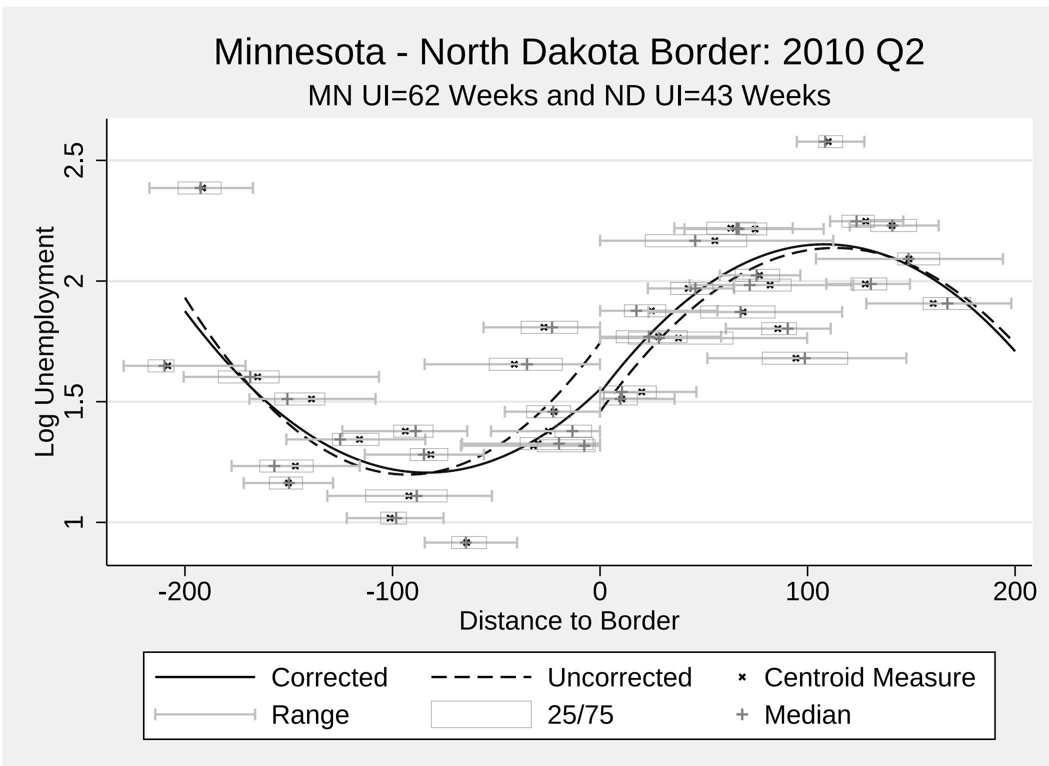
account the fact that the population distribution is skewed toward the state border for many of these counties lowering the estimated intercept on the North Dakota side. On the Minnesota side, the effect is the opposite, raising the estimated intercept. Combined, this reduces the estimated treatment effect.

6 Conclusion

RD designs have become increasingly popular in empirical studies, but researchers often face situations where there are several types of group-specific measurement errors in the forcing variable. In order to accommodate these situations, we propose a new procedure that utilizes auxiliary information to correct for the bias induced by group-specific measurement error. We propose a valid estimator of the parameter of interest and derive its asymptotic distribution which takes into account both the variability introduced by the measurement error correction and the use of multiple data sets in estimation. This method complements previous work on measurement error in RD designs by allowing more flexible forms of the measurement error, including measurement error that is potentially non-classical and discontinuous at the cutoff. Furthermore, the approach is effective regardless of whether treatment is assigned based on the "true" or mismeasured running variable.

Simulation evidence is presented supporting the theoretical results proposed on the paper and its superior performance relative to "naive" estimators. In our empirical illustrations, we demonstrate that correcting for measurement error can provide a new empirical perspective on the data.

Figure 5.2



Source: LAUS and TIGER Geographic Shapefiles. $N_p = 38; N_a = 93,530; \min(N_{a,g}) = 845; \max(N_{a,g}) = 7,921$. N_p is the primary sample size, N_a is the total auxiliary sample size, and $N_{a,g}$ is the auxiliary sample size for group g. Lines represent local polynomial regressions with order chosen using AICc.

References

Almond, Douglas, Joseph J Doyle, Amanda E Kowalski, and Heidi Williams. 2010. "Estimating Marginal Returns to Medical Care: Evidence from At-risk Newborns." *Quarterly Journal of Economics* 125 (2):591–634.

