I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 16508

Embrace the Noise: It Is OK to Ignore Measurement Error in a Covariate, Sometimes

Hao Dong
Daniel L. Millimet

OCTOBER 2023

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 16508

Embrace the Noise: It Is OK to Ignore Measurement Error in a Covariate, Sometimes

Hao Dong
Southern Methodist University

Daniel L. Millimet
Southern Methodist University and IZA

OCTOBER 2023

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

IZA DP No. 16508 OCTOBER 2023

ABSTRACT

Embrace the Noise: It Is OK to Ignore Measurement Error in a Covariate, Sometimes*

In linear regression models, measurement error in a covariate causes Ordinary Least Squares (OLS) to be biased and inconsistent. Instrumental Variables (IV) is a common solution. While IV is also biased, it is consistent. Here, we undertake an asymptotic comparison of OLS and IV in the case where a covariate is mismeasured for $\lfloor N^{\delta} \rfloor$ of N observations with $\delta \in [0, 1]$. We show that OLS is consistent for $\delta < 1$ and is asymptotically normal and more efficient than IV for $\delta < 0.5$. Simulations and an application to the impact of body mass index on family income demonstrate the practical usefulness of this result.

JEL Classification: C13, C26, C52

Keywords: errors-in-variables, measurement error, asymptotics

Corresponding author:
Daniel L. Millimet
Department of Economics
Box 0496
Southern Methodist University
Dallas, TX 75275-0496
USA
E-mail: millimet@smu.edu

* The authors thank Badi Baltagi, Esfandiar Maasoumi, John Mullahy, and Chris Parmeter for helpful comments.

To do nothing is sometimes a good remedy.
– Hippocrates

Don't underestimate the value of Doing Nothing, of just going along, listening to all the things you can't hear, and not bothering.
– Winnie the Pooh (A.A. Milne)

1. Introduction

Conventional wisdom in empirical research seems to be that the appropriate statistical or econometric estimator is dictated by the most problematic observations. For example, if any covariate in a regression model is correlated with the error term for any subset of the sample, then Ordinary Least Squares (OLS) should be abandoned in favor of, say, Instrumental Variables (IV). Or, in a linear panel regression model, if there exists relevant unobserved, time invariant heterogeneity for at least some units, then the Fixed Effects (FE) estimator should be used in lieu of Random Effects (RE) or Pooled OLS (POLS). Or, if the variance of the regression error for any subset of the data differs from that for the remainder of the sample, then Generalized Least Squares (GLS) or heteroskedasticity-robust standard errors should be used. In the first two cases, the conventional wisdom is to abandon OLS and RE/POLS as the presence of problematic observations causes these estimators to be biased. In third case, the problematic observations cause the OLS standard errors to be biased and, hence, abandoned.

This conventional wisdom needs to be rethought as bias is not the only relevant criteria. As data become more plentiful, consistency is perhaps more relevant than (finite sample) bias. Moreover, often the estimator being abandoned is more efficient than the alternative. This point is cleverly illustrated in Pesaran and Zhou (2018) in the context of linear panel data models. The authors show under what conditions POLS is consistent and has a smaller mean squared error (MSE) than FE when only *some* units have unobserved, time invariant heterogeneity. Specifically, the authors allow for unobserved, time invariant heterogeneity in $\lfloor N^{\delta} \rfloor$ of N units, where $\lfloor \cdot \rfloor$ is the floor function representing the largest integer less than the argument, for different values of δ in the unit interval. Of course, such insights have a long history. Fisher (1961) shows that simultaneous equation estimators need not be abandoned if the restrictions embedded in the model hold approximately. The key, according to Fisher (1961, p. 148) is "deciding whether the particular approximations are 'good enough"'.

Here, we perform a similar analysis as in Pesaran and Zhou (2018) in the context of a linear regression model with a covariate suffering from measurement error for *some* observations. Analogous to Pesaran and Zhou (2018), we consider cases where the covariate suffers from measurement error in $\lfloor N^{\delta} \rfloor$ of N units for $\delta \in [0, 1]$. We derive the theoretical conditions under which OLS remains consistent and has a smaller MSE than IV in this situation. We show the empirical relevance of this analysis through simulations and an application assessing the impact of body mass index (BMI) on income. Thus, as Fisher (1961) emphasizes, we are able to derive theoretically and show empirically how a researcher might decide if assuming no measurement error in a covariate is a "good enough" approximation.

The results are striking. Specifically, simulations illustrate the smaller asymptotic MSE of the OLS estimator as N gets large and $\lfloor N^{0.5} \rfloor$ observations suffer from measurement error even if the instrument is very strong (first-stage F-statistic ≈ 100). Importantly, this result continues to hold even if the remaining $N - \lfloor N^{0.5} \rfloor$ observations contain modest measurement error. The simulations also show that as the first-stage F-statistic declines (even to ≈ 20, which exceeds the usual rule-of-thumb), the relative performance of IV deteriorates rapidly. In this case, OLS has a smaller MSE even in samples as small as 500. Finally, despite substantial measurement error in *self-reported* BMI and the existing literature focused on correcting for this measurement error (often by IV), we find that the estimated effects on monthly family income are virtually identical to those obtained using *measured* BMI and confirm a negative obesity penalty for young women in the United States. In contrast, the IV estimates are predominantly statistically insignificant and much less precise despite the use of a strong instrument based on genetics. While the theoretical results provided here may be "obvious" to some, showing that they have practical relevance in *small* samples, in cases when a *strong* instrument is available, and in a *practical* application where IV is the dominant choice should cause empirical researchers to take note.

The remainder of the paper is organized as follows. Section 2 lays out the standard errors-in-variables framework. Section 3 derives the asymptotic properties of OLS and IV when $\lfloor N^{\delta} \rfloor$ of N observations suffer from measurement error in a covariate for $\delta \in [0, 1]$. Sections 4 and 5 present the results of our simulations and illustrate the usefulness of our analysis in practice. Section 6 concludes.

2. Linear errors-in-variables regression

Consider the linear regression model with errors-in-variables as

$$y_i = \alpha + \beta x_i^* + \epsilon_i, \tag{2.1}$$

$$x_i = x_i^* + \eta_i, \tag{2.2}$$

$$x_i^* = \psi + \pi z_i + \omega_i, \tag{2.3}$$

where x_i^* is the exogenous variable that is unobserved due to measurement error and z_i is a valid instrument for x_i.[1] To keep things simple, we focus on the case where both x^* and z are scalar.[2]

We make use of the following assumptions.

Assumption 1. *(x_i^*, y_i, z_i) are independently and identically distributed for $i = 1, \ldots, N$, $Cov[x_i^*, \epsilon_i] = 0$, $Cov[x_i^*, z_i] \neq 0$, and $Cov[z_i, \epsilon_i] = 0$.*

Assumption 2. *The measurement errors η_i are independently distributed across i and satisfy*

$$\frac{1}{N}\sum_{i=1}^{N} E[|\eta_i|^s] = O(N^{\delta-1}) \tag{2.4}$$

for $s = 1, 2$ and some $0 \leq \delta \leq 1$.

To understand Assumption 2, consider the following example. If there exists an ordering of the individual units such that

$$\eta_i = \begin{cases} u_i & i = 1, 2, \ldots, \lfloor N^{\delta} \rfloor \\ 0 & i = \lfloor N^{\delta} \rfloor + 1, \ldots, N, \end{cases}$$

[1](2.3) is based on (38) in Schennach (2016) or (8) in Schennach (2020). If we change the definition of ω_i, this is equivalent to modeling x_i as a function of z_i.

[2]It is inconsequential to view y, x^*, x, and z as net of other correctly measured determinants of y according to the Frisch-Waugh-Lovell theorem as long as Assumptions 1 and 2 hold. Moreover, allowing x^* and z to be vectors does not add any new insights to the analysis.

where $\{u_i : i = 1, 2, \ldots, N\}$ is a sequence of random variables with zero mean and finite variances such that

$$\frac{1}{M}\sum_{i=1}^{M} E[|u_i|^s] = O(1)$$

for $s = 1, 2$ as $M \to \infty$, then Assumption 2 is satisfied as

$$\frac{1}{N}\sum_{i=1}^{N} E[|\eta_i|^s] = N^{\delta-1}\left(N^{-\delta}\sum_{i=1}^{\lfloor N^\delta \rfloor} E[|u_i|^s]\right) = O(N^{\delta-1}).$$

In this case, $\lfloor N^\delta \rfloor / N$ is the share of observations suffering from measurement error; for the remainder, $x_i = x_i^*$. In general, δ as in (2.4) can be understood as the strength of the measurement error.

3. Least squares and instrumental variables estimators

Our goal is to compare the OLS and IV estimators of β in the model given by (2.1)-(2.3) under (2.4). The OLS and IV estimators of β, $\hat{\beta}_{LS}$ and $\hat{\beta}_{IV}$, respectively, can be written as

$$\hat{\beta}_{LS} = \hat{M}_{xx}^{-1}\hat{M}_{xy},$$
$$\hat{\beta}_{IV} = \hat{M}_{zx}^{-1}\hat{M}_{zy},$$

where

$$\hat{M}_{xx} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2, \qquad \hat{M}_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}),$$
$$\hat{M}_{zx} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(z_i - \bar{z}), \quad \hat{M}_{zy} = \frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})(y_i - \bar{y}),$$

and

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i, \quad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i, \quad \bar{z} = \frac{1}{N}\sum_{i=1}^{N} z_i.$$

3.1. **Least squares estimator.**

For the OLS estimator $\hat{\beta}_{LS}$, we have

$$\begin{aligned}
\hat{\beta}_{LS} - \beta &= \left\{\frac{1}{N}\sum_{i=1}^{N}(x_i^* - \bar{x}^* + \eta_i - \bar{\eta})^2\right\}^{-1}\left\{\frac{1}{N}\sum_{i=1}^{N}(x_i^* - \bar{x}^* + \eta_i - \bar{\eta})(y_i - \bar{y})\right\} - \beta \\
&= \left\{\underbrace{\frac{1}{N}\sum_{i=1}^{N}(x_i^* - \bar{x}^*)^2}_{\hat{M}_{x^*x^*}} + \underbrace{\frac{2}{N}\sum_{i=1}^{N}(x_i^* - \bar{x}^*)(\eta_i - \bar{\eta})}_{2\hat{M}_{\eta x^*}} + \underbrace{\frac{1}{N}\sum_{i=1}^{N}(\eta_i - \bar{\eta})^2}_{\hat{M}_{\eta\eta}}\right\}^{-1} \\
&\times \left\{\beta\hat{M}_{x^*x^*} + \underbrace{\frac{1}{N}\sum_{i=1}^{N}(x_i^* - \bar{x}^*)(\epsilon_i - \bar{\epsilon})}_{\hat{M}_{x^*\epsilon}} + \underbrace{\frac{1}{N}\sum_{i=1}^{N}(\eta_i - \bar{\eta})(x_i^* - \bar{x}^*)}_{\hat{M}_{\eta x^*}}\beta + \underbrace{\frac{1}{N}\sum_{i=1}^{N}(\eta_i - \bar{\eta})(\epsilon_i - \bar{\epsilon})}_{\hat{M}_{\eta\epsilon}}\right\} - \beta \\
&= (\hat{M}_{x^*x^*} + 2\hat{M}_{\eta x^*} + \hat{M}_{\eta\eta})^{-1}(\hat{M}_{x^*\epsilon} + (\beta - 2)\hat{M}_{\eta x^*} + \hat{M}_{\eta\epsilon} - \hat{M}_{\eta\eta}) \qquad (3.1)
\end{aligned}$$

To study the properties of $\hat{\beta}_{LS}$, we add the following assumption.

Assumption 3. *$Var[x_i^*] > 0$, $E[\epsilon_i^2|x^*] = \sigma_\epsilon^2$, $\max_{1\leq i\leq N}\{\sup_a E[|x_i^*||\eta_i = a] \vee \sup_a E[|\epsilon_i||\eta_i = a]\} < \infty$.*

Assumption 3 imposes some regularity conditions. It requires the regressor x^* to be non-constant and have a finite second order moment, the regression error ϵ_i to be homoskedastic, and the conditional expectations of x_i^* and ϵ_i given η_i to be bounded. Thus, η_i is not required to be classical measurement error. Under random sampling—as required in Assumption 1—the boundedness of the conditional expectations of x_i^* and ϵ_i given η_i would degenerate to the existence of the unconditional first order moment of x^* and ϵ_i if the measurement error is assumed to be classical, which would automatically follow from the rest of Assumption 3.

To derive the asymptotic distribution of $\hat{\beta}_{LS}$, note that Assumption 2 implies

$$\hat{M}_{\eta\eta} = O_p(N^{\delta-1}). \tag{3.2}$$

Since $\hat{M}_{\eta x^*} = \frac{1}{N}\sum_{i=1}^N(\eta_i - \bar{\eta})x_i^*$, we have

$$E[|\hat{M}_{\eta x^*}|] \leq 2\max_{1\leq i\leq N}\sup_a E[|x_i^*||\eta_i = a]\left\{\frac{1}{N}\sum_{i=1}^N E[|\eta_i|]\right\} = O(N^{\delta-1}), \tag{3.3}$$

where the first step follows from the triangular inequality and the second step follows from Assumption 2 and the boundedness of $E[|x_i^*||\eta_i]$ under Assumption 3. By similar arguments, if we assume the boundedness of $E[|\epsilon_i||\eta_i]$, we have

$$\hat{M}_{\eta\epsilon} = O_p(N^{\delta-1}). \tag{3.4}$$

Using (3.2), (3.3), and $\hat{M}_{x^*x^*} \xrightarrow{p} Var[x^*] > 0$, $\hat{M}_{x^*\eta}$ and $\hat{M}_{\eta\eta}$ are asymptotically negligible compared to $Var[x^*]$ if $\delta < 1$ and

$$\hat{M}_{xx} \xrightarrow{p} Var[x^*]. \tag{3.5}$$

Moreover, since $\hat{M}_{x^*\epsilon} \xrightarrow{p} 0$ as $N \to \infty$ under Assumption 1, (3.2), (3.3), (3.4), and (3.5) together imply that $\hat{\beta}_{LS}$ is a consistent estimator of β if $\delta < 1$. Lastly, we have

$$\sqrt{N}(\hat{\beta}_{LS} - \beta) = \{M_{x^*x^*} + o_p(1)\}^{-1}\{\sqrt{N}\hat{M}_{x^*\epsilon} + O_p(N^{\delta-1/2})\}, \tag{3.6}$$

according to (3.1), (3.2), (3.3), (3.4), and (3.5) which leads to the following result.

Theorem 1. *Under Assumptions 1, 2 and 3, if $0 \leq \delta < 1$, $\hat{\beta}_{LS}$ is a consistent estimator of β. If $\delta < 1/2$, we have*

$$\sqrt{N}(\hat{\beta}_{LS} - \beta) \xrightarrow{d} N\left(0, \frac{\sigma_\epsilon^2}{Var[x^*]}\right).$$

Theorem 1 shows that the limiting distribution of the OLS estimator $\hat{\beta}_{LS}$ is contingent upon the strength of measurement error δ. If the measurement errors are not too strong, specifically when $\delta < 1$, $\hat{\beta}_{LS}$ can still be consistent in the presence of a contaminated regressor. If the measurement errors are weak enough, specifically when $\delta < 0.5$, the asymptotic distribution of $\hat{\beta}_{LS}$ is identical to that of the error-free case. Moreover, note that the asymptotic variance of $\hat{\beta}_{LS}$ depends on the unobserved x^*. For the standard error of $\hat{\beta}_{LS}$, we can use the common OLS standard error based on x. In particular, (3.5) shows that $Var[x^*]$ can be consistently estimated by $\frac{1}{N}\sum_{i=1}^N(x_i - \bar{x})^2$ if $\delta < 1$. By similar arguments, we can show that σ_ϵ^2 can be consistently estimated by $\frac{1}{N}\sum_{i=1}^N(y_i - \bar{y} - \hat{\beta}_{LS}(x_i - \bar{x}))^2$ if the measurement errors are not too strong.

3.2. **Instrumental variables estimator.** For the IV estimator $\hat{\beta}_{IV}$, we have

$$\begin{aligned}
\hat{\beta}_{IV} - \beta &= \left\{ \frac{1}{N} \sum_{i=1}^{N} (z_i - \bar{z})(x_i^* - \bar{x}^* + \eta_i - \bar{\eta}) \right\}^{-1} \left\{ \frac{1}{N} \sum_{i=1}^{N} (z_i - \bar{z})(y_i - \bar{y}) \right\} - \beta \\
&= \left\{ \underbrace{\frac{1}{N} \sum_{i=1}^{N} (z_i - \bar{z})(x_i^* - \bar{x}^*)}_{\hat{M}_{zx^*}} + \underbrace{\frac{1}{N} \sum_{i=1}^{N} (z_i - \bar{z})(\eta_i - \bar{\eta})}_{\hat{M}_{\eta z}} \right\}^{-1} \left\{ \beta \hat{M}_{zx^*} + \underbrace{\frac{1}{N} \sum_{i=1}^{N} (z_i - \bar{z})(\epsilon_i - \bar{\epsilon})}_{\hat{M}_{z\epsilon}} \right\} - \beta \\
&= \{\hat{M}_{zx^*} + \hat{M}_{\eta z}\}^{-1} \{\hat{M}_{z\epsilon} - \beta \hat{M}_{\eta z}\}. \qquad (3.7)
\end{aligned}$$

To study properties of $\hat{\beta}_{IV}$, we add the following assumption.

Assumption 4. *$Var[x_i^*] > 0$, $E[z_i^2] < \infty$, and $E[\epsilon_i^2 | z_i] = \sigma_\epsilon^2$.*

If $\eta_i \sim IID(0, \sigma_\eta^2)$ and $Cov[z_i, \eta_i] = 0$, then we have the standard results for the asymptotic properties of $\hat{\beta}_{IV}$ under certain regularity assumptions as this is a special case of $\delta = 1$. If η_i is not identically distributed, by similar arguments as for (3.3), Assumption 2 and the boundedness of $E[|z_i||\eta_i]$ imply

$$\hat{M}_{\eta z} = O_p(N^{\delta - 1}). \qquad (3.8)$$

Then (3.7) and (3.8) imply

$$\sqrt{N}\{\hat{\beta}_{IV} - \beta\} = \{M_{zx^*} + o_p(1)\}^{-1}\{\sqrt{N} M_{z\epsilon} + O_p(N^{\delta - 1/2})\},$$

which gives the following result.

Theorem 2. *Under Assumptions 1, 2 and 4, if η_i is identically distributed, $Cov[z_i, \eta_i] = 0$, and $E[\eta_i^2 | z_i] = \sigma_\eta^2$, we have*[3]

$$\sqrt{N}(\hat{\beta}_{IV} - \beta) \xrightarrow{d} N\left(0, \frac{\sigma_\epsilon^2 + \beta^2 \sigma_\eta^2}{Var[x^*]\rho_{z,x^*}^2}\right),$$

where ρ_{z,x^*} denotes the correlation coefficient between z and x^*. If the distributions of η_i are not identical, but satisfy (2.4) with $0 \le \delta < 1$ and $\max_{1 \le i \le N} \sup_a E[|z_i||\eta_i = a] < \infty$, $\hat{\beta}_{IV}$ remains a consistent estimator of β. If $\delta < 1/2$, we have

$$\sqrt{N}(\hat{\beta}_{IV} - \beta) \xrightarrow{d} N\left(0, \frac{\sigma_\epsilon^2}{Var[x^*]\rho_{z,x^*}^2}\right).$$

Theorem 2 shows that the properties of the IV estimator $\hat{\beta}_{IV}$ are also contingent upon the strength of measurement error δ. If all observations suffer from measurement error such that $\delta = 1$, where the errors are drawn from identical distributions, then uncorrelatedness between the instrument z_i and the measurement error η_i is necessary for the consistency of $\hat{\beta}_{IV}$. If the measurement errors are not too strong, specifically when $\delta < 1$, and the conditional expectation of z_i given η_i is bounded, $\hat{\beta}_{IV}$ can still be consistent even if there is a non-zero correlation between the instrument z_i and the measurement error η_i. If the measurement errors are weak enough, specifically when $\delta < 1/2$,

[3]Note, the variance can also be expressed as $\frac{\sigma_\epsilon^2 + \beta^2 \sigma_\eta^2}{Var[x]\rho_{z,x}^2}$ under $Cov[z_i, \eta_i] = 0$. Here, since we do not require $Cov[z, \eta] = 0$ when $\delta < 1$, we retain the most general case as presented above.

the asymptotic distribution of $\hat{\beta}_{IV}$ is identical to that of the error-free case. Moreover, note that the asymptotic variance of $\hat{\beta}_{IV}$ depends on the unobserved x^*. For the standard error of $\hat{\beta}_{IV}$, we can use the common IV standard error based on x. The validity of such standard error is clear when η_i is identically distributed because the asymptotic variance can be expressed as $\frac{\sigma_\epsilon^2+\beta^2\sigma_\eta^2}{Var[x]\rho_{z,x}^2}$ under $Cov[z_i,\eta_i]=0$. When the measurement errors are not too strong, by similar arguments as in (3.5), even if $Cov[z_i,\eta_i]\neq 0$, we can show that $Cov[z,x^*]$ and σ_ϵ^2 can be consistently estimated by $\frac{1}{N}\sum_{i=1}^N(z_i-\bar{z})(x_i-\bar{x})$ and $\frac{1}{N}\sum_{i=1}^N(y_i-\bar{y}-\hat{\beta}_{IV}(x_i-\bar{x}))^2$, respectively.

