I Z A

IZA DP No. 5140

The Elephant in the Corner: A Cautionary Tale about Measurement Error in Treatment Effects Models

Daniel L. Millimet

August 2010

The Elephant in the Corner:
A Cautionary Tale about Measurement Error in Treatment Effects Models

Daniel L. Millimet
*Southern Methodist University
and IZA*

ABSTRACT

The Elephant in the Corner: A Cautionary Tale about Measurement Error in Treatment Effects Models[*]

Researchers in economics and other disciplines are often interested in the causal effect of a binary treatment on outcomes. Econometric methods used to estimate such effects are divided into one of two strands depending on whether they require the conditional independence assumption (i.e., independence of potential outcomes and treatment assignment conditional on a set of observable covariates). When this assumption holds, researchers now have a wide array of estimation techniques from which to choose. However, very little is known about their performance – both in absolute and relative terms – when measurement error is present. In this study, the performance of several estimators that require the conditional independence assumption, as well as some that do not, are evaluated in a Monte Carlo study. In all cases, the data-generating process is such that conditional independence holds with the 'real' data. However, measurement error is then introduced. Specifically, three types of measurement error are considered: (i) errors in treatment assignment, (ii) errors in the outcome, and (iii) errors in the vector of covariates. Recommendations for researchers are provided.

JEL Classification: C21, C52

Keywords: treatment effects, propensity score, unconfoundedness,
              selection on observables, measurement error

Corresponding author:

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

---

1   Introduction

Empirical researchers in economics and other disciplines are often interested in the causal effect of a binary treatment on an outcome of interest. Often randomization is used to ensure comparability (at least in expectation) across the treatment and control groups. However, when randomization is not feasible – either due to ethical considerations or cost – researchers must rely on non-experimental or observational data. In such situations, nonrandom selection of subjects into the treatment group becomes a paramount concern and the demands placed on the data are heightened.

Econometric methods used to address nonrandom selection in observational data are divided into two strands depending on whether they require the conditional independence assumption (i.e., independence of potential outcomes and treatment assignment conditional on a set of observable covariates). If subjects self-select into the treatment group on the basis of attributes observable to the researcher, this is referred to as the case of *selection on observables*. On the other hand, if subjects self-select into the treatment group on the basis of attributes unobserved to the researcher, but correlated with the outcome of interest, this is known as the case of *selection on unobservables*.

The econometric and statistics literature on program evaluation in the case of selection on observables has witnessed profound growth over the past few decades.[1] Researchers now have at their disposal an array of statistical methods appropriate for the estimation of the causal effect(s) of the treatment, the most popular of which include parametric regression methods, semiparametric methods utilizing the propensity score, and combinations of the two. Despite this growth, our understanding of the effects of measurement error on the performance of these methods is woefully inadequate. In particular, very little is known about the performance – both in absolute and relative terms – of these methods when measurement error is present. Moreover, as discussed in Section 3, this lack of attention has occurred alongside a bevy of recent examples of just how unreliable data may be at times, particularly at the micro-level where program evaluation methods are most often applied. This is perhaps not too surprising in light of research on the impact of recall window, social norms, and familiarity with subject matter on the accuracy of survey responses.[2]

In this study, the performance of several estimators that require the conditional independence assumption, as well as some that do not, are evaluated in a Monte Carlo study. In all cases, the data-generating process is such that conditional independence holds with the 'real' data, but varying degrees of measurement error are introduced into the observed data along various dimensions. Because the data-generating process imposes independence of *treatment assignment* and potential *outcomes* conditional on observable

---

[1] See D'Agostino (1998), Imbens (2004), and Imbens and Wooldridge (2009) for excellent surveys.
[2] Bound et al. (2001) provide a thorough review of the literature.

covariates when all data are measured accurately, measurement error along each of these three dimensions is considered. The goal is to provide researchers some guidance concerning how much measurement error is 'too much' and whether some estimators perform better than others in the presence of measurement error. Thus, this study is similar in spirit to Almeida et al. (2010), Kreider (2010), and Basu et al. (2008). Almeida et al. (2010) assess the sensitivity to measurement error of various estimators commonly applied in the corporate finance literature. Kreider (2010) examines the width of the worst case bounds for the coefficient on a mismeasured binary covariate in a linear regression framework and probit specification with modest, arbitrary misclassification under different data structures. Basu et al. (2008) is not concerned with measurement error, but does assess the performance of several program evaluation methods when the data-generating process is non-linear.

The remainder of the paper is organized as follow. Section 2 provides a brief overview of the literature on measurement error, focusing on cross-sectional empirical methods common in the program evaluation literature. Section 3 discusses empirical evidence on the magnitude of measurement errors for many variables commonly used by empirical researchers. Section 4 begins by providing a quick overview of the potential outcomes framework and parameter of interest. In addition, it outlines the estimators considered in this study. Section 5 contains the Monte Carlo study. Section 6 concludes.

2 Consequences of Measurement Error

The existing literature on the consequences of measurement error in the program evaluation literature is relatively sparse. Rigorous examination of measurement error in treatment assignment dates to the seminal work in Aigner (1973). Aigner (1973), and subsequent work in Bollinger (1996), Black et al. (2000, 2003), Frazis and Lowenstein (2003), Hu (2006), Kreider (2010), and others, considers the case of misclassification of a binary covariate in a regression context. The primary result is that measurement error in a binary – in fact, any *bounded* – variable must be non-classical except in degenerate cases. Specifically, the measurement error must be (negatively) correlated with the truth. As a consequence, it is possible for measurement error to not only result in attenuation bias, but also to cause the estimated treatment effect to be of the wrong sign. Kreider (2010, p. 2) emphasizes the importance of not ignoring measurement error in this case: "What may not be fully appreciated, however, is that the extreme nature of the measurement error in a binary regressor can result in severe identification deterioration of regression coefficients in the presence of very few classification errors. For a binary regressor, measurement error implies that the variable's true value must be the polar opposite of its reported value." In fact, Kreider (2010) notes that simple examples with misclassification rates less than two percent can lead to confidence intervals around the coefficient

estimates obtained from the mismeasured data and the true data (if this were known) that do not overlap.

The issue is further complicated by the fact that the usual solution to a mismeasured covariate, Instrumental Variable (IV) estimation, does not in general yield a consistent estimate of the treatment effect (Bound et al. 2001). This arises from the fact that any instrument correlated with the observed treatment indicator is likely to be correlated with the measurement error since the measurement error is correlated with the true value. Thus, these papers focus on using Ordinary Least Squares (OLS) and IV to bound the treatment effect. Battistin and Sianesi (2010) and Kreider and Pepper (2007) also discuss methods to bound the effect of a binary treatment, but in a semiparametric (propensity score) and nonparametric context, respectively. Point identification is possible in a Generalized Method of Moments (GMM) framework, as discussed in Black et al. (2000), Bound et al. (2001), and Lewbel (2007). Keane and Sauer (2009) utilize Simulated Maximum Likelihood (SML) to estimate the effect of past employment on current employment in a dynamic probit model when employment is potentially misclassified.

Measurement error in an observed covariate required for the conditional independence assumption to hold has a lengthy history in a regression context, particularly under classical measurement error (see, e.g., Frisch 1934; Koopmans 1937; Reiersøl 1950). In this case, it is well known that the OLS estimate of the coefficient on the mismeasured regressor suffers from attenuation bias. However, the estimated treatment effect will also be biased if treatment assignment is correlated with the true value of the mismeasured covariate (Bound et al. 2001). Moreover, the sign of the bias depends on the sign of this covariance. If the measurement error is nonclassical in that it is correlated with treatment assignment, then the bias depends on the sign of the partial correlation between the measurement error and treatment assignment (Bound et al. 2001). Finally, if more than one covariate is measured with error, the bias on any single coefficient is complex and difficult to sign, even if the measurement errors are classical (Bound et al. 2001). Understanding these implications is vital since many researchers ignore measurement error in covariates since these variables are not the focus of the investigation.

Possible solutions to measurement error in an observed covariate began with the bounding approach proposed in Gini (1921). Here, the coefficient is bounded using the direct and reverse regressions, each estimable by OLS. Klepper and Leamer (1984) extend this approach to the case of multiple mismeasured covariates when the errors are classical and independent of each other. Typically, however, point identification is achieved via IV estimation, where identification is obtained using an external variable as an exclusion restriction or using third or higher moments of variables included in the model as instruments (e.g., Lewbel 1997). Nonetheless, it is important to realize that the estimated treatment effect will still be biased if treatment assignment is correlated with the measurement error (Bound et al. 2001).

Beyond the regression context, Battistin and Chesher (2009) derive the bias of various treatment effect

parameters estimated using semiparametric (propensity score) methods as a function of the variance of the measurement error. The bias may be in either direction. Their proposed solution entails estimating the bias under various assumptions about the reliability of the data and forming bias-corrected estimates of the treatment effects. Also noteworthy for the discussion of measurement error in the program evaluation context is the fact that nonlinearities can accentuate the bias caused by measurement error. Specifically, higher order and interaction terms involving mismeasured covariates tend to suffer from increased bias (Griliches and Ringstad 1970; Hausman et al. 1991). This is relevant since specification of the propensity score model typically entails including non-linear terms to ensure balancing of the covariates (e.g., Millimet and Tchernis 2009).

Finally, measurement error in the outcome of interest has a lengthy history, but predominantly in a regression context. Here, with classical measurement error, OLS estimates of the treatment effects remain unbiased, but efficiency is reduced. However, in many instances, measurement error in the outcome is more consequential. If the measurement error is correlated with the true value of the outcome, correlated with the treatment assignment and/or covariates, the dependent variable is a non-linear transformation of a mismeasured continuous outcome, or the outcome is discrete or categorical, estimates of the regression coefficients are no longer consistent (e.g., Chua and Fuller 1987; Whittemore and Gong 1991; Hausman et al. 1998; Bound et al. 2001; Li et al. 2003; Abrevaya and Hausman 2004). To my knowledge, no studies have considered the impact of mismeasured outcomes on the performance of other estimators common to the program evaluation literature such as estimators based on the propensity score.