Almond, Douglas, Joseph J. Doyle, Jr., Amanda E. Kowalski, and Heidi Williams. 2011. "The Role of Hospital Heterogeneity in Measuring Marginal Returns to Medical Care: A Reply to Barreca, Guldi, Lindo, and Waddell*." *The Quarterly Journal of Economics* 126 (4):2125–2131. URL +http://dx.doi.org/10.1093/qje/qjr037.

Armstrong, Timothy and Michal Kolesár. 2018b. "Simple and honest confidence intervals in nonparametric regression." .

Armstrong, Timothy B. and Michal Kolesár. 2018. "Optimal Inference in a Class of Regression Models." *Econometrica* 86 (2):655–683. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA14434.

Barreca, Alan I, Melanie Guldi, Jason M Lindo, and Glen R Waddell. 2011. "Saving Babies? Revisiting the Effect of Very Low Birth Weight Classification." *Quarterly Journal of Economics* 126 (4):2117–2123.

Barreca, Alan I, Jason M Lindo, and Glen R Waddell. 2016. "Heaping-Induced Bias in Regression-Discontinuity Designs." *Economic Inquiry* 54 (1):268–293.

Bayer, Patrick, Fernando Ferreira, and Robert McMillan. 2007. "A Unified Framework for Measuring Preferences for Schools and Neighborhoods." *Journal of Political Economy* 115 (4):588–638.

Black, Sandra E. 1999. "Do Better Schools Matter? Parental Valuation of Elementary Education." *Quarterly Journal of Economics* 114 (2):577–599.

Calonico, Sebastian, Matias D. Cattaneo, and Rocio Titiunik. 2014. "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs." *Econometrica* 82 (6):2295–2326.

Chen, Xiaohong, Han Hong, and Elie Tamer. 2005. "Measurement Error Models with Auxiliary Data." *Review of Economic Studies* 72 (2):343–366.

Davezies, Laurent and Thomas Le Barbanchon. 2017. "Regression Discontinuity Design with Continuous Measurement Error in the Running Variable." *Journal of Econometrics* 200 (2):260–281.

Dell, Melissa. 2010. "The Persistent Effects of Peru's Mining Mita." *Econometrica* 78 (6):1863–1903.

Dieterle, Steven, Otavio Bartalotti, and Quentin Brummet. 2018. "Revisiting the Effects of Unemployment Insurance Extensions on Unemployment: A Measurement Error-Corrected Regression Discontinuity Approach." IZA Working Paper No. 11496.

Dong, Hao. 2017. "Sharp Regression-discontinuity Design with a Mismeasured Running Variable." Working Paper.

Dong, Yingying. 2015. "Regression Discontinuity Applications with Rounding Errors in the Running Variable." *Journal of Applied Econometrics* 30 (3):422–446.

Eugster, Beatrix, Rafael Lalive, Andreas Steinhauer, and Josef Zweimüller. 2011. "The Demand for Social Insurance: Does Culture Matter?" *Economic Journal* 121 (556):F413–F448.

Falk, Oliver, Robert Gold, and Stephan Heblich. 2014. "E-lections: Voting Behavior and the Internet." *American Economic Review* 104 (7):2238–2265.

Fan, Jianqing and Irene Gijbels. 1996. *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, vol. 66. CRC Press.

Frölich, Markus and Martin Huber. 2018. "Including covariates in the regression discontinuity design." *Journal of Business & Economic Statistics* (just-accepted).

Gibbons, Stephen, Stephen Machin, and Olmo Silva. 2013. "Valuing School Quality Using Boundary Discontinuities." *Journal of Urban Economics* 75:15–28.

Hagedorn, Marcus, Fatih Karahan, Iourii Manovskii, and Kurt Mitman. 2015. "Unemployment Benefits and Unemployment in the Great Recession: The Role of Macro Effects." Federal Reserve Bank of New York Staff Reports No. 646.

Hausman, Jerry A., Whitney K. Newey, Hidehiko Ichimura, and James L. Powell. 1991. "Identification and estimation of polynomial errors-in-variables models." *Journal of Econometrics* 50 (3):273 – 295. URL http://www.sciencedirect.com/science/article/pii/0304407691900226.

Jeanty, P.W. 2010. "nearstat: Stata Module to Calculate Distances, Generate Distance-based Variables, and Export Distance to Text Files." URL `http://ideas.repec.org/c/boc/bocode/s457110.html`.