3.3. **Asymptotic comparison.** Under Theorems 1 and 2, both estimators are consistent when the measurement error is not too strong, specifically when $\delta<1$. In addition, $\hat{\beta}_{LS}$ is asymptotically more efficient than $\hat{\beta}_{IV}$ when the measurement error is weak enough, specifically when $\delta<1/2$. Thus, the asymptotic MSE of the OLS estimator is smaller in this case.

3.4. **Specification Testing.** With large N, it is advisable to use OLS when $\delta<1/2$. To test this condition, we proceed as in Pesaran and Zhou (2018) and specify a Hausman type test of

$$\mathtt{H_o}:\delta=1/2-\varepsilon$$

for $\varepsilon\in(0,1/2]$ against

$$\mathtt{H_a}:\delta\geq 1/2.$$

As OLS is asymptotically efficient under $\mathtt{H_o}$ given Assumption 3, the standard Hausman test statistic

$$H_N=\frac{\left(\hat{\beta}_{LS}-\hat{\beta}_{IV}\right)^2}{\widehat{Var}\left(\hat{\beta}_{IV}\right)-\widehat{Var}\left(\hat{\beta}_{LS}\right)}\sim\chi_1^2, \tag{3.9}$$

can be used. In addition, the Hausman test can be used to devise a Pre-test estimator given by

$$\hat{\beta}_{Pre}=\hat{\beta}_{LS}+\left(\hat{\beta}_{IV}-\hat{\beta}_{LS}\right)\mathtt{I}\left[H_N>\chi^2_{1,1-\tau}\right], \tag{3.10}$$

where $\mathtt{I}(\cdot)$ is an indicator function taking on the value one if the argument is true, zero otherwise, and τ is te significance level.

4. Simulations

All experimental designs considered are nested in the following data-generating process.

$$\begin{aligned}
y &= \beta x_i^*+\epsilon_i, \qquad i=1,...,N\\
x_i &= x_i^*+\eta_i\\
x_i^* &= \pi z_i+\omega_i\\
\eta_i &\sim \begin{cases} N(0,\sigma_1^2) & \text{if } i=1,...,\lfloor N^\delta\rfloor\\ N(0,\sigma_2^2) & \text{if } i=\lfloor N^\delta\rfloor+1,...,N\end{cases}\\
\epsilon_i &\sim N(0,1)\\
\omega_i &\sim N(0,2)\\
z_i &\sim N(0,1)
\end{aligned}$$

For each design, we conduct 1,000 simulations and assess the median absolute bias and the root MSE (RMSE) for the OLS, IV, and Pre-test estimates of β, where the true value of β is set to one.

We consider values of $\delta = \{0.25, 0.45, 0.475, 0.50, 0.525, 0.55, 0.75\}$. We vary the sample sizes, $N = \{500, 1000, 5000, 10000, 50000, 100000\}$. We vary σ_1^2 such that the reliability ratio (RR) of x in the subsample of $i = 1, ..., \lfloor N^\delta \rfloor$ is $\{0.25, 0.50, 0.75\}$ and σ_2^2 such that the reliability ratio of x in the subsample of $i = \lfloor N^\delta \rfloor + 1, ..., N$ is $\{1, 0.99, 0.975, 0.95\}$. Thus, we consider cases where the restriction of zero measurement error in the subsample of $i = \lfloor N^\delta \rfloor + 1, ..., N$ is relaxed. Finally, we vary π such that the first-stage F-statistic $\approx \{10, 20, 100\}$.[4] Thus, we analyze 36 experimental designs, for each combination of δ and N.

Figures 1 – 3 plot the *difference* in RMSE between IV and OLS.[5] Positive values indicate a higher RMSE for IV. Within each figure, Panels (A) – (F) correspond to different sample sizes. The figures differ in terms of the expected strength of the instrument. In Figure 1, $F \approx 10$, which is considered to be the rule-of-thumb (Stock *et al.*, 2002). In Figures 2 – 3, $F \approx 20$ and 100, respectively, which are considered to be sufficient in Stock *et al.* (2002). However, Lee *et al.* (2022) consider $F > 100$ as necessary for valid inference. Figures A.1 – A.3 in Appendix A display the corresponding graphs for the difference in median absolute bias. Figures A.4 – A.15 in Appendix A display RMSE and median absolute bias comparing OLS and IV to the Pre-test estimator.

In Figure 1 where $F \approx 10$, OLS has a lower RMSE for all values of δ, reliability ratios, and sample sizes considered. Moreover, the RMSE of IV is highly volatile, consistent with the poor performance of IV with (moderately) weak instruments. OLS also has a lower median absolute bias in all cases and for all sample sizes when $\delta \leq 0.45$, in all cases and for all $N \geq 1,000$ when $\delta \leq 0.5$, and in all cases and for all $N \geq 5,000$ when $\delta \leq 0.75$ (Figure A.1). Thus, OLS outperforms IV even when the sample size is small and the instrument meets the conventional rule-of-thumb to not be considered weak unless the measurement error is both very strong (high δ) and very severe (low reliability ratios).

In Figures 2 and A.2, OLS has a lower RMSE than IV when $\delta \leq 0.5$ for all sample sizes and all reliability ratios despite the instrument being relatively strong by conventional standards ($F \approx 20$). OLS also has a smaller median absolute bias when $\delta \leq 0.5$ for all reliability ratios if $N \geq 10,000$; or, $N \geq 500$ if $\delta \leq 0.25$. Moreover, OLS has a lower RMSE (median absolute bias) for all reliability ratios if $N \geq 1,000$ ($N \geq 50,000$) even if $\delta = 0.55$.

The experiments to this point show that theoretical result favoring OLS is practical as well, applying to situations likely to be common in applied research. Specifically, OLS outperforms IV in terms of RMSE even when all observations suffer from at least some measurement error, the instrument is strong according to the conventional rule-of-thumb, and the sample size is as small as $N = 500$ if $\delta \leq 0.50$. Thus, at least in these simulations, the assumption of no measurement error in the remainder of the sample is not a necessary condition for OLS to have a lower RMSE in practice; nor is a very large sample size. The practical performance of OLS relative to IV differs from the theoretical results in Section 3 due to the fact that convergence for $\hat{\beta}_{IV}$ is slower and a larger sample is needed for the normal approximation to be accurate when the first-stage parameter π is small relative to its sampling variability (e.g., Andrews *et al.*, 2019).

[4]Formally, we set $\pi = \left[(F-1)(\sigma_\omega^2 + \sigma_\varphi^2)/\sum_i z_i^2\right]^{1/2}$, where F is the desired F-statistic, $\sigma_\varphi^2 = \lambda\sigma_1^2 + (1-\lambda)\sigma_2^2$, and $\lambda = \lfloor N^\delta \rfloor / N$. Note, Millimet (2015) shows that the (absolute) finite sample bias of IV is not monotonically decreasing in the reliability ratio. But, the finite sample performance of IV does improve monotonically with instrument strength. The asymptotic distribution of IV may differ from Theorem 2 under weak IV asymptotics (Staiger and Stock, 1997).

[5]Note, the figures censor the difference in RMSE values at ten. When this occurs, the vertical axis is labelled "> 10".

In Figures 3 and A.3, $F \approx 100$, close to the suggested benchmark in Lee *et al.* (2022). Here, the relative performance of OLS depends on δ, N, and the reliability ratios. When $\delta < 0.50$, OLS has a lower RMSE than IV when $N = 100,000$ in all cases except when $RR = \{0.95, 0.25\}$; a lower median absolute bias except when $RR = 0.95$ in the subsample of observations $i = \lfloor N^{\delta} \rfloor + 1, ..., N$. Not unsurprisingly, the relative performance of IV improves as N declines. However, OLS still has a lower RMSE than IV when $N = 10,000$ as long as $RR \geq 0.975$ in the subsample of observations $i = \lfloor N^{\delta} \rfloor + 1, ..., N$; a lower median absolute bias in seven (of 12) cases with less severe measurement error. Even with $N = 1,000$ OLS has a lower RMSE in all but five (of 12) cases ($RR = \{1, 0.25\}, \{0.99, 0.25\}, \{0.975, 0.25\}, \{0.95, 0.25\}$, and $\{0.95, 0.50\}$); a lower median absolute bias in four (of 12) cases with less severe measurement error. Thus, even with a very strong instrument, OLS should not be automatically abandoned in many practical situations.

For the Pre-test estimator, we find little advantage relative to OLS. When $F \approx 10$ or 20, OLS has a smaller RMSE and median absolute bias than the Pre-test estimator in all cases with $\delta \leq 0.55$ (Figures A.4 – A.5 and A.7 – A.8). When $F \approx 100$, OLS has a smaller RMSE and median absolute bias in nearly all cases with $\delta \leq 0.5$ and $N \geq 5,000$ (Figures A.6 and A.9). However, there is some advantage of the Pre-test estimator relative to IV. The Pre-test estimator has a lower RMSE and median absolute bias with $\delta < 0.5$ for all reliability ratios when $F \lesssim 20$ and $N \geq 1,000$ (Figures A.10 – A.11 and A.13 – A.14). When $F \approx 100$, the relative performances are less sensitive to δ and more sensitive to the severity of the measurement error (Figures A.12 and A.15). As a result, there is some benefit to considering the Pre-test estimator over IV (although obtaining appropriate standard errors is more difficult). But, in these cases, OLS is preferable to the Pre-test estimator in moderate or large samples.

Lastly, Tables 1 – 3 display empirical rejection rates at the $p < 0.05$ significance level for the Hausman test comparing OLS and IV for select values of δ. A few results stand out. First, when $\delta < 0.5$ and $RR = 1$ in the subsample of observations $i = \lfloor N^{\delta} \rfloor + 1, ..., N$, then OLS and IV are both consistent and we would expect the null of equality to be rejected in 5% of cases as $N \to \infty$ if the test is appropriately sized. In these cases, when the instrument is strong ($F \approx 100$), the rejection rates are roughly 4%. When $\delta = 0.5$ in these cases, the rejection rates are very close to 5%. However, the test becomes under-sized with rejection rates roughly 2% and 1% as the instrument weakens ($F \approx 20$ and 10, respectively).

Second, for the cases where OLS and IV are both consistent, but N is small, the test over-rejects when the instrument is strong (F-statistic is approximately 100) and under-rejects in most other instances. The over-rejection when the instrument is strong primarily occurs when the measurement error is severe ($RR = 0.25$ in the subsample of observations $i = 1, ..., \lfloor N^{\delta} \rfloor$). This is not surprising as the finite sample bias of OLS in this situation is large and the small N leads to results that differ from what one expects asymptotically.

Third, as we deviate from the conditions required by the asymptotic theory and allow for some measurement error in the entire sample, but $\delta < 0.5$ continues, OLS is no longer consistent and the rejection rate represents the power of the test. In these cases, the rejection rates increase two to threefold regardless of the F-statistic relative to the cases where $RR = 1$ in the subsample of observations $i = \lfloor N^{\delta} \rfloor + 1, ..., N$, but this corresponds to relatively low power. However, as discussed previously, OLS does have a lower RMSE in the vast majority of these cases.

Finally, when $\delta \geq 0.5$, the rejection rates also correspond to the power of the test. When the instrument is strong ($F \approx 100$) and $RR = 1$ in the subsample of observations $i = \lfloor N^\delta \rfloor + 1, ..., N$, power declines with N and is only relatively high when the measurement error is severe ($RR = 0.25$ in the subsample of observations $i = 1, ..., \lfloor N^\delta \rfloor$). When $\delta = 0.75$, the rejection rate is 99.6% in this case when $N = 500$ and 67.8% when $N = 100,000$. As the instrument weakens, power diminishes but the general pattern remains. When $\delta = 0.75$ and $RR < 1$ in the subsample of observations $i = \lfloor N^\delta \rfloor + 1, ..., N$, then power increases in all cases. For instance, the rejection rate is 83.7% when $N = 100,000$, $RR = (0.95, 0.25)$, and the instrument is strong ($F \approx 100$). This falls to 15.7% and 4.7% when the F-statistic is roughly 20 and 10, respectively.

Overall, these results suggests that the Hausman test can be a useful specification test in the context considered here. It has roughly correct size in the presence of a large sample and strong instrument. The test also has good power when the instrument is strong and the measurement error is strong and severe. In other cases, the power is low. However, the empirical RMSE of OLS is lower than IV in the majority of these situations.

5. Application

To provide an example where the above theoretical and simulation results may be relevant, we re-visit the literature on the labor market effects of obesity. Since Register and Williams (1990) researchers have examined the empirical relationship between measures of *body composition* such as weight, height, body mass index (BMI), obesity, and body fat and *economic outcomes* such as hourly wages, employment, and household income. Additional studies have examined the effects of body composition on health outcomes and health care utilization. While these studies are often undertaken using data from the United States, current research includes data from other countries as well.

Interest in this topic, particularly in the United States, stems from the dramatic rise in obesity over the past few decades. Using data from the National Health and Nutrition Examination Survey (NHANES) over the 2017-2020 pre-Covid period, 41.9% of adults age 20 and over are classified as obese (defined as BMI, equal to weight in kilograms divided by height in meters squared, exceeding $30\mathtt{km/m^2}$) and 9.2% as severely obese (defined as BMI exceeding $40\mathtt{km/m^2}$), up from 22.9% and 2.9% in 1988-1994, respectively (Stierman *et al.*, 2021; National Center for Health Statistics, 2019). In light of this, understanding the economic consequences of obesity is critical.

However, identifying the causal effect of body composition on labor market outcomes is not trivial for several reasons. First, and most important for our purposes here, information on body composition is self-reported in most data sources. As a result, measurement error is well known to exist (Cawley, 2004; Stommel and Schoenborn, 2002; O'Neill and Sweetman, 2013; Cawley *et al.*, 2015; Flegal *et al.*, 2019). Second, body composition may be endogenous for reasons other than measurement error. Two possibilities are reverse causation and omitted heterogeneity. Reverse causation may arise if individuals with poor labor market outcomes cannot afford a healthy diet. Omitted heterogeneity may arise, for example, due to unobserved individual discount rates. The concern is that myopic individuals invest less in their health as well as other forms of human capital, leading to a spurious correlation. If there is a causal effect of body composition on labor market outcomes, it is hypothesized to be due to the negative effects of obesity on productivity, employer or customer discrimination, or employers recovering some of the higher cost of providing health insurance (e.g., Cawley, 2004).

Existing studies attempt to overcome these challenges primarily using several tactics. First, use self-reported data on height and weight to predict 'true' body composition using auxiliary data containing both self-reported and measured values (Plankey *et al.*, 1997; Cawley, 2004; Wada and Tekin, 2010; O'Neill and Sweetman, 2013). Second, use lagged (over many years) measures of body composition (Cawley, 2004; Wada and Tekin, 2010). Third, use panel data over time and/or across siblings to include individual and/or family fixed effects (Baum II and Ford, 2004; Cawley, 2004; Han *et al.*, 2009; Wada and Tekin, 2010). Fourth, use instrumental variables exploiting the large genetic component of obesity (Cawley, 2004; Cawley and Lillard, 2005; Brunello and D'Hombres, 2007; Greve, 2008; Norton and Han, 2008; Atella *et al.*, 2008; Kline and Tobias, 2008; Lindeboom *et al.*, 2010; Wada and Tekin, 2010; O'Neill and Sweetman, 2013).

At the risk of oversimplifying, these prior studies establish the following results. First, self-reported body composition contains measurement error. Data containing both self-reported and measured body composition confirm the existence of errors. Second, using lagged measures of body composition does not qualitatively alter the association between body composition and labor market outcomes. This is consistent with the reverse causation concern being unwarranted.[6] Third, fixed effects models generally lead to smaller associations between body composition and labor market outcomes. This is consistent with the presence of omitted heterogeneity. It is also consistent with greater attenuation bias from classical measurement error in fixed effects models (Griliches, 1979). Fourth, IV estimates are mixed and volatile. In these studies, the first-stage F-statistics range from approximately 5 to 45 and standard errors are typically 5 to 10 times larger than for the corresponding OLS estimates. As a result, many of these studies fail to find statistical evidence of endogeneity. Finally, there seems to be robust evidence of a wage penalty for obese women, particularly for white women in the United States. Evidence for other groups is less conclusive.

Interpreting these findings in light of our theoretical and simulation results suggests that measurement error should perhaps be the primary econometric concern given the availability of a large set of demographic and economic controls in most data sources. However, IV may not be the preferred solution. One might have more confidence in the OLS estimates if measurement error is strong only for a subsample of $\lfloor N^{\delta} \rfloor$ observations, where $\delta \lesssim 0.50$. Prior studies suggest that self-reported BMI fits into the framework discussed in Section 3. For example, using data from NHANES 2001–2006, Stommel and Schoenborn (2002, p. 4) concludes, "Generally, deviations of BMI values based on self-reported height and weight from BMI values based on measured height and weight are moderate: an estimated 56% have self-reported BMI values within a one-unit interval of their measured BMI, and 81.5% have self-reported BMI values within two units of their measured BMI." O'Neill and Sweetman (2013) Cawley *et al.* (2015) use data from NHANES 1988–1994 and 2003–2010, respectively, and show similar results.

To undertake our own analysis, we use data from the NHANES 2017–2018 survey. NHANES is conducted under the auspices of the National Center for Health Statistics (NCHS) which is housed in the Centers for Disease Control and Prevention (CDC). While NHANES began in the 1960s, the survey has been administered continuously since 1999. Data on demographic, socioeconomic, dietary, and health-related topics are collected from a nationally representative sample of roughly 5,000 individuals each year. NHANES also contains a medical examination and laboratory tests. Information on weight and height are self-reported in the survey, but also collected as part of the

[6] It is also consistent with a high degree of persistence in both body composition and labor market outcomes.

medical examination. Thus, NHANES is one of the few data sets with both *self-reported* and *measured* body composition.

Panels A and C in Figure 4 show the distribution of the measurement errors and its relation to measured BMI for the full sample of individuals aged 16 and older with information on both self-reported and measured BMI ($N = 5,498$). The distribution is reasonably clustered around zero, with the center of the distribution shifted to the left indicative of small under-reporting of BMI on average. If we split the sample into the $\lfloor N^{0.5} \rfloor$ observations with the smallest errors in absolute value and the remainder, reliability ratio is 0.95 in the former and 0.37 in the latter. Panels B and D in Figure 4 restrict the data to our estimation sample. This includes individuals aged 25 to 80 with complete information on both self-reported and measured BMI and covariates (discussed below). We exclude women who are pregnant as in Cawley (2004) and Cawley *et al.* (2015). The final sample size is 3,008. Again splitting the sample into the $\lfloor N^{0.5} \rfloor$ observations with the smallest errors in absolute value and the remainder, the reliability ratio is 0.95 in the former and 0.51 in the latter.

This illustrates the main point we wish to make with this example. The distribution of errors for BMI in the United States aligns with cases in the simulations where OLS is found to be superior to IV. With $\delta \approx 0.5$, $RR = \{0.95, 0.51\}$, $N = 3,000$, and first-stage F-statistics for common instruments in the literature around 20, OLS has a smaller median absolute bias and RMSE than IV. There are surely many applications where measurement error is suspected and likely to match this distribution. For example, Auffhammer (2018) discusses measurement error in weather conditions, where the errors are likely small (and classical) near monitoring stations but become large at greater distances. Zivin and Neidell (2013) discusses a similar problem when individual exposure to pollution is obtained using the nearest monitoring station.