In sum, measurement error in treatment assignment or covariates necessary for the conditional independence assumption generally precludes the possibility of obtaining point estimates of the treatment effect(s). Non-classical measurement error in outcomes, or non-linear transformations of classical measurement error, is considerably more complex and typically entails strong parametric assumptions to overcome. Even in the case of classical measurement error, it is not obvious if semiparametric estimators, such as those based on the propensity score, continue to perform well. In light of these implications, it is not surprising that many researchers treat measurement error like an elephant in the corner and ignore its existence.

3   Evidence of Measurement Error

Despite the lack of frequent discussion of measurement error in program evaluation studies, missing data due to measurement error is likely be the norm, not the exception. Many great minds have espoused the difficulty of accurate measurement of quantities of interest. Albert Einstein famously quipped: "Not everything that can be counted counts, and not everything that counts can be counted." The English writer

Jeanette Wintersten opined: "Any measurement must take into account the position of the observer. There is no such thing as measurement absolute, there is only measurement relative." Alvin Toffler, a writer and former associate editor at *Fortune*, argued: "You can use all the quantitative data you can get, but you still have to distrust it and use your own intelligence and judgment."

Within economics, Griliches (1985, p. 197-198) states:

> "Economic data tend to be collected (or often more correctly 'reported') by firms and persons who are not professional observers and who do not have any stake in the correctness and precision of the observations they report... The encounters between econometricians and data are frustrating and ultimately unsatisfactory, both because econometricians want too much from the data and hence tend to be disappointed by the answers, and because the data are incomplete and imperfect... [M]easurement errors which tend to cancel out when averaged over thousands or even millions of respondents, loom much larger when the individual is the unit of analysis... Thus any serious data analysis has to consider at least two data generation components: the economic behavior model describing the stimulus-response behavior of the economic actors and the measurement model describing how and when this behavior was recorded and summarized. While it is usual to focus our attention on the former, a complete analysis must consider them both."

More recently, Hausman (2001, p. 57) states: "The effect of mismeasured variables in statistical and econometric analysis is one of the oldest known problems..." Hyslop and Imbens (2001, p. 475) argue: "Many variables used in econometric analyses are recorded with error. These errors may have occurred at various stages of the data collection. They may be the result of misreporting by subjects, miscoding by the collectors of the data, or incorrect transformation from initial reports into a form ready for analysis. Often such errors are ignored." Similarly, Schennach (2004, p. 33) states: "The assumption that the regressors are measured without error is often made on the basis of convenience, rather than being based on a formal justification."

Indeed, for much of the data used by researchers applying program evaluation methods, there is little justification for the lack of attention given to measurement error. There are many instances of well-documented inaccuracy in micro-level data. Consider the following examples of common binary treatments. Barron et al. (1997) use data from a 1993 survey administered by the Upjohn Institute to compare employee and employer responses on worker training, perhaps the canonical example of a binary treatment in labor economics. With respect to on-site (off-site) formal job training, the responses differ in 28.2% (11.6%) of the

cases; responses differ 11.0% of the time with respect to management training.[3] Mellow and Sider (1983), Freeman (1984), Card (1996), and Barron et al. (1997) examine union coverage, finding disagreement between management and worker responses in the range of 3.5-7%. Hausman et al. (1998) estimate the probability of misclassification of job mobility using a parametric probit model for identification. The authors estimate that the common method of coding job changers as individuals who report tenure less than 12 months misclassifies workers as job changers in 25% of cases; workers are misclassified as job stayers in more than 1% of cases.

Black et al. (2003) use matched data from the 1990 U.S. Census and the 1993 National Survey of College Graduates (NSCG) to analyze consistency in self-reported education levels. The authors find that probability of reporting having a Bachelor's (Master's) [Doctorate] in the NSCG conditional on reporting a Bachelor's (Master's) [Doctorate] in the Census is 91.3% (87.4%) [82.3%]. Moreover, the extent of disagreement varies on the basis of observable demographic attributes. Bound et al. (2001, p. 3708) survey other studies, stating: "Empirical work in economics depends crucially on the use of survey data. The evidence we have, however, makes it clear that survey responses are not perfectly reliable. Even such salient features of an individual's life as years of schooling seem to be reported with some error." For example, using a sample of twins, Ashenfelter and Krueger (1994) estimate the reliability rato – the ratio of the record variance to the survey response variance – of years of schooling to be around 0.9.

Self-reported participation in transfer programs is also highly susceptible to misclassification. Bound et al. (2001) provide an extensive review of studies assessing measurement error in participation in programs such as Food Stamps, Unemployment Insurance (UI), Social Security, and Aid for Families with Dependent Children (AFDC). These studies compare survey responses with administrative records to compute false negative and positive rates. These studies typically find a false negative (positive) rate of about 12% (1%), although a few studies report false negative rates above 40% for UI or AFDC. Aside from the classification errors being correlated with true participation (since participation is a bounded variable), Bound et al. (2001) also note that individuals with only intermittent usage of transfer programs are more likely to incorrectly report no participation. Thus, misclassification is likely to positively correlated with attributes that reduce the likelihood of prolonged participation.

As a final example, Black et al. (2003) re-visit the 1993 survey administered by the Upjohn Institute to compare employee and employer responses on worker eligibility for group health insurance. The authors find that workers report being eligible (ineligible) for insurance when, in fact, management reports they are not (are) in over 15% (5%) of the cases. Bound et al. (2001) also discuss the agreement rates between

---

[3]The important point is not which source – employer or employee – is more accurate, but rather that the disagreements suggest errors in both sources.

worker and firm responses regarding other fringe benefits.

Examples of measurement error in commonly analyzed outcomes abound, particularly outcomes related to income or wealth, labor market outcomes, or measures of health status. Bound et al. (2001) provide an extensive review of studies assessing measurement error in earnings and assets. With respect to earnings, the assessments typically entail comparison of survey responses to records obtained from either the Social Security Administration or employers. The general conclusions of these studies are that measurement error is mean-reverting, correlated over time, and less severe for annual earnings than monthly, weekly, or hourly earnings (with hourly earnings measured the least precisely). The correlation between survey responses and records is typically between 0.8 and 0.9 for annual earnings and around 0.6 for hourly earnings. With classical measurement error, the reliability ratio is equal to this correlation squared; with mean-reverting measurement error, the reliability ratio may be greater or less than the squared correlation. Finally, there is some evidence that measurement error in earnings is correlated with demographic and human capital variables.

Poterba and Summers (1986) use data from the Current Population Survey (CPS) to analyze measurement error in current employment. The authors report that individuals misreport being employed (not employed) in 4.0% (1.5%) of cases. Keane and Sauer (2009) use data from the Panel Survey of Income Dynamics (PSID). The authors report that individuals misreport being employed (not employed) in 8.1% (1.0%) of cases when employment is measured as annual employment. Reviewing the literature, Bound et al. (2001) conclude that 11-16% of individuals classified as unemployed are in error; most of these individuals are in fact out of the labor force. In addition to overstating unemployment, Hausman (2001) states that 37% of workers overstate the duration of unemployment spells. Relatedly, Bound et al. (2001) note that workers consistently overestimate their hours worked. Studies validating reported hours worked using employers' records typically find a correlation between records and survey responses between 0.6 and 0.8.

In terms of health, Strauss and Thomas (1995) discuss the imperfections in commonly used anthropometric measures of health status. Kreider and Pepper (2007) review the existing literature on the shortcomings associated with self-reported measures of disability status. Bound et al. (2001) provide a thorough review of self-reported assessments of health status, as well as health care utilization and expenditures.

A final interesting study examines self-reported emissions by firms in the U.S. Toxic Release Inventory (TRI), a popular data source in environmental research. de Marchi and Hamilton (2006) compare self-reported emissions by firms to measured chemical concentration levels by nearby monitors operated by the U.S. Environmental Protection Agency (EPA). The authors find that large decreases in self-reported

emissions in an area are not matched by similar declines in measured concentration levels. In addition, the authors apply Benford's Law and document a difference in the distribution of the first digit of measured concentration levels and self-reported emissions of lead and nitric acid.

Given the extent of measurement error in commonly used treatments and outcomes, one should also expect measurement error in many observed covariates utilized by applied researchers. While the covariates utilized by researcher are too vast to discuss here, a few examples should illustrate the problem. In labor economics, most empirical research assessing the effect of various treatments on income or employment include marital status and industry or occupation as covariates. Mitchell (2010) matches data from the 1995 Life Events and Satisfaction Study to administrative data to assess the accuracy of one component of marital status: divorce. The author finds that 3.6% of divorced individuals fail to report this in the survey. Bound et al. (2001) review studies assessing the reliability of industry or occupation reports by comparing survey responses to employers' records or reports. In general, the agreement rate is higher using 1-digit classification schemes than 3-digit schemes. However, the rate is still only about 90% for industry and 80% for occupation at the 1-digit level. These rates drop to roughly 80% and 55%, respectively, at the 3-digit level. Lastly, in the empirical finance literature, the use of program evaluation methods to assess the effects of treatments such as a merger or spin-off or type of corporate governance structure has grown rapidly of late. However, Almeida et al. (2010, p. 1) state that "it is hard to think of any empirical proxies in corporate research whose measurement is not a concern."