Keele, Luke and Rocío Titiunik. 2014. "Natural Experiments Based on Geography." *Political Science Research and Methods* :1–31.

Lalive, Rafael. 2008. "How Do Extended Benefits Affect Unemployment Duration? A Regression Discontinuity Approach." *Journal of Econometrics* 142 (2):785–806.

Lavy, Victor. 2006. "From Forced Busing to Free Choice in Public Schools: Quasi-Experimental Evidence of Individual and General Effects." NBER Working Paper No. 11969.

Lee, David S. 2008. "Randomized Experiments from Non-Random Selection in U.S. House Elections." *Journal of Econometrics* 142 (2):675–697.

Lee, David S. and David Card. 2008. "Regression Discontinuity Inference with Specification Error." *Journal of Econometrics* 142 (2):655–674.

Lee, David S. and Thomas Lemieux. 2010. "Regression Discontinuity Designs in Economics." *Journal of Economic Literature* 48 (2):281–355.

Lee, Lung-fei and Jungsywan H Sepanski. 1995. "Estimation of Linear and Nonlinear Errors-in-variables Models Using Validation Data." *Journal of the American Statistical Association* 90 (429):130–140.

Pei, Zhuan and Yi Shen. 2017. "The Devil is in the Tails: Regression Discontinuity Design with Measurement Error in the Assignment Variable." In *Regression Discontinuity Designs: Theory and Applications (Advances in Econometrics)*, vol. 38, edited by Matias D. Cattaneo and Juan Carlos Escanciano. 455–502.

Rothstein, Jesse. 2011. "Unemployment Insurance and Job Search in the Great Recession." *Brookings Papers on Economic Activity* 43 (2):143–213.

Yu, Ping. 2012. "Identification of Treatment Effects in Regression Discontinuity Designs with Measurement Error." Working Paper, University of Auckland.

A Proofs

Proof of Theorem 2.1

Proof. First note that, under Assumption A2, we can rewrite the conditional expectation $f_t(x)$ using a polynomial of order J for $t = 0, 1$ with unknown coefficients b_{jt}.

$$E\left[Y|x, D = t\right] = f_t(x) = \sum_{j=0}^{J} b_{jt} x^j \tag{A.1}$$

Let these coefficients be collected in the column vector $B'_t = [b_{0t}, b_{1t}, \ldots, b_{Jt}]$. Then, since for each observation $x = \tilde{x} + e$,

$$f_t(x) = \sum_{j=0}^{J} b_{jt}(\tilde{x} + e)^j = \sum_{k=0}^{J}\sum_{j=k}^{J} \binom{j}{k} b_{jt} e^{j-k} \tilde{x}^k \tag{A.2}$$

where $\binom{j}{k}$ is the binomial coefficient $\frac{j!}{k!(j-k!)}$. Then,

$$\begin{aligned}
E\left[Y|\tilde{x}, D = t, G\right] =& E\left[E\left[Y|x, D = t\right]|\tilde{x}, D = t, G\right] = E\left[f_t(x)|\tilde{x}, G\right] && \text{(A.3)}\\
=& E\left[\sum_{k=0}^{J}\sum_{j=k}^{J} \binom{j}{k} b_{jt} e^{j-k} \tilde{x}^k |\tilde{x}, G\right] && \text{(A.4)}\\
=& \sum_{k=0}^{J}\sum_{j=k}^{J} \binom{j}{k} b_{jt} E\left[e^{j-k}|\tilde{x}, G\right] \tilde{x}^k && \text{(A.5)}\\
=& \sum_{k=0}^{J}\sum_{j=k}^{J} \binom{j}{k} b_{jt} \mu_g^{(j-k)}(\tilde{x}) \tilde{x}^k && \text{(A.6)}\\
=& \sum_{j=0}^{J} b_{jt} \left[\sum_{k=0}^{j} \binom{j}{k} \mu_g^{(j-k)}(\tilde{x}) \tilde{x}^k\right] && \text{(A.7)}
\end{aligned}$$