The prior literature finds that measurement error in BMI is nonclassical in that it is negatively correlated with measured BMI. Indeed, in the full sample, the (weighted) measurement errors and measured BMI negatively correlated (correlation = -0.35). In the estimation sample, the (weighted) measurement errors and measured BMI remain negatively correlated (correlation = -0.32). The distributions are similar in Figure 5 when we split the estimation sample by gender and age. Importantly, the theoretical derivations in Section 3 show that the asymptotic results do not depend on $Cov[x^*, \eta] = 0$.

Interestingly, while the asymptotic results do not depend on $Cov[x^*, \eta] = 0$, the finite sample properties of OLS do. However, the negative correlation between measured BMI and the measurement error *reduces* the finite sample bias of OLS, providing an additional reason to favor OLS over IV in this application. After partialling out the other covariates listed above,

$$E[\hat{\beta}_{LS}] = \beta \left(1 - \frac{\hat{M}_{\tilde{x}^*\tilde{\eta}} + \hat{M}_{\tilde{\eta}\tilde{\eta}}}{\hat{M}_{\tilde{x}\tilde{x}}} \right), \tag{5.1}$$

where $\tilde{x}^*$ is measured BMI net of other covariates, $\tilde{\eta}$ is the difference between self-reported BMI net of other covariates and measured BMI net of other covariates, and $\hat{M}$ is defined in Section 3. In the estimation sample, $\hat{M}_{\tilde{x}^*\tilde{\eta}} = -4.17$, $\hat{M}_{\tilde{\eta}\tilde{\eta}} = 3.36$, and $\hat{M}_{\tilde{x}\tilde{x}} = 41.68$, implying that the term in parentheses equals 1.02. It would decline to 0.92 if $\hat{M}_{\tilde{x}^*\tilde{\eta}} = 0$ *ceteris paribus*, leading to a larger bias in absolute value. Thus, the nonclassical nature of the measurement error *strengthens* our conclusion that OLS is preferable.[7]

[7] Wile it is not relevant for the analysis here, it is noteworthy that any dependence between the measurement error and the other covariates included in the model is small. An OLS regression of the error on a vector of covariates

To examine regression results with this data, we assess the effect of BMI on log monthly family income as in O'Neill and Sweetman (2013). As the responses are coded in intervals, we use the mid-point of each interval and 1.5 times the lower boundary of the upper interval and treat income as continuous. For comparison, we also draw income values from a uniform distribution within each interval.[8] We estimate the effects of *self-reported* and *measured* BMI, along with the covariates listed in footnote 7, on monthly family income using OLS. We also estimate the effect of self-reported BMI using IV. As an instrument, we follow Greve (2008) and use a measure of family diabetes history. Specifically, we generate a binary variable equal to one if the sample individual indicates that they are at an elevated risk of diabetes due to family history, zero otherwise. As noted in Stommel and Schoenborn (2002) and elsewhere, the risk of diabetes rises with BMI.

Table 4 presents the results for the pooled estimation sample. Table 5 gives the results for subsamples based on gender and age.[9] In the pooled sample, the first-stage F-statistic is 76.8. In the subsamples delineated by gender and age, the first-stage F-statistics range from 1.2 to 58.1. In both tables, the OLS estimates using self-reported *and* measured BMI are virtually identical, indicating a statistically significant wage penalty for women aged 25 to 54. In contrast, the IV results are never statistically significant except for males aged 25 to 54, where the coefficient is positive. The IV standard errors are between roughly 3 and 18 times larger than the OLS standard errors. Moreover, Hausman tests comparing the OLS and IV estimates fail to reject the null hypothesis of $\delta < 1/2$ in all cases, with the p-values approximately one in all cases. These findings are consistent with previous papers deriving instruments for BMI based on genetics. However, our analysis gives us more confidence in the OLS estimates even if we did not have a benchmark using measured BMI in which to compare them.

6. Conclusion

Issues of mis-specification in econometric models are not black and white. There are degrees of mis-specification. However, the current practice in empirical research is to chooses the econometric estimator appropriate for the most problematic of observations. That 'appropriate' estimator, at least in the contexts considered here and in Pesaran and Zhou (2018), requires more of the data such that the cure may be worse than disease. This is precisely the case when only a portion of the sample suffers from substantial classical measurement error. IV, the typical solution when *any* measurement error is suspected, performs much worse than OLS as measured by mean squared error and median absolute bias as the sample size grows. But, this also holds for sample sizes now commonplace in empirical research, cases when instruments are (very) strong by conventional standards, and in situations where all observations suffer from at least some measurement error. Examining data where both self-reported and measured BMI are available, this bears out in practice. While OLS estimates using self-reported BMI are virtually identical to those using measured BMI, IV estimates using

including a quadratic in age, indicators for different education categories, indicators for different marital statuses, gender, indicators for race, indicators for different family sizes, indicators for different numbers of children less than six years old in the household, and indicators for different numbers of children between six and 17 years old in the household, R^2 is only 0.03. A similar result is found in Stommel and Schoenborn (2002) where the R^2 is only 0.045. The authors control for household income in addition to the covariates included here, as well as pregnancy status since they do not exclude pregnant women.

[8] In this case, we use the lower boundary and two times the lower boundary as the upper interval.

[9] We only show the results using monthly income created using the mid-points of the intervals given the similarity of the results.

self-reported BMI and a strong instrument based on genetics are imprecise and often very different. As such, researchers would do well to approach potential mis-specification in a more nuanced way.

References

Andrews, I., Stock, J. H. and Sun, L. (2019). Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics*, **11** (1), 727–753.

Atella, V., Pace, N. and Vuri, D. (2008). Are employers discriminating with respect to weight?: European evidence using quantile regression. *Economics Human Biology*, **6** (3), 305–329.

Auffhammer, M. (2018). Quantifying economic damages from climate change. *The Journal of Economic Perspectives*, **32** (4), 33–52.

Baum II, C. L. and Ford, W. F. (2004). The wage effects of obesity: a longitudinal study. *Health Economics*, **13** (9), 885–899.

Brunello, G. and D'Hombres, B. (2007). Does body weight affect wages?: Evidence from europe. *Economics Human Biology*, **5** (1), 1–19.

Cawley, J. (2004). The impact of obesity on wages. *Journal of Human Resources*, **39** (2), 451–474.

—, Maclean, J., Hammer, M. and Wintfeld, N. (2015). Reporting error in weight and its implications for bias in economic models. *Economics & Human Biology*, **19**, 27–44.

Cawley, M. M., John Grabka and Lillard, D. R. (2005). A comparison of the relationship between obesity and earnings in the U.S. and Germany. *Schmollers Jahrbuch*, **125**, 119–129.

Fisher, F. M. (1961). On the cost of approximate specification in simultaneous equation estimation. *Econometrica*, **29** (2), 139–170.

Flegal, K. M., Ogden, C. L., Fryar, C., Afful, J., Klein, R. and Huang, D. T. (2019). Comparisons of self-reported and measured height and weight, bmi, and obesity prevalence from national surveys: 1999-2016. *Obesity*, **27** (10), 1711–1719.

Greve, J. (2008). Obesity and labor market outcomes in denmark. *Economics & Human Biology*, **6** (3), 350–362.

Griliches, Z. (1979). Sibling models and data in economics: Beginnings of a survey. *Journal of Political Economy*, **87** (5), S37–S64.

Han, E., Norton, E. C. and Stearns, S. C. (2009). Weight and wages: fat versus lean paychecks. *Health Economics*, **18** (5), 535–548.

Kline, B. and Tobias, J. L. (2008). The wages of bmi: Bayesian analysis of a skewed treatment–response model with nonparametric endogeneity. *Journal of Applied Econometrics*, **23** (6), 767–793.

Lee, D. S., McCrary, J., Moreira, M. J. and Porter, J. R. (2022). Valid t-ratio inference for IV. *American Economic Review*, **112** (10), 3260–3290.

Lindeboom, M., Lundborg, P. and van der Klaauw, B. (2010). Assessing the impact of obesity on labor market outcomes. *Economics & Human Biology*, **8** (3), 309–319.

Millimet, D. L. (2015). Covariate measurement and endogeneity. *Economics Letters*, **136** (C), 59–63.

National Center for Health Statistics (2019). *Health, United States, 2019: Table 026*. Tech. rep.

Norton, E. C. and Han, E. (2008). Genetic information, obesity, and labor market outcomes. *Health Economics*, **17** (9), 1089–1104.

O'Neill, D. and Sweetman, O. (2013). The consequences of measurement error when estimating the impact of obesity on income. *IZA Journal of Labor Economics*, **2** (1), 3–22.

Pesaran, M. H. and Zhou, Q. (2018). To pool or not to pool: revisited. *Oxford Bulletin of Economics and Statistics*, **80** (2), 185–217.

Plankey, M. W., Stevens, J., Fiegal, K. M. and Rust, P. F. (1997). Prediction equations do not eliminate systematic error in self-reported body mass index. *Obesity Research*, **5** (4), 308–314.

Register, C. A. and Williams, D. R. (1990). Wage effects of obesity among young workers. *Social Science Quarterly*, **71** (1), 130–141.

Schennach, S. M. (2016). Recent advances in the measurement error literature. *Annual Review of Economics*, **8** (1), 341–377.

— (2020). Mismeasured and unobserved variables. In S. N. Durlauf, L. P. Hansen, J. J. Heckman and R. L. Matzkin (eds.), *Handbook of Econometrics*, vol. 7A, Elsevier, pp. 487–565.

Staiger, D. and Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica*, **65** (3), 557–586.

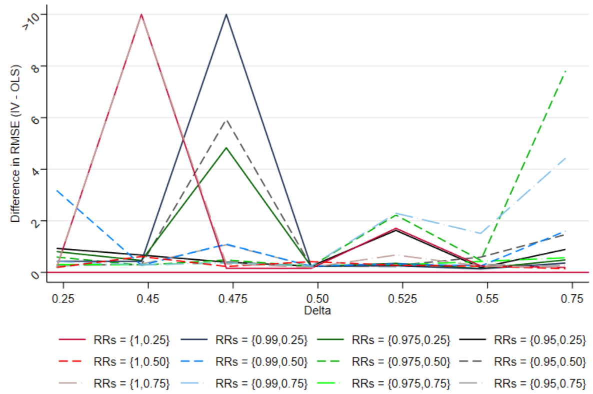
Stierman, B., Afful, J., Carroll, M. D., Chen, T.-C., Davy, O., Fink, S., Fryar, C. D., Gu, Q., Hales, C. M., Hughes, J. P., Ostchega, Y., Storandt, R. J. and Akinbami, L. J. (2021). *National Health and Nutrition Examination Survey 2017–March 2020 Prepandemic Data Files Development of Files and Prevalence Estimates for Selected Health Outcomes*. Tech. rep., National Center for Health Statistics (U.S.).

Stock, J. H., Wright, J. H. and Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, **20** (4), 518–529.

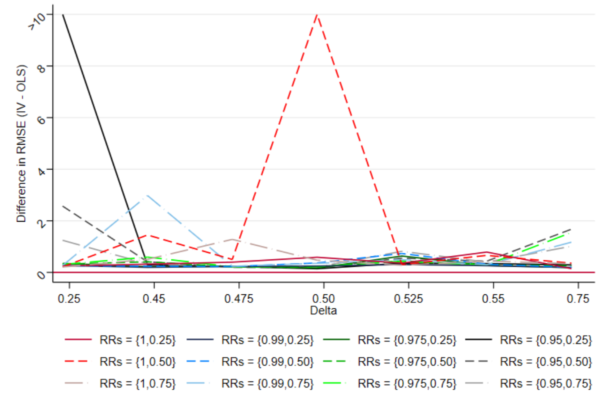
Stommel, M. and Schoenborn, C. A. (2002). Accuracy and usefulness of bmi measures based on self-reported weight and height: findings from the nhanes & nhis 2001-2006. *BMC Public Health*, **9** (1), 421–430.

Wada, R. and Tekin, E. (2010). Body composition and wages. *Economics & Human Biology*, **8** (2), 242–254.

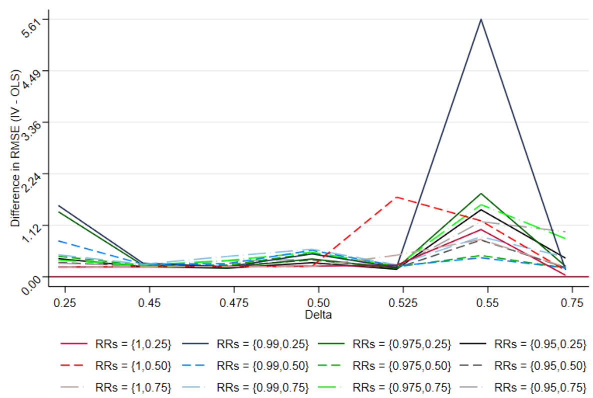
Zivin, J. G. and Neidell, M. (2013). Environment, health, and human capital. *Journal of Economic Literature*, **51** (3), 689–730.



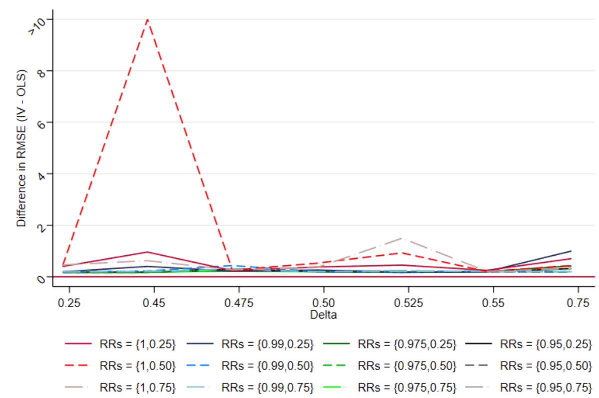
(A) $N = 500$



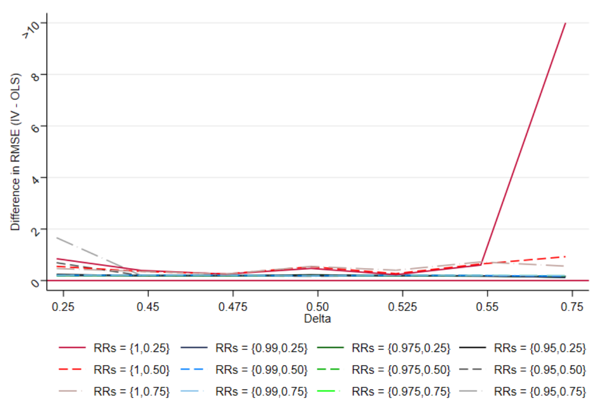
(B) $N = 1,000$



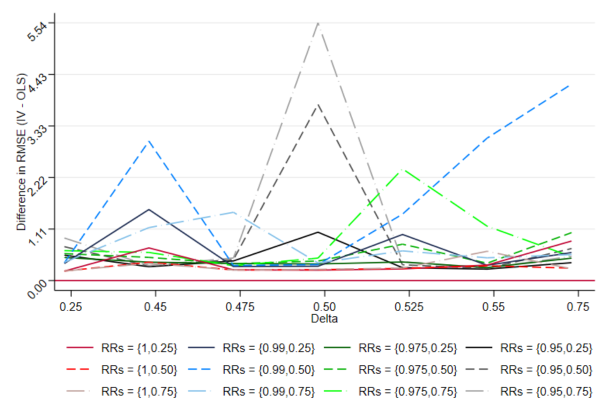
(C) $N = 5,000$



(D) $N = 10,000$



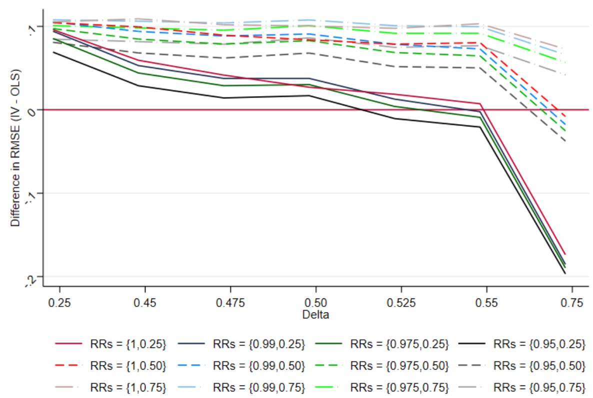
(E) $N = 50,000$



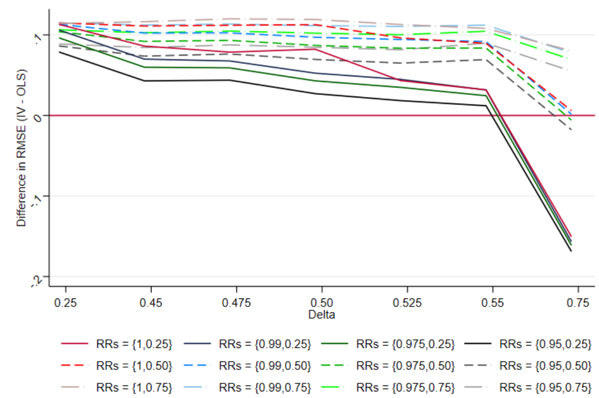
(F) $N = 100,000$

Figure 1. Simulation Results: Root Mean Squared Error (F-statistic ≈ 10).

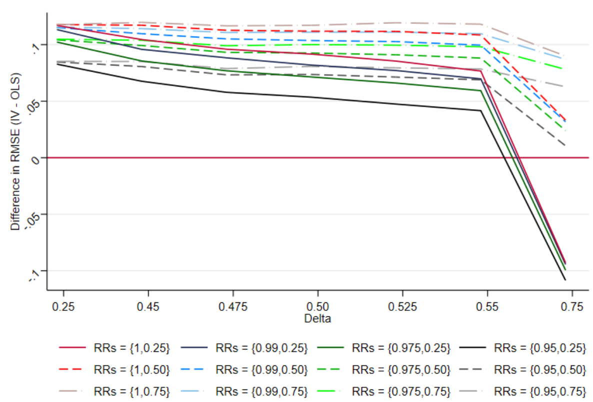
Notes: RR = reliability ratio. First value applies to observations $i = \lfloor N^{\delta} \rfloor + 1, ..., N$; second value applies to observations $i = 1, ..., \lfloor N^{\delta} \rfloor$. See text for more details.