4 Causal Inference Under Conditional Independence

4.1 Potential Outcomes Framework

Consider a random sample of N individuals from a large population indexed by $i = 1, ..., N$. Utilizing the potential outcomes framework (see, e.g., Neyman 1923; Fisher 1935; Roy 1951; Rubin 1974), let $Y_i(T)$ denote the potential outcome of individual i under treatment T, $T \in \mathcal{T}$.[4] Here, the focus is on binary treatments: $\mathcal{T} = \{0, 1\}$. The observation-specific causal effect of the treatment ($T = 1$) relative to the control ($T = 0$) is defined as the difference between the corresponding potential outcomes. Formally,

$$\tau_i = Y_i(1) - Y_i(0). \tag{1}$$

[4]Implicit in this specification is the Stable Unit Treatment Value Assumption (SUTVA), whereby the potential outcomes of observation i do not depend on the treatment assignment of other observations in the population (Neyman 1923; Rubin 1986).

In the evaluation literature, several population parameters are of potential interest. Here, for brevity, attention is given to the average treatment effect (ATE), defined as $\tau = \mathbb{E}[\tau_i] = \mathbb{E}[Y_i(1) - Y_i(0)]$. The ATE is the expected treatment effect of an observation chosen at random from the population.

In the absence of measurement error, the triple $\{Y_i, T_i, X_i\}$ is observed for each observation, where Y_i is the observed outcome, T_i is a binary indicator of the treatment received, and X_i is a vector of covariates. The only requirement of the covariates included in X_i is that they are pre-determined (that is, they are unaffected by T_i) and do not perfectly predict treatment assignment. The relationship between the potential and observed outcomes is given by

$$Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0) \tag{2}$$

which makes clear that only one potential outcome is observed for any individual. As such, estimating τ is not trivial as there is an inherent missing data problem.

Aside from random experiments, the methods utilized by researchers to circumvent this missing data problem are classified into two groups: selection on observables estimators and selection on unobservable estimators. The distinction lies in whether a method consistently estimates the causal effect of the treatment in the presence of unobservable attributes of subjects that are correlated with both treatment assignment and the outcome of interest conditional on the set of observable variables. Assuming a lack of such unobservables is referred to as *conditional independence* or *unconfoundedness* assumption (Rubin 1974; Heckman and Robb 1985). Formally, under the conditional independence assumption (CIA), treatment assignment is said to be independent of potential outcomes conditional on the set of covariates, X, and is expressed as

$$Y(0), Y(1) \perp T \mid X. \tag{3}$$

As a result, selection into treatment is random conditional on X and the average effect of the treatment can be obtained by comparing outcomes of individuals in different treatment states with identical values of the covariates. To solve the dimensionality problem that is likely to arise if X is multi-dimensional, Rosenbaum and Rubin (1983) propose using the propensity score (PS), $P(X_i) = \Pr(T_i = 1|X_i)$, instead of X as the conditioning variable. The authors prove that (3) implies

$$Y(0), Y(1) \perp T \mid P(X). \tag{4}$$

4.2 Estimation In The Absence of Measurement Error

4.2.1 Under CIA

Given a random sample from the population, no measurement error, and conditional independence, several estimators of the ATE are available to researchers. Estimators employed in this study include OLS, OLS in combination with the propensity score (OLS-PS), stratification, a doubly robust (DR) estimator, inverse propensity score weighting (IPW), and propensity score matching (PSM).

To derive the estimating equation for the OLS estimator, the following assumptions are invoked:

(A1) Potential outcomes and latent treatment assignment are additively separable in observables and unobservables

$$
\begin{aligned}
Y_i(0) &= \alpha_0 + X_i\beta + \varepsilon_{0i} \\
Y_i(1) &= \alpha_1 + X_i\beta + \varepsilon_{1i} \\
T_i &= \mathbb{I}[X_i\gamma + \eta_i > 0]
\end{aligned}
$$

where $\mathbb{I}[\cdot]$ is the indicator function; and,

(A2) $\varepsilon_0, \varepsilon_1, \eta \sim \mathbb{N}_3(0, \Sigma)$, where

$$
\Sigma = \begin{bmatrix} \sigma_0^2 & \rho & 0 \\ & \sigma_1^2 & 0 \\ & & 1 \end{bmatrix}.
$$

(A1) implies that the treatment effect is heterogeneous since $\varepsilon_{0i} \neq \varepsilon_{1i}$, but assumes there is no heterogeneity due to differential effects of X on the two potential outcomes.[5] (A2) implies conditional independence in that there is no selection into the treatment on the basis of unobservables. Together, these assumptions imply $\tau = \mathbb{E}[\alpha_1 - \alpha_0 + \varepsilon_{1i} - \varepsilon_{0i}] = \alpha_1 - \alpha_0$.

Substituting these functional form assumptions into (2) yields

$$
Y_i = \alpha_0 + (\alpha_1 - \alpha_0)T_i + X_i\beta + [\varepsilon_{0i} + (\varepsilon_{1i} - \varepsilon_{0i})T_i] \tag{5}
$$

Under (A1) and (A2), OLS estimation of (5) provides an unbiased of the ATE; the estimate is given by $\hat{\tau}_{OLS} = \widehat{\alpha_1 - \alpha_0}$.

The next three estimators make use of both OLS and the propensity score. The OLS-PS estimator maintains the linear regression framework, but replaces the functional form assumption in (A1) with the

---

[5]Future may wish to consider estimation of treatment effects when the effect varies on the basis of observables.

assumption that the potential outcomes are a polynomial function of the propensity score (Rosenbaum and Rubin 1983; Wooldridge 2002). The estimating equation is now

$$Y_i = \alpha_0 + (\alpha_1 - \alpha_0)T_i + \sum_{s=1}^{3} \beta_s \left[\widehat{P}(X_i)\right]^s + [\varepsilon_{0i} + (\varepsilon_{1i} - \varepsilon_{0i})T_i] \tag{6}$$

where $\widehat{P}(X_i)$ is the estimated propensity score obtained from a first-stage probit model. The OLS-PS estimator of the ATE is given by $\hat{\tau}_{OLS-PS} = \widehat{\alpha_1 - \alpha_0}$.

The stratification estimator is one of the more popular estimators relying on the propensity score (Basu et al. 2008). After estimating the propensity score, ten indicator variables are created indicating which decile of the empirical distribution of the propensity score each observation lies. Then, the following equation is estimated via OLS

$$Y_i = \sum_{q=1}^{10} \beta_q D_{qi} + \sum_{q=1}^{10} \alpha_q T_i D_{qi} + \upsilon_i \tag{7}$$

where $D_{qi} = 1$ if $\widehat{P}(X_i)$ lies in the qth decile and zero otherwise. The stratification estimate of the ATE is given by $\hat{\tau}_S = \sum_{q=1}^{10}(\hat{\alpha}_q/10)$.

The DR estimator in Scharfstein et al. (1999) and Bang and Robins (2005) entails OLS estimation of the following specification

$$Y_i = \alpha_0 + (\alpha_1 - \alpha_0)T_i + X_i\beta + \theta_0(1 - T_i)\{1 - \widehat{P}(X_i)\}^{-1} + \theta_1 T_i\{\widehat{P}(X_i)\}^{-1} + \xi_i.$$

The DR estimate of the ATE is given by $\hat{\tau}_{OLS} = \widehat{\alpha_1 - \alpha_0} + (1/N)\sum_i[\hat{\theta}_1\{\widehat{P}(X_i)\}^{-1} - \hat{\theta}_0\{1 - \widehat{P}(X_i)\}^{-1}]$. In practice, to avoid excessively large values when the propensity score approaches a boundary, the sample is trimmed by including only observations with $\hat{P}(X_i) \in [0.02, 0.98]$.

The final two estimators rely solely on the propensity score. The IPW estimator requires CIA, but replaces a functional form assumption, like that in (A1), with a common support condition to ensure sufficient overlap in the distribution of the propensity score across the treatment and control groups (Dehejia and Wahba 1999; Smith and Todd 2005). To implement the IPW estimator, the normalized estimator of Hirano and Imbens (2001) is utilized. The IPW estimator originates in Horvitz and Thompson (1952) who show that the ATE may be expressed as

$$\tau = \mathbb{E}\left[\frac{Y \cdot T}{P(X)} - \frac{Y \cdot (1 - T)}{1 - P(X)}\right],$$

11

with the sample analogue given by

$$\hat{\tau}_{HT} = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{Y_i T_i}{\hat{P}(X_i)} - \frac{Y_i(1 - T_i)}{1 - \hat{P}(X_i)} \right]. \tag{8}$$

The estimator in (8) is the *unnormalized* estimator as the weights do not necessarily sum to unity. To circumvent this issue, Hirano and Imbens (2001) propose an alternative estimator, referred to as the *normalized* or HI estimator, which is given by

$$\hat{\tau}_{HI} = \left[\sum_{i=1}^{N} \frac{Y_i T_i}{\hat{P}(X_i)} \bigg/ \sum_{i=1}^{N} \frac{T_i}{\hat{P}(X_i)} \right] - \left[\sum_{i=1}^{N} \frac{Y_i(1 - T_i)}{1 - \hat{P}(X_i)} \bigg/ \sum_{i=1}^{N} \frac{(1 - T_i)}{1 - \hat{P}(X_i)} \right]. \tag{9}$$

In practice, to avoid giving too much weight to observations with an estimated propensity score near the boundary, the sample is trimmed by including only observations with $\hat{P}(X_i) \in [0.02, 0.98]$. Millimet and Tchernis (2009) provide evidence of the superiority of the normalized estimator.

