Online Appendix: Assessing the Impact of Non-Random Measurement Error on Inference: A Sensitivity Analysis Approach

November 3, 2016

Contents

A Consequences of Non-Random Measurement Error

A.1 Analytical Proof of Equation (2) of the Manuscript

Suppose we have n observations $Y = (y_1, \ldots, y_n)$, for which the following relationship is true:

$$Y = \beta_0 + \beta_1 T + \beta_2 C + \beta_3 W + \epsilon \tag{1}$$

where $T = (t_1, \ldots, t_n)$ is the treatment, $C = (c_1, \ldots, c_n)$ and $W = (w_1, \ldots, w_n)$ are two other variables, and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ is a random error term such that $\epsilon_i \sim N(0, \sigma^2)$. Denote the design matrix $\mathbf{Z_1} = [1, T, C, W]$. The OLS estimate is

$$\hat{\boldsymbol{\beta}} = (\mathbf{Z_1}^T \mathbf{Z_1})^{-1} \mathbf{Z_1}^T Y \tag{2}$$

which is an unbiased estimate of $\boldsymbol{\beta}$. However, instead of T we observe $X = f(T|C)$. If we denote $\mathbf{Z_2} = [1, X, C, W]$, the OLS estimate then is

$$\hat{\boldsymbol{\delta}} = (\mathbf{Z_2}^T \mathbf{Z_2})^{-1} \mathbf{Z_2}^T Y \tag{3}$$

Is $\hat{\boldsymbol{\delta}}$ an unbiased estimate of $\boldsymbol{\beta}$?

$$\begin{aligned}
\mathbb{E}(\hat{\boldsymbol{\delta}}) &= \mathbb{E}((\mathbf{Z_2}^T \mathbf{Z_2})^{-1} \mathbf{Z_2}^T Y) && (4) \\
&= (\mathbf{Z_2}^T \mathbf{Z_2})^{-1} \mathbf{Z_2}^T \mathbb{E}(Y) && (5) \\
&= (\mathbf{Z_2}^T \mathbf{Z_2})^{-1} \mathbf{Z_2}^T \mathbf{Z_1} \boldsymbol{\beta} && (6)
\end{aligned}$$

We can define a matrix $\mathbf{R} = \mathbf{Z_1} - \mathbf{Z_2}$:

$$\mathbf{R} = \begin{bmatrix} 0 & t_1 - x_1 & 0 & 0 \\ 0 & t_2 - x_2 & 0 & 0 \\ & \vdots & & \\ 0 & t_n - x_n & 0 & 0 \end{bmatrix} \tag{7}$$

We thus have:

$$\mathbb{E}(\hat{\boldsymbol{\delta}}) = (\mathbf{Z_2}^T\mathbf{Z_2})^{-1}\mathbf{Z_2}^T\mathbf{Z_1}\boldsymbol{\beta} \tag{8}$$

$$= (\mathbf{Z_2}^T\mathbf{Z_2})^{-1}\mathbf{Z_2}^T(\mathbf{R} + \mathbf{Z_2})\boldsymbol{\beta} \tag{9}$$

$$= \boldsymbol{\beta} + (\mathbf{Z_2}^T\mathbf{Z_2})^{-1}\mathbf{Z_2}^T\mathbf{R}\boldsymbol{\beta} \tag{10}$$

The last line makes it clear that the amount of bias in $\hat{\boldsymbol{\delta}}$ is $(\mathbf{Z_2}^T\mathbf{Z_2})^{-1}\mathbf{Z_2}^T\mathbf{R}\boldsymbol{\beta}$. To evaluate what this means for the coefficients $\hat{\delta}_0$, $\hat{\delta}_1$, $\hat{\delta}_2$, and $\hat{\delta}_3$, define the 4×4 matrix $\mathbf{D} \equiv (\mathbf{Z_2}^T\mathbf{Z_2})^{-1}$. Multiplying out $\mathbf{D}\mathbf{Z_2}^T\mathbf{R}\boldsymbol{\beta}$ results in the 4×1 vector shown in Equation (2) of the manuscript, showing the amount of bias in $\hat{\delta}_0$, $\hat{\delta}_1$, $\hat{\delta}_2$, and $\hat{\delta}_3$.

A.2 Additional Results Monte Carlo Analysis

In Figure 1 of the manuscript, we simulate the bias in $\hat{\delta}_1$, $\hat{\delta}_2$, and $\hat{\delta}_3$, for a specific situation. In particular, we set $c_i \sim \mathcal{B}(0.2)$ and $\eta = 2$. Here, we vary the proportion of observations affected by measurement error as well as the divergence between true and observed values.

While we show above that systematic measurement error will lead to bias in the coefficient estimates for all variables, regardless of the type of error, knowing the extent of the error depends on the matrix $\mathbf{D} \equiv (\mathbf{Z_2}^T\mathbf{Z_2})^{-1}$, which itself depends on the values of each variable. Thus, to show the effect of measurement error on bias, we conduct a simple Monte Carlo analysis using different levels and types of measurement error. In particular, we set:

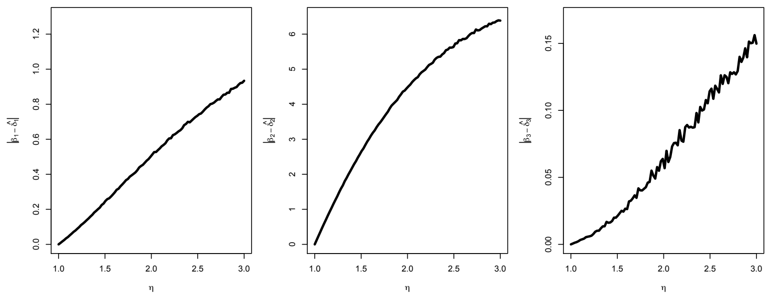
$$Y_i = 1 + 2t_i - 3c_i + 1w_i + \epsilon_i \tag{11}$$

where we simulate $t_i \sim \mathcal{N}(3,2)$, $c_i \sim \mathcal{B}(p)$, $w_i \sim \mathcal{N}(1,1)$, and $\epsilon_i \sim \mathcal{N}(0,1)$, with $p \in \{0.2, 0.4, 0.6, 0.8\}$. Again, we allow for a correlation of 0.3 between T and W.[1] Rather than observing t_i we observe x_i:

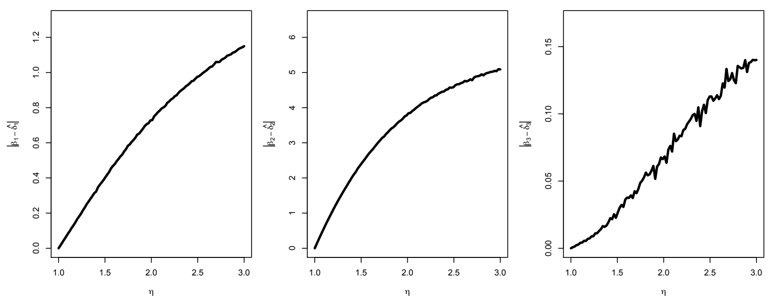
$$x_i = \begin{cases} t_i, & \text{if } c_i = 0 \\ \eta t_i, & \text{if } c_i = 1 \end{cases} \tag{12}$$

with $\eta \in [1,3]$. For each value of p and η we simulate 100 datasets with 1000 observations each, and compare the coefficient estimated with t_i to those estimated by x_i. The results are depicted in Figure 1.

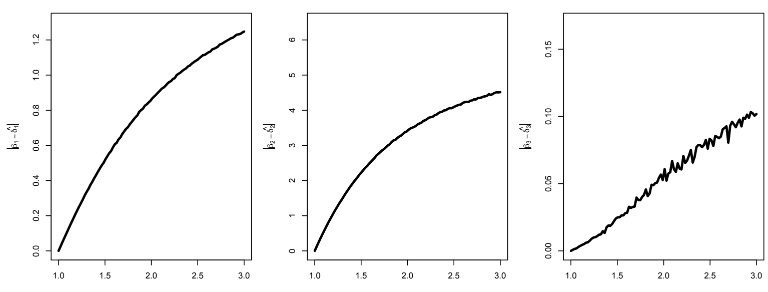
[1] If we set this correlation to zero, i.e. T and W are independent, then $\hat{\delta}_3$ will be an unbiased estimate of β_3 no matter how much measurement error there is in T.



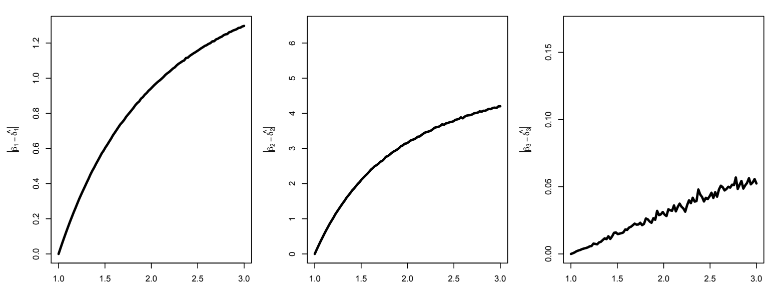
(a) $P(c_i = 1) = 0.2$



(b) $P(c_i = 1) = 0.4$



(c) $P(c_i = 1) = 0.6$



(d) $P(c_i = 1) = 0.8$

Figure 1: The effect of systematic measurement error on bias for coefficients $\beta_0, \beta_1, \beta_2$. The true data generating process is given by $y_i = 1 + 2x_i - 3c_i + 1w_i + \epsilon$. The four rows show the results when $c_i = 1$ in 20%, 40%, 60%, and 80% of cases. Bias is averaged over 100 simulated datasets, each with 1000 observations.

A.3 Derivation of Bias if Measurement Error in the Dependent Variable

Suppose again that we have n observations $Y = (y_1, \ldots, y_n)$, for which the following relationship is true:

$$Y = \beta_0 + \beta_1 T + \beta_2 C + \beta_3 W \epsilon \tag{13}$$

where $T = (t_1, \ldots, t_n)$ is the treatment, $C = (c_1, \ldots, c_n)$ and $W = (w_1, \ldots, w_n)$ are controls, and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ is a random error term such that $\epsilon_i \sim N(0, \sigma^2)$. Denote the design matrix $\mathbf{Z} = [1, T, C, W]$. The OLS estimate is $\hat{\boldsymbol{\delta}} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T Y$.