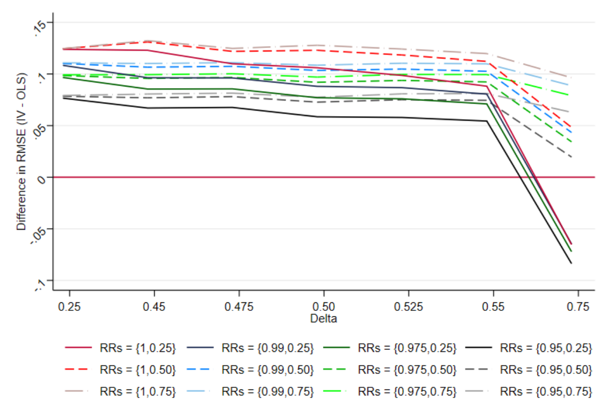
(A) $N = 500$



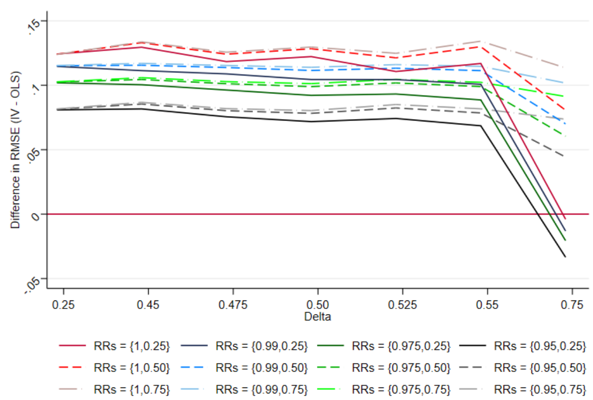
(B) $N = 1,000$



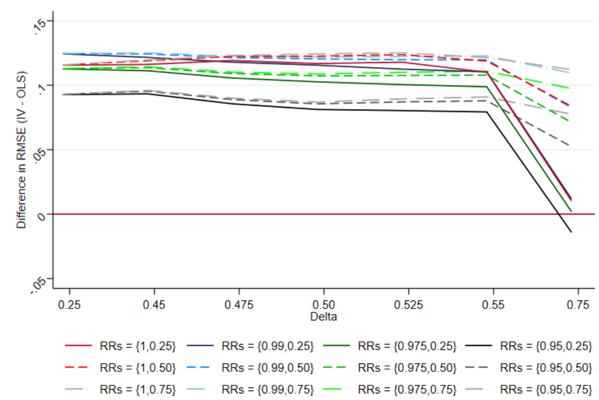
(C) $N = 5,000$



(D) $N = 10,000$



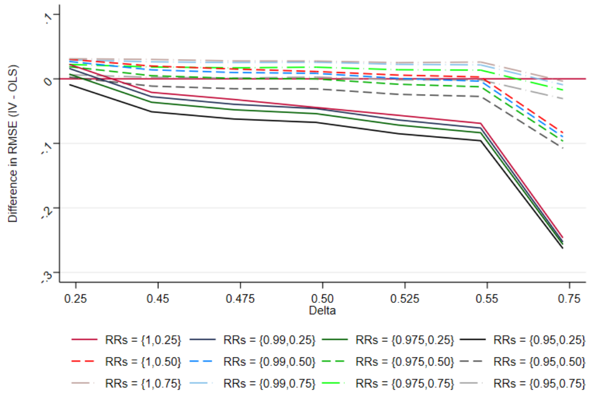
(E) $N = 50,000$



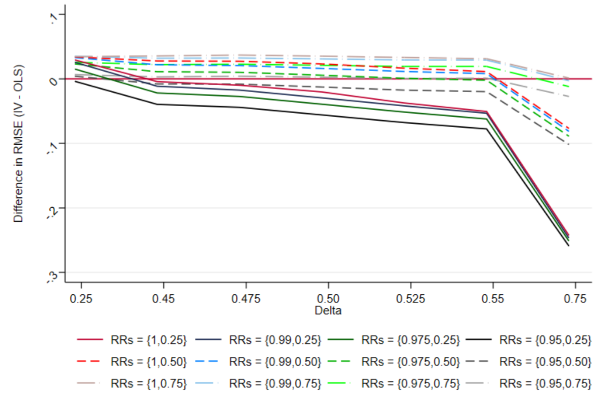
(F) $N = 100,000$

FIGURE 2. Simulation Results: Root Mean Squared Error (F-statistic ≈ 20).

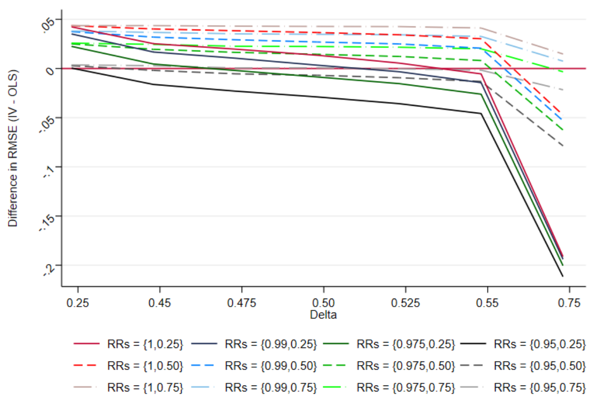
Notes: See Figure 1 for details.



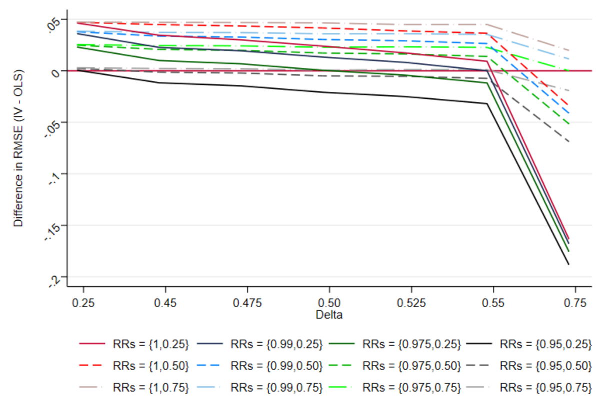
(A) $N = 500$



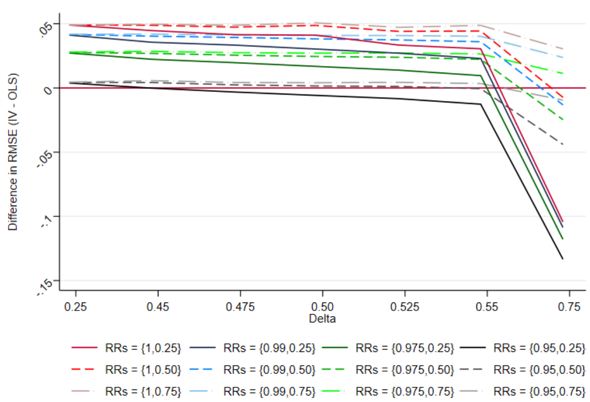
(B) $N = 1,000$



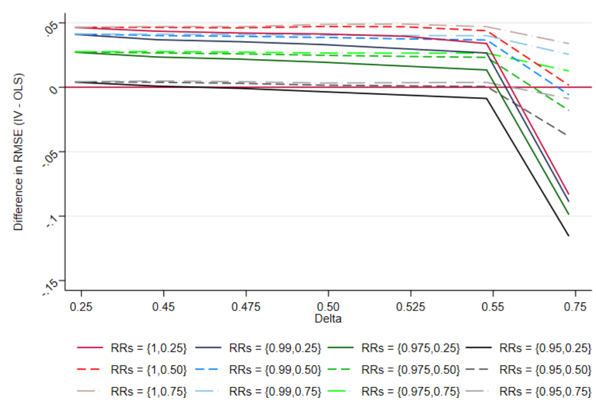
(C) $N = 5,000$



(D) $N = 10,000$



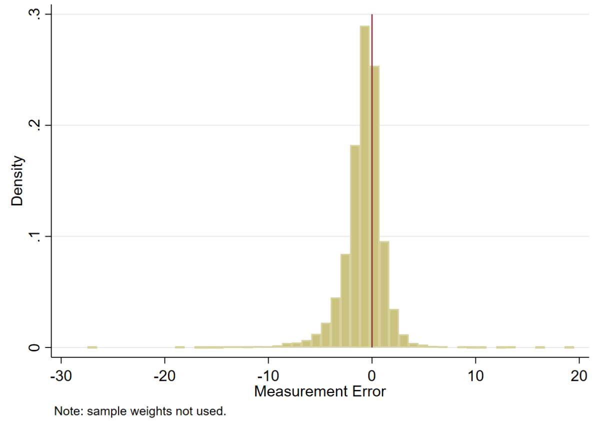
(E) $N = 50,000$



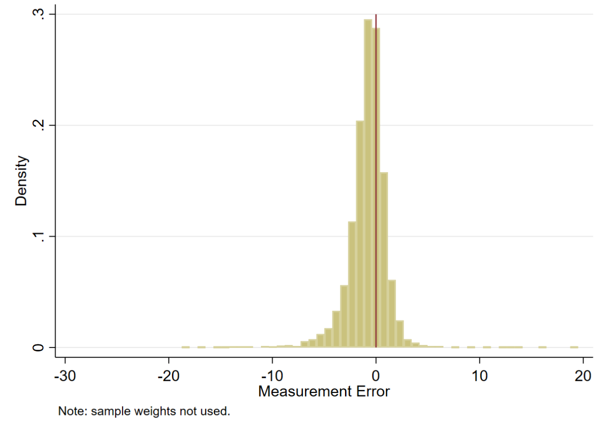
(F) $N = 100,000$

Figure 3. Simulation Results: Root Mean Squared Error (F-statistic ≈ 100).

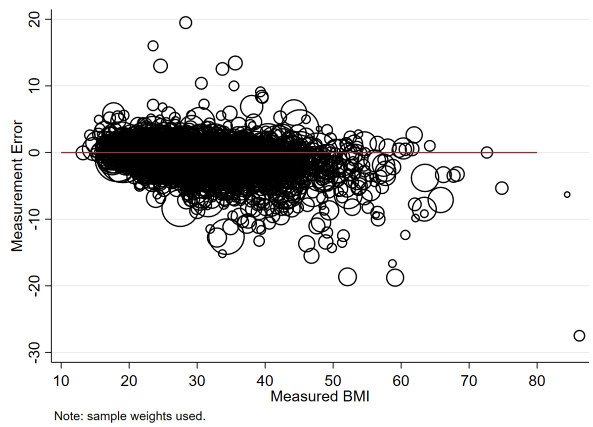
Notes: See Figure 1 for details.



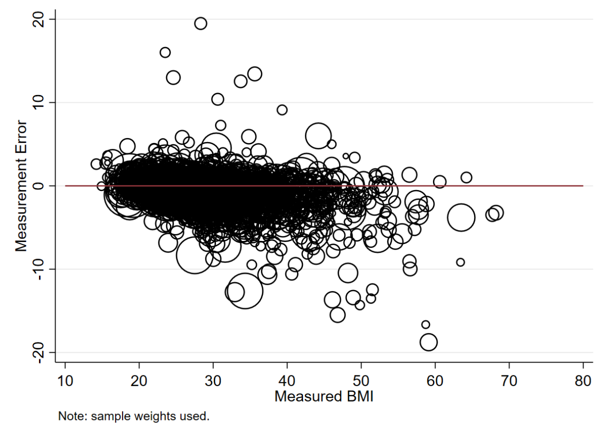
(A) Full Sample



(B) Estimation Sample



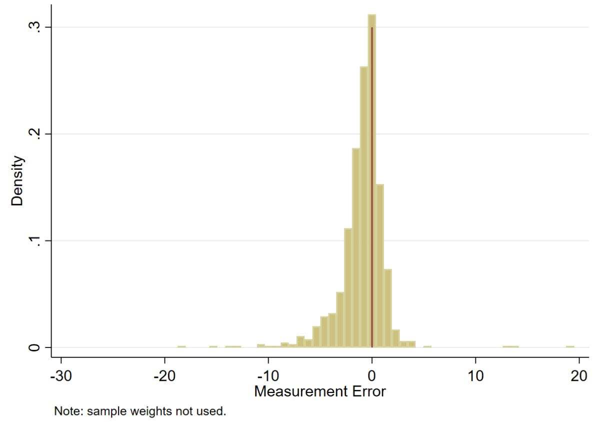
(C) Full Sample



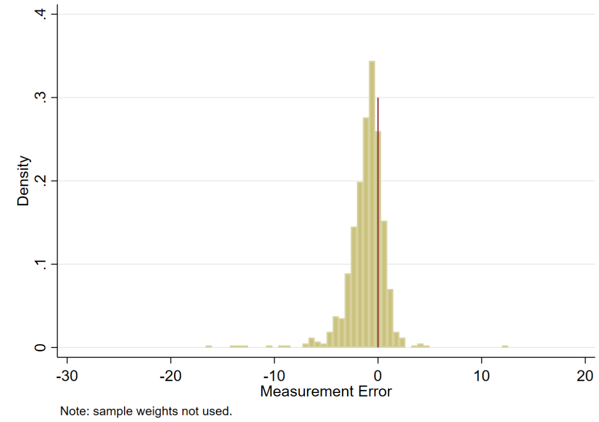
(D) Estimation Sample

FIGURE 4. Measurement Error Distribution for Self-Reported BMI: NHANES 2017-2018.

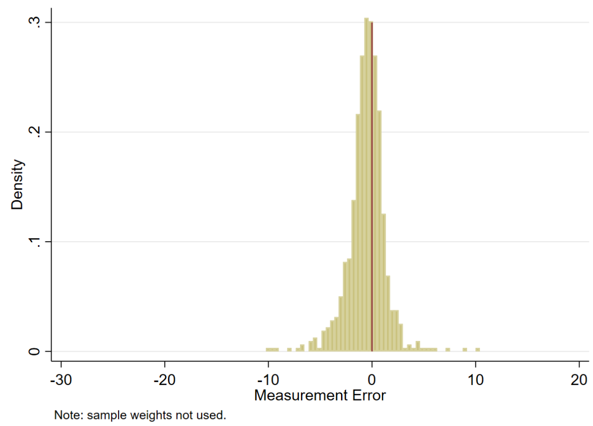
Notes: In Panels (A) and (C) $N = 5,498$. In Panels (B) and (D) $N = 3,008$.



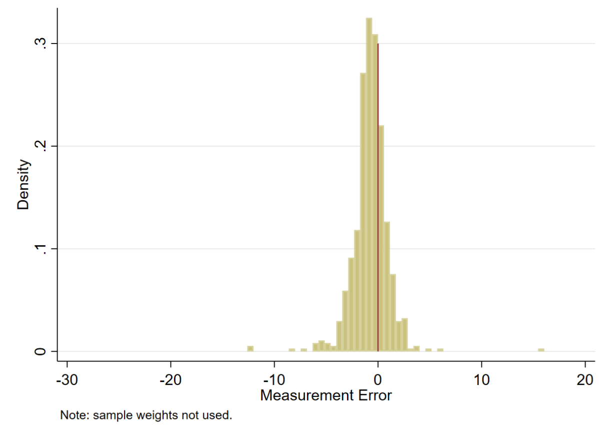
(A) Female, Ages 25-55



(B) Female, Ages 55-80



(C) Male, Ages 25-55



(D) Male, Ages 55-80

FIGURE 5. Measurement Error Distribution for Self-Reported BMI by Gender & Age: NHANES 2017-2018

Notes: See text for details.

TABLE 1. Empirical Hausman Rejection Rates: IV (F-statistic ≈ 10)

		RR = 0.95			RR = 0.975			RR = 0.99			RR = 1		
N	δ	RR = 0.25	RR = 0.50	RR = 0.75	RR = 0.25	RR = 0.50	RR = 0.75	RR = 0.25	RR = 0.50	RR = 0.75	RR = 0.25	RR = 0.50	RR = 0.75
500	0.25	0.016	0.013	0.012	0.013	0.013	0.013	0.013	0.013	0.013	0.015	0.015	0.018
	0.45	0.034	0.020	0.018	0.028	0.018	0.016	0.026	0.018	0.013	0.026	0.013	0.013
	0.50	0.041	0.023	0.019	0.037	0.020	0.015	0.035	0.018	0.016	0.035	0.019	0.019
	0.75	0.191	0.057	0.024	0.185	0.052	0.021	0.179	0.048	0.018	0.188	0.041	0.015
1,000	0.25	0.014	0.015	0.015	0.015	0.011	0.012	0.015	0.012	0.011	0.007	0.008	0.009
	0.45	0.032	0.024	0.021	0.027	0.022	0.017	0.023	0.017	0.017	0.017	0.012	0.010
	0.50	0.037	0.025	0.025	0.033	0.021	0.019	0.032	0.018	0.019	0.025	0.022	0.017
	0.75	0.184	0.050	0.022	0.177	0.042	0.019	0.173	0.038	0.018	0.146	0.032	0.014
5,000	0.25	0.009	0.008	0.008	0.008	0.007	0.006	0.007	0.008	0.008	0.017	0.018	0.017
	0.45	0.017	0.014	0.014	0.014	0.012	0.011	0.012	0.010	0.010	0.013	0.013	0.011
	0.50	0.012	0.009	0.008	0.009	0.009	0.007	0.009	0.009	0.006	0.012	0.011	0.011
	0.75	0.094	0.021	0.013	0.084	0.017	0.009	0.086	0.016	0.008	0.092	0.026	0.008
10,000	0.25	0.012	0.011	0.011	0.008	0.008	0.008	0.009	0.009	0.010	0.012	0.012	0.012
	0.45	0.015	0.013	0.013	0.012	0.010	0.010	0.010	0.008	0.010	0.012	0.011	0.012
	0.50	0.015	0.012	0.012	0.014	0.012	0.012	0.012	0.008	0.007	0.013	0.013	0.016
	0.75	0.080	0.026	0.013	0.071	0.023	0.010	0.067	0.022	0.010	0.069	0.018	0.012
50,000	0.25	0.013	0.013	0.013	0.012	0.011	0.011	0.020	0.019	0.020	0.006	0.006	0.006
	0.45	0.014	0.017	0.017	0.013	0.016	0.016	0.015	0.017	0.018	0.009	0.009	0.009
	0.50	0.020	0.018	0.016	0.016	0.018	0.018	0.019	0.019	0.018	0.026	0.025	0.024
	0.75	0.059	0.029	0.018	0.051	0.029	0.019	0.050	0.026	0.015	0.040	0.018	0.015
100,000	0.25	0.011	0.011	0.011	0.013	0.013	0.013	0.013	0.013	0.013	0.005	0.006	0.006
	0.45	0.018	0.018	0.018	0.015	0.016	0.016	0.015	0.016	0.016	0.008	0.009	0.009
	0.50	0.014	0.013	0.014	0.009	0.012	0.012	0.011	0.014	0.014	0.012	0.013	0.012
	0.75	0.047	0.019	0.019	0.040	0.017	0.015	0.040	0.015	0.015	0.034	0.013	0.011

Notes: Empirical rejection rates at the $p < 0.05$ significance level for select values of δ.

TABLE 2. Empirical Hausman Rejection Rates: IV (F-statistic ≈ 20)

		RR = 0.95			RR = 0.975			RR = 0.99			RR = 1		
N	δ	RR = 0.25	RR = 0.50	RR = 0.75	RR = 0.25	RR = 0.50	RR = 0.75	RR = 0.25	RR = 0.50	RR = 0.75	RR = 0.25	RR = 0.50	RR = 0.75
500	0.25	0.046	0.038	0.035	0.036	0.030	0.023	0.032	0.029	0.029	0.029	0.034	0.037
	0.45	0.101	0.061	0.047	0.081	0.044	0.033	0.075	0.034	0.026	0.064	0.039	0.030
	0.50	0.131	0.071	0.048	0.115	0.056	0.041	0.107	0.049	0.030	0.093	0.040	0.030
	0.75	0.676	0.190	0.062	0.656	0.170	0.046	0.650	0.162	0.043	0.616	0.168	0.044
1,000	0.25	0.049	0.042	0.042	0.036	0.036	0.034	0.039	0.034	0.034	0.018	0.019	0.018
	0.45	0.077	0.050	0.039	0.069	0.034	0.030	0.055	0.036	0.030	0.037	0.029	0.026
	0.50	0.112	0.067	0.057	0.093	0.054	0.040	0.085	0.043	0.036	0.066	0.038	0.030
	0.75	0.588	0.178	0.076	0.569	0.152	0.061	0.551	0.138	0.050	0.538	0.124	0.030
5,000	0.25	0.029	0.027	0.025	0.019	0.019	0.021	0.018	0.020	0.021	0.024	0.025	0.025
	0.45	0.053	0.036	0.033	0.034	0.029	0.029	0.033	0.024	0.022	0.031	0.030	0.029
	0.50	0.047	0.032	0.028	0.031	0.022	0.021	0.025	0.022	0.015	0.035	0.029	0.029
	0.75	0.406	0.121	0.049	0.380	0.096	0.035	0.363	0.083	0.029	0.359	0.071	0.034
10,000	0.25	0.039	0.040	0.040	0.023	0.022	0.021	0.020	0.021	0.020	0.032	0.032	0.032
	0.45	0.044	0.038	0.037	0.035	0.027	0.027	0.031	0.025	0.022	0.032	0.029	0.029
	0.50	0.047	0.040	0.037	0.038	0.032	0.029	0.033	0.031	0.026	0.029	0.029	0.029
	0.75	0.330	0.090	0.044	0.293	0.069	0.034	0.270	0.064	0.027	0.248	0.065	0.023
50,000	0.25	0.034	0.035	0.035	0.027	0.026	0.026	0.032	0.032	0.033	0.020	0.020	0.020
	0.45	0.050	0.046	0.047	0.035	0.036	0.034	0.033	0.033	0.033	0.023	0.021	0.020
	0.50	0.048	0.042	0.040	0.042	0.036	0.035	0.037	0.033	0.033	0.042	0.039	0.038
	0.75	0.228	0.087	0.058	0.195	0.069	0.048	0.176	0.062	0.040	0.152	0.043	0.035
100,000	0.25	0.038	0.038	0.036	0.026	0.026	0.026	0.022	0.022	0.022	0.017	0.017	0.018
	0.45	0.039	0.038	0.042	0.028	0.031	0.031	0.024	0.023	0.025	0.017	0.020	0.020
	0.50	0.034	0.034	0.032	0.028	0.026	0.027	0.030	0.024	0.024	0.022	0.020	0.022
	0.75	0.157	0.067	0.042	0.132	0.054	0.030	0.117	0.041	0.028	0.122	0.041	0.029

Notes: Empirical rejection rates at the $p < 0.05$ significance level for select values of δ.