The final estimator considered is PSM; specifically, a kernel matching (KM) estimator. As with IPW, this estimator requires CIA and the common support condition. To proceed, τ_i is estimated for all i by replacing the missing counterfactual and then averaging $\hat{\tau}_i$ over the sample. The missing counterfactual is estimated by

$$
\begin{aligned}
\widehat{Y_i(0)} &= \frac{1}{\displaystyle\sum_{l \in \{T_l=0\}} \omega_{il}} \sum_{l \in \{T_l=0\}} \omega_{il} Y_l(0) \quad \text{if } i \in \{T_i = 1\} \\
\widehat{Y_i(1)} &= \frac{1}{\displaystyle\sum_{l \in \{T_l=1\}} \omega_{il}} \sum_{l \in \{T_l=1\}} \omega_{il} Y_l(1) \quad \text{if } i \in \{T_i = 0\}
\end{aligned}
$$

where ω_{il} is the weight given by observation i to observation l when estimating i's missing counterfactual. With KM, these weights have the form

$$\omega_{il} = \max \left\{ 0, \frac{G\left(\frac{\hat{P}(X_l) - \hat{P}(X_i)}{a_N} \right)}{\displaystyle\sum_{l' \in \{T_{l'}=0\}} G\left(\frac{\hat{P}(X_{l'}) - \hat{P}(X_i)}{a_N} \right)} \right\}$$

where $G(\bullet)$ is the kernel function and a_N is the bandwidth. In the Monte Carlo study, the Epanechnikov kernel is used along with two bandwidths: 0.05 and 0.25. The smaller the bandwidth, the more weight is concentrated on observations with very similar propensity scores.

4.2.2 Without CIA

Each of the preceding estimators will be biased if CIA fails to hold. While the Monte Carlo study only considers data-generating processes where CIA holds, it may not hold in the sample once measurement error is introduced. Thus, three estimators that do not require CIA are also examined for comparison.

The most common approach for dealing with a failure of the CIA is to utilize IV methods. Here, we do not wish to assume the existence of a standard exclusion restriction or make use of higher moments (e.g., Lewbel 1997).[6] Instead, three estimators are utilized that do not require CIA and also do not require an exclusion restriction for identification.

The first estimator is based on Heckman's bivariate normal (BVN) selection model. The model relies on a relaxed version of (A2):

(A2') $\varepsilon_0, \varepsilon_1, \eta \sim \mathbb{N}_3(0, \Sigma)$, where

$$\Sigma = \begin{bmatrix} \sigma_0^2 & \rho & \rho_{0\eta} \\ & \sigma_1^2 & \rho_{1\eta} \\ & & 1 \end{bmatrix}.$$

Under (A1) and (A2'), OLS applied to the following augmented estimating equation

$$y_i = \alpha_0 + X_i\beta + (\alpha_1 - \alpha_0)T_i + \beta_{\lambda 0}(1 - T_i)\left[\frac{\phi(X_i\gamma)}{1 - \Phi(X_i\gamma)}\right] + \beta_{\lambda 1}T_i\left[\frac{-\phi(X_i\gamma)}{\Phi(X_i\gamma)}\right] + \varsigma_i \qquad (10)$$

where $\phi(\cdot)/\Phi(\cdot)$ is the inverse Mills' ratio, ς is a well-behaved error term,

$$\begin{aligned} \beta_{\lambda 0} &= \rho_{0\eta}\sigma_0 & (11) \\ \beta_{\lambda 1} &= \rho_{0\eta}\sigma_0 + \rho_{\delta u}\sigma_\delta, \end{aligned}$$

and $\delta = \varepsilon_1 - \varepsilon_0$ will yield an unbiased estimate of the ATE; the estimate is given $\hat{\tau}_{BVN} = \widehat{\alpha_1 - \alpha_0}$. In practice, OLS estimation of (10) occurs after replacing γ with an estimate obtained from a first-stage probit model.

The second estimator is the bias-corrected IPW estimator proposed in Millimet and Tchernis (2010). Following Black and Smith (2004) and Heckman and Navarro-Lozano (2004), the bias of the ATE at a particular value of the propensity score (i.e., $\mathbb{E}[\tau \mid P(X)]$) when CIA does not hold is given by

$$B[P(X)] = \left\{\rho_{0\eta}\sigma_0 + [1 - P(X)]\rho_{\delta\eta}\sigma_\delta\right\}\left\{\frac{\phi(\Phi^{-1}(P(X)))}{P(X)[1 - P(X)]}\right\}. \qquad (12)$$

---

[6]Future work may certainly wish to assess the performance of such estimators.

By integrating (12) over the distribution of $P(X)$, a bias-corrected version of the normalized IPW estimator is given by

$$\tau_{BC-HI} = \tau_{HI} - \int B[P(X)] f_P(P(X)) dP(X) \tag{13}$$

with a sample analogue given by

$$\hat{\tau}_{BC-HI} = \hat{\tau}_{HI} - \frac{1}{N} \sum_i B[\widehat{P(X_i)}] \tag{14}$$

where the propensity score continues to be estimated by a first-stage probit model and $\rho_{0\eta}\sigma_0$ and $\rho_{\delta\eta}\sigma_\delta$ are estimated using (10).

The final estimator considered comes from Klein and Vella (2009). The parametric implementation of this estimator relies on a similar functional form assumption to the BVN estimator in the absence of heteroskedasticity, but effectively induces a valid exclusion restriction in the presence of heteroskedasticity. To proceed, suppose that latent treatment assignment is now given by

$$T^* = X\gamma + \eta^*$$

where $\eta^* = S(X)\eta$ and η is drawn from a standard normal density. In this case, the probability of receiving the treatment conditional on X is given by

$$\Pr(T = 1|X) = \Phi\left(\frac{X}{S(X)}\gamma\right). \tag{15}$$

Assuming $S(X) = \exp(X\theta)$, the parameters of (15) are estimable by maximum likelihood (ML), with the log-likelihood function given by

$$\ln \mathcal{L} = \sum_i \left[\ln \Phi\left(\frac{X\gamma}{\exp(X\theta)}\right)\right]^{T_i} \left\{\ln\left[1 - \Phi\left(\frac{X\gamma}{\exp(X\theta)}\right)\right]\right\}^{1-T_i} \tag{16}$$

where the element of θ corresponding to the intercept is normalized to zero for identification.

The ML estimates are then used to obtain the predicted probability of treatment, $\hat{P}(X)$, which may be used as an instrument for T in equation (5). In the Monte Carlo study, the ATE is estimated via two-stage least squares (KV-TSLS). Note, even if $S(X) = 1$, $\hat{P}(X)$ remains a valid instrument since it is non-linear in X. However, since the non-linearity arises mostly in the tails, identification typically relies on a small fraction of the sample.[7] On the other hand, if $S(X) \neq 1$, then the Klein and Vella (2009) approach

---

[7]Mroz (1999) analyzes this case and finds that while an IV strategy using the predicted probability obtained from a homoskedastic probit performs reasonably well in terms of bias, it does very poorly in terms of mean squared error.

effectively induces a valid exclusion restriction as $Z \equiv X/S(X)$ is frequently linearly independent of X.

5   Monte Carlo Study

5.1   Setup

To assess the performance of the various estimators, data are simulated using a data-generating process (DGP) that imposes (A1) and (A2); thus, CIA holds in the population and the functional form for the observed outcome is known. This allows one to assess the impact of measurement error without confounding the effects of model misspecification.

Specifically, 1000 data sets are simulated, each with 1000 observations, containing

$$
\begin{aligned}
Y(0) &= 1 + g(X) + \varepsilon_0 \\
Y(1) &= 2 + g(X) + \varepsilon_1 \\
T &= \mathbb{I}[-0.25 + g(X) + \eta > 0] \\
Y &= TY(1) + (1 - T)Y(0)
\end{aligned}
$$

$$
g(X) = 0.5(X_1 - X_2) + 0.25(X_1^2 - X_2^2) + 0.25X_1X_2
$$

$$
X_1, X_2 \overset{\text{iid}}{\sim} \mathbb{N}(-1, 1)
$$
$$
\varepsilon_0, \varepsilon_1, \eta \overset{\text{iid}}{\sim} \mathbb{N}(0, 1)
$$

implying that $\mathbb{E}[\tau] = \mathbb{E}[1 + \varepsilon_1 - \varepsilon_0] = 1$. As a result, all the estimators considered here are consistent if $\{Y, T, X\}$ are observed; OLS is the most efficient. However, varying amounts of measurement error are introduced to explore the absolute and relative performance of the various estimators.

Initially, measurement error is introduced one dimension at a time. Thus, when measurement error in, say, T is introduced, Y and X are measured correctly. Similarly, when measurement error in Y or X is introduced, the other two aspects are accurately observed. In the final simulation, measurement error in all three dimensions are introduced simultaneously. This is a particularly interesting case since not only is it the most plausible scenario in most applications – in light of the discussion in Section 3 – but the vast majority of the existing methodological literature on addressing measurement error only considers measurement error along a single dimension.

To begin, two types of classification errors in treatment assignment are considered. In the first case,

15

the *observed* treatment assignment, T^o, is given by

$$T^o = \begin{cases} T & \text{if } u > c \\ 1 - T & \text{otherwise} \end{cases}$$

where $u \overset{\text{iid}}{\sim} \mathbb{U}[0,1]$ and $c = 0$, 0.01, 0.025, 0.05, and 0.10. Thus, c is the misclassification rate; when $c = 0$, $T^o = T$ and there is no misclassification. Note, the potential outcomes depend on T, not T^o. As such, the measurement error is non-differential and, in fact, completely arbitrary.

In the second case, the probability of misclassification depends on X_1 and X_2. Specifically, the *observed* treatment assignment is now given by

$$T^o = \begin{cases} T & \text{if } F(Z) < 1 - c \\ 1 - T & \text{otherwise} \end{cases}$$

where $Z = X_1 - X_2$, $F(Z)$ is the empirical CDF of Z, and $c = 0$, 0.01, 0.025, 0.05, and 0.10. This amounts to misclassifying observations with values of Z above the $(1 - c)^{th}$ quantile of its empirical distribution. As in the first case, c is the misclassification rate; $T^o = T$ and there is no misclassification when $c = 0$.