By Assumption $A5$, the first J uncentered moments of the measurement error distribution for each group, $\mu_g^{(j-k)}(\tilde{x})$, are known or estimable. The last equality is simply rewriting the sum for convenience. Let $x^*_j = \sum_{k=0}^{j} \binom{j}{k} \mu_g^{(j-k)}(\tilde{x}) \tilde{x}^k$ and $X^{*\prime} = [1, x^*_1, \ldots, x^*_J]$, then the expectation of the outcome Y conditional on the observed $\tilde{x}$ and treatment status can be written as

$$E\left[Y|\tilde{x}, D = t, G\right] = \sum_{j=0}^{J} b_{jt} x^*_j = X^{*\prime} B_t \tag{A.8}$$

For which B_t is identified under the conditions imposed. Hence, $f_0(0)$ and $f_1(0)$ can be identified through b_{0t} for $t = 0, 1$ and $\tau = f_1(0) - f_0(0)$. $\square$

Proof of Theorem 2.2

Proof. To establish the asymptotic properties of the proposed estimator, it is useful to write the vector of transformed running variable used in the estimation. For unit i, associated with a measurement error group g, let:

$$X_i^* = \mu_g(\tilde{x}_i)\Gamma_i' \tag{A.9}$$

where, $X_i^* = [1, x_{i,1}^*, \ldots, x_{i,J}^*]$, $\mu_g(\tilde{x}_i) = [1, \mu_g^{(1)}(\tilde{x}_i), \ldots, \mu_g^{(J)}(\tilde{x}_i)]$ and $\Gamma_i = L_{J+1} \circ Q$ is the Hadamard product of the lower diagonal Pascal matrix, L_{J+1}, and the matrix Q_i, where $Q_{(b.c)} = \tilde{x}_i^{c-b}$. For concreteness, if $J = 3$,

$$\Gamma_i = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1\tilde{x}_i & 1 & 0 & 0 \\ 1\tilde{x}_i^2 & 2\tilde{x}_i & 1 & 0 \\ 1\tilde{x}_i^3 & 3\tilde{x}_i^2 & 3\tilde{x}_i & 1 \end{bmatrix} \tag{A.10}$$

Since $\mu_g(\tilde{x})$ is not observed, but can be consistently estimated from the auxiliary data, let the feasible transformed running variable used in estimation be given by $\hat{X}_i^* = \hat{\mu}_g(\tilde{x}_i)\Gamma_i'$ Then the feasible estimator for the vector $B_+' = [\beta_+, \beta_+^{(1)}, \ldots, \beta_+^{(J)}]$ (and equivalently for B_-), is given by

$$\hat{B}_+ = \left[\hat{X}^{*\prime}\hat{X}^*\right]^{-1} \hat{X}^{*\prime}Y \tag{A.11}$$

and

$$\sqrt{N_p}(\hat{B}_+ - B_+) = \hat{A}^{-1}N_p^{-\frac{1}{2}} \sum_{i=1}^{N_p} \hat{x}_i^{*\prime} \left[(x_i^* - \hat{x}_i^*) B_+ + \varepsilon_i\right] \tag{A.12}$$

$$= \hat{A}^{-1}N_p^{-\frac{1}{2}} \sum_{i=1}^{N_p} \hat{x}_i^{*\prime} \left[(\mu_g(\tilde{x}_i) - \hat{\mu}_g(\tilde{x}_i)) \Gamma_i' B_+ + \varepsilon_i\right] \tag{A.13}$$

with $\hat{A} = N_p^{-1} \sum_{i=1}^{N_p} \hat{x}_i^{*\prime}\hat{x}_i^*$ and $\varepsilon = Y - E\left[Y|\tilde{x}, D, G\right]$ Then we can rewrite

$$\sqrt{N_p}(\hat{B}_+ - B_+) = \hat{A}^{-1}N_p^{-\frac{1}{2}} \sum_{i=1}^{N_p} \hat{x}_i^{*\prime}\varepsilon_i + \hat{A}^{-1} \left[N_p^{-1} \sum_{i=1}^{N_p} \hat{x}_i^{*\prime} \left[N_p^{\frac{1}{2}} \left(\mu_g(\tilde{x}_i) - \hat{\mu}_g(\tilde{x}_i)\right)\right] \Gamma_i' B_+\right] \tag{A.14}$$