Table 3. Empirical Hausman Rejection Rates: IV (F-statistic ≈ 100)

		RR = 0.95			RR = 0.975			RR = 0.99			RR = 1		
N	δ	RR = 0.25	RR = 0.50	RR = 0.75	RR = 0.25	RR = 0.50	RR = 0.75	RR = 0.25	RR = 0.50	RR = 0.75	RR = 0.25	RR = 0.50	RR = 0.75
500	0.25	0.211	0.147	0.136	0.122	0.082	0.074	0.084	0.056	0.051	0.079	0.064	0.064
	0.45	0.499	0.251	0.171	0.407	0.159	0.088	0.350	0.112	0.061	0.316	0.085	0.057
	0.50	0.656	0.293	0.171	0.558	0.187	0.098	0.501	0.139	0.076	0.447	0.114	0.056
	0.75	1.000	0.849	0.380	0.999	0.808	0.273	0.999	0.773	0.210	0.996	0.737	0.193
1,000	0.25	0.181	0.154	0.138	0.104	0.083	0.076	0.082	0.064	0.063	0.046	0.037	0.035
	0.45	0.417	0.213	0.162	0.306	0.134	0.092	0.228	0.095	0.062	0.191	0.064	0.052
	0.50	0.530	0.246	0.166	0.404	0.158	0.100	0.332	0.125	0.075	0.309	0.093	0.060
	0.75	1.000	0.798	0.337	1.000	0.737	0.235	1.000	0.701	0.170	0.999	0.672	0.160
5,000	0.25	0.144	0.133	0.127	0.065	0.062	0.058	0.038	0.039	0.039	0.044	0.042	0.040
	0.45	0.252	0.174	0.150	0.134	0.086	0.070	0.090	0.052	0.047	0.074	0.051	0.052
	0.50	0.289	0.188	0.156	0.195	0.092	0.067	0.128	0.049	0.043	0.115	0.050	0.045
	0.75	0.996	0.659	0.266	0.993	0.551	0.169	0.987	0.481	0.123	0.981	0.447	0.098
10,000	0.25	0.141	0.138	0.136	0.076	0.073	0.073	0.042	0.040	0.041	0.056	0.057	0.055
	0.45	0.207	0.155	0.142	0.105	0.074	0.068	0.067	0.051	0.051	0.062	0.051	0.050
	0.50	0.255	0.168	0.147	0.150	0.090	0.070	0.093	0.056	0.048	0.075	0.048	0.044
	0.75	0.989	0.597	0.248	0.979	0.468	0.146	0.968	0.394	0.093	0.961	0.337	0.085
50,000	0.25	0.155	0.156	0.154	0.074	0.075	0.074	0.053	0.054	0.055	0.040	0.039	0.040
	0.45	0.195	0.172	0.164	0.094	0.081	0.080	0.062	0.062	0.060	0.059	0.051	0.049
	0.50	0.214	0.172	0.161	0.102	0.081	0.076	0.072	0.056	0.053	0.060	0.060	0.061
	0.75	0.919	0.454	0.240	0.858	0.330	0.136	0.813	0.273	0.091	0.810	0.185	0.058
100,000	0.25	0.141	0.140	0.140	0.079	0.078	0.077	0.044	0.042	0.041	0.039	0.039	0.039
	0.45	0.151	0.140	0.134	0.081	0.078	0.076	0.052	0.051	0.049	0.037	0.041	0.041
	0.50	0.156	0.130	0.120	0.085	0.067	0.064	0.052	0.051	0.049	0.048	0.044	0.047
	0.75	0.837	0.367	0.201	0.771	0.247	0.107	0.719	0.173	0.067	0.678	0.143	0.061

Notes: Empirical rejection rates at the $p < 0.05$ significance level for select values of δ.

Table 4. Effect of BMI on Log Monthly Family Income

	Income Mid-Point			Uniform Income		
	OLS		IV	OLS		IV
	(1)	(2)	(3)	(4)	(5)	(6)
BMI	-0.002	-0.002	0.001	-0.002	-0.002	-0.000
	(0.002)	(0.002)	(0.013)	(0.002)	(0.002)	(0.014)
Observations	3008	3008	3008	3008	3008	3008
F-stat			76.831			76.831
Hausman			$p = 1.000$			$p = 1.000$
Measure of BMI	Measured	Self-Reported	Self-Reported	Measured	Self-Reported	Self-Reported

The dependent variable in Columns 1-3 is log family income using the mid-points of incomes intervals. The dependent variable in Columns 4-6 is log family income using a random draw from a uniform distribution within the appropriate income interval. Sample weights used. Standard errors in parentheses. * p <.10, ** p< .05, *** p<.01.

TABLE 5. Effect of BMI on Log Monthly Family Income by Gender & Age

	Age 25-80			Age 25-54			Age 55-80		
	OLS		IV	OLS		IV	OLS		IV
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A. Females									
BMI	-0.006***	-0.006**	-0.021	-0.007**	-0.008**	-0.015	-0.006	-0.005	-0.077
	(0.002)	(0.003)	(0.015)	(0.003)	(0.003)	(0.012)	(0.004)	(0.004)	(0.123)
Observations	1586	1586	1586	854	854	854	732	732	732
F-stat			45.352			58.116			1.232
Hausman			$p = 1.000$			$p = 1.000$			$p = 1.000$
Panel B. Males									
BMI	0.005	0.005	0.035	0.005	0.005	0.047*	0.004	0.007	-0.065
	(0.003)	(0.004)	(0.024)	(0.004)	(0.004)	(0.025)	(0.006)	(0.007)	(0.128)
Observations	1422	1422	1422	772	772	772	650	650	650
F-stat			32.037			23.437			1.954
Hausman			$p = 1.000$			$p = 1.000$			$p = 1.000$
Measure of BMI	Measured	Self-Reported	Self-Reported	Measured	Self-Reported	Self-Reported	Measured	Self-Reported	Self-Reported

The dependent variable is log family income using the mid-points of incomes intervals. Sample weights used. Standard errors in parentheses. * p <.10, ** p< .05, *** p<.01.

Embrace the Noise:
It is OK to Ignore Measurement Error in a Covariate, Sometimes

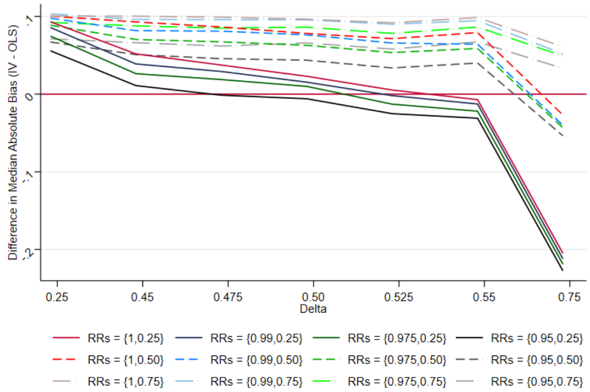
Supplemental Appendix

Hao Dong & Daniel L. Millimet

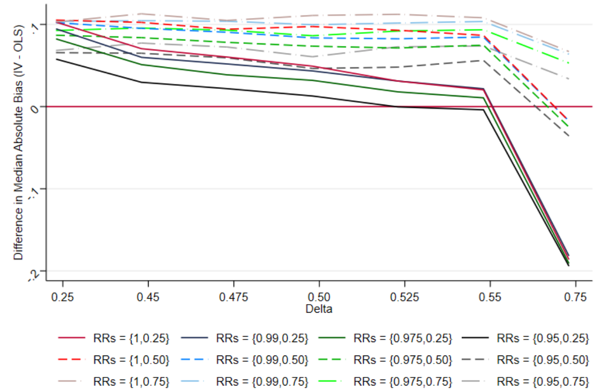
October 6, 2023

Appendix A. Additional Simulation Results

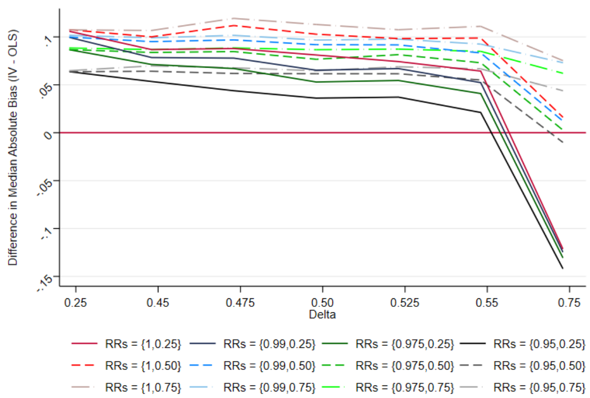
A.1. Median Absolute Error: OLS & IV.



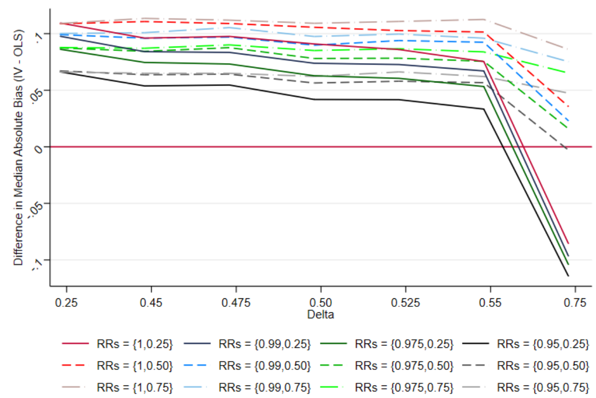
(A) $N = 500$



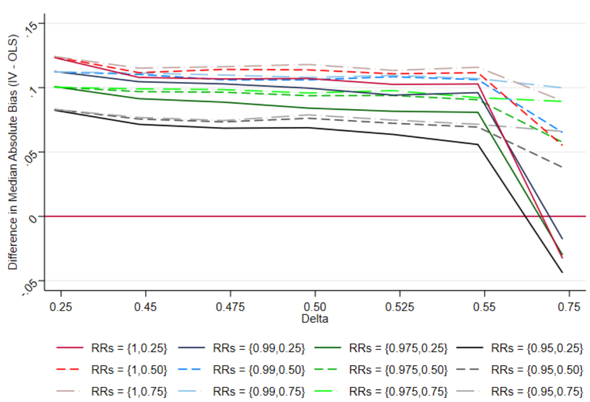
(B) $N = 1,000$



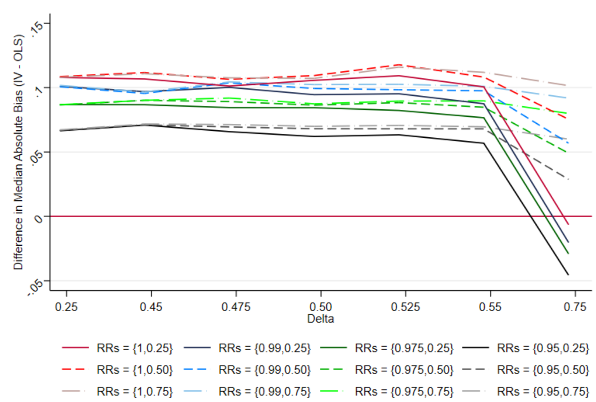
(C) $N = 5,000$



(D) $N = 10,000$

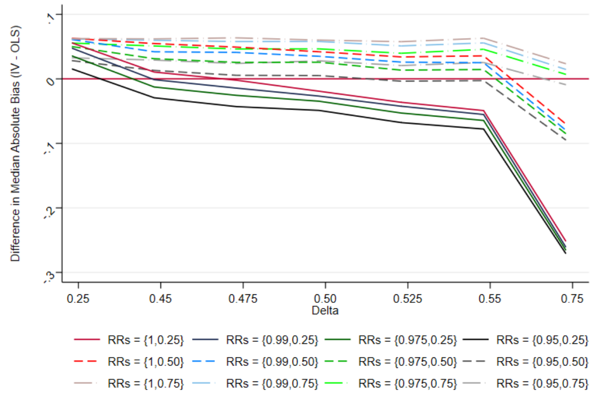


(E) $N = 50,000$

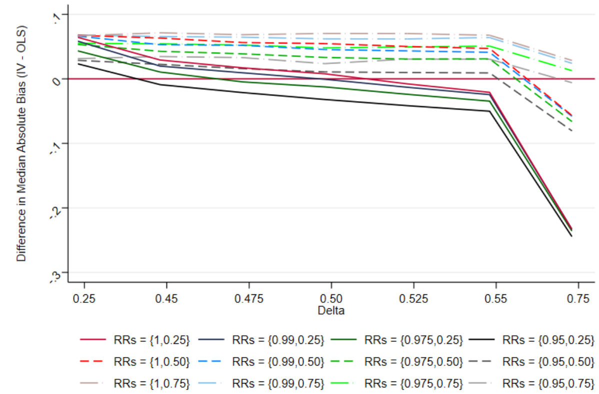


(F) $N = 100,000$

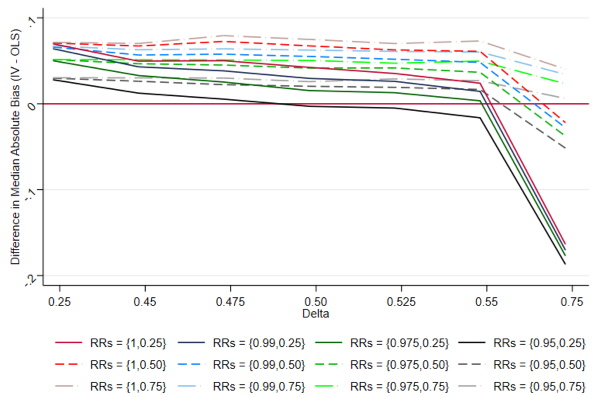
Figure A.1. Simulation Results: Median Absolute Error (F-statistic ≈ 10).



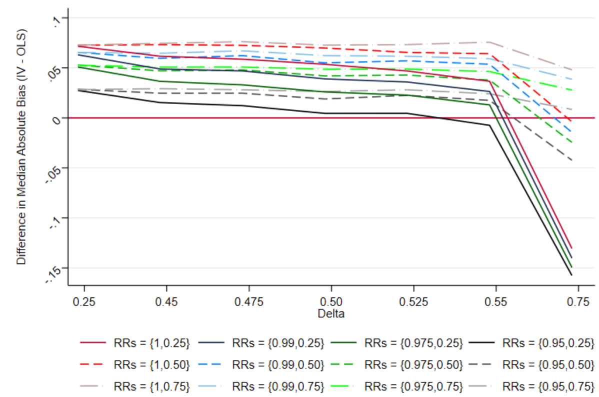
(A) $N = 500$



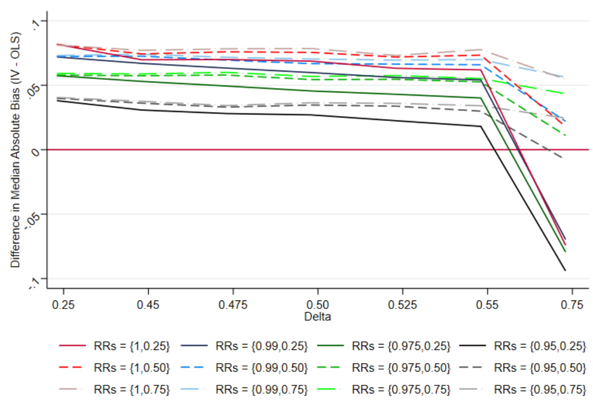
(B) $N = 1,000$



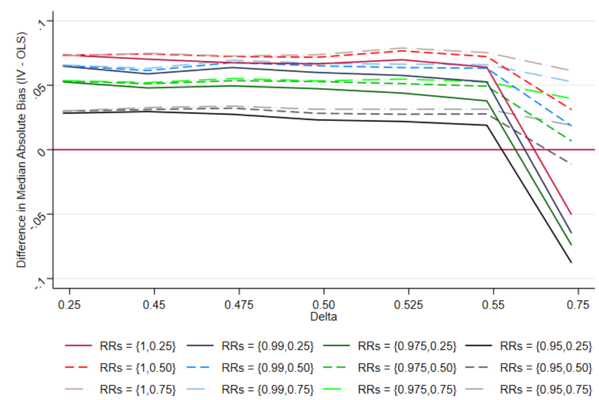
(C) $N = 5,000$



(D) $N = 10,000$



(E) $N = 50,000$



(F) $N = 100,000$

FIGURE A.2. Simulation Results: Median Absolute Error (F-statistic $\approx$ 20).

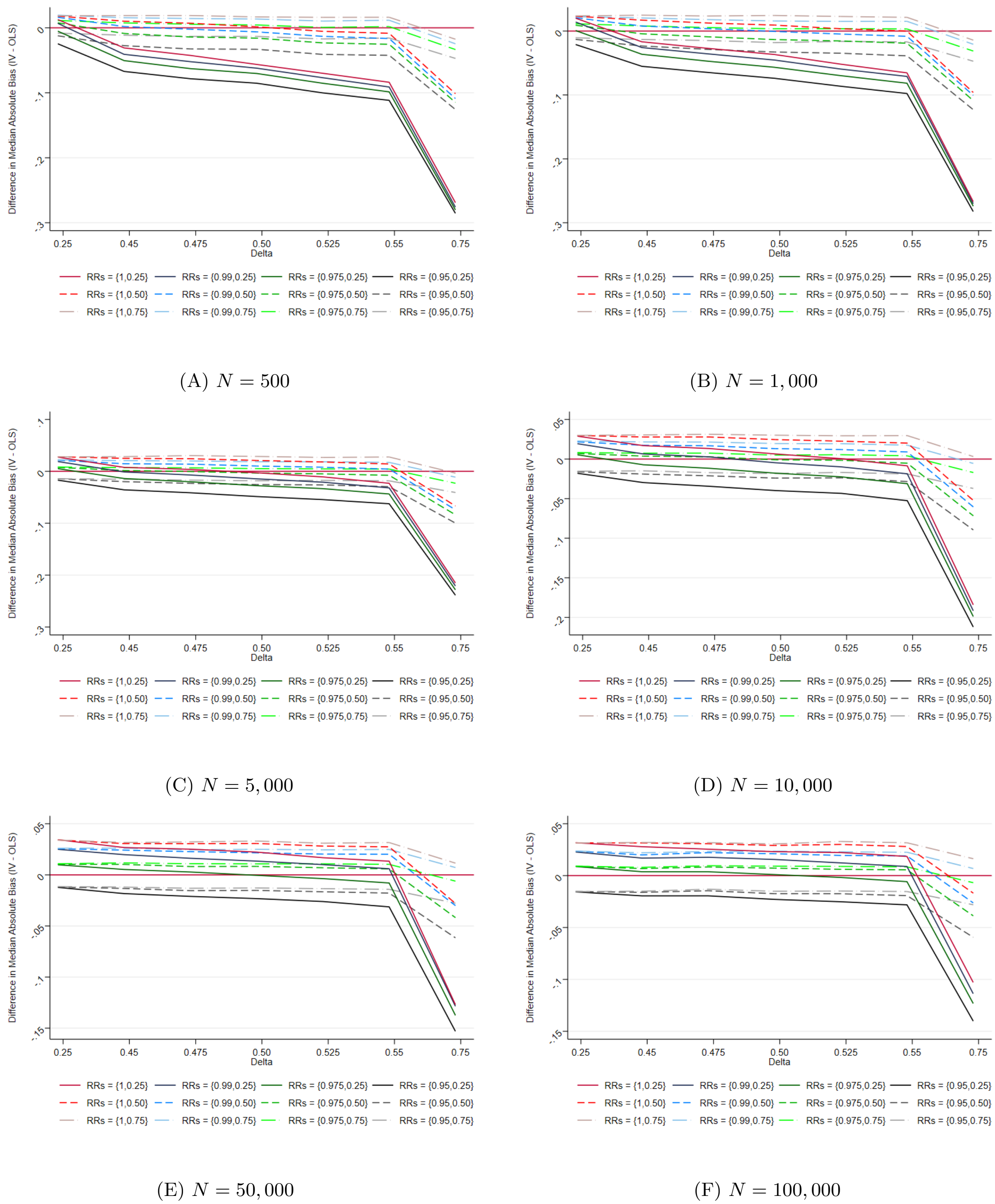
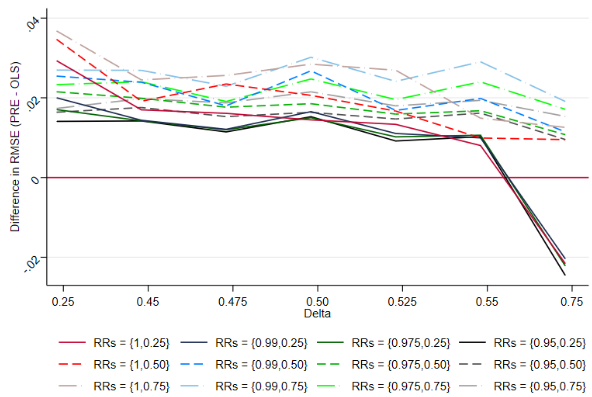
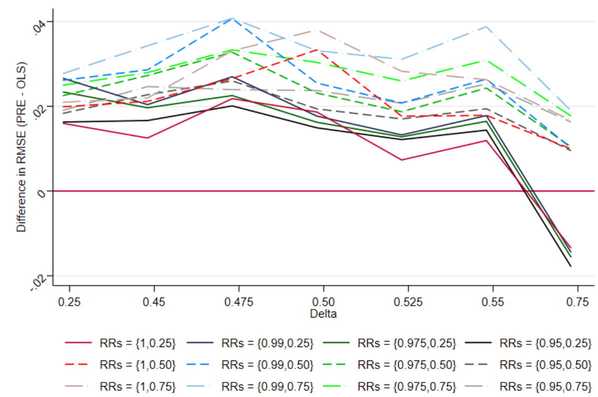


FIGURE A.3. Simulation Results: Median Absolute Error (F-statistic ≈ 100).