Instead of Y we observe $X = f(Y|C)$. This means that the OLS estimate is

$$\hat{\boldsymbol{\delta}} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T X \tag{14}$$

We define a vector $R = Y - X$ and write

$$\mathbb{E}(\hat{\boldsymbol{\delta}}) = \mathbb{E}((\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T X) \tag{15}$$

$$= \mathbb{E}((\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T (Y - R)) \tag{16}$$

$$= \beta - \mathbb{E}((\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T R) \tag{17}$$

The bias if the measurement error is in the dependent variable is therefore $\mathbb{E}((\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T R)$.

B Guide to Sensitivity Analysis for Non-Random Measurement Error

In this section, we provide a detailed step-by-step guide to conduct a sensitivity analysis for non-random measurement error. The first step is to quantify the measurement error. This is arguably the most difficult part of the process, and how to specify the relation between the corrupted variable X and the simulated true variable T' depends on whether the corrupting variable C and X are binary, categorical, or continuous. In Section B.1, we provide guidance for all nine possible combinations. The second step is to simulate what T' looks like given different levels of measurement error. In Section B.2, we show how to summarize and display the observed and simulated data. Finally, the third step is to re-estimate the original model using T' instead of X. In Section B.3, we provide several ways to summarize how inference changes given different levels of non-random measurement error.

B.1 Step 1: Quantifying Measurement Error

How to model the relation between the corrupted variable X and the simulated true variable T' depends on the measurement level of X as well as of the corrupting variable C. Table 1 shows the nine possible combinations and directs to the subsection discussing each.

Table 1: Overview for Step 1: Sections that describe how to quantify the measurement error in detail.

		X		
		binary	categorical	continuous
	binary	B.1.1.1	B.1.1.2	B.1.1.3
C	categorical	B.1.2.1	B.1.2.2	B.1.2.3
	continuous	B.1.3.1	B.1.3.2	B.1.3.3

B.1.1 Binary C

B.1.1.1 Binary C, Binary X

If both the corrupting variable C and the corrupted variable X are binary we need to specify a matrix of transition probabilities, as displayed in Table 2. For each of the four possible combinations of C and X, a transition probability is specified. For combinations where no measurement error is suspected, these transition probabilities are zero. For example, if we only suspect measurement error when $C = 1$, then $P_0(0 \to 1) = P_0(1 \to 0) = 0$. The two transition probabilities for $C = 1$ depend on the direction of the measurement error. If we suspect that there is systematic underreporting, then $P_1(1 \to 0) = 0$ and $P_1(0 \to 1) > 0$. That is, all cases where $X = 1$ was reported are correctly measured, whereas when we observe $X = 0$, the true value is one with probability $P_1(0 \to 1)$. If we suspect systematic overreporting, then $P_1(1 \to 0) > 0$ and $P_1(0 \to 1) = 0$.

Table 2: Transition probabilities for binary C and X.

		X = 0	X = 1
C	0	$P_0(0 \to 1)$	$P_0(1 \to 0)$
	1	$P_1(0 \to 1)$	$P_1(1 \to 0)$

Consider an example where we suspect systematic underreporting when $C = 1$. In this case, $P_0(0 \to 1) = P_0(1 \to 0) = P_1(1 \to 0) = 0$, so $t'_i = x_i$ in all of these cases. But if $c_i = 1$ and $x_i = 0$, $t'_i = 1$ with probability $P_1(0 \to 1) = \eta$. If η is close to zero the underreporting is small. If $\eta = 1$, then every single observation where $c_i = 1$ and $x_i = 0$ is misreported. For details on the simulation of T', see Section B.2.1.

B.1.1.2 Binary C, Categorical X

The extension from a binary corrupted variable X to a categorical one is straightforward. Instead of specifying only the transition probabilities from 0 to 1 (or vice versa), the researcher now needs to specify transition probabilities to all categories. Table 3 below illustrates the simplest case where X has three categories. In most circumstances many of these transition probabilities will be set to zero.

Table 3: Transition probabilities for binary C and categorical X (3 categories).

C		$X = 0$	$X = 1$	$X = 2$
	0	$P_0(0 \to 1)$	$P_0(1 \to 0)$	$P_0(2 \to 0)$
		$P_0(0 \to 2)$	$P_0(1 \to 2)$	$P_0(2 \to 1)$
	1	$P_1(0 \to 1)$	$P_1(1 \to 0)$	$P_1(2 \to 0)$
		$P_1(0 \to 2)$	$P_1(1 \to 2)$	$P_1(2 \to 1)$

B.1.1.3 Binary C, Continuous X

If the corrupted variable is continuous, we need to specify a transition function instead of a transition probability (see Table 4). Suppose again that we only suspect measurement error for $C = 1$. It follows that $f_0(x_i) = x_i$. For the set of observations where measurement error is suspected, we need to map the observed values X to the "true" values T' using a linear function.

Table 4: Transition functions for binary C and continuous X.

C		
	0	$f_0(X)$
	1	$f_1(X)$

Specifying this function is somewhat more complicated than when X is binary or categorical. One of the simplest ways to do so is $f_1(x_i) = \eta\ x_i$. This means that the simulated true value is η times the value of x_i. This means that the simulated true value is η times the value of x_i. For example, in our first empirical application, we model true settler mortality derived from campaign rates this way, motivated by the statement that they were between 66 and 2000 percent higher than barracks rates. If necessary, the values can be rounded to the closest integer, e.g. if the measurement error is in the dependent variable and it is a count variable. No measurement error is $\eta = 1$. If $\eta < 1$ then we suspect that the true values are lower than the reported one's, and if $\eta > 1$ we suspect that they are higher.

The functional form of the transition functions can be more complex. It might be useful to specify a step function where there is no measurement error for some values of X, but above a certain threshold measurement error sets in. One could also use higher-order polynomials, logarithms, or even condition the relation between X and T' on another variable. Such more complex specifications, however, are only recommended when good information about the form of the error is available, in particular because the results of such specifications will be harder to communicate. In many circumstances, a simpler approach is sufficient to analyze the robustness of a finding to non-random measurement error.

B.1.2 Categorical C

The extension from a binary corrupting variable C to a categorical one with K categories is straightforward. For all tables in the previous section, we need to simply have K rows instead of only two (see the following tables). Everything else is equivalent.

B.1.2.1 Categorical C, Binary X

Table 5: Transition probabilities for categorical C (3 categories) and binary X.

		X	
		0	1
	0	$P_0(0 \to 1)$	$P_0(1 \to 0)$
C	1	$P_1(0 \to 1)$	$P_1(1 \to 0)$
	2	$P_2(0 \to 1)$	$P_2(1 \to 0)$

B.1.2.2 Categorical C, Categorical X

Table 6: Transition probabilities for categorical C (3 categories) and categorical X (3 categories).

		X = 0	X = 1	X = 2
C	0	$P_0(0 \to 1)$ $P_0(0 \to 2)$	$P_0(1 \to 0)$ $P_0(1 \to 2)$	$P_0(2 \to 0)$ $P_0(2 \to 1)$
	1	$P_1(0 \to 1)$ $P_1(0 \to 2)$	$P_1(1 \to 0)$ $P_1(1 \to 2)$	$P_1(2 \to 0)$ $P_1(2 \to 1)$
	2	$P_2(0 \to 1)$ $P_2(0 \to 2)$	$P_2(1 \to 0)$ $P_2(1 \to 2)$	$P_2(2 \to 0)$ $P_2(2 \to 1)$

B.1.2.3 Categorical C, Continuous X

Table 7: Transition functions for categorical C (3 categories) and continuous X.

C	0	$f_0(X)$
	1	$f_1(X)$
	2	$f_2(X)$

B.1.3 Continuous C

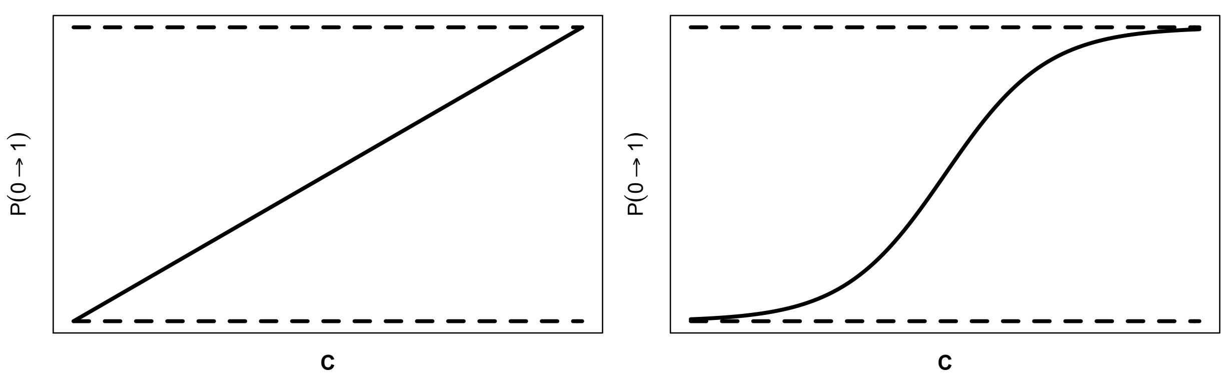
B.1.3.1 Continuous C, Binary X

There are many occasions in the social science where the corrupting variable C is continuous. For example, measurement error may decrease (or increase) in GDP per capita, the rate of urbanization, or time. If the corrupted variable X is binary, we again have to specify transition probabilities (see Table 8). But instead of setting different probabilities for the different values of C, we now specify the transition probability as a continuous function of C.

Table 8: Transition probabilities for continuous C and binary X.

	X	
	0	1
C	$P_1(0 \to 1 \mid C)$	$P_1(1 \to 0 \mid C)$

Figure 2 plots two possibilities for the functional form. The first one is a linear function, in which the probability of misclassification increases in C. This simple form can be augmented by making it a step function. For example, we might suspect that all observations below a certain value of C have zero probability of being misclassified, above a certain value they are misclassified for sure, and in between the probability increases linearly from zero to one. Of course, we can also impose floors from above and/or below, for example if we suspect the maximal probability of misclassification to be smaller than one. Alternatively, the second panel shows the logistic function, which is naturally bounded and therefore does not require the specification of complicated boundaries. The researcher has to specify $\bar{p}$, the maximum transition probability, the steepness of the curve η_n, and the x-value of the curve's midpoint c_0.