Next, measurement error in the covariates, X, is introduced. Again, two cases are considered. First, varying degrees of classical measurement error are simultaneously introduced in X_1 and X_2. Formally, the *observed* covariates are given by

$$\begin{aligned} X_1^o &= X_1 + u_1 \\ X_2^o &= X_2 + u_2 \end{aligned}$$

where $u_1, u_2 \overset{\text{iid}}{\sim} \mathbb{N}(0, \sigma_u^2)$ and σ_u^2 is chosen such that the reliability ratio – the ratio of the variance of X_k to the variance of X_k^o, $k = 1, 2$ – of X_1^o and X_2^o is equal to 0.90, 0.95, 0.975, 0.99, and 1 (in expectation). The baseline case of no measurement error corresponds to a reliability ratio of one for both covariates. In the second case, u_1 and u_2 are correlated, with a correlation coefficient of 0.5, and they are also correlated with treatment assignment. The DGP yields an expected correlation coefficient between u_1 (u_2) and T of roughly 0.8 (0.4), and sets σ_u^2 to roughly achieve the desired reliability ratio.[8]

---

[8]Specifically, the data are generated by first drawing N realizations, u_{1i} and u_{2i}, from a bivariate normal distributon, $\mathbb{N}_2(0, 0, \sigma_u^2, \sigma_u^2, 0.5)$. The data are sorted in ascending order by u_1. The measurement errors are then merged with the 'true' data, $\{Y, T, X\}$, sorted in ascending order by T. As a result, observations with $T = 1$ are merged with higher values of u_1 (and hence higher values of u_2 on average). The variance of the measurement errors, σ_u^2, is chosen to come sufficiently close to the desired reliability ratio. Based on simulations using five million observations, the reliability ratios of X_1 and X_2 are 0.992 and 0.962, respectively, in the case where the reliability ratio is said to be 0.99. These ratios are 0.982 and 0.950 in the 0.975 case; 0.966 and 0.929 in the 0.95 case; and, 0.928 and 0.886 in the 0.90 case. Thus, X_1 is measured more reliably than X_2, but both are close to the desired values.

Next, four cases of measurement error in the outcome, Y, are considered. In the first case, classical measurement is introduced. Specifically, the *observed* outcome is given by

$$Y^o = Y + u$$

where $u \overset{iid}{\sim} \mathbb{N}(0, \sigma_u^2)$ and σ_u^2 is chosen such that the reliability ratio of Y^o is equal to 0.90, 0.95, 0.975, 0.99, and 1 (in expectation). Again, no measurement error corresponds to a reliability ratio of one.

In the second case, the measurement error is mean-reverting. The motivation for this follows from the fact that earnings is frequently the outcome of interest and, as detailed earlier, mean-reverting measurement error is frequently observed at the individual-level. In this case, the *observed* outcome is generated as

$$Y^o = \kappa Y$$

where $\kappa = cu$, $u \overset{iid}{\sim} \mathbb{U}[0,1]$ and c is chosen such that the reliability ratio of Y^o is 1, 1.01, 1.025, 1.05, and 1.10 (in expectation).[9] As before, no measurement error correspond to a reliability ratio of one. This procedure yields a correlation coefficient between the deviations from the truth, $Y^o - Y$, and the truth, Y, of roughly -0.4.

The third and fourth cases consider a binary outcome. The motivation here is that outcomes are often binary, such as employment or a binary measure of health status such as obesity. To proceed, the prior DGP is amended; potential outcomes are now simulating according to

$$
\begin{aligned}
Y(0) &= \mathbb{I}[-1.5 + g(X) + \varepsilon_0 > 0] \\
Y(1) &= \mathbb{I}[0.5 + g(X) + \varepsilon_1 > 0].
\end{aligned}
$$

The remainder of the DGP is identical to above.

In the third case, the *observed* outcome is now given by

$$
Y^o = \begin{cases} Y & \text{if } u > c \\ 1 - Y & \text{otherwise} \end{cases}
$$

where $u \overset{iid}{\sim} \mathbb{U}[0,1]$ and $c = 0$, 0.01, 0.025, 0.05, and 0.10. When $c = 0$, $Y^o = Y$ and there is no misclassification. Note, $\mathbb{E}[\tau] = \mathbb{E}[Y(1) - Y(0)] \approx 0.599$.[10] In the fourth case, the probability of misclassification is

---

[9]Note, this corresponds to mean-reverting measurement error since the true values, Y, take on both positive and negative values.

[10]The 'true' value of the ATE is obtained using a simulation of five million observations.

correlated with X_1 and X_2. Specifically, the *observed* outcome is now generated as

$$Y^o = \begin{cases} Y & \text{if } F(Z) < 1 - c \\ 1 - Y & \text{otherwise} \end{cases}$$

where $Z = X_1 - X_2$, $F(Z)$ is the empirical CDF of Z in the sample, and $c = 0, 0.01, 0.025, 0.05$, and 0.10. Again, this induces measurement error in observations with values of Z above the $(1 - c)^{th}$ quantile of the empirical distribution. As in the prior case, $Y^o = Y$ and there is no misclassification when $c = 0$.

In the final scenario, measurement errors in Y, T, X_1, and X_2 are introduced simultaneously. Two cases are considered. First, mean-reverting measurement error in Y is introduced such that the reliability ratio is 1.01, one percent of observations have treatment assignment misclassified, and classical measurement errors in X_1 and X_2 are introduced such that the reliability ratio for each is 0.99. Second, mean-reverting measurement in Y is introduced such that the reliability ratio is 1.05, five percent of observations have treatment assignment misclassified, and classical measurement error in X_1 and X_2 are introduced such that the reliability ratio for each is 0.95. The first case is referred to as the 'mild' case and the second case as 'more severe' case.

5.2   Results

The results from each scenario are displayed in Figures 1-9. Each figure contains four panels. The top left panel displays the average bias of each estimator across the different degrees of measurement error. The top right panel displays the mean absolute percentage error (MAPE), defined as

$$MAPE_s = \frac{1}{R} \sum_{r=1}^{R} \left| \frac{\widehat{\tau}_{sr} - \tau}{\tau} \right|$$

where s indexes a particular estimator, $r = 1, ..., R$ indexes the number of data replications ($R = 1000$), and τ is the population ATE according to the DGP. The bottom two panels display box and whisker plots to illustrate the distribution of point estimates. The bottom right plot simply excludes the three estimators that do not require CIA since these estimators tend to have much wider distributions and distort the scaling of the graphs.

Errors in Treatment Assignment   Figure 1 displays the results with varying degrees of random misclassification in treatment assignment. A few observations stand out. First, with no misclassification, all the estimators have a bias near zero, as one would expect. However, the three estimators that do not

require CIA – BVN, BC-HI, and KV-TSLS – are very inefficient.[11] Second, even with relatively infrequent classification errors (one percent of the sample), the bias of the most efficient estimators – OLS, OLS-PS, and DR – increase more than *sixfold*, although the MAPE remains essentially unaffected. When the rate of misclassification increases to five (ten) percent, the MAPE doubles (quadruples) relative to no misclassification for most of the estimators that require CIA. Misclassification also has a sizeable deleterious effect on BVN and BC-HI, particularly in terms of bias. Third, and most importantly, not all estimators are equally adversely affected by random misclassification. KM with a *large* bandwidth performs marginally worse in the absence of misclassification (due to the bias introduced by giving more weight to observations with large differences in propensity scores), but its performance does not deteriorate – in terms of bias or precision – with misclassification. KV-TSLS also remains unbiased in the presence of misclassification, but is very imprecise despite the fact that its precision actually improves modestly with misclassification.

In Figure 2, the classification errors are correlated with the covariates in the model. A few interesting differences emerge relative to the prior case of random misclassification. First, KV-TSLS is no longer unbiased; its performance is dramatically affected even when the misclassification rate is only one percent (bias increases roughly *eightfold*). Second, the other two estimators that do not require CIA – BVN and BC-HI – do not perform quite as poorly as in the prior case, but they still do not perform well in terms of either bias or precision. Finally, KM with a *large* bandwidth continues to perform best in the face of misclassification. However, even so, researchers need to be wary as the MAPE increases by more than 30% moving from no misclassification to a ten percent misclassification rate.

Errors in Covariates   Figures 3 and 4 display the results with measurement error in the covariates; Figure 3 corresponds to classical measurement error in each of the two covariates, while Figure 4 corresponds to nonclassical measurement error (measurement errors are positively correlated with each other as well as actual treatment assignment).

In Figure 3, several observations stand out. First, with the exception of KM with a large bandwidth, the bias of the estimators that require CIA roughly double *each time* as the reliability ratio of the covariates falls from one to 0.99 to 0.975 to 0.95 to 0.90. KM with a large bandwidth has a relatively sizeable bias even with no measurement error in the covariates, and the bias nearly doubles as the reliability ratio of the covariates drops to 0.90. Second, in terms of the MAPE, there is a less noticeable deterioration in performance as the measurement error worsens. The MAPE roughly doubles for each of the estimators that require CIA with the exception of HI, which only increases by about 30%, as the reliability ratio falls from one to 0.90. In addition, as with bias, the MAPE for KM with a large bandwidth is above that of

---

[11]Note, the DGP imposes homoskedasticity in the treatment assignment equation. Thus, the imprecision of KV-TSLS reflects, at least in part, this fact. See footnote 7.

the other estimators that require CIA. Finally, the three estimators that do not require CIA continue to be very imprecise. However, KV-TSLS does relatively well in terms of bias, particularly as the reliability ratio falls. Specifically, whereas the bias more than doubles in (absolute) magnitude – going from -0.040 to 0.099 – as the reliability ratio of the covariates falls from one to 0.99, the bias is only 0.092 when the reliability ratio falls further to 0.90.

Figure 4 presents a fairly stark contrast to Figure 3 in terms of cost of measurement error in the covariates. With nonclassical measurement error, a reliability ratio of roughly 0.99 severely impacts the performance of all the estimators considered with the possible exception of KV-TSLS. For example, the bias of OLS, OLS-PS, DR, and KM with a small bandwidth increase roughly *fortyfold* when the reliability ratio falls to 0.99. The bias of the other estimators at least double in magnitude. Performance as measured by bias continues to deteriorate as the reliability ratio falls, but at a slower rate. A similar picture emerges when analyzing performance in terms of MAPE as well. Finally, KV-TSLS is most robust in terms of bias once measurement error is introduced, but it is much less precise than the estimators requiring CIA.

