When Can We Ignore Measurement Error in the Running Variable?

Yingying Dong[1], Michal Kolesár[2]

[1] University of California Irvine; [2] Princeton

@ASSA 2023

Motivation - Why should We Care about Measurement Error?

Measurement error is common in empirical applications of RDD:

- Surveyed 154 RDD papers published in 2005 - 2020 in the leading journals - 27 feature a running variable measured with error; 23 use a rounded or coarsened running variable.
- In addition, many RDD applications use survey data, and hence are subject to reporting/recall error (Pei and Shen, 2017; Davezies and Le Barbanchon, 2017; Yu, 2013, Yanagi, 2015).
    - E.g. self-reported income used as a running variable.

Contributions

- Show that under weak and interpretable conditions, *RDD with a mismeasured running variable is a valid design - it can provide inference for a well-defined causal object.*
- Illustrate the proposed approach in an empirical application.

Sharp Design - Notation

- True running variable X^* is not observed; only observe $X = X^* - e$, where e is the measurement error.
  - E.g., X^* is birthdate; only observe X, month or year of birth. So $X = \lfloor X^* \rfloor$ and $e = X^* - X$.
  - e can be discrete, continuous, or mixed; e may correspond to a classical or Berkson measurement error.
- $Z = 1\{X^* \geq 0\}$. Treatment $T = Z$.
- Y is observed outcome; $Y(t)$, $t = 0, 1$ are potential outcomes.

Sharp Design - Assumptions

Assumption C1 $1\{X \geq 0\} = Z$ almost surely.

- X correctly classifies the treatment assignment w. p. 1.
- E.g., C1 holds when $X = \lfloor X^* \rfloor$ e.g., month or year of birth, with an integer RDD threshold.
- More generally, C1 requires donut trimming, i.e., removing observations with $X \in [-s_1, s_0)$ when e has a bounded support $[-s_0, s_1]$.

Sharp Design - Assumptions

Assumption C2 $g_t(x) := E[Y(t)|X = x]$, $t = 0, 1$, are continuous at 0.

- C2 is commonly assumed in practice.
- C2 is weaker than the smoothness of $g_t^*(x) := E[Y(t)|X^* = x]$, when **a)** $F_{e|X}(e|x)$ is smooth in x; and **b)** $g_t^*(X^*, e) := E[Y(t)|X^*, e]$ is smooth in e; Note $g_t(x) = E[g_t^*(X^*, e)|X = x]$.
  - a) and b) hold in the case of rounding error, but fail for heaping error.

Sharp Design - Assumptions

Assumption C3 $E\left[Y\left(1\right)-Y\left(0\right)|X^{*}=x,X=0\right]=$
$E\left[Y\left(1\right)-Y\left(0\right)|X^{*}=x\right]$.

- C3 requires that once conditional on the true X^{*}, rounding error e does not affect the ATE at $X=0$.
- C3 holds when e is rounding error, or more generally whenever e is non-differential, i.e., $e\perp\left\{Y\left(0\right),Y\left(1\right)\right\}|X^{*}$, which is commonly assumed in the literature.

Sharp Design - Main Result

Let $\tau^*(x) := E[Y(1) - Y(0) \mid X^* = x]$.

Lemma (2.1)

Suppose that $T = Z$ and conditions C1 and C2 hold. Then the jump in the conditional mean function $g(x) := E[Y \mid X = x]$ at 0 identifies the ATE for units with $X = 0$,

$$\tau := E[Y(1) - Y(0) \mid X = 0] = \lim_{x \downarrow 0} g(x) - \lim_{x \uparrow 0} g(x). \tag{1}$$

If, in addition, condition C3 holds, then

$$\tau = \int \tau^*(e) dF_{e|X}(e \mid 0),$$

where $F_{e|X}(e \mid x)$ is the conditional distribution of $e = X^* - X$ given $X = x$.

Sharp Design - Comparison of Causal Parameters

- τ^*:=$E[Y(1) - Y(0) \mid X^* = 0]$ is the ATE for units with $X^* = 0$.
- τ:=$E[Y(1) - Y(0) \mid X = 0]$ is the ATE for units with $X = 0$, or a weighted average of the ATEs given X^* for units at the mismeasured cell at $X = 0$.