(a) Linear Function: $p_i = \eta_l \; c_i$ (b) Logistic Function: $p_i = \frac{\bar{p}}{1+\exp(-\eta_n \; (c_i - c_0))}$

Figure 2: Two examples for transition probabilities when C is continuous. The horizontal axis shows the values of a corrupting variable C, and the vertical axis the transition probability.

B.1.3.2 Continuous C, Categorical X

If C is continuous and X is categorical, the approach in the previous section can be extended in a straightforward manner.

Table 9: Transition probabilities for continuous C and categorical X (3 categories).

	X = 0	X = 1	X = 2
C	$P_1(0 \to 1 \vert C)$	$P_1(1 \to 0 \vert C)$	$P_1(2 \to 0 \vert C)$
	$P_1(0 \to 2 \vert C)$	$P_1(1 \to 2 \vert C)$	$P_1(2 \to 1 \vert C)$

B.1.3.3 Continuous C, Continuous X

Finally, the most complex case is the one where both the corrupting variable C and the corrupted variable X are continuous. While only one function needs to be specified (see Table 10), it needs to relate both C and X to T' at the same time.

Table 10: Transition function for continuous C and continuous X.

C	$f_1(X \vert C)$

Figure 3 plots the following example:

$$f_1(x_i) = \eta_1 \, x_i + \eta_2 \, c_i + \eta_3 \, x_i \, c_i \tag{18}$$

This looks like an interaction effect in a regression setting, and the interpretation is analogous. The first parameter η_1 gives the slope that quantifies the measurement error in x_i when $c_i = 0$. If $\eta_1 = 1$ there is no measurement error for $c_i = 0$. The second parameter η_2 gives the slope that quantifies the measurement error in x_i as a function of c_i when $x_i = 0$. Finally, η_3 quantifies the measurement error as both x_i and c_i increase. Again, it is possible to specify more complex funciton containing non-linearities or higher-order interactions. However, those will be harder to communicate and are thus only recommended if the researcher has good reason to believe that a simple specification is unable to capture the measurement error.

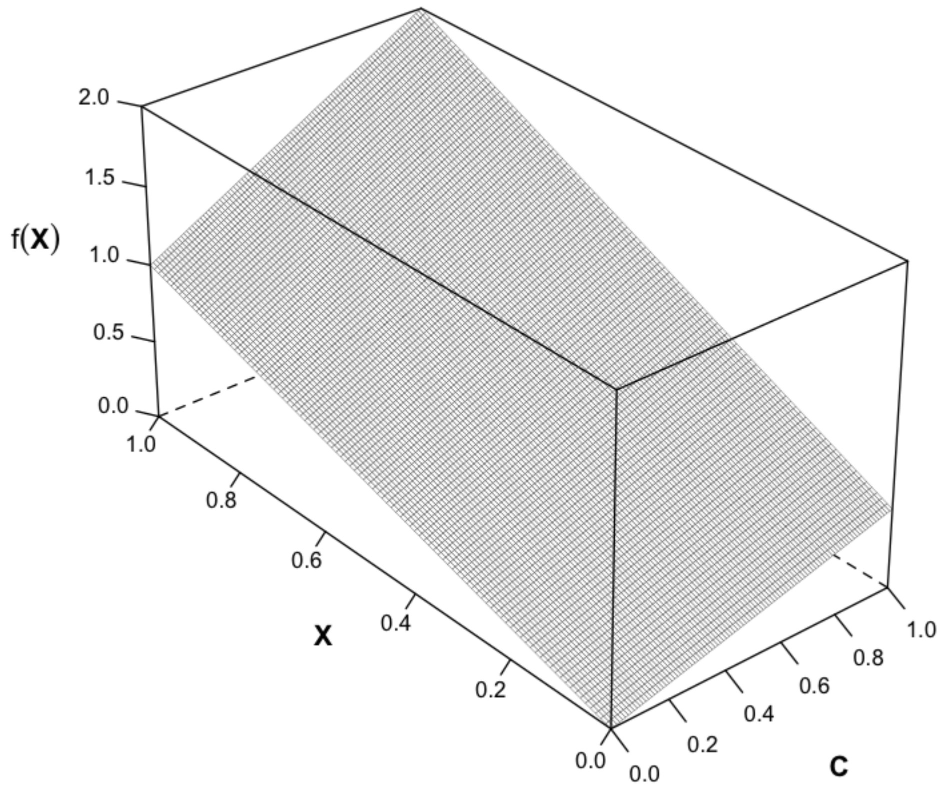


Figure 3: Examples for transition function when both C and X are continuous: $f_1(x_i) = \eta_1\ x_i + \eta_2\ c_i + \eta_3\ x_i\ c_i$

B.2 Step 2: Simulation of Data

B.2.1 Binary X

Based on Step 1, we now have a set of transition probabilities that depend on X, C, and η. First, we take m equally spaced values from $\eta \in [\underline{\eta}, \overline{\eta}]$ (see the main manuscript for a discussion of how to chose $\underline{\eta}$ and $\overline{\eta}$). For each η, we then simulate $T'(\eta)$. For binary X, it is necessary to simulate $T'(\eta)$ s times, wherein each observation has a probability of changing from 0 to 1 or vice versa as determined by a Bernoulli distribution with parameter η. The goal is to have a large number of simulated datasets for each value of η, and then take the median of those s simulations to net out the effect of the stochastic draws. The sensitivity analysis will be more accurate if we perform a larger number of draws for each value.

Once the variables are simulated, it is useful to visualize them to gain a firm understanding of what $T'(\eta)$ looks for different values of η. Figure 4 provides an example. It plots the probability that $T'(\eta)$ is equal to one against η for the subset of the data that is thought to be affected by measurement error (say $C = 1$). In this example, slightly more than 60 percent of observations for which $c_i = 1$ have a value of one in the observed data ($\eta = 0$). Because we suspect underreporting in this case, $P(T'(\eta))$ increases in η. When $\eta = 1$, all observations for which $c_i = 1$ take on the value of one.

B.2.2 Categorical X

Simulating new data with a categorical X is a straightforward extension of the binary case.

B.2.3 Continuous X

Dealing with a continuous X is comparatively simple. Based on the work in Step 1, we have a transition function depending on X, C, and η which yields a set of variables $T'(\eta)$ for each of m equally spaced values from $\eta \in [\underline{\eta}, \overline{\eta}]$ (again, see the main manuscript for a discussion

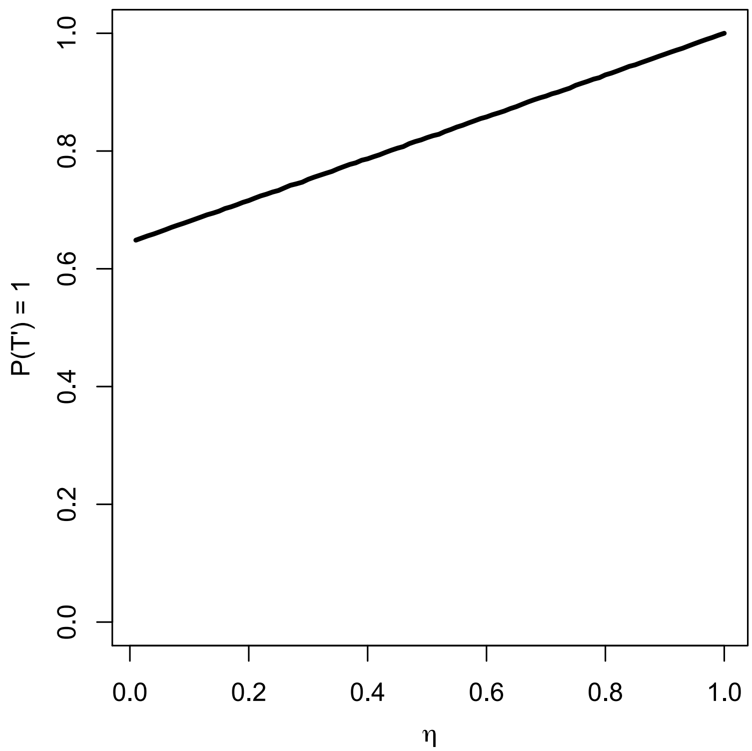


Figure 4: Probability of $T'(\eta) = 1$ for the subset of data potentially affected by measurement error.

of how to select $\underline{\eta}$ and $\overline{\eta}$). Because this step lacks a stochastic component, there is only one variable $T'(\eta)$ for each value of η.

To visualize the data, we suggest plotting X against $T^{'}$ for different values of η to assess how much they diverge. It is also useful to compare histograms of $T^{'}$ at various values of η to determine how the distribution of the data changes. For an example of both of these, see Figure 1 in the main manuscript as well as Figure 5 of this Online Appendix.

B.3 Step 3: Model Re-Estimation with Simulated Data

The final step is to re-estimate the original model using the simulated $T'(\eta)$ instead of the observed X for each of the m values of $\eta \in [\underline{\eta}, \overline{\eta}]$. If $T'(\eta)$ is binary, the median of the s simulations for each step should be taken. This provides the inferential quantity of interest, conditional on a given level of non-random measurement error, which can then be compared to the effect assuming no measurement error.

We recommend to always show a graphical representation of how the effect of interest (on the vertical axis) changes for different values of η (on the horizontal axis). For examples of such a plot, see Figures 3 and 5 of the main manuscript as well as Figures 6, 12, 15, 18, and 20 in this Online Appendix. The effect of interest can be a regression coefficient, a marginal effect, the ATE or ATT, or any other quantity. If η is easily interpretable (as in our examples on settler mortality and democratic deliberation), it can be directly plotted on the x-axis. If there are more intuitive ways to communicate the degree of measurement error, they can be placed on the x-axis instead. For example, in our application on the effect of cellphone coverage and political violence (see below), we express measurement error as the percent of additional cells without cellphone coverage that have a violent event. Not only is this more intuitive, it also allows us to compare the results of the sensitivity analysis to the amount of measurement error detected by Weidmann (2016) in Afghanistan.

If appropriate, the results of the sensitivity analysis can additionally be presented in numerical form. It is straightforward to calculate η^*, the value of η at which the effect can no longer be supported by the data (e.g. before it is no longer statistically different from zero at a given confidence level). Again, it is possible to transform η^* into more easily interpretable quantities if appropriate. One way to compare the sensitivity to measurement error across different data sources is to express them in terms of the difference between the observed variable X and $T'(\eta^*)$, the simulated variable at which the effect of interest does not find support anymore. To do so, the researcher can report the value k^* which solves the following equation:

$$\mathbb{E}(T'(\eta^*)) = \mathbb{E}(X) + k^* \, \sigma(X) \tag{19}$$

where σ stands for the standard deviation. The value of k^* tells us how many standard deviations larger or smaller the true variable needs to be compared to the observed variable before the effect of interest can no longer be supported by the data.