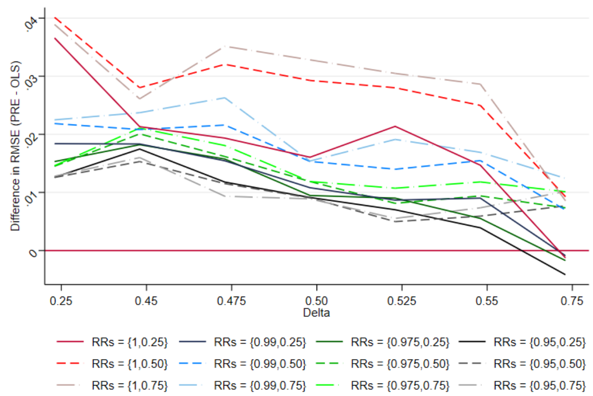
A.2. **Root Mean Squared Error: OLS & Pre-test Estimator.**



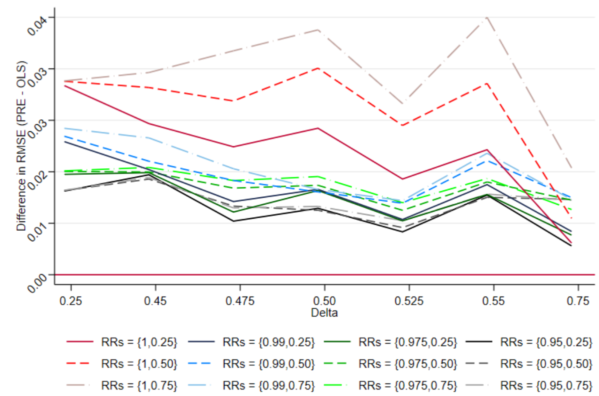
(A) $N = 500$



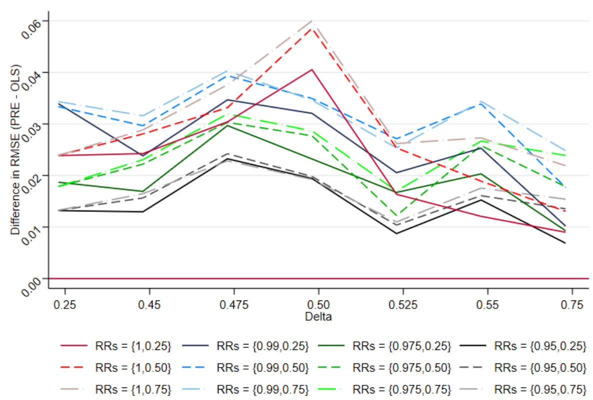
(B) $N = 1,000$



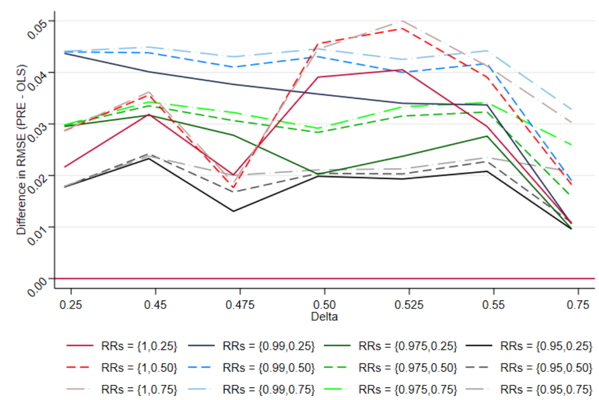
(C) $N = 5,000$



(D) $N = 10,000$



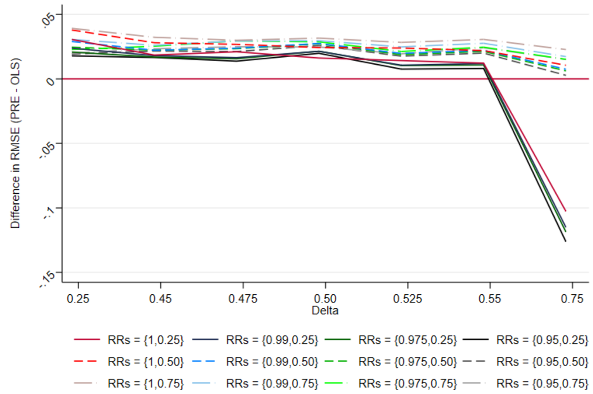
(E) $N = 50,000$



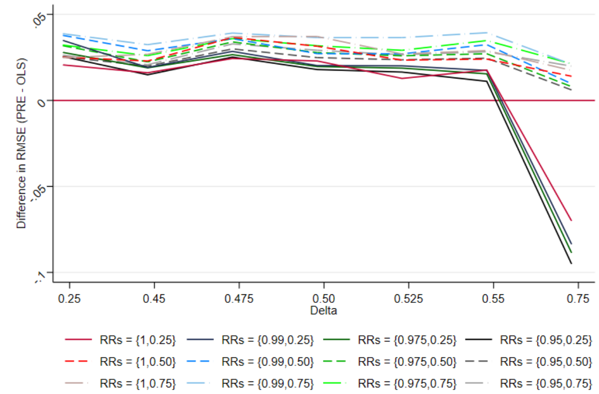
(F) $N = 100,000$

FIGURE A.4. Simulation Results: Root Mean Squared Error (F-statistic ≈ 10).

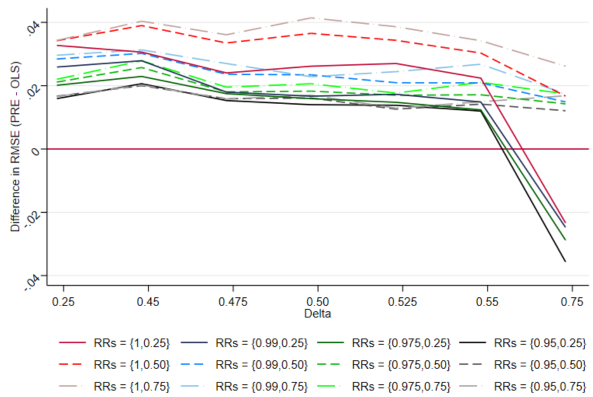
Notes: RR = reliability ratio. First value applies to observations $i = \lfloor N^{\delta} \rfloor + 1, ..., N$; second value applies to observations $i = 1, ..., \lfloor N^{\delta} \rfloor$. See text for more details.



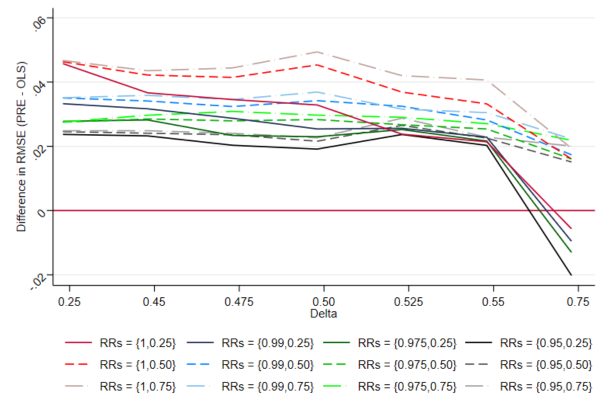
(A) $N = 500$



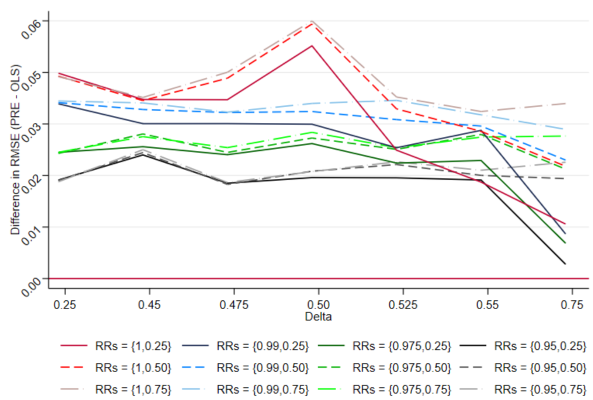
(B) $N = 1,000$



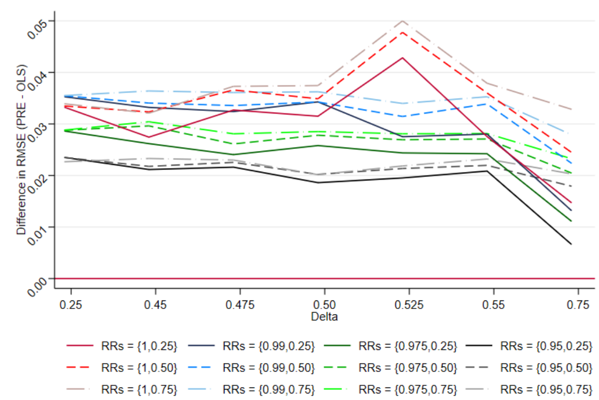
(C) $N = 5,000$



(D) $N = 10,000$



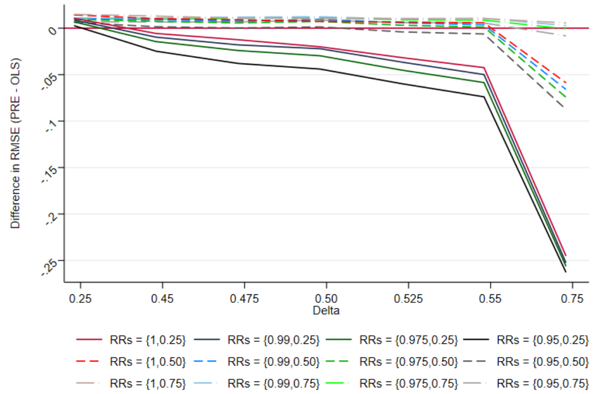
(E) $N = 50,000$



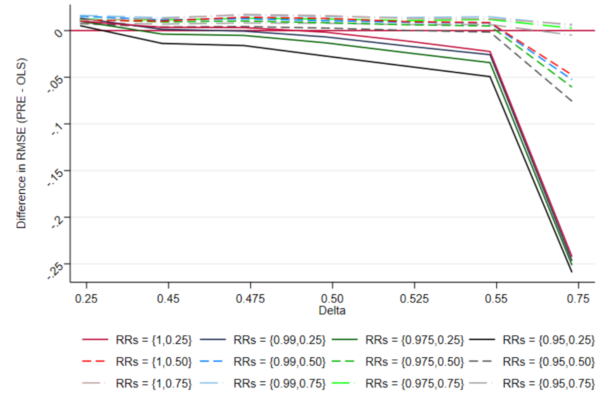
(F) $N = 100,000$

FIGURE A.5. Simulation Results: Root Mean Squared Error (F-statistic $\approx$ 20).

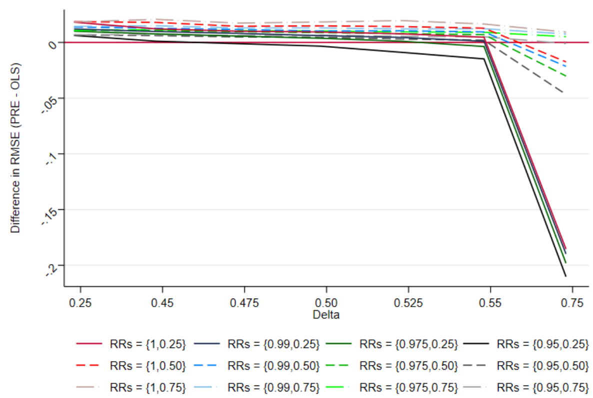
Notes: See Figure A.4 for details.



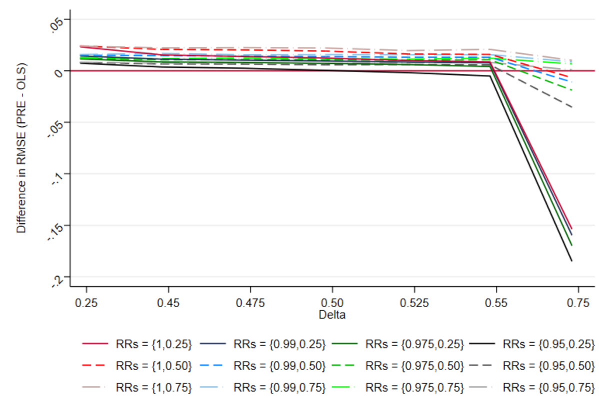
(A) $N = 500$



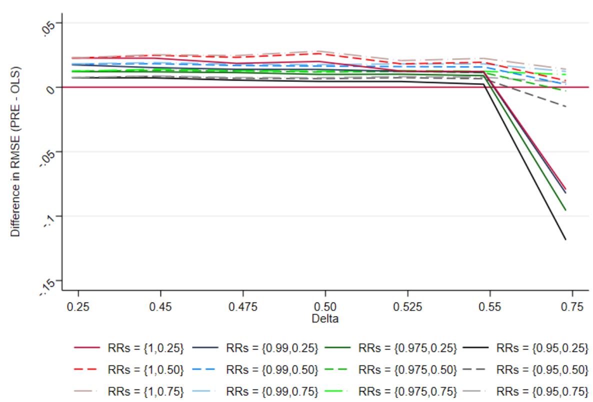
(B) $N = 1,000$



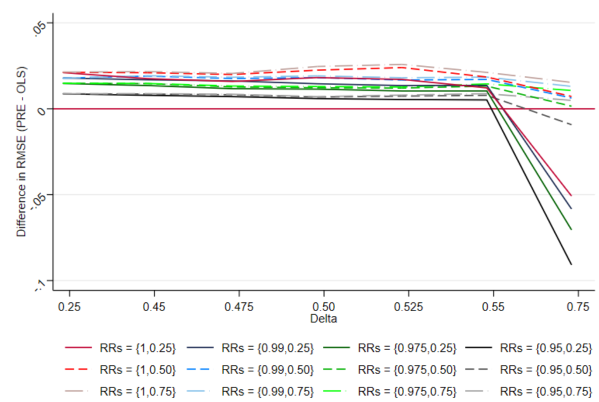
(C) $N = 5,000$



(D) $N = 10,000$



(E) $N = 50,000$

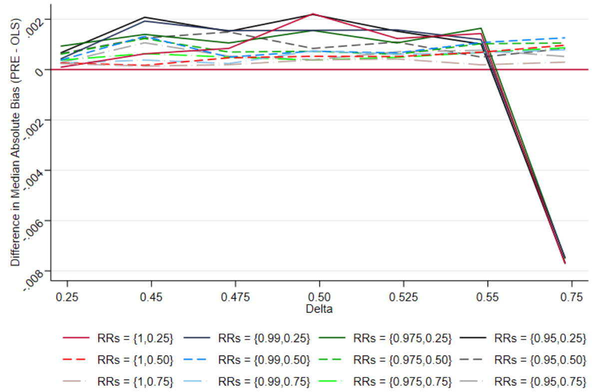


(F) $N = 100,000$

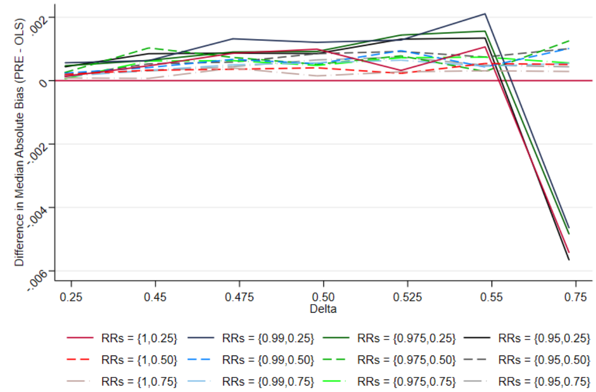
FIGURE A.6. Simulation Results: Root Mean Squared Error (F-statistic ≈ 100).

Notes: See Figure A.4 for details.

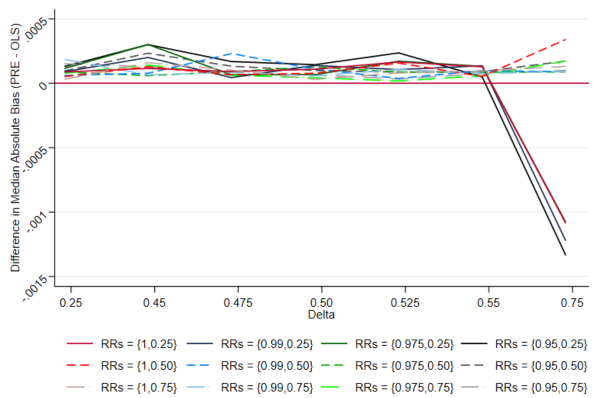
A.3. **Median Absolute Error: OLS & Pre-test Estimator.**



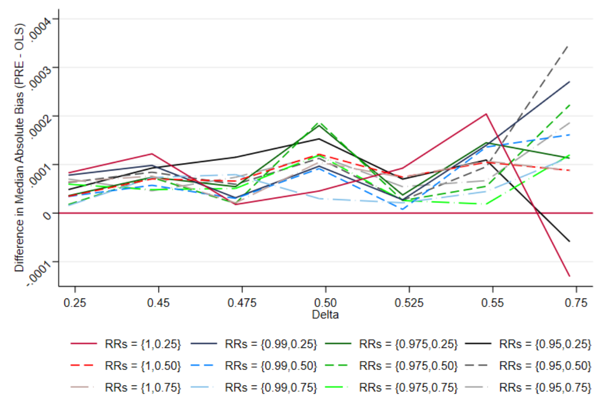
(A) $N = 500$



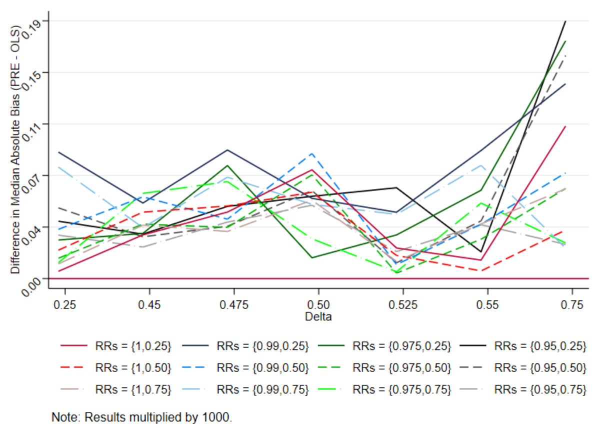
(B) $N = 1,000$



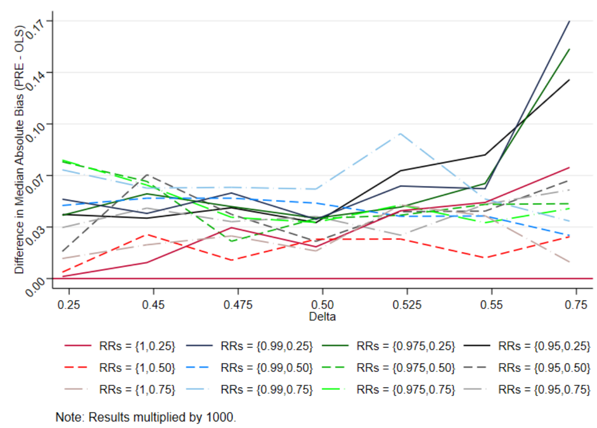
(C) $N = 5,000$



(D) $N = 10,000$



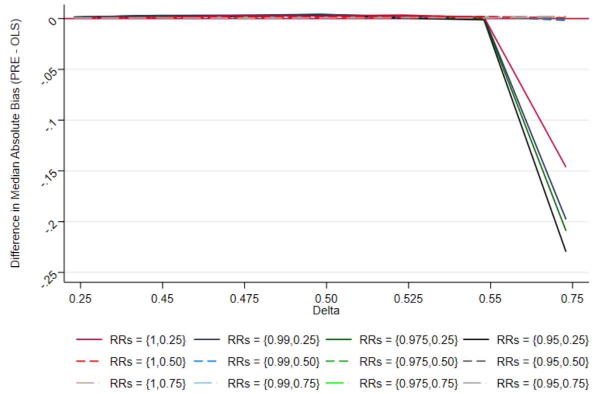
(E) $N = 50,000$



(F) $N = 100,000$

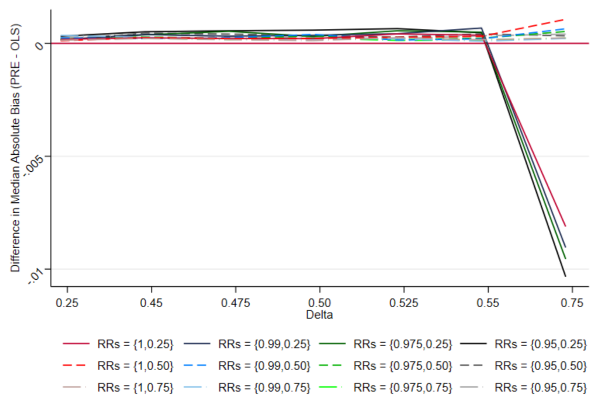
FIGURE A.7. Simulation Results: Median Absolute Bias (F-statistic ≈ 10).