Errors in Outcome   The next set of figures, Figures 5 and 6, display the results when measurement error is induced in a continuous outcome. Figure 5 corresponds to the case of classical measurement error; Figure 6 to nonclassical (mean-reverting) measurement error. The results in Figure 5 are as expected: measurement error has no impact on the bias of the estimators, but all estimators suffer a loss in precision as the reliability ratio of the outcome decreases. It does not appear that classical measurement error has a differential effect on performance across the estimators.

Figure 6 reveals that nonclassical measurement error in the outcome is not nearly as benign as classical measurement error. With even a small amount of mean-reverting measurement error – yielding a reliability ratio of 1.01 – the bias of all the estimators considered increase at least *fivefold* with the exception of KM with a large bandwidth. In fact, the bias of OLS, OLS-PS, DR, and KM with a small bandwidth increases by roughly *fiftyfold*. In terms of precision, the estimators requiring CIA – again, with the exception of KM with a large bandwidth – suffer a *four* to *fivefold* increase when the reliability ratio increases to 1.01. However, the bias and MAPE do not deteriorate further as the reliability ratio deviates further from one. Lastly, as noted, KM with a large bandwidth is the most robust under mean-reverting measurement error in the outcome, although the improvement over the other estimators requiring CIA is not dramatic. KM with a large bandwidth still has a MAPE of roughly 20% when the reliability ratio is 1.01; twice as large as under no misclassification.

Figures 7 and 8 display the results using a binary outcome subject to classification errors. Figure 7 corresponds to random misclassification; Figure 8 reflects classification errors correlated with the covariates

20

in the model. A few findings stand out. First, with the exception of the estimators that do not require CIA, the performances of the estimators is relatively consistent across the two cases. Moreover, for the estimators requiring CIA, the estimators do not suffer as a result of the misclassification until the misclassification rate reaches ten percent. With a misclassification rate of ten percent, the MAPE increases *two* to *threefold* (relative to a misclassification rate of only five percent) for this set of estimators. Second, there is little difference in the performance across the various estimators requiring CIA. In contrast, the three estimators that do not require CIA perform significantly worse in terms of bias and precision, although misclassification in the outcome does not further deteriorate their performance.[12] Finally, while the performance of the estimators requiring CIA is unaffected by the change from random to nonrandom misclassification, the three estimators that do not require CIA perform worse – particularly in terms of precision – when the classification errors are correlated with the covariates in the model. For example, with a misclassification rate of ten percent, the MAPE for KV-TSLS is more than twice as large when the classification errors are nonrandom.

Simultaneous Errors   In the final figure, Figure 9, simultaneous measurement errors in the outcome, treatment assignment, and covariates are introduced. Specifically, the data contain mean-reverting measurement error in a continuous outcome, classification errors in treatment assignment correlated with the true values of the covariates, and classical measurement error in the covariates. In the mild case – a reliability ratio of the outcome of 1.01, a misclassification rate for treatment assignment of one percent, and a reliability ratio of the covariates of 0.99 – the performance of all the estimators suffers fairly dramatically relative to no measurement errors. For the estimators requiring CIA, the bias increases from close to zero (except for KM with a large bandwidth) to around 0.3 in absolute value; the MAPE increases roughly *four* to *fivefold*. The exception is KM with a large bandwidth. This estimator, not surprisingly, fares worse than the other estimators requiring CIA in the baseline case of no measurement error. However, in the mild case, the bias and MAPE deteriorate less rapidly relative to the other estimators; this estimator achieves the lowest bias and MAPE among the estimators requiring CIA. In terms of the three estimators that do not require CIA, the bias is actually relatively small for BVN and BC-HI; the bias of KV-TSLS is quite sizeable. However, as in all the previous scenarios, all three estimators are very imprecise. Lastly, in the more severe case – a reliability ratio of the outcome of 1.05, a misclassification rate for treatment assignment of five percent, and a reliability ratio of the covariates of 0.95 – the results are qualitatively similar across the estimators and the additional measurement error has only a marginal detrimental effect

---

[12]In a linear probability model, it is well known that the error term is heteroskedastic. As such, the parametric assumptions of the standard BVN model do not hold, explaining the considerable bias in the BVN and BC-HI estimators even under no misclassification.

relative to the mild case.

Summary   While a number of scenarios have been considered in the Monte Carlo study, a few conclusions can be drawn. First, some types of measurement error lead to an immediate deterioration in performance of the various estimators, while other types of errors do not adversely affect performance until the errors are relatively severe. On the one hand, classification errors – random or correlated with covariates – in treatment assignment, classical measurement error in the covariates, and classification errors – random or correlated with covariates – in a binary outcome do not significantly harm the performance of the various estimators until the measurement error is relatively large. Note, however, that 'relatively large' in the Monte Carlo study is still fairly small for some commonly used variables for which validation studies are available, as detailed in Section 3. On the other hand, nonclassical measurement error in the covariates, mean-reverting measurement error in the outcome, and simultaneous measurement errors in the outcome, treatment assignment, and covariates have a stark, deleterious effect on the performance of the various estimators even with relatively small ad infrequent errors. Unfortunately, as discussed in Section 3, these cases are likely to be frequently encountered in empirical work.

Second, there is no clear winner if one views the Monte Carlo study as a horse race among the various estimators. Nonetheless, KM with a large bandwidth does prove to be more robust to measurement error in many of the scenarios considered. Specifically, with random and nonrandom classification errors in treatment assignment or a binary outcome, mean-reverting measurement error in the outcome, or measurement error in all aspects of the DGP (outcome, treatment assignment, and covariates), KM with a large bandwidth performs as well or better than the remaining estimators. However, with just classical or nonclassical measurement error in the covariates, KM with a large bandwidth is outperformed by other estimators not requiring the CIA.

Finally, KV-TSLS does well in the cases of classical and nonclassical measurement error in the covariates in terms of bias, but not precision. This is not surprising as the instrument is linearly independent of the covariates in the model and, hence, the measurement error in the error term. Thus, while the TSLS procedure does not eliminate the bias of the estimated coefficients on the covariates, it does not eliminate the bias in the estimated treatment effect. Moreover, KV-TSLS performs well in terms of bias, but not precision, with random classification errors in treatment assignment. As stated previously, the relative imprecision of KV-TSLS is at least partly attributable to the fact that the DGPs considered here impose homoskedasticity of the errors in the treatment assignment equation. However, in many of the other scenarios, KV-TSLS does very poorly in terms of both bias and precision. Specifically, it performs particularly poorly – relative to estimators requiring CIA – with nonrandom classification errors

22

in treatment assignment, random and nonrandom classification errors in a binary outcome, or measurement error in all aspects of the DGP (outcome, treatment assignment, and covariates).

6   Conclusion

The program evaluation literature has expanded rapidly over the past few decades. While our knowledge concerning methods that are designed to provide consistent estimates of some measure of the causal effect of a binary treatment under conditional independence is relatively well developed, the consequences of measurement error on the performance of these methods is not. In this study, a fairly extensive Monte Carlo study is undertaken to examine the absolute and relative performance of many estimators under various degrees of measurement, entering through different channels. Overall, the results suggest a cautionary tale to researchers tempted to treat measurement error like the elephant in the corner and simply ignore it. In particular, nonclassical measurement error in the covariates, mean-reverting measurement error in the outcome, and simultaneous measurement errors in the outcome, treatment assignment, and covariates have a dramatic, adverse effect on the performance of the various estimators even with relatively small and infrequent errors.

Unfortunately, no single estimator performs best across the various scenarios considered. Thus, applied researchers ought to utilize a number of methods to assess sensitivity of the estimated treatment effects. That said, kernel matching with a relatively *large* bandwidth does outperform the other estimators in a number of situations frequently encountered. Specifically, with random and nonrandom classification errors in treatment assignment or a binary outcome, mean-reverting measurement error in the outcome, or simultaneous measurement errors in the outcome, treatment assignment, and covariates, this estimator does outperform the others considered. Nonetheless, much work is needed to develop estimators that can address not just measurement error in one aspect of the data-generating process, but multiple aspects. Until then, researchers need to be extremely wary of the consequences of ignoring measurement error.

References

[1] Abrevaya, J. and J.A. Hausman (2004), "Response Error in a Transformation Model with an Application to Earnings-Equation Estimation," *Econometrics Journal,*.7, 366-388.

[2] Aigner, D.J. (1973), "Regression with a Binary Independent Variable Subject to Errors of Observation," *Journal of Econometrics*, 1, 49-60.

[3] Almeida, H., M. Campello, and A.F. Galvao Jr. (2010), "Measurement Errors in Investment Equations," NBER Working Paper No. 15951.

[4] Ashenfelter, O. and A. Krueger (1994), "Estimates of the Economic Return to Schooling from a New Sample of Twins," *American Economic Review*, 84, 1157-1173.

[5] Bang, H. and J.M. Robins (2005), "Doubly Robust Estimation in Missing Data and Causal Inference Models," *Biometrics*, 61, 962-972.

[6] Barron, J.M., M.C. Berger, and D.A. Black (1997), "How Well Do We Measure Training?" *Journal of Labor Economics*, 15, 507-528.

[7] Basu, A., D. Polsky, and W.G. Manning (2008), "Use of Propensity Scores in Non-Linear Response Models: The Case of Health Care Expenditures," NBER Working Paper No. 14086.

[8] Battistin, E. and B. Sianesi (2010), "Misclassified Treatment Status and Treatment Effects: An Application to Returns to Education in the UK," *Review of Economics and Statistics*, forthcoming.