C Settler Mortality and Comparative Development: Additional Material

In the manuscript, we focus on potential non-random measurement error in the mortality data due to the assignment of rates from soldiers in campaign and in barracks. In addition, Albouy (2012) argues that other data points may also suffer from measurement error. In particular, data for Latin American countries are taken from mortality rates of bishops. These are benchmarked to the mortality rate in Mexico (which is taken from French soldiers in campaign) by multiplying them by 4.25. Again, this potentially induces systematic measurement error, as “there are many other alternatives to benchmark the data” (Albouy, 2012, 3063). In this section, we conduct a sensitivity analysis focusing on this source of measurement error. As one of the extensions in the manuscript, we then look at the two potential sources of measurement error together.

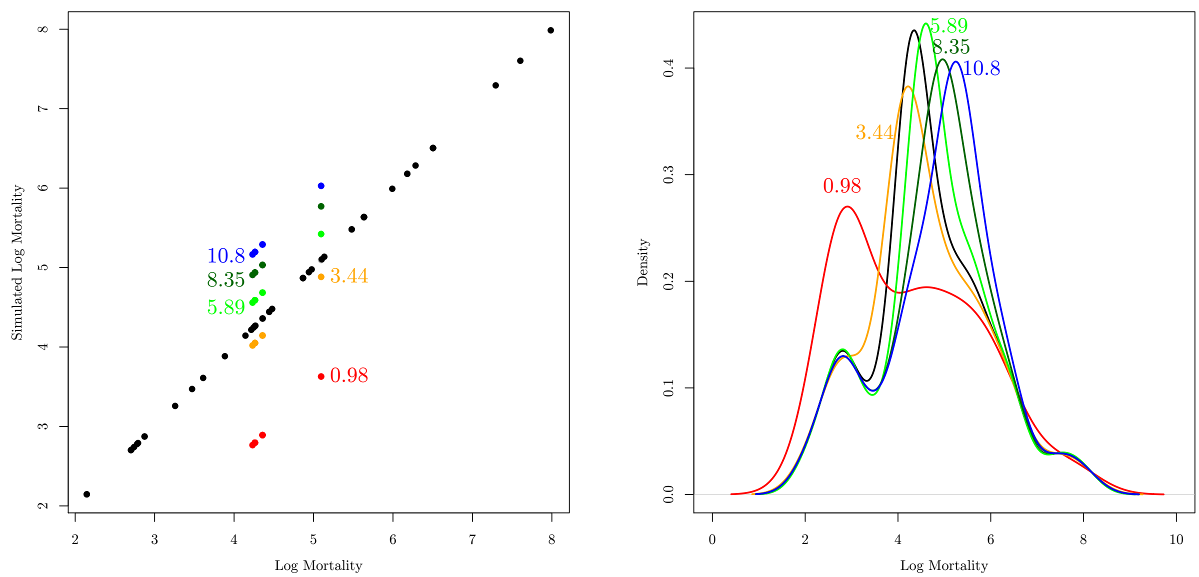
Step 1 – Quantifying Measurement Error

This section concerns the procedure AJR use to estimate settler mortality in much of Latin America. In the absence of soldier mortality rates, AJR use rates for bishops. They benchmark them using Mexico, a country where data on both soldiers and bishops are available. There, bishop mortality is 16.7 and soldier mortality is 71 out of a thousand, so 4.25 times as much. This factor is applied to the other countries for which bishop data is used. Albouy notes that this leads to mortality rates that are much higher than those in the United States, despite the fact that many countries in Latin America have similar disease environments. Given the relatively poor economic outcomes in much of Latin America, such potential non-random measurement error may have severe consequences for inference about the institution-growth link.

Using other available rates, Albouy states that “the ratio of actual soldier to bishop mortality rates varies from 0.98 to 10.80” (Albouy, 2012, 3064). Again, this information about the direction and magnitude of the non-random measurement error is not used any further in his analysis.[2] Let x_i be the observed mortality rate of country i, and let $d_i = 1$ if x_i the mortality rate comes from bishops. Table 11 specifies the transition probabilities.

Table 11: Transition functions for settler mortality rates X depending on the source of the data D.

D	X
0	$f_0(x_i) = x_i$
1	$f_1(x_i) = \frac{1}{4.25}\,\eta_2\,x_i$



(a) (b)

Figure 5: Effect of η_2 on log mortality rates. Panel (a): Original log mortality data on the horizontal axis and log mortality given a certain value of η_2 on the vertical axis. Black dots: Cases do not use benchmarked data. Gray dots: Cases do use benchmarked data. Colored dots: What cases using benchmarked data would be for given η_2. Panel (b): Densities of original variable and those implied different levels of η_2.

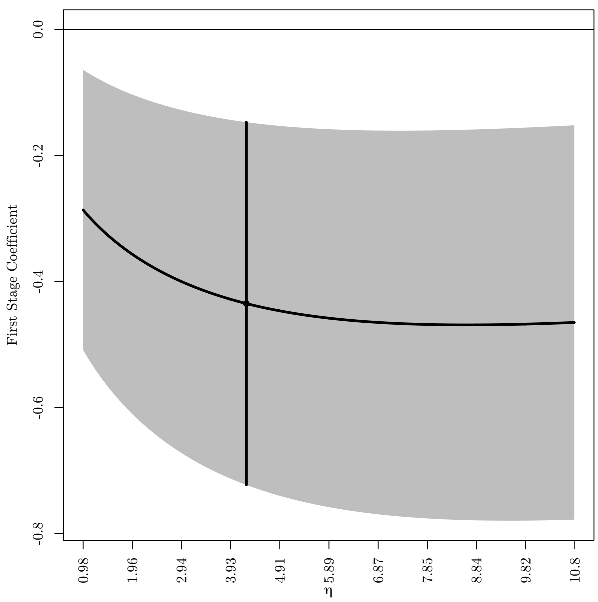
[2] Albouy also notes that the relationship between bishop and soldier mortality rates varies over time. Since there is not sufficient information on the direction of this effect, we do not incorporate it into our analysis.

Step 2 – Simulation of Data

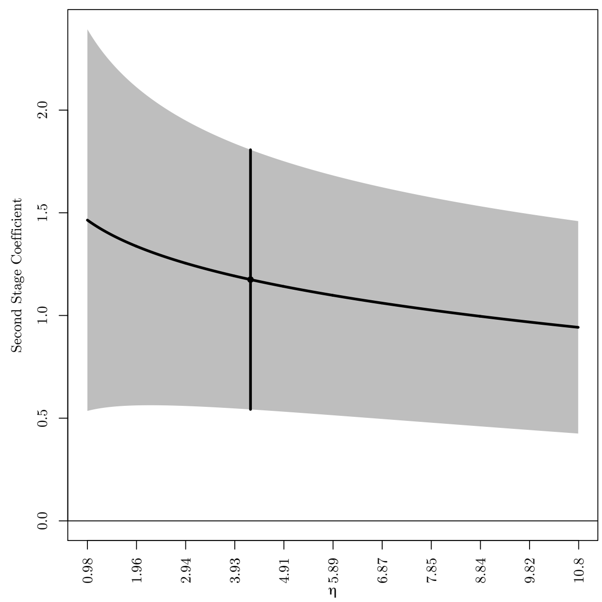
Given Table 11, we simulate what the data would look like with η_2 varying from 0.98 to 10.8. Figure 5 plots how this affects the data. There are two differences compared to the section addressing rates from soldiers. First, fewer cases are affected. Second, the measurement error can now go in both directions, so AJR's benchmarking process may have overestimated or underestimated the true rates.

Step 3 – Model Re-Estimation with Simulated Data

Figure 6 shows the results of the re-estimated model using the simulated data. Panel (a) shows the point estimates and 95 percent confidence intervals of the first stage coefficient of log settler mortality (given η_2) on institutional quality. With a benchmarking factor of less than 4.25, the instrument is slightly weaker, but the confidence interval never includes zero. With a benchmarking factor of higher than 4.25, the first stage coefficients actually increases in magnitude. Panel (b) gives the second stage coefficient of instrumented institutions. The variance is slightly increased for lower benchmarking factors, but again the coefficient is statistically different from zero over the entire range. The major takeaway is that in isolation, accounting for measurement error due to choosing the benchmark for bishop rates within a plausible range derived from other sources does not seriously impact AJR's findings.



(a) First Stage



(b) Second Stage

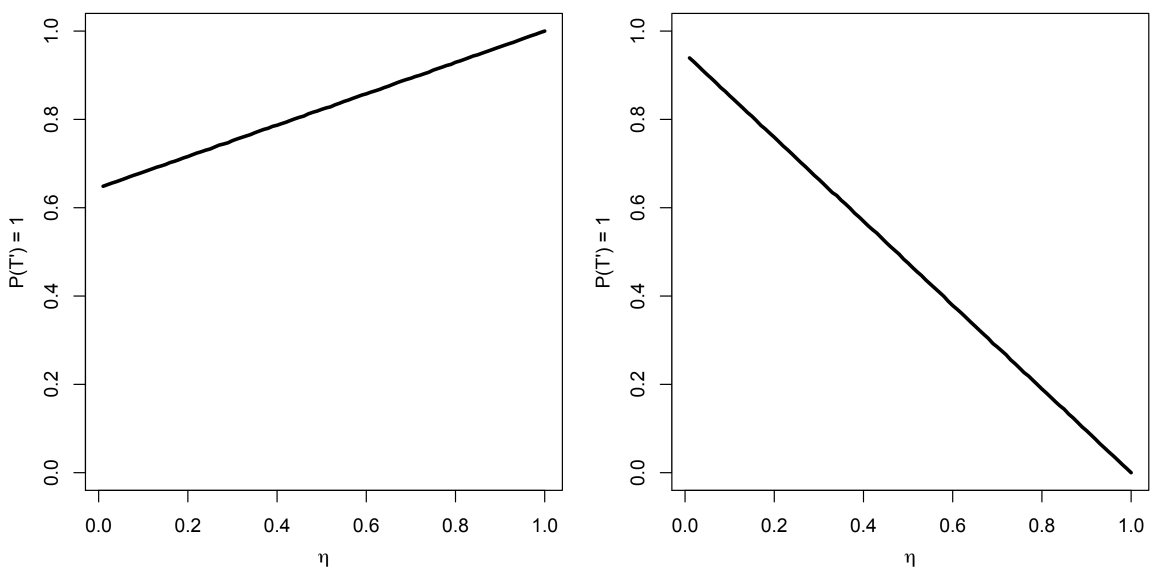
Figure 6: Effect of η_2 on inference. Panel (a): First stage coefficient of log settler mortality (given η_2) on institutions. Panel (b): Second stage coefficient of institutions (instrumented by log settler mortality) on economic development. Point estimates and 95 percent confidence intervals. The black vertical lines show the coefficients assuming AJR's benchmarking factor of 4.25.