Notes: RR = reliability ratio. First value applies to observations $i = \lfloor N^{\delta} \rfloor + 1, ..., N$; second value applies to observations $i = 1, ..., \lfloor N^{\delta} \rfloor$. See text for more details.

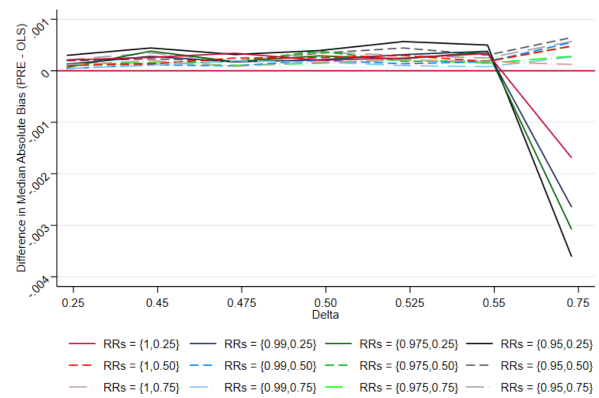


(A) $N = 500$

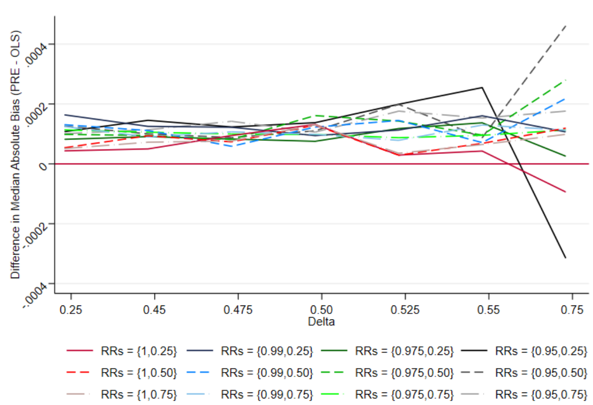
(B) $N = 1,000$



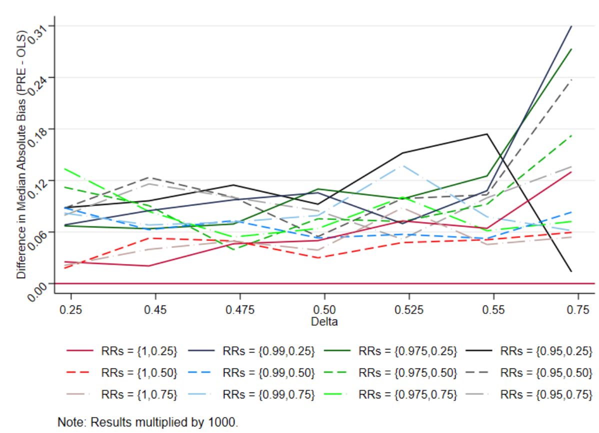
(C) $N = 5,000$



(D) $N = 10,000$



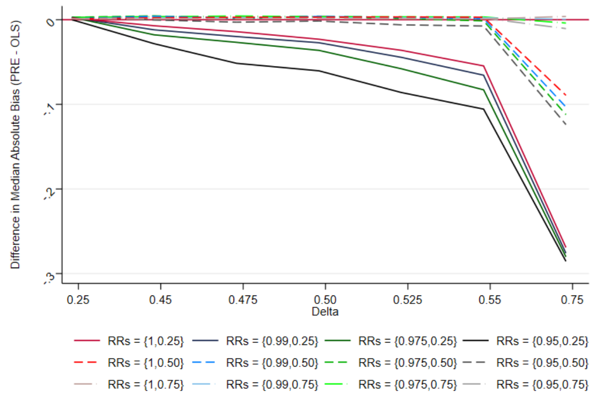
(E) $N = 50,000$



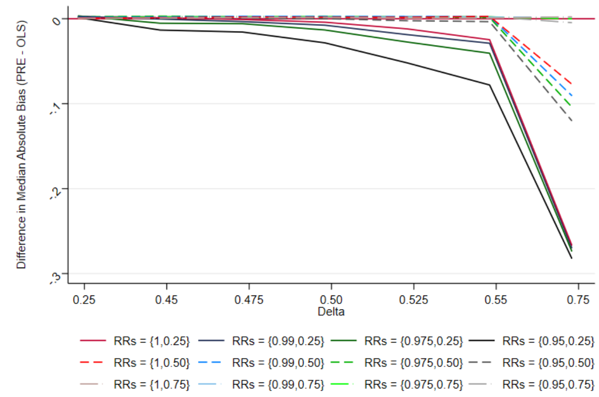
(F) $N = 100,000$

FIGURE A.8. Simulation Results: Median Absolute Bias (F-statistic ≈ 20).

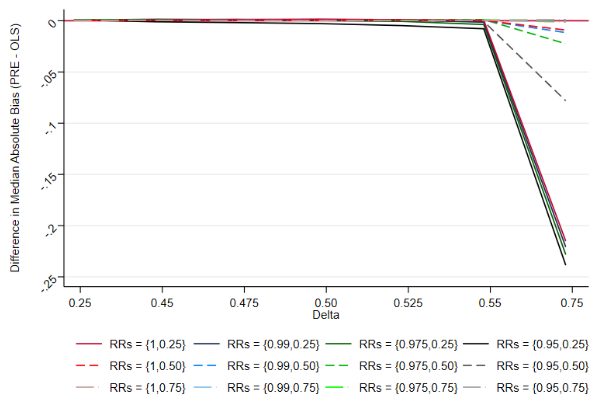
Notes: See Figure A.7 for details.



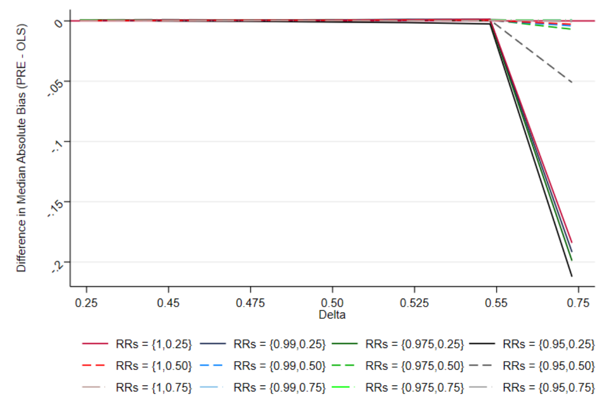
(A) $N = 500$



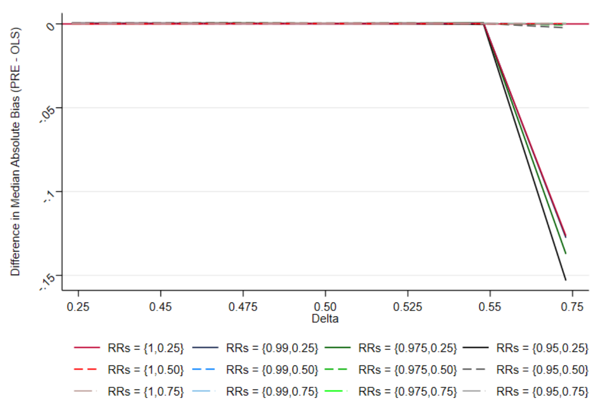
(B) $N = 1,000$



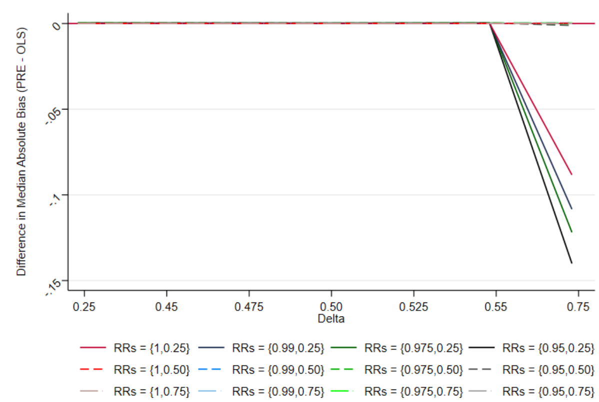
(C) $N = 5,000$



(D) $N = 10,000$



(E) $N = 50,000$

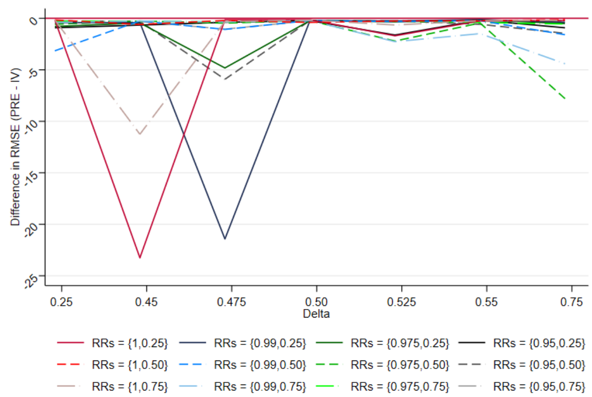


(F) $N = 100,000$

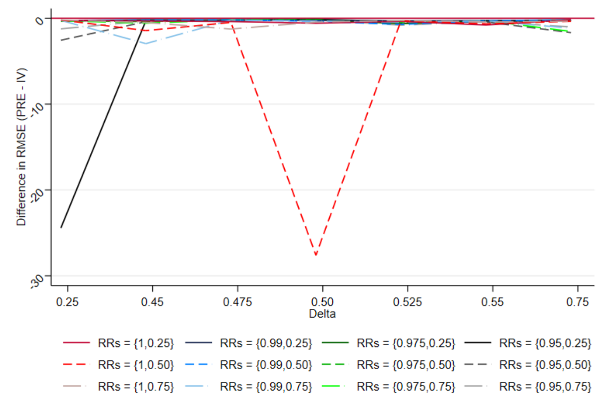
FIGURE A.9. Simulation Results: Median Absolute Bias (F-statistic ≈ 100).

Notes: See Figure A.7 for details.

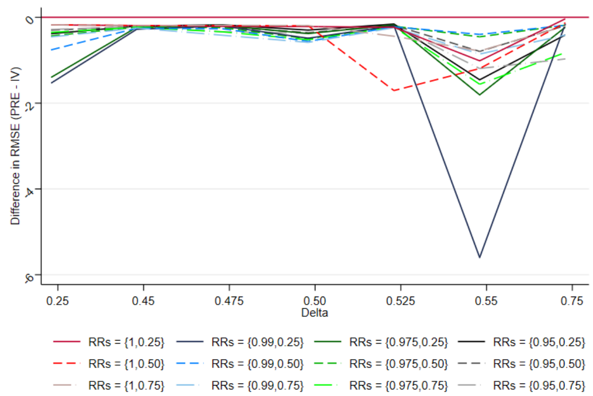
A.4. **Root Mean Squared Error: IV & Pre-test Estimator.**



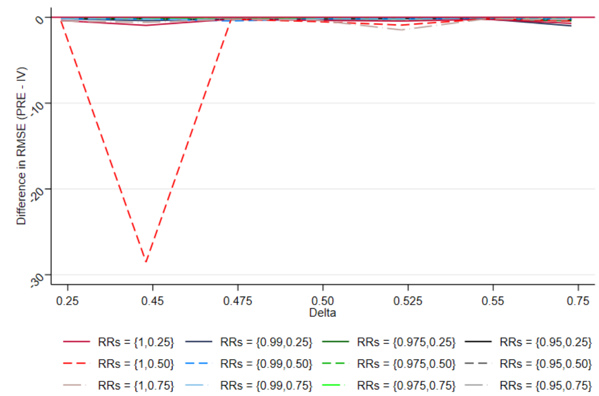
(A) $N = 500$



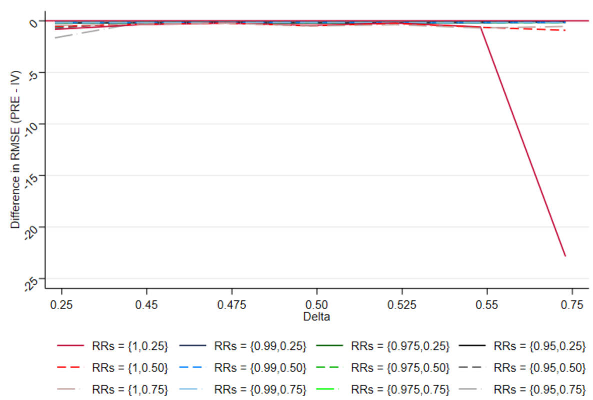
(B) $N = 1,000$



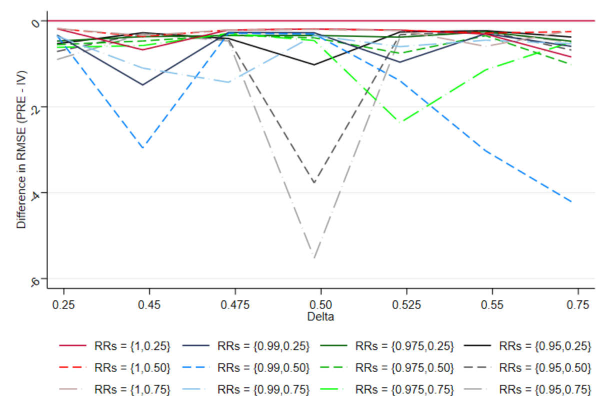
(C) $N = 5,000$



(D) $N = 10,000$



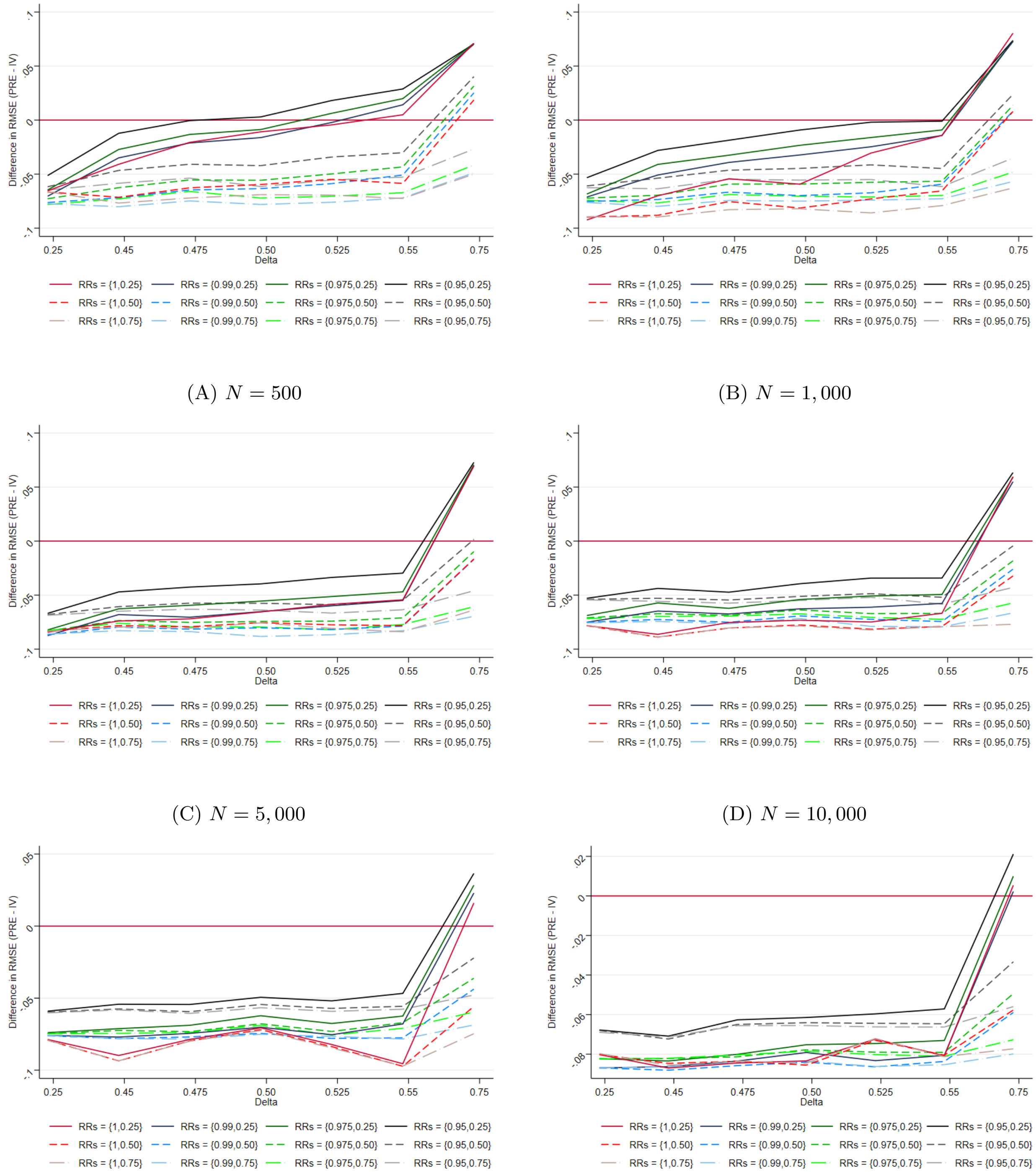
(E) $N = 50,000$



(F) $N = 100,000$

FIGURE A.10. Simulation Results: Root Mean Squared Error (F-statistic ≈ 10).

Notes: RR = reliability ratio. First value applies to observations $i = \lfloor N^{\delta} \rfloor + 1, ..., N$; second value applies to observations $i = 1, ..., \lfloor N^{\delta} \rfloor$. See text for more details.