[9] Battistin, E. and A. Chesher (2009), "Treatment Effect Estimation with Covariate Measurement Error," CEMMAP Working Paper 25/09.

[10] Black, D.A., M.C. Berger, and F.A. Scott (2000), "Bounding Parameter Estimates with Nonclassical Measurement Error," *Journal of the American Statistical Association*, 95, 739-748.

[11] Black, D., S. Sanders, and L. Taylor (2003), "Measurement of Higher Education in the Census and Current Population Survey," *Journal of the American Statistical Association*, 98, 545-554.

[12] Black, D.A. and J.A. Smith (2004), "How Robust is the Evidence on the Effects of College Quality? Evidence from Matching," *Journal of Econometrics*, 121, 99-124.

[13] Bollinger, C.R. (1996), "Bounding Mean Regressions When a Binary Regressor is Mismeasured," *Journal of Econometrics*, 73, 387-399.

[14] Bound, J., C. Brown, and N.A. Mathiowetz (2001), "Measurement Error in Survey Data," in J.J. Heckman and E.E. Leamer (eds.) *Handbook of Econometrics*, Vol. 5, 3705-3843.

[15] Card, D. (1996), "The Effect of Unions on the Structure of Wages: A Longitudinal Analysis," *Econometrica*, 64, 957-979.

[16] Chua, T.C. and W.A. Fuller (1987), "A Model for Multinomial Response Error Applied to Labor Flows," *Journal of the American Statistical Association*, 82, 46-51.

[17] de Marchi, S. and J.T. Hamilton (2006), " Assessing the Accuracy of Self-Reported Data: An Evaluation of the Toxics Release Inventory," *Journal of Risk and Uncertainty*, 32, 57-76.

[18] Dehejia, R. H., and S. Wahba (1999), "Casual Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 94, 1053-1062.

[19] Fisher, R.A. (1935), *The Design of Experiments*, Edinburgh: Oliver & Boyd.

[20] Frazis, H. and M.A. Loewenstein (2003), "Estimating Linear Regressions with Mismeasured, Possibly Endogenous, Binary Explanatory Variables," *Journal of Econometrics*, 117, 151-178.

[21] Freeman, R.B. (1984), "Longitudinal Analyses of the Effects of Trade Unions," *Journal of Labor Economics*, 2, 1-26.

[22] Frisch, R. (1934), *Statistical Confluence Analysis By Means of Complete Regression Systems*, Oslo, University Institute for Economics.

[23] Gini, C. (1921), "Sull'interpolazione di una retta quando i valori della variabile indipendente sono affetti da errori accidntali," *Metron*, 1, 63-82.

[24] Griliches, Z. (1985), "Data and Econometricians–The Uneasy Alliance," *American Economic Review*, 75, 196-200.

[25] Griliches, Z. and V. Ringstad (1970), "Error in Variables Bias in Nonlinear Contexts," *Econometrica*, 42, 971-998.

[26] Hausman, J.A. (2001), "Mismeasured Variables in Econometric Analysis: Problems from the Right and Problems from the Left," *Journal of Economic Perspectives*, 15, 57-67.

[27] Hausman, J.A., J. Abrevaya, and F.M. Scott-Morton (1998), "Misclassification of the Dependent Variable in a Discrete-Response Setting," *Journal of Econometrics*, 87, 239-269.

[28] Hausman, J.A., W.K. Newey, H. Ichimura, and J.L. Powell (1991), "Identification and Estimation of Polynomial Errors-in-Variables Models," *Journal of Econometrics*, 50, 273-295.

[29] Heckman, J. and S. Navarro-Lozano (2004), "Using Matching, Instrumental Variables, and Control Functions to Estimate Economic Choice Models," *Review of Economics and Statistics*, 86, 30-57.

[30] Hirano, K. and Imbens, G.W. (2001), "Estimation of Causal Effects using Propensity Score Weighting: An Application to Data on Right Heart Catheterization," *Health Services and Outcomes Research Methodology*, 2, 259-278.

[31] Hirano, K., G.W. Imbens, and G. Ridder (2003), "Efficient Estimation of Average Treatment Effects using the Estimated Propensity Score," *Econometrica*, 71, 1161-1189.

[32] Horvitz, D.G. and D.J. Thompson (1952), "A Generalization of Sampling Without Replacement from a Finite Universe," *Journal of the American Statistical Association*, 47, 663-685.

[33] Hu, Y. (2006), "Bounding Parameters in a Linear Regression Model with a Mismeasured Regressor Using Additional Information," *Journal of Econometrics*, 133, 51-70.

[34] Hyslop, R. and G.W. Imbens (2001), "Bias from Classical and Other Forms of Measurement Error," *Journal of Business & Economic Statistics*, 19, 475-481.

[35] Imbens, G.W. (2004), "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," *Review of Economics and Statistics*, 86, 4-29.

[36] Imbens, G.W. and J.M. Wooldridge (2009), "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47, 5-86.

[37] Keane, M.P. and R.M. Sauer (2009), "Classification Error in Dynamic Discrete Choice Models: Implications for Female Labor Supply Behavior," *Econometrica*, 77, 975-991.

[38] Klepper, S. and E.E. Leamer (1984), "Consistent Sets of Estimates for Regressions With Errors in All Variables," *Econometrica*, 52, 163-184.

[39] Klein, R. and F. Vella (2009), "A Semiparametric Model for Binary Response and Continuous Outcomes Under Index Heteroskedasticity," *Journal of Applied Econometrics*, 24, 735-762.

[40] Koopmans, T. (1937), *Linear Regression Analysis of Economic Time Series*, Amsterdam, Netherlands Econometric Institute, Harrlem-de Erwen F Bohn N.V.

[41] Kreider, B. and J.V. Pepper (2007), "Disability and Employment: Reevaluating the Evidence in Light of Reporting Errors," *Journal of the American Statistical Association*, 102, 432-441.

[42] Kreider, B. (2010), "Regression Coefficient Identification Decay in the Presence of Infrequent Classification Errors," *Review of Economics and Statistics*, forthcoming.

[43] Lewbel, A. (1997), "Constructing Instruments For Regressions With Measurement Error When No Additional Data Are Available, With an Application to Patents and R&D" *Econometrica*, 65, 1201-1213.

[44] Lewbel, A. (2007), "Estimation of Average Treatment Effects with Misclassification," *Econometrica*, 75, 537–551.

[45] Li, T., P.K. Trivedi, and J. Guo (2003), "Modeling Response Bias in Count: A Structural Approach with an Application to the National Crime Victimization Survey Data," *Sociological Methods & Research*, 31, 514-544

[46] Mellow, W. and H. Sider (1983), "Accuracy of Response to Labor Market Surveys: Evidence and Implications," *Journal of Labor Economics*, 1, 331-344.

[47] Millimet, D.L. and R. Tchernis (2009), "On the Specification of Propensity Scores: with Applications to the Analysis of Trade Policies," *Journal of Business & Economic Statistics*, 27, 397-415.

[48] Millimet, D.L. and R. Tchernis (2010), "Estimating Treatment Effects Without an Exclusion Restriction: With an Application to the School Breakfast Program," NBER WP No. 15539.

[49] Mitchell, C. (2010), "Are Divorce Studies Trustworthy? The Effects of Survey Nonresponse and Response Errors," *Journal of Marriage and Family*, 72, 893-905.

[50] Mroz, T. (1999), "Discrete Factor Approximations for Use in Simultaneous Equation Models: Estimating the Impact of a Dummy Endogenous Variable on a Continuous Outcome," *Journal of Econometrics*, 92, 233-274.

[51] Neyman, J. (1923), "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9," translated in *Statistical Science*, (with discussion), 5, 465-480, (1990).

[52] Poterba, J.M. and L.H. Summers (1986), "Reporting Errors and Labor Market Dynamics," *Econometrica*, 54, 1319-1338.

[53] Reiersøl, O. (1950), "Identifiability of a Linear Relation Between Variables Which Are Subject to Error," *Econometrica*, 18, 375-389.

[54] Rosenbaum, P.R. and D.B. Rubin (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41-55.

[55] Roy, A.D. (1951), "Some Thoughts on the Distribution of Income," *Oxford Economic Papers*, 3, 135-146.

[56] Rubin, D. (1974), "Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies," *Journal of Educational Psychology*, 66, 688-701.

[57] Rubin, D. (1986), "Statistics and Causal Inference: Which Ifs Have Causal Answers," *Journal of the American Statistical Association*, 81, 961-962.

[58] Scharfstein, D.O., A. Rotnitzky, and J.M. Robins (1999), "Adjusting for Nonignorable Dorp-Out Using Semiparametric Nonresponse Models," *Journal of the American Statistical Association*, 94, 1096-1120.

[59] Schennach, S.M. (2004), "Estimation of Nonlinear Models with Measurement Error," *Econometrica*, 72, 33-75.

[60] Smith, J.A. and P.E. Todd (2005), "Does Matching Overcome LaLonde's Critique?" *Journal of Econometrics*, 125, 305-353.

[61] Whittemore, A.S. and G. Gong (1991), "Poisson Regression with Misclassified Counts: Application to Cervical Cancer Mortality Rates," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 40, 81-93.