D Leadership and Democratic Deliberation: Additional Material

D.1 Additional Analyses for Clinics or Hospitals

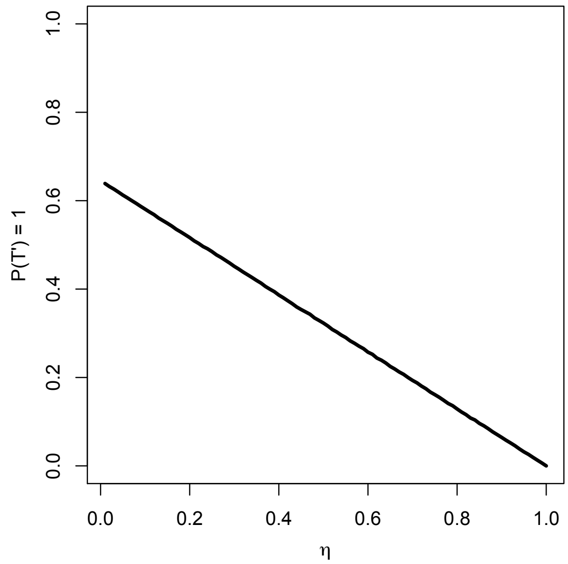
In the article, we apply the sensitivity analysis to one of the questions examined in Humphreys, Masters and Sandbu (2006), namely whether money should be invested in clinics ($y_i = 0$) or hospitals ($y_i = 1$). Here, we first describe the simulated data for different levels of η for Scenarios (a) and (b), as described in the main manuscript. Then, we describe results from sensitivity analyses in which the motives from the two scenarios are simultaneously present.

D.1.1 Description of the Simulated Data

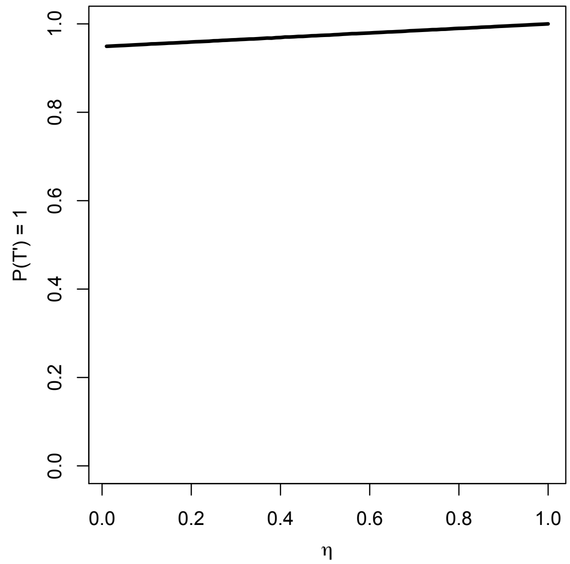


(a) $Y = 0$ (b) $Y = 1$

Figure 7: Probability of $T'(\eta) = 1$ for Scenario (a) from Table 3 of article for Question 3 (clinics or hospitals).



(a) $Y = 0$



(b) $Y = 1$

Figure 8: Probability of $T'(\eta) = 1$ for Scenario (b) from Table 3 of article for Question 3 (clinics or hospitals).

D.1.2 Sensitivity Analysis with Both Motives

In the article, we conduct the sensitivity analysis for two extreme scenarios. In Scenario (a), there was no measurement error when the group outcome and the stated leader preference diverge. When they are the same, the stated answer may not accurately reflect the pre-treatment preference. In Scenario (b), there is no measurement error when the group outcome and the stated leader preference are the same, but there is measurement error when they diverge. In this extension, we examine what happens if both motives are present simultaneously, as displayed in Table 12.

Table 12: Transition probabilities for stated leader preference X and group discussion outcome Y.

		X = 0	X = 1
Y	0	$P_0(0 \to 1) = \eta_1$	$P_0(1 \to 0) = \eta_2$
	1	$P_1(0 \to 1) = \eta_2$	$P_1(1 \to 0) = \eta_1$

Figure 9 plots the p-values of the ATE depending on η_1 and η_2. $\eta_2 = 0$ corresponds to scenario (a) in the manuscript, and $\eta_1 = 0$ corresponds to scenario (b). In the upper left of the plot, very few leaders falsely report their policy preference as similar to the preference of the group. Here, the effect of the discussion leader on the outcome is positive. In the lower right, few leaders falsely report that they had the opposite of the group's preference. Here, the ATE is negative.

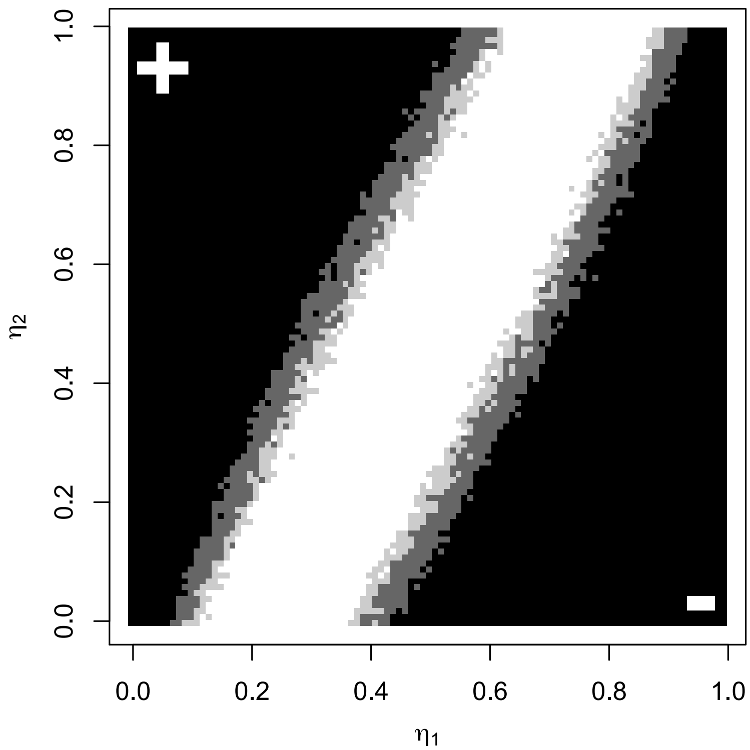
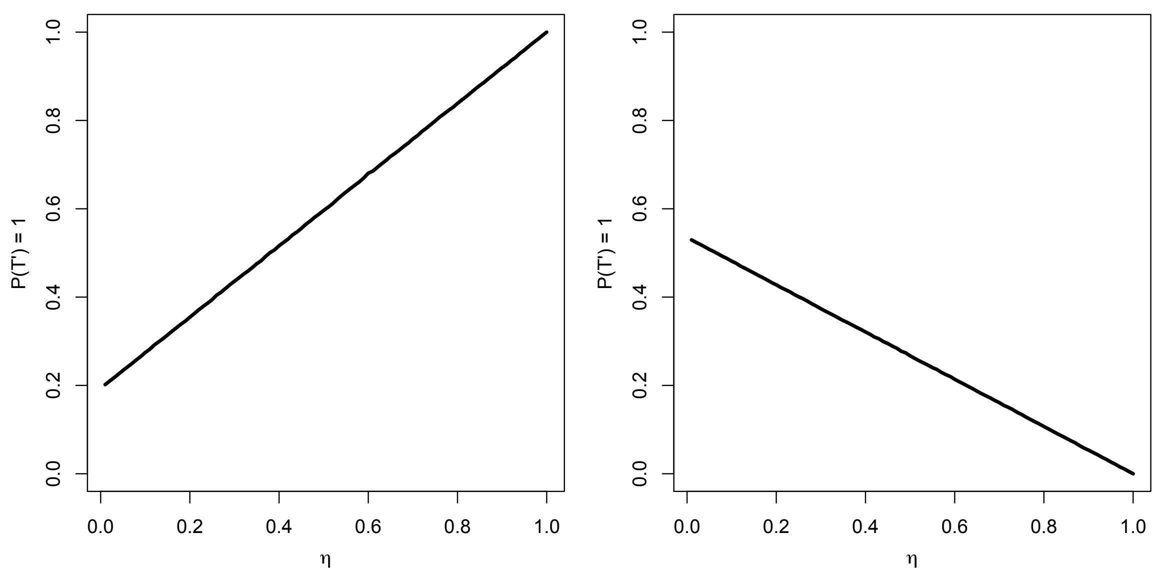


Figure 9: p-value of the average treatment effect conditional on η_1 and η_2. Black: $p < 0.01$, dark gray: $p < 0.05$, light gray: $p < 0.10$. white: $p > 0.10$.

D.2 Sensitivity Analyses for Other Dependent Variables

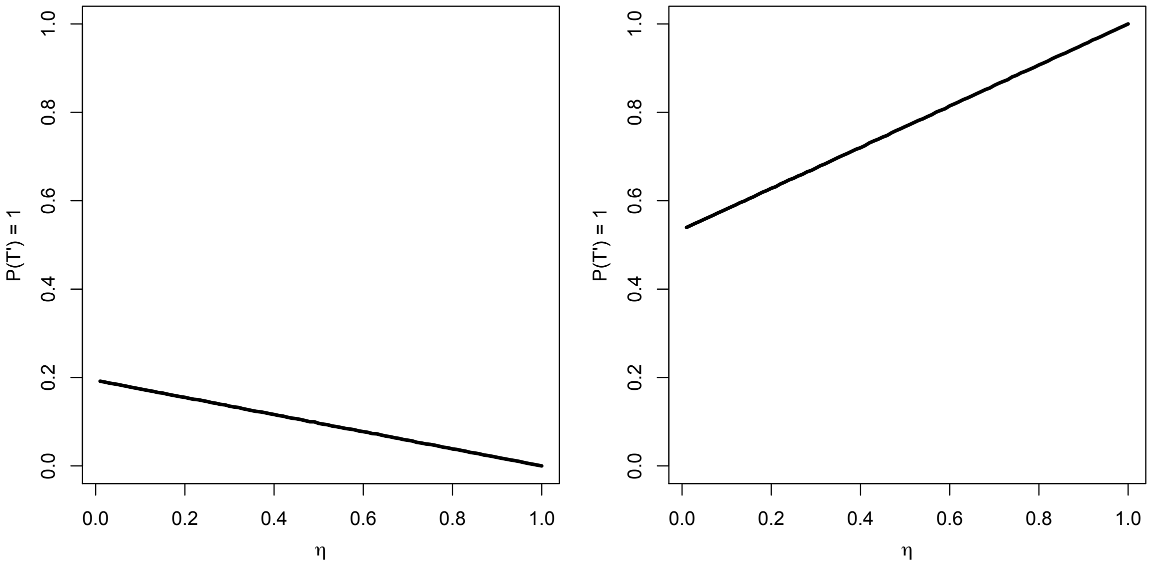
In addition to choosing between clinics and hospitals, citizens made decisions between other pairs of potential areas of spending. Here, we present the simulated data and results for three other questions which are also discussed in Imai and Yamamoto (2010). The procedure is exactly the same as described in the article.