(A) $N = 500$

(B) $N = 1,000$

(C) $N = 5,000$

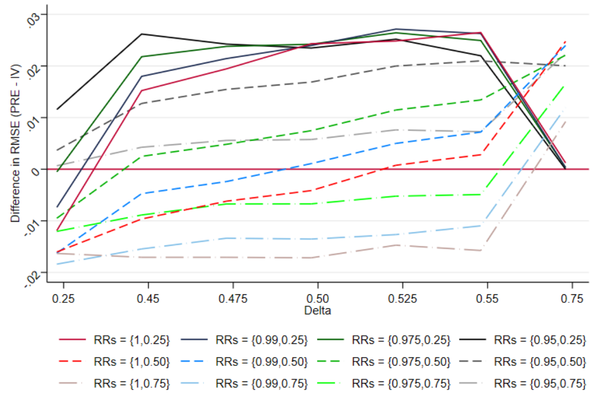
(D) $N = 10,000$

(E) $N = 50,000$

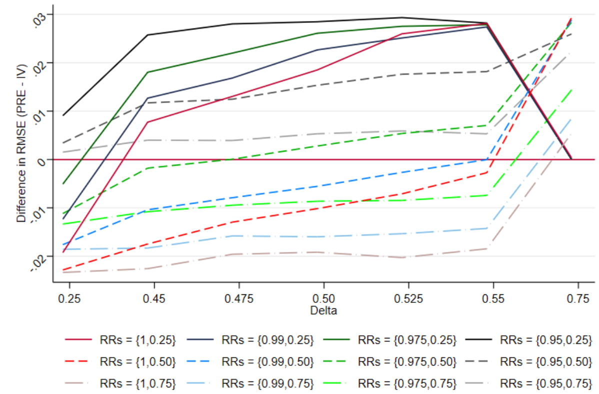
(F) $N = 100,000$

FIGURE A.11. Simulation Results: Root Mean Squared Error (F-statistic $\approx$ 20).

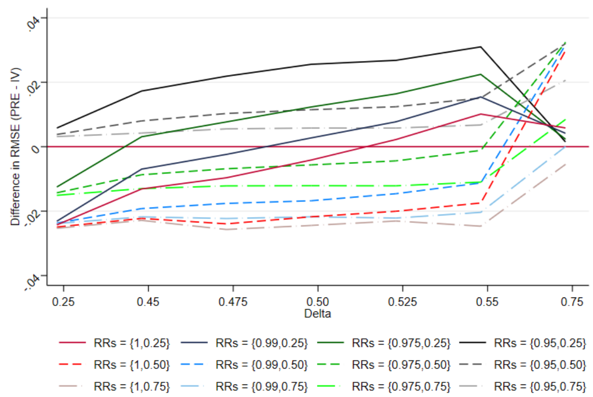
Notes: See Figure A.10 for details.



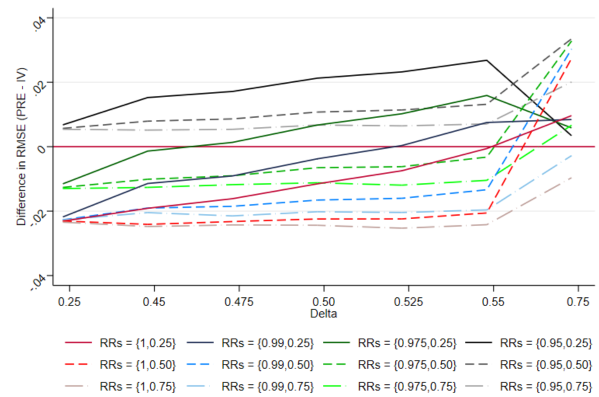
(A) $N = 500$



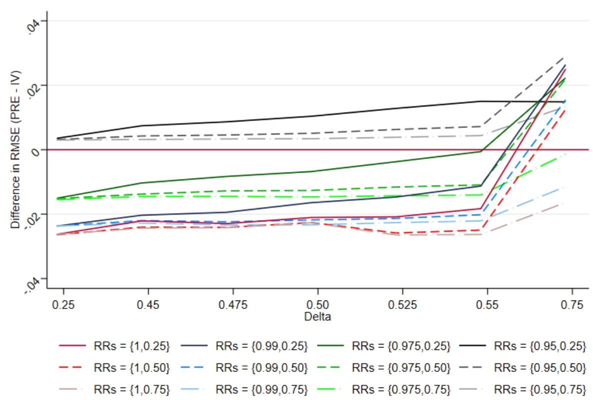
(B) $N = 1,000$



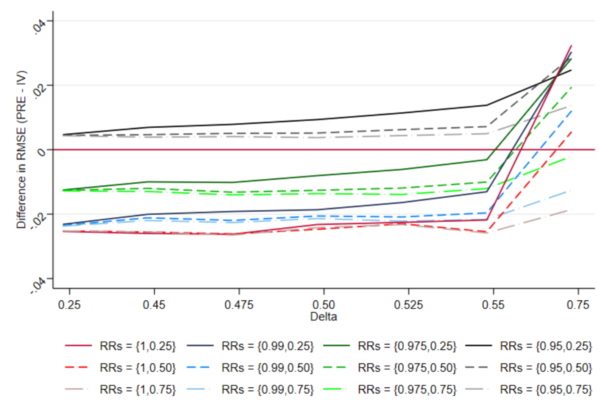
(C) $N = 5,000$



(D) $N = 10,000$



(E) $N = 50,000$

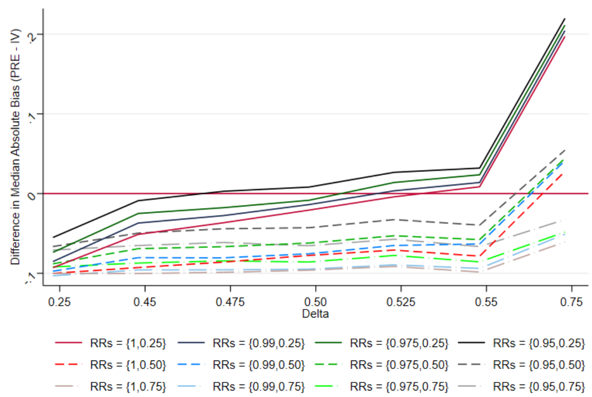


(F) $N = 100,000$

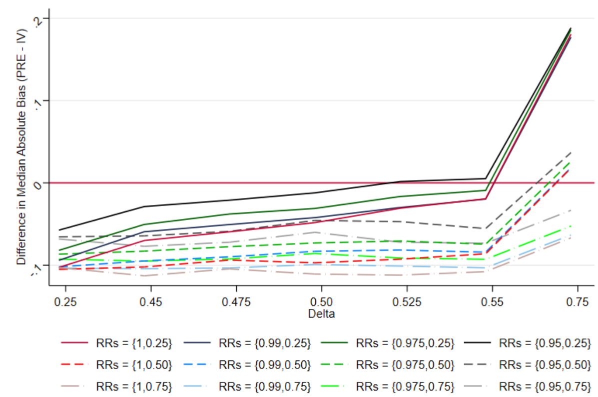
Figure A.12. Simulation Results: Root Mean Squared Error (F-statistic ≈ 100).

Notes: See Figure A.10 for details.

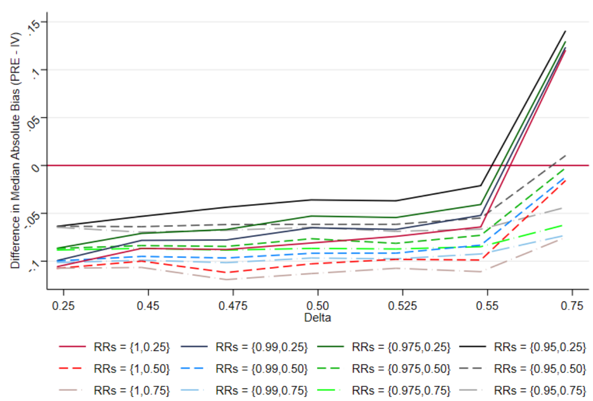
A.5. **Median Absolute Error: IV & Pre-test Estimator.**



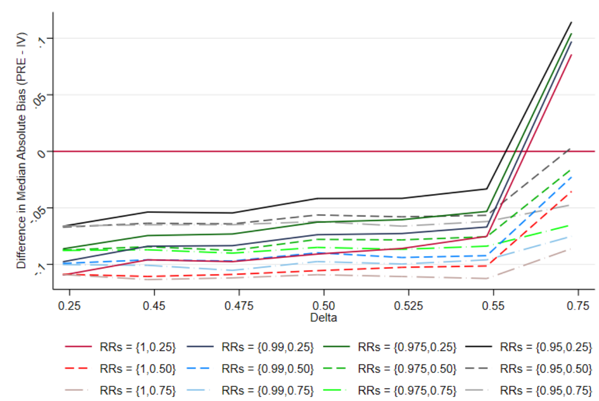
(A) $N = 500$



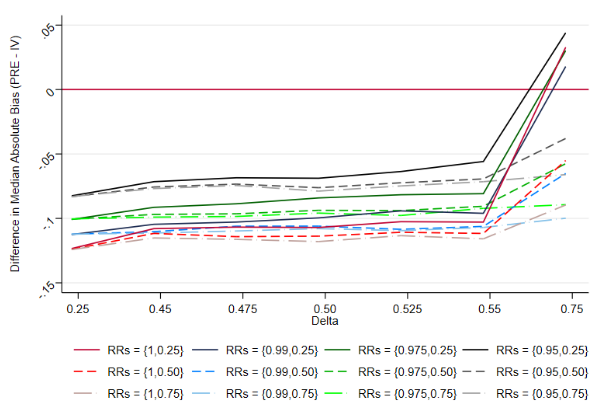
(B) $N = 1,000$



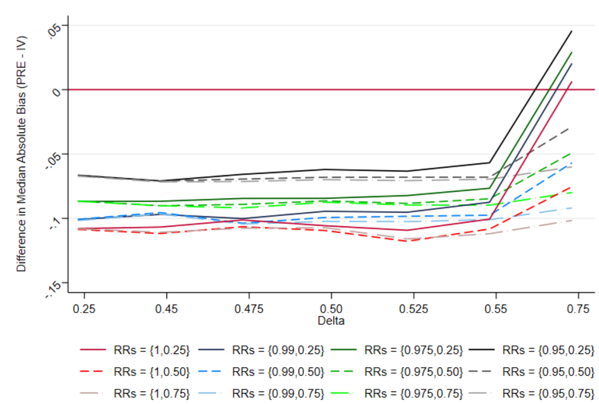
(C) $N = 5,000$



(D) $N = 10,000$



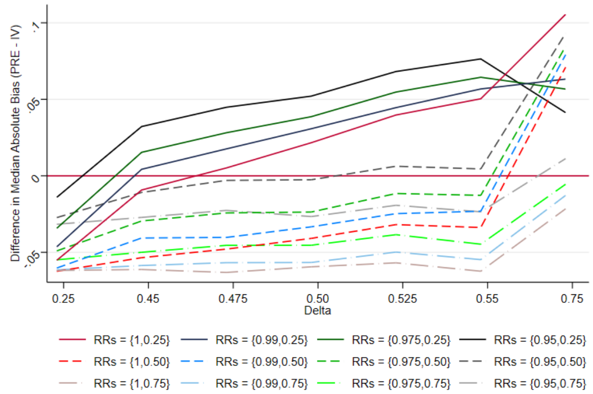
(E) $N = 50,000$



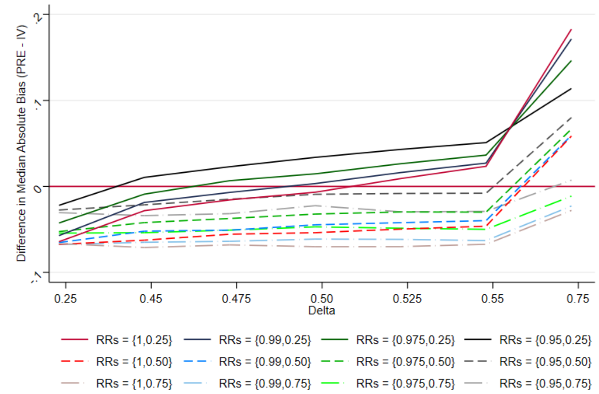
(F) $N = 100,000$

FIGURE A.13. Simulation Results: Median Absolute Bias (F-statistic ≈ 10).

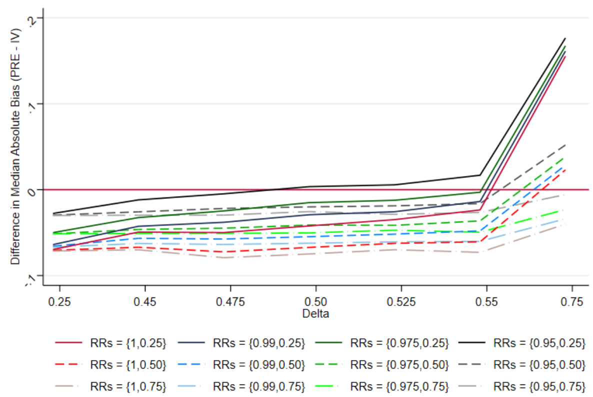
Notes: RR = reliability ratio. First value applies to observations $i = \lfloor N^{\delta} \rfloor + 1, ..., N$; second value applies to observations $i = 1, ..., \lfloor N^{\delta} \rfloor$. See text for more details.



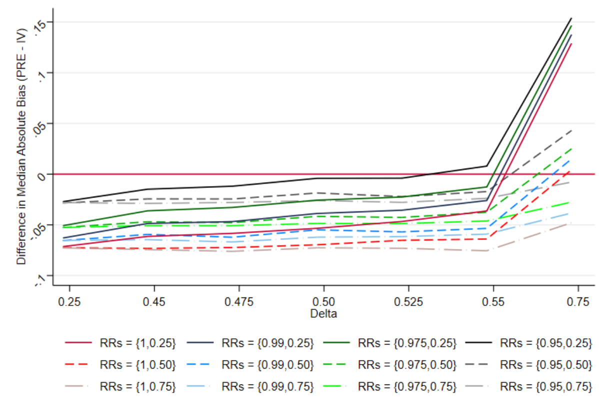
(A) $N = 500$



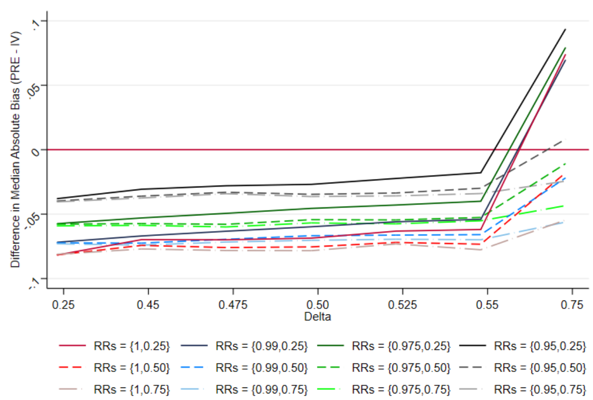
(B) $N = 1,000$



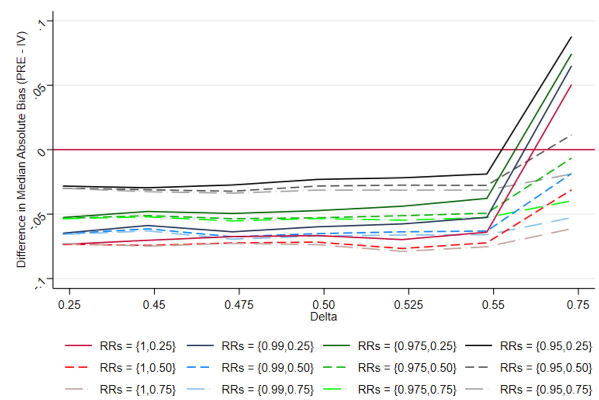
(C) $N = 5,000$



(D) $N = 10,000$



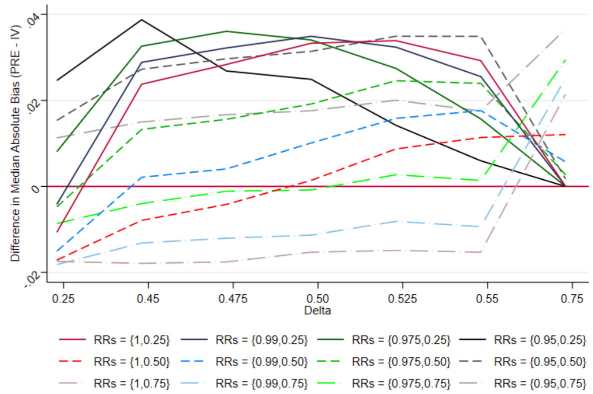
(E) $N = 50,000$



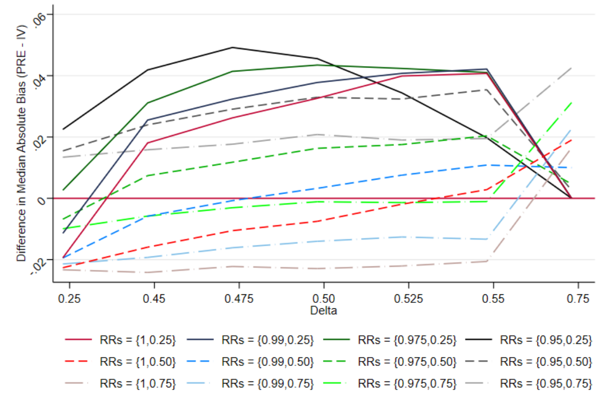
(F) $N = 100,000$

FIGURE A.14. Simulation Results: Median Absolute Bias (F-statistic ≈ 20).

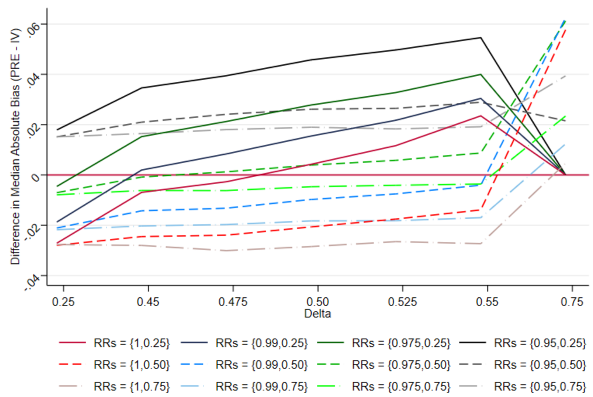
Notes: See Figure A.13 for details.



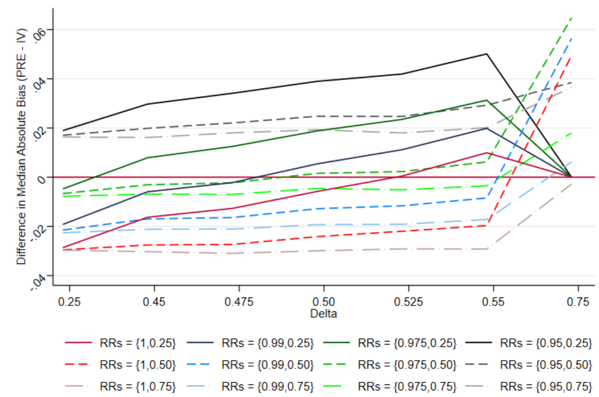
(A) $N = 500$



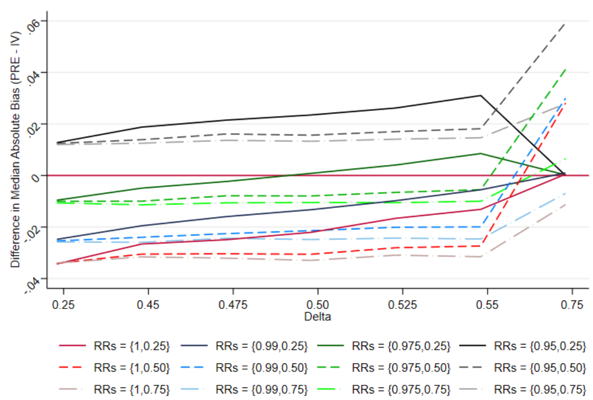
(B) $N = 1,000$



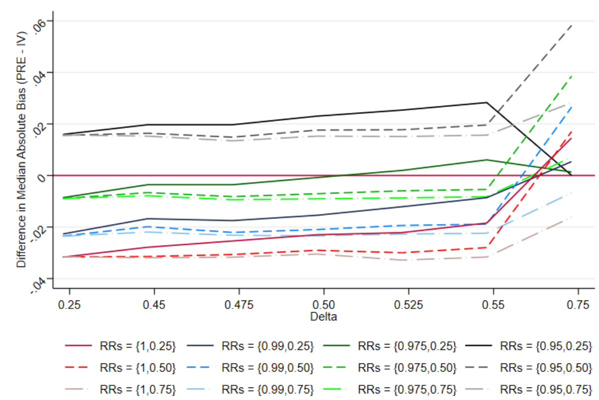
(C) $N = 5,000$



(D) $N = 10,000$



(E) $N = 50,000$



(F) $N = 100,000$

FIGURE A.15. Simulation Results: Median Absolute Bias (F-statistic ≈ 100).

Notes: See Figure A.13 for details.