[62] Wooldridge, J.M. (2002), *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Cambridge, MA.
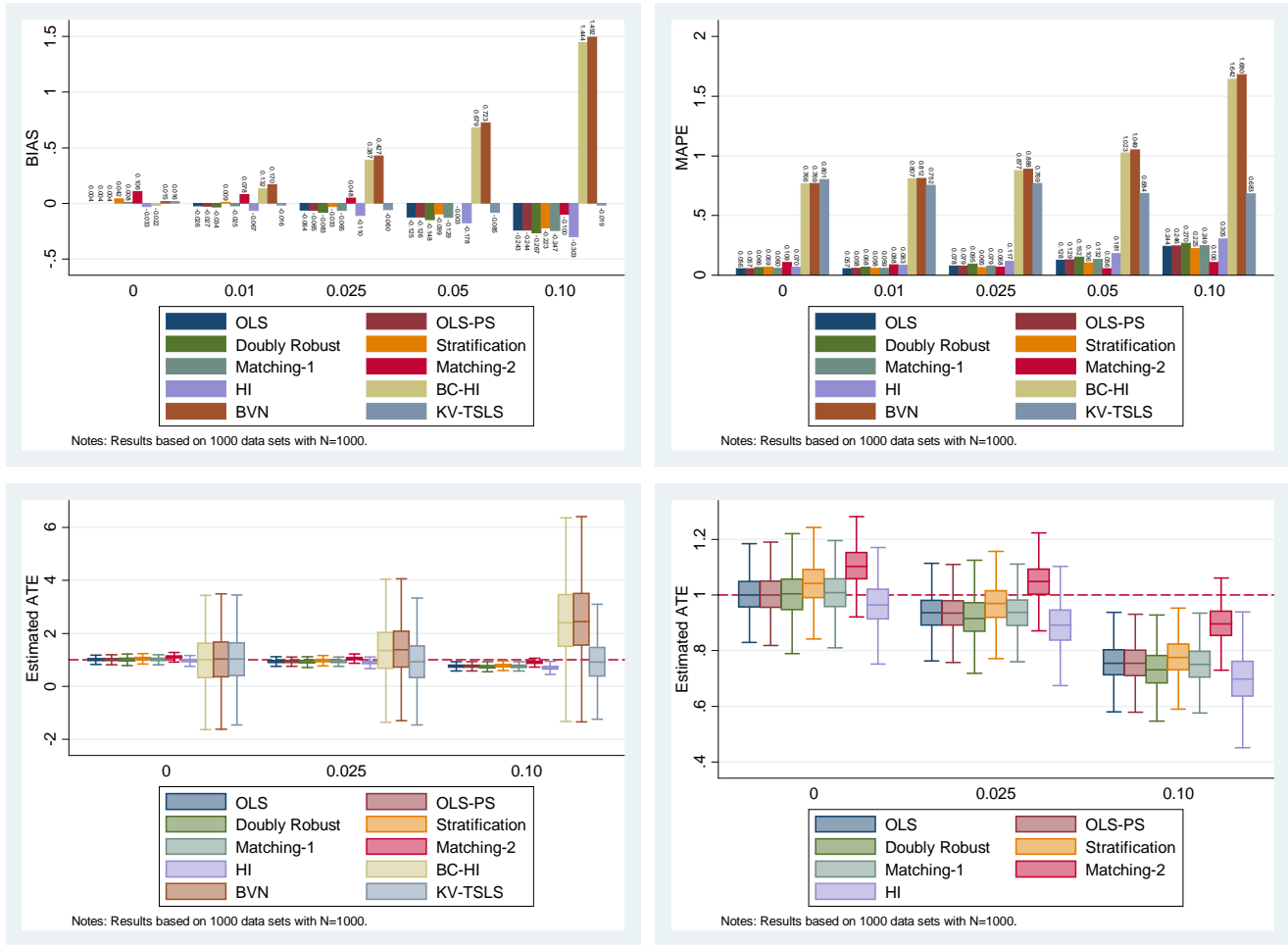
Figure 1. Random Classification Error in Treatment Assignment.

Notes: Numbers on horizontal axis refer to the misclassification rate. MAPE = Mean Absolute Percentage Error. In the bottom panel, box and whisker plots provide boxes that span the 25th and 75th percentiles of the estimates with the middle line at the median estimate; the whiskers extend to 1.5 times the interquartile range. The dashed line in this panel represents the true value of the ATE.
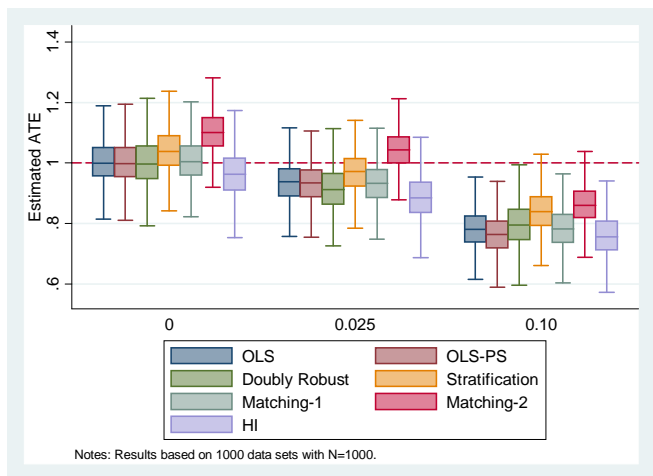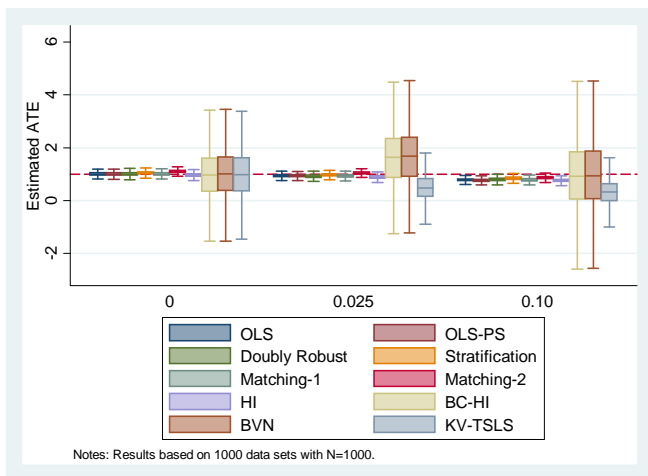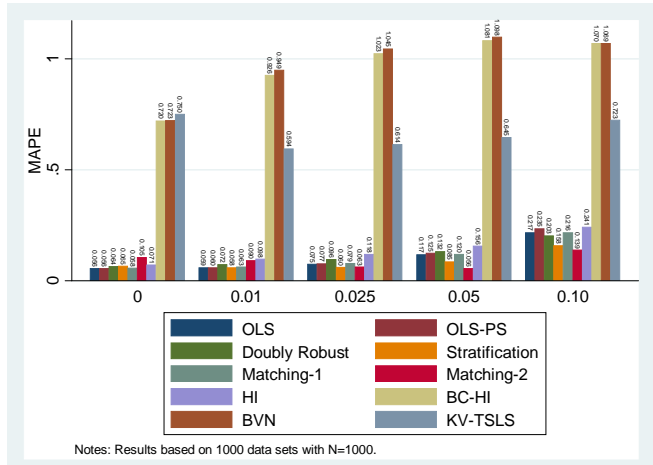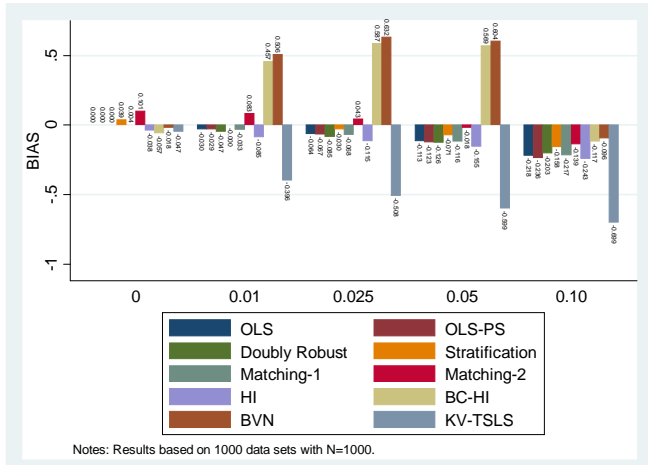
Figure 2. Correlated Classification Error in Treatment Assignment.

Notes: Classification errors are correlated with covariates in the model. See Figure 1 for additional details.
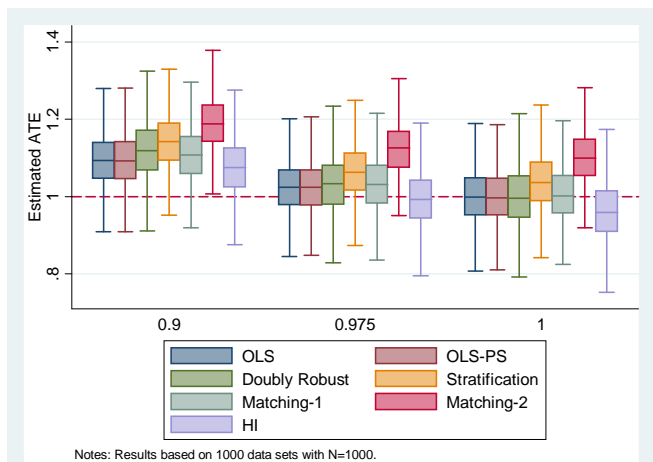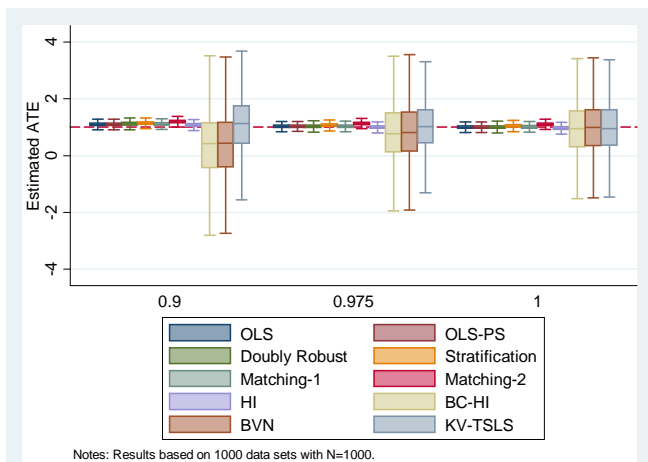