D.2.1 Primary or Secondary Education?



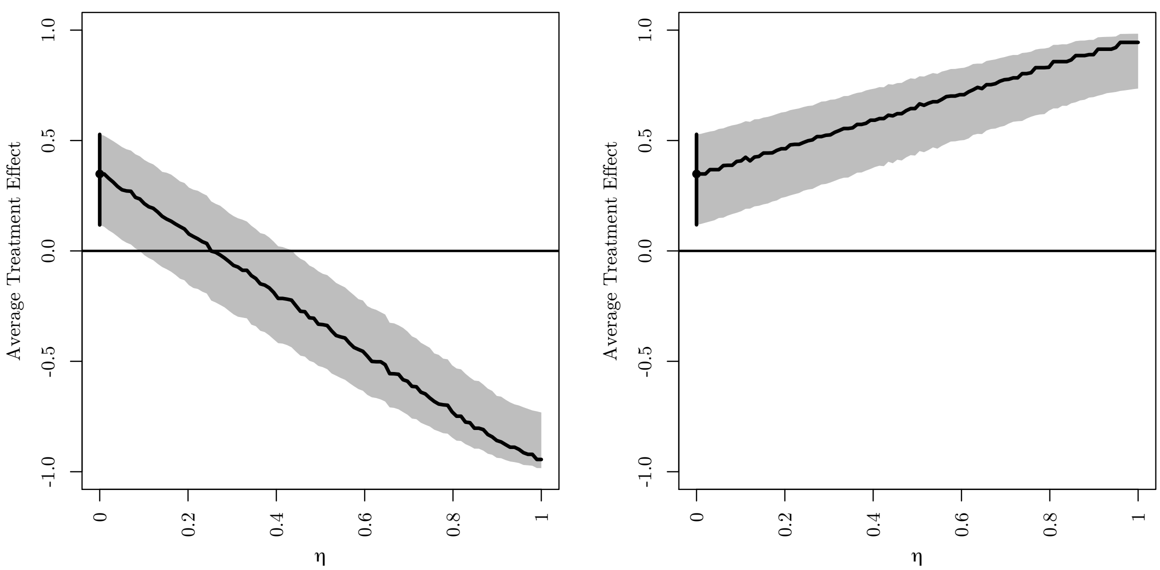
(a) $Y = 0$ (b) $Y = 1$

Figure 10: Probability of $T'(\eta) = 1$ for Scenario (a) from Table 3 of article for Question 4a (primary or secondary education).



(a) $Y = 0$ (b) $Y = 1$

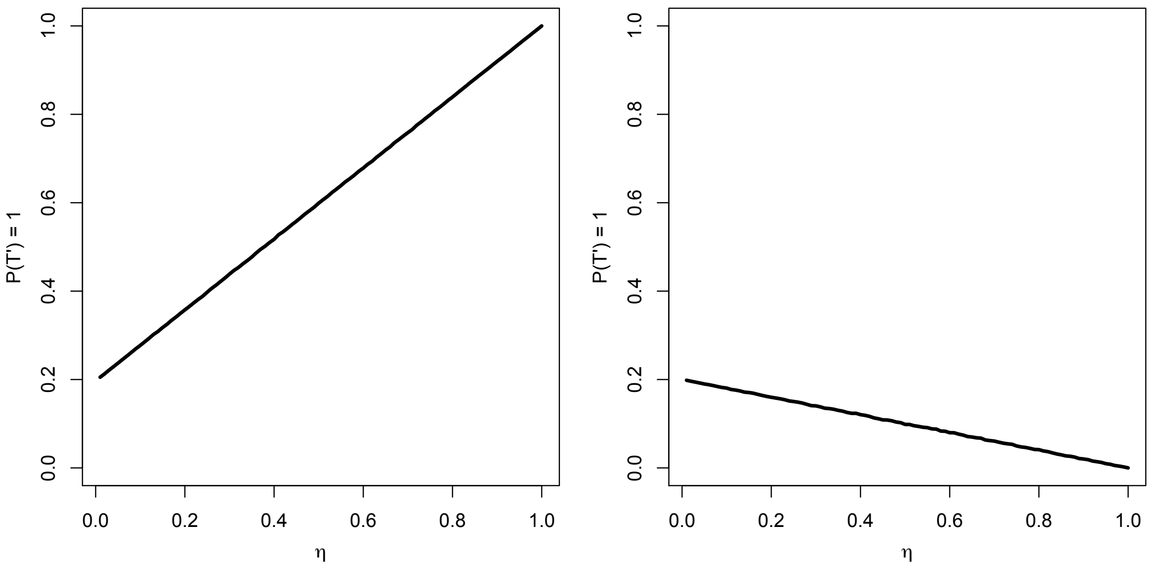
Figure 11: Probability of $T'(\eta) = 1$ for Scenario (b) from Table 3 of article for Question 4a (primary or secondary education).



(a) Scenario (a) from Table 3 of article (b) Scenario (b) from Table 3 of article

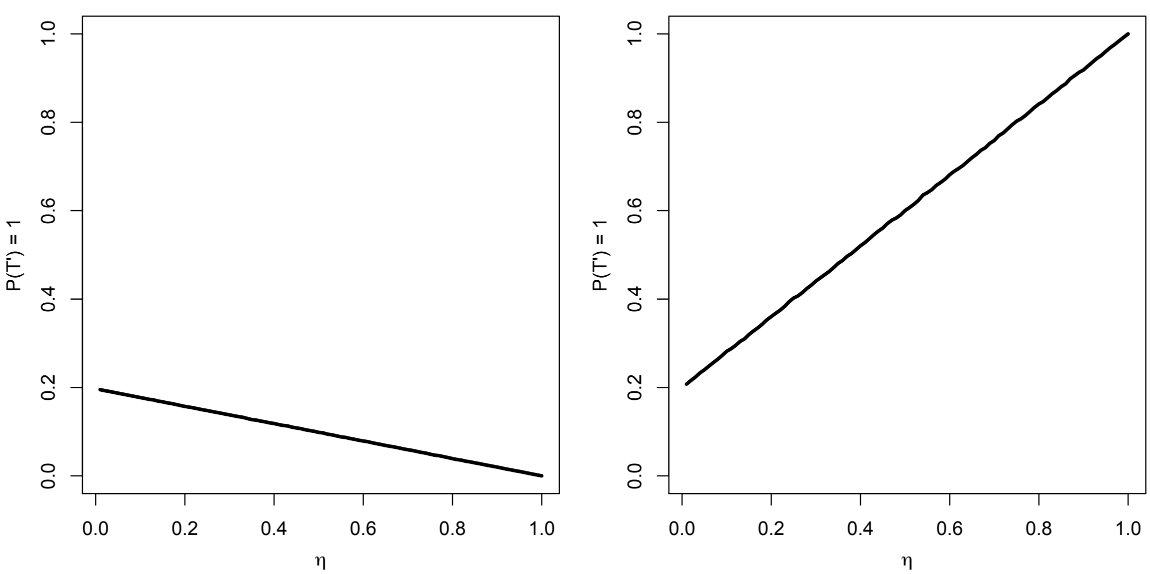
Figure 12: Effect of η on the average treatment effect for Question 4a (primary or secondary education). Point estimates with 95 percent confidence intervals. Black vertical lines: no measurement error.

D.2.2 Roads or Public Transportation?



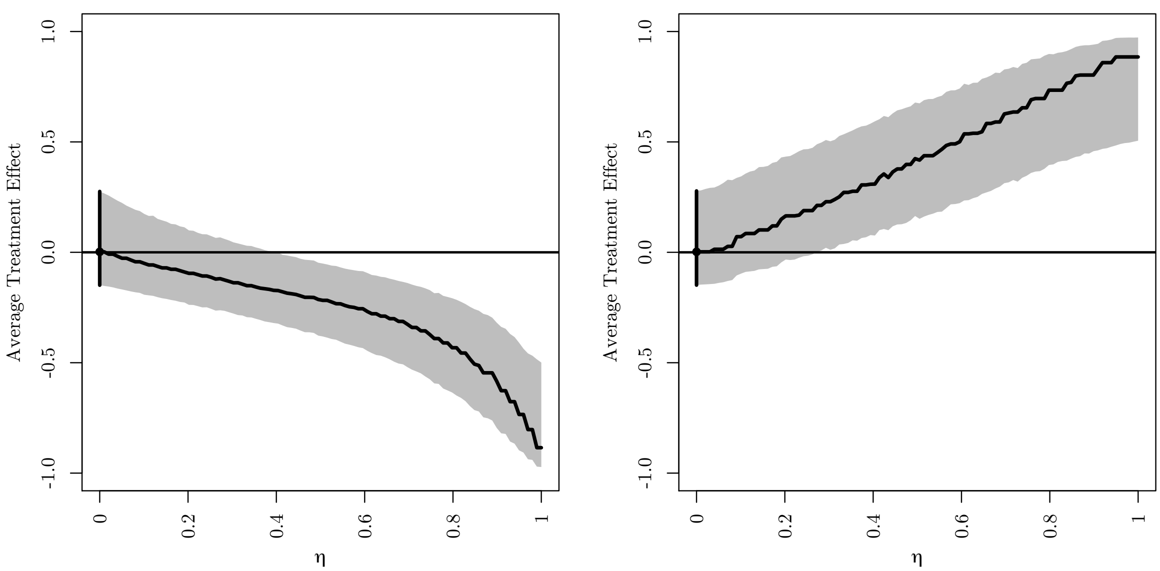
(a) $Y = 0$ (b) $Y = 1$

Figure 13: Probability of $T'(\eta) = 1$ for Scenario (a) from Table 3 of article for Question 7b (roads or public transportation).