Example: X^* = birthdate, observe X = year of birth.

- τ^*: the ATE for individuals born on the cutoff date;
- τ: the ATE for individuals born in the cutoff year, or a weighted average of the ATEs for individuals born on each day within the cutoff year; the weights depend on the birthdate distribution in the year.

Sharp Design - Estimation and Inference

- Measurement error may lead to irregular support of X, e.g., discreteness, and loss of point identification.
- Use bias-aware or "honest" inference (Armstrong and Kolesár, 2018, 2020 and Kolesár and Rothe, 2018):
  - CI with level $1-\alpha$:

$$\hat{\tau} \pm cv_\alpha\left(B(\hat{\tau})/\hat{\sigma}(\hat{\tau})\right) \cdot \hat{\sigma}(\hat{\tau}),$$

    where $B(\hat{\tau})$ *is the bound on the conditional (on* X*) finite-sample bias of* $\hat{\tau}$, and $cv_\alpha(t)$ is the $1-\alpha$ quantile of a folded normal distribution $|N(t,1)|$.
  - $B(\hat{\tau})$ is obtained by assuming g has a second derivative bounded in absolute value by M.
  - Valid for any sequence of bandwidths (incl. fixed ones), partial identification of τ, and irregular support of X.

Fuzzy Design - Assumptions

$T(z)$, $z = 0, 1$, are the potential treatments. $\mathfrak{C}$ denotes the event $T(1) > T(0)$.

Assumption F1 $P(T(1) \geq T(0)|X = 0) = 1$ and $P(\mathfrak{C}|X = 0) > 0$

Assumption F2 $E[T(z)|X = x]$ and $E[Y(T(z))|X = x]$, $z = 0, 1$, are continuous at 0.

- F1 imposes monotonicity and non-zero discontinuity in the first stage.
- F2 imposes smoothness on the conditional means of potential treatments and *reduced-form* potential outcomes given X.

Fuzzy Design - Assumptions

Assumption F3 $E\left[Y\left(1\right)-Y\left(0\right)|\mathfrak{C},X^{*}=x,X=0\right]=$
$E\left[Y\left(1\right)-Y\left(0\right)|\mathfrak{C},X^{*}=x\right]$ and
$\Pr\left(\mathfrak{C}|X^{*}=x,X=0\right)=\Pr\left(\mathfrak{C}|X^{*}=x\right).$

- F3 requires that once conditional on the true X^{*}, e has no effects on the compliance probability and the ATE for compliers.
- F3 holds if e is rounding error or more generally e is differentiable, i.e., $e\perp\left\{Y\left(1\right),Y\left(0\right),T\left(1\right),T\left(0\right)\right\}|X^{*}$.

Fuzzy Design - Main Result

Let $p(x) = E[T|X = x]$.

Lemma (2.2)

Suppose that conditions C1, F1, and F2 hold. Then

$$\tau_F := E[Y(1) - Y(0) \mid \mathfrak{C}, X = 0] = \frac{\lim_{x \downarrow 0} g(x) - \lim_{x \uparrow 0} g(x)}{\lim_{x \downarrow 0} p(x) - \lim_{x \uparrow 0} p(x)}.$$

If, in addition, condition F3 holds, then $\tau_F = \int \tau_F^*(e)\omega(e)dF_{e|X}(e \mid 0)$, *where* $\tau_F^*(x) := E[Y(1) - Y(0) \mid \mathfrak{C}, X^* = x]$, *and*
$\omega(e) = \frac{P(\mathfrak{C}|X^*=e)}{\int P(\mathfrak{C}|X^*=e)dF_{e|X}(e|0)}$.

Estimation and Inference - Fuzzy Design

Use the bias-aware AR inference (Noack and Rothe,2022).

- Analogous to the AR confidence set in the IV literature.
- Can test the hypothesis $H_0 : \tau_F = \tau_{F,0}$ by checking whether 0 is in *the sharp RDD honest CI* using $Y - \tau_{F,0} T$ as the outcome. The confidence set for τ_F is constructed by collecting all values of $\tau_{F,0}$ that are not rejected.

Empirical Application

Estimate the impact of preregistration on youth turnout in an election (Holbein and Hillygus, 2016).

T = an indicator for preregistration or not in 2008.

Y = an indicator for voting in the 2012 election.

Z = whether one was born after Nov. 4, 1990.

X^* = proximity to the eligibility cutoff in days (based on birth date).

X = proximity to the eligibility cutoff in months (based on birth month).

Sample: individuals born within six months of the eligibility cutoff; 186,575 observations.

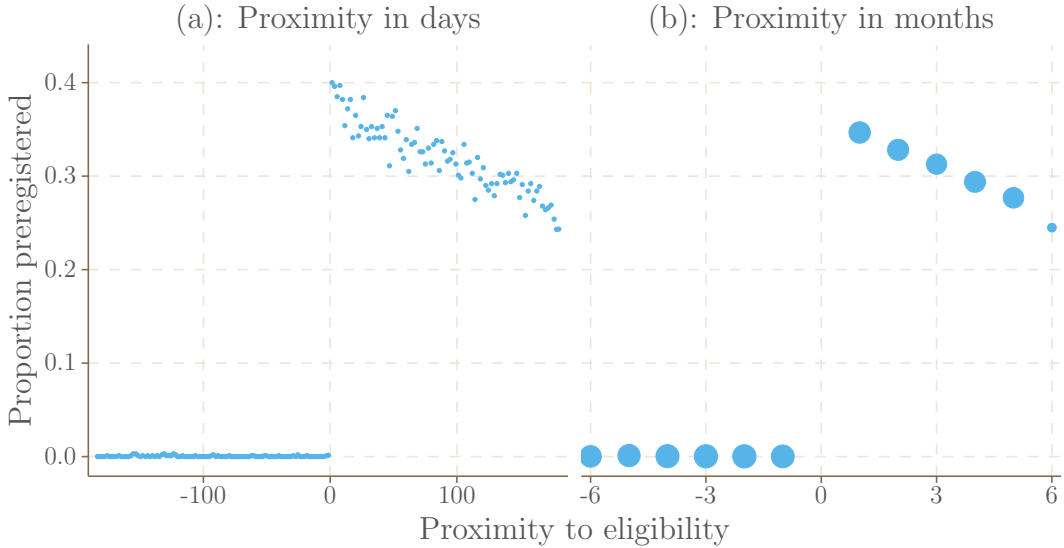

Figure 1: Effect of proximity on preregistering.

Notes: In panel (a), proximity is measured in days, and each point corresponds to an average of 1,000 individuals. In panel (b), proximity is measured in months, and each point corresponds to an average across all individuals born in a given month.