Then,

$$\sqrt{N_p}(\hat{B}_+ - B_+) = \hat{A}^{-1} N_p^{-\frac{1}{2}} \sum_{i=1}^{N_p} \hat{x}_i^{*\prime} \varepsilon_i - \hat{A}^{-1} \left[N_p^{-1} \sum_{i=1}^{N_p} (\Gamma_i' B_+ \otimes \hat{x}_i^*)' \left[N_p^{\frac{1}{2}} \left(\hat{\mu}_g(\tilde{x}_i) - \mu_g(\tilde{x}_i) \right) \right] \right] \tag{A.15}$$

$$\sqrt{N_p}(\hat{B}_+ - B_+) = \hat{A}^{-1} N_p^{-\frac{1}{2}} \sum_{i=1}^{N_p} \hat{x}_i^{*\prime} \varepsilon_i - \hat{A}^{-1} \left[N_p^{-1} \sum_{i=1}^{N_p} \left(\hat{x}_i^{*\prime} \otimes B_+' \Gamma_i \right) \left[N_p^{\frac{1}{2}} \left(\hat{\mu}_g(\tilde{x}_i) - \mu_g(\tilde{x}_i) \right) \right] \right] \tag{A.16}$$

Let h_g be the tuning parameter of a kernel based nonparametric estimator of $\mu_g(\tilde{x}_j)$ such that $h_g \to 0$ and $N_{a,g} h_g \to \infty$ as the sample sizes increase for all g. Finally, let $\lambda_g = \lim_{N_p \to \infty} \left(\frac{N_p}{N_{a,g} h_g} \right)$, for all g.

$$\sqrt{N_p}(\hat{B}_+ - B_+) = \hat{A}^{-1} N_p^{-\frac{1}{2}} \sum_{i=1}^{N_p} \hat{x}_i^{*\prime} \varepsilon_i - \hat{A}^{-1} \left[N_p^{-1} \sum_{i=1}^{N_p} \left(\hat{x}_i^{*\prime} \otimes B_+' \Gamma_i \right) \lambda_g^{\frac{1}{2}} \left[(N_{a,g} h_g)^{\frac{1}{2}} \left(\hat{\mu}_g(\tilde{x}_i) - \mu_g(\tilde{x}_i) \right) \right] \right] \tag{A.17}$$

.

In a mild abuse of notation, let the units that are part of a group g be indexed by a "group unit" denomination l such that $N_p = \sum_{i=1}^{N_p} w_i = \sum_{g=1}^{G} \sum_{l=1}^{N_g} w_{lg}$ for any variable w, with N_g the number of observations in group g in our primary sample.

$$\sqrt{\bar{N}_p}(\hat{B}_+ - B_+) = \hat{A}^{-1} N_p^{-\frac{1}{2}} \sum_{i=1}^{N_p} \hat{x}_i^{*\prime} \varepsilon_i - \hat{A}^{-1} \left[N_p^{-1} \sum_{g=1}^{G} \sum_{l=1}^{N_g} \left(\hat{x}_{lg}^{*\prime} \otimes B_+' \Gamma_{lg} \right) \lambda_g^{\frac{1}{2}} \left[(N_{a,g} h_g)^{\frac{1}{2}} \left(\hat{\mu}_g(\tilde{x}_l) - \mu_g(\tilde{x}_l) \right) \right] \right] \tag{A.18}$$

Let $F_{+,g} = E\left[\left(x_{lg}^{*\prime} \otimes B_+' \Gamma_{lg} \right) \right]$. Also, assume that a CLT holds for the measurement error moment estimator using the auxiliary data such that $\sqrt{N_{a,g} h_g} \left(\hat{\mu}_g(\tilde{x}) - \mu_g(\tilde{x}) \right) \to N(0, V_g(\tilde{x}))$. Then the asymptotic variance of $\sqrt{N_p}(\hat{B}_+ - B_+)$ is given by:

$$\Omega_+ = A^{-1} E\left[X^{*\prime} \varepsilon_i \varepsilon_i' X^* \right] A^{-1} + A^{-1} \left[\sum_{g=1}^{G} \lambda_g F_{+,g}' V_g(\tilde{x}) F_{+,g} \right] A^{-1} \tag{A.19}$$

Asymptotic normality is achieved by combining the assumption of independence between auxiliary and primary datasets, the asymptotoc normality of $N_p^{-\frac{1}{2}} \sum_{i=1}^{N_p} \hat{x}_i^{*\prime} \varepsilon_i$ and $(N_{a,g} h_g)^{\frac{1}{2}} \left(\hat{\mu}_g(\tilde{x}_j) - \mu_g(\tilde{x}_j) \right)$, and the definitions of λ_g for every $g = 1, ..., G$. Similarly for the left side of the cutoff with Ω_-. Combining the results for both sides of the cutoff, under random sampling gives the result. □