(a) $Y = 0$ (b) $Y = 1$

Figure 14: Probability of $T'(\eta) = 1$ for Scenario (b) from Table 3 of article for Question 7b (roads or public transportation).

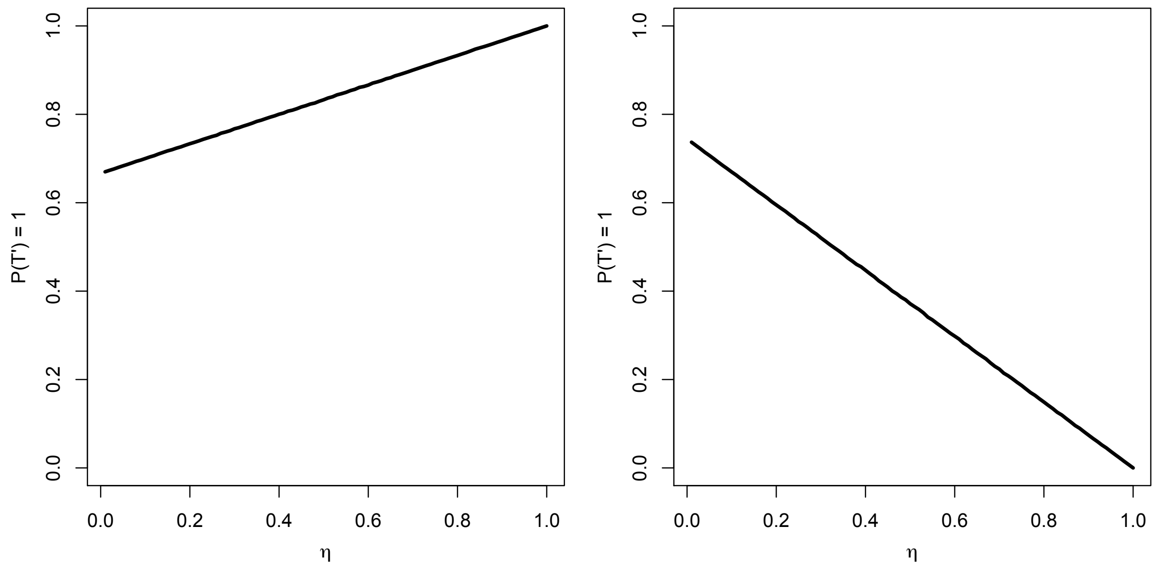


(a) Scenario (a) from Table 3 of article

(b) Scenario (b) from Table 3 of article

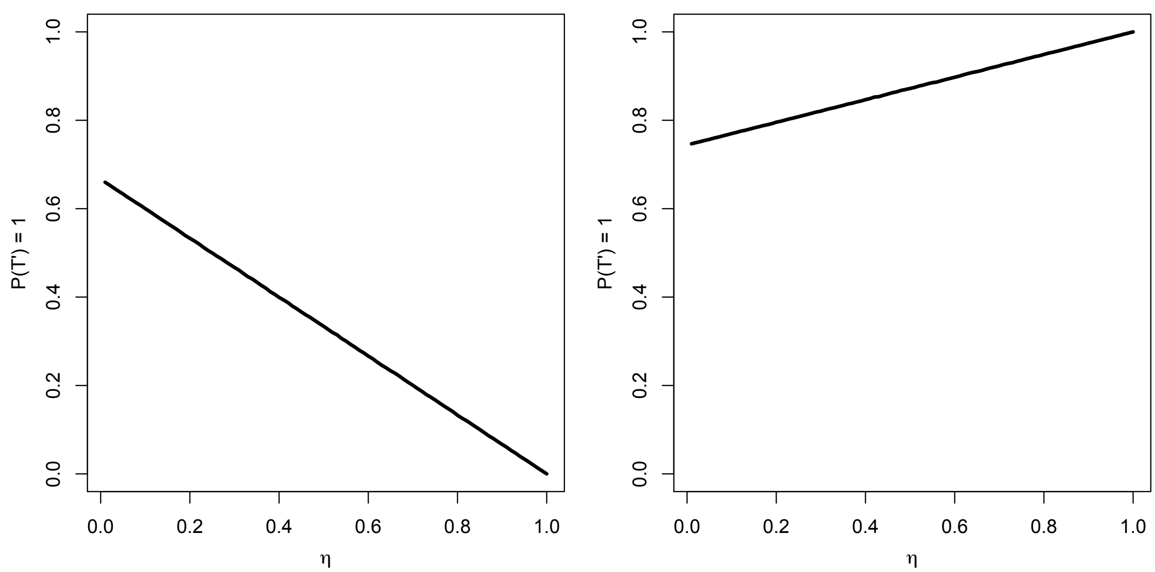
Figure 15: Effect of η on the average treatment effect for Question 7b (roads or public transportation). Point estimates with 95 percent confidence intervals. Black vertical lines: no measurement error.

D.2.3 Consume or Invest Windfalls?



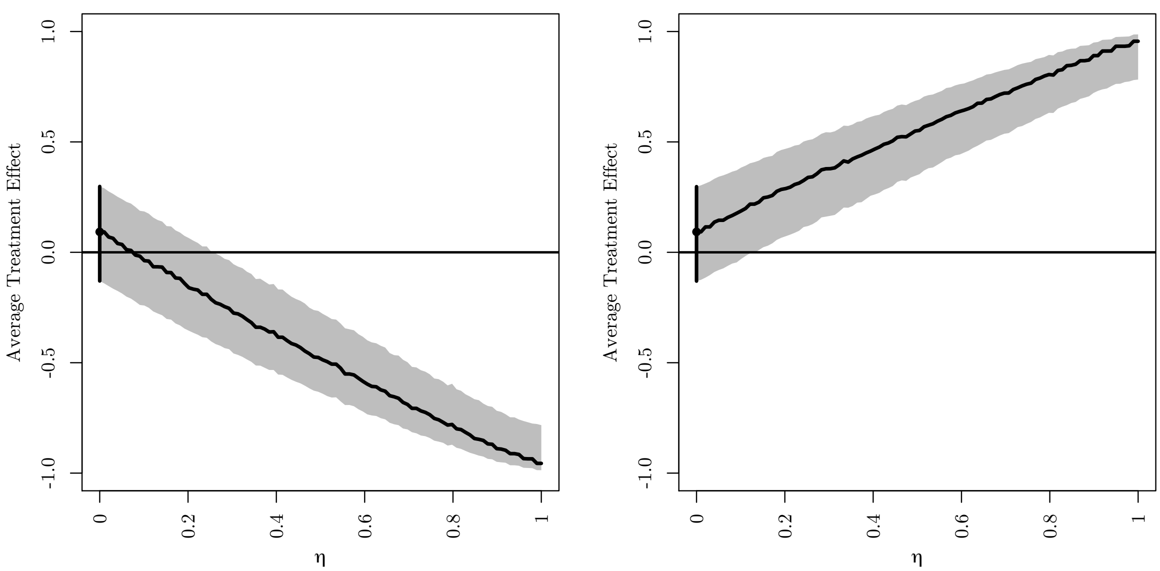
(a) $Y = 0$ (b) $Y = 1$

Figure 16: Probability of $T'(\eta) = 1$ for Scenario (a) from Table 3 of article for Question 11a (consume or invest windfalls).



(a) $Y = 0$ (b) $Y = 1$

Figure 17: Probability of $T'(\eta) = 1$ for Scenario (b) from Table 3 of article for Question 11a (consume or invest windfalls).



(a) Scenario (a) from Table 3 of article

(b) Scenario (b) from Table 3 of article

Figure 18: Effect of η on the average treatment effect for Question 11a (consume or invest windfalls). Point estimates with 95 percent confidence intervals. Black vertical lines: no measurement error.

E Application: Cellphone Coverage and Political Violence

Finally, we present a third application of the sensitivity analysis that we were unable to include in the manuscript due to space constraints. We address a recent debate about the effect of information and communication technologies on political violence. Pierskalla and Hollenbach (2013) compare violent events in areas with and without cellphone coverage and find that the former are more likely to see organized political violence. Both Dafoe and Lyall (2015) and Weidmann (2016) raise the possibility that the dependent variable suffers from reporting bias, that is non-random measurement error. Pierskalla and Hollenbach (2013) use geo-coded event data of political violence provided by the UCDP GED project (Sundberg and Melander, 2013). This dataset is partly based on reports from non-governmental organizations, but relies heavily on international media reports as sources. Having cellular coverage in an area may make it more likely that information about a given violent event is related to the media. In other words, the positive association between cellphone coverage and political violence may not be causal, but instead a result of the fact that events in areas with coverage are more likely to be reported.

Weidmann (2016) compares geo-referenced UCDP data for Afghanistan to SIGACTS data, which were collected by US and Coalition forces. He shows that a SIGACT event which occurred in an area without cellphone coverage is significantly less likely to be recorded in the media-based data. While cellphone coverage has a positive effect on violence in Afghanistan when using the UCDP data, the coefficient is not significant when using the SIGACT data. This suggests that reporting bias may be present in Pierskalla and Hollenbach's study, raising the concern that non-random measurement error is driving the association between cellphone coverage and organized political violence in Africa. What we do not know is whether the result is very fragile and is not supported by the data when they contain even a small amount of measurement error, or whether the error needs to be more severe. Again, the sensitivity analysis approach brings clarity to this question.

Table 13: Transition probabilities for observed violent events Y given presence or absence of cellphone coverage C.

		Y = 0	Y = 1
C	0	$P_0(0 \to 1) = \eta$	$P_0(1 \to 0) = 0$
	1	$P_1(0 \to 1) = 0$	$P_1(1 \to 0) = 0$

The unit of analysis are grid cells of approximately 55km by 55km size. Let $y_i \in \{0, 1\}$ be the observed dependent variable, which takes on a value of one if a violent conflict event for cell i is reported in the UCDP GED data. Cellphone coverage is denoted by $c_i \in \{0, 1\}$. Note that the measurement error is suspected to be in the dependent variable. Figure 13 shows the transition probabilities. A cell that has no cellphone coverage and no violent event according to UCDP GED has an event with probability η. Observations for cells with cellphone coverage but no reported event are considered accurate. Cells that have a violent event reported in the UCDP GED data cannot suffer from under-reporting.