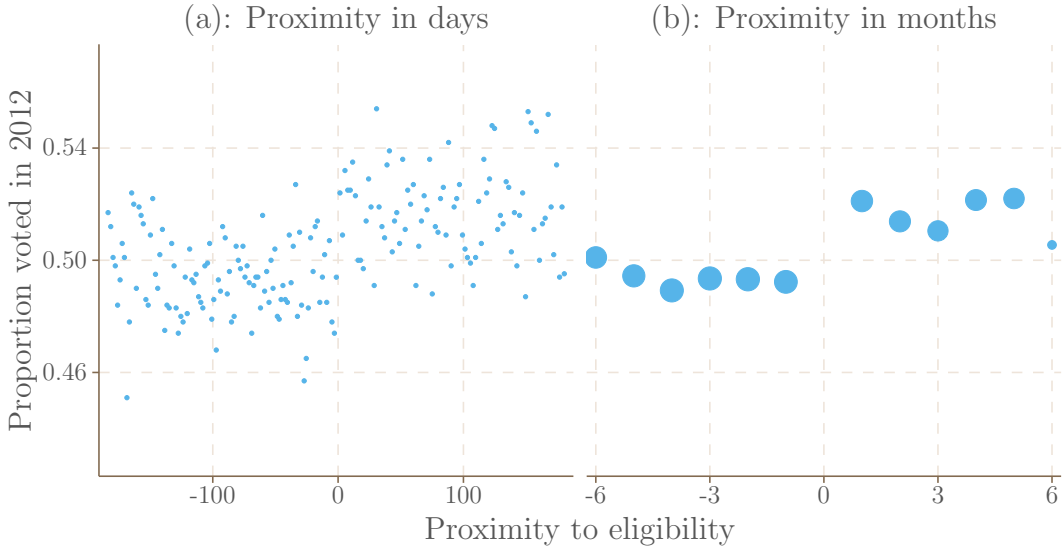

Figure 2: Effect of proximity on voting.

Notes: In panel (a), proximity is measured in days, and each point corresponds to an average of 1,000 individuals. In panel (b), proximity is measured in months, and each point corresponds to an average across all individuals born in a given month.

Table 1: First stage estimates: effect of eligibility on preregistration.

	OLS		RBC	Bias-aware inference	
	(1)	(2)	(3)	(4)	(5)
Panel A: Proximity in days					
Estimate	0.384	0.379	0.393	0.396	0.381
SE	0.006	0.005	0.008	0.009	0.005
95% CI	(0.373, 0.396)	(0.370, 0.388)	(0.378, 0.409)	(0.377, 0.414)	(0.371, 0.391)
Bandwidth	60	90	43	28	86
Eff. obs.	63,220	94,118	43,538	28,274	89,776
Rescaled M_t				1.015	0.061
Panel B: Proximity in months					
Estimate	0.365	0.363	0.368	0.365	0.365
SE	0.009	0.006	0.017	0.009	0.009
95% CI	(0.348, 0.382)	(0.351, 0.375)	(0.335, 0.402)	(0.303, 0.428)	(0.345, 0.385)
Bandwidth	2	3	3	2	2
Eff. obs.	64,011	94,662	94,662	64,011	64,011
Rescaled M_t				0.865	0.092

Table 3: Fuzzy RD estimates of the effect of preregistration on voting.

	OLS		RBC	Bias-aware inference	
	(1)	(2)	(3)	(4)	(5)
Panel A: Proximity in days					
Estimate	0.073	0.072	0.076	0.080	0.074
SE	0.021	0.018	0.032	0.031	0.018
95% CI	$(0.031, 0.114)$	$(0.037, 0.106)$	$(0.013, 0.139)$	$(0.012, 0.143)$	$(0.034, 0.122)$
Bandwidth	60	90	36	29	83
Eff. obs.	$63,220$	$94,118$	$35,785$	$29,285$	$86,881$
Rescaled M_y				1.401	0.099
Rescaled M_t				1.015	0.061
Panel B: Proximity in months					
Estimate	0.101	0.094	0.113	0.101	0.094
SE	0.035	0.024	0.068	0.035	0.024
95% CI	$(0.034, 0.169)$	$(0.047, 0.141)$	$(-0.020, 0.246)$	$(-0.268, 0.505)$	$(0.023, 0.180)$
Bandwidth	2	3	3	2	3
Eff. obs.	$64,011$	$94,662$	$94,662$	$64,011$	$94,662$
Rescaled M_y				1.818	0.121
Rescaled M_t				0.865	0.092

Empirical Application - Summary

1. RDD estimates using date vs. month of birth yield different estimates, reflecting the impact of the rounding error on the estimand.
    . The former captures the ATE of compliers born on the cutoff date, while the latter captures the ATE of compliers born in the cutoff month.
    . The jump in the preregistration probability is smaller when using birth month, capturing the lower preregistration probabilities among those born further away from the eligibility cutoff; this further translates into a slightly larger fuzzy RDD estimate.
2. Using month of birth generally leads to wider CIs.

Conclusion

- Measurement error - especially in the form of coarsening or rounding - is common in RDD applications.
- RDD with a mismeasured running variable can provide inference for a well-defined causal object.
- Care needs to be taken when interpreting the estimand.
- Inference methods need to account for the potentially irregular support of the running variable, and possible loss of point identification - we recommend bias-aware inference.