The lower boundary with no measurement error is $\underline{\eta} = 0$, and for the upper boundary we are again able to use auxiliary information. For Afghanistan, the SIGACTS data used in Weidmann (2016) show that of the 72 districts that have no cellphone coverage, 17 have at least one reported violent event in the UCDP GED data. When using the SIGACTS data, 30 do. This corresponds to an increase of about 76 percent. However, the UCDP GED data for Afghanistan are entirely based on media reports, whereas the one's for Africa used in Pierskalla and Hollenbach (2013) also utilize non-media sources less susceptible to reporting bias (see Hollenbach and Pierskalla, 2016). The measurement error found in Weidmann (2016) is thus likely an upper boundary.

In the cross-sectional specification of Pierskalla and Hollenbach (2013), there are 6725 cells without cellphone coverage. In 182 of them, UCDP GED reports at least one event. The

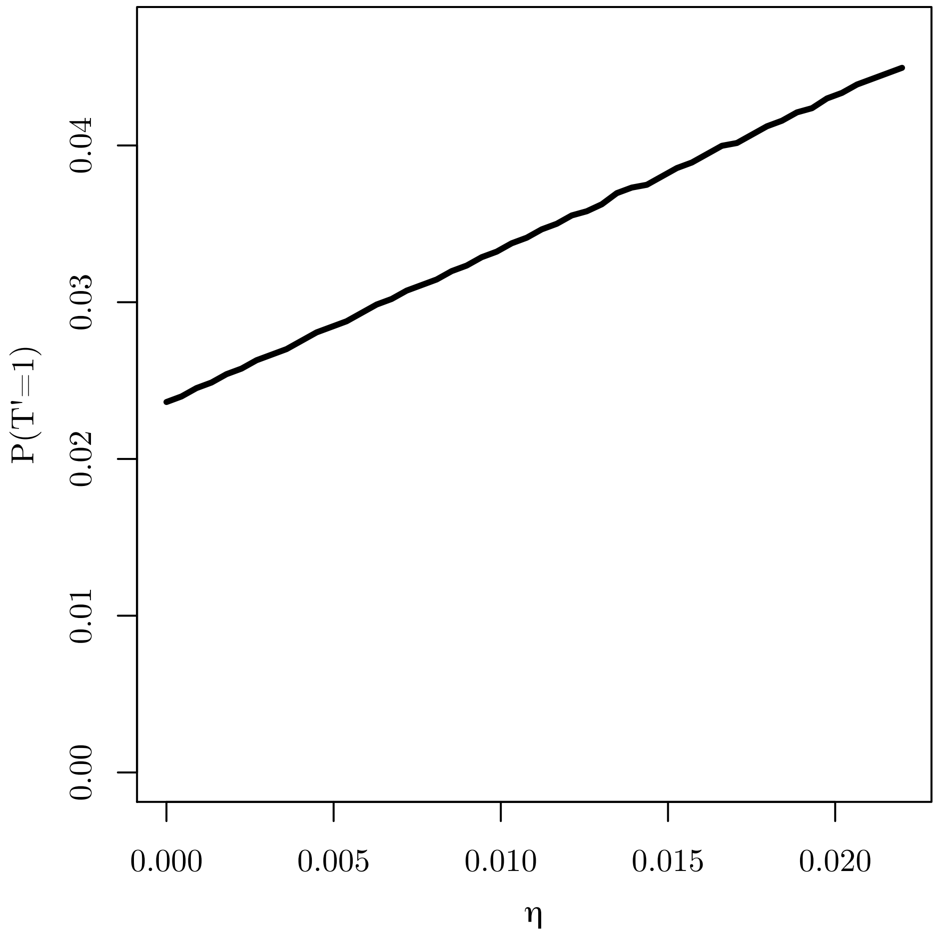


Figure 19: Mean of simulated $T'(\eta)$ for $C = 0$.

other 6543 cells are potentially subject to underreporting and receive $t_i' = 1$ with probability η. We chose $\bar{\eta} = 0.022$, which corresponds to (in expectation) about 80 percent more cells with at least one violent event, so approximately the amount of measurement error found by Weidmann (2016). Figure 19 shows the probability that a cell without cellphone coverage has a violent event depending on η. Because the sampling of cells with unreported violence is a probabilistic process, we simulate 500 draws for each value of η and take the median point estimates and standard errors.

Figure 20 shows how the coefficient of the cellphone coverage variable changes for different levels of systematic measurement error. Panel (a) uses the basic logit model from Table 1 in Pierskalla and Hollenbach (2013). When there is no reporting bias, the point

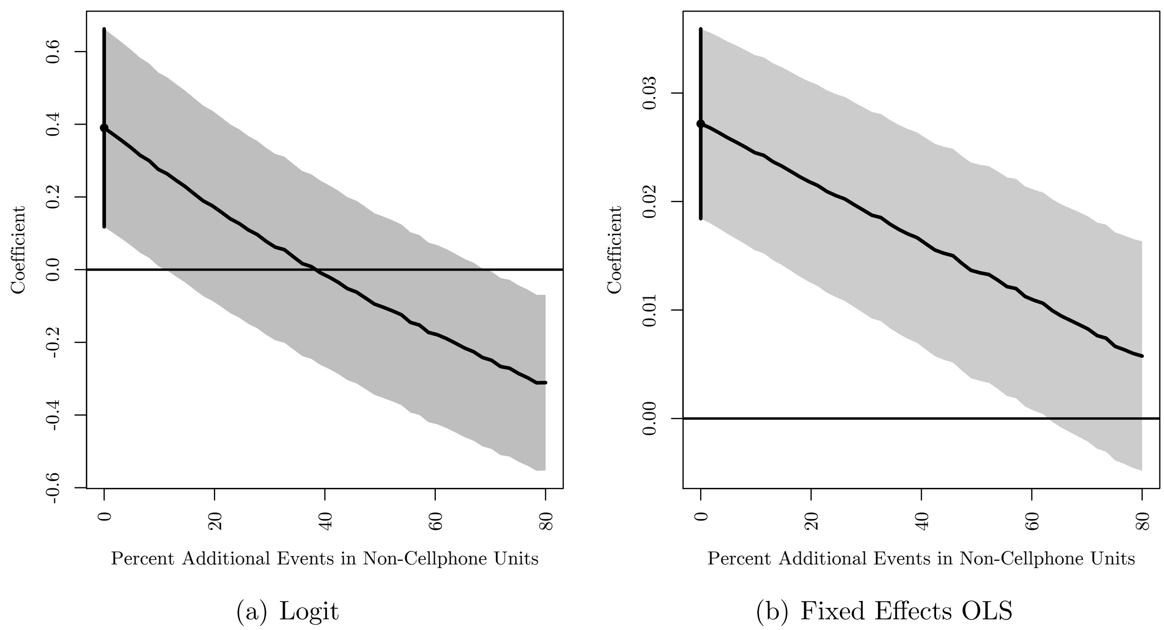


Figure 20: Effect of adding violence events in non-cellphone units for two different model specifications. Point estimates with 95 percent confidence intervals. Black vertical lines: no measurement error.

estimate is about 0.4 and statistically different from zero. As more cells without cellphone coverage are transitioned towards having a violent event, the coefficient goes towards zero. It ceases to be significant for η^* of around 0.003, which corresponds to adding about 10 percent more cells with a violent event. Once it exceeds around 67 percent, the effect of cellphone coverage on organized violence is in fact negative.

Because unobserved country characteristics may confound the results, Pierskalla and Hollenbach (2013) also estimate a linear probability model with country fixed effects. As Panel (b) shows, the argument that mobile communication facilitates violent conflict is much more robust to measurement error in this case. The coefficient ceases to be significant for η^* of around 0.0175, so when the true number of cells without coverage but with violent events is about 63 percent higher than the reported one.

How robust the finding that cellphone coverage increases organized violence is to systematic measurement error thus depends on the type of model employed. In estimations that pool all cells together, cellphone coverage does not have an effect on organized political violence when around 10 percent of the cells without cellphone coverage have events not reported in UCDP GED. The pooled model specifications are appropriate if there are no relevant unobserved country characteristics that might confound the results, so if there is no omitted variable bias. If this is the case, it is possible that the findings of Pierskalla and Hollenbach (2013) are driven by reporting bias. However, if there are unobserved differences between countries, the fixed effects model is more appropriate. In this case, the positive effect of cellphone coverage on organized political violence is spurious only if reporting bias is relatively close to the one found by Weidmann for Afghanistan – where, in contrast to Africa, the UCDP data rely purely on media reports and are not supplemented by other sources less likely to suffer from reporting bias.

References

Albouy, David Y. 2012. "The Colonial Origins of Comparative Development: An Empirical Investigation: Comment." *American Economic Review* 102(6):3059–3076.

Dafoe, Allan and Jason Lyall. 2015. "From Cell Phones to Conflict? Reflections on the Emerging ICT–Political Conflict Research Agenda." *Journal of Peace Research* 52(3):401–413.

Hollenbach, Florian M. and Jan H. Pierskalla. 2016. "A Re-Assessment of Reporting Bias in Event-Based Violence Data with Respect to Cell Phone Coverage." `https://www.dropbox.com/s/4shzby96qodl4jl/Response_ReportingBias_V3.pdf?dl=0`.

Humphreys, Macartan, William A Masters and Martin E Sandbu. 2006. "The Role of Leaders in Democratic Deliberations: Results from a Field Experiment in São Tomé and Príncipe." *World Politics* 58(4):583–622.

Imai, Kosuke and Teppei Yamamoto. 2010. "Causal Inference with Differential Measurement Error: Nonparametric Identification and Sensitivity Analysis." *American Journal of Political Science* 54(2):543–560.

Pierskalla, Jan H. and Florian M. Hollenbach. 2013. "Technology and Collective Action: The Effect of Cell Phone Coverage on Political Violence in Africa." *American Political Science Review* 107(2):207–224.

Sundberg, Ralph and Erik Melander. 2013. "Introducing the UCDP Georeferenced Event Dataset." *Journal of Peace Research* 50(4):534–532.

Weidmann, Nils B. 2016. "A Closer Look at Reporting Bias in Conflict Event Data." *Americal Journal of Political Science* 60(1):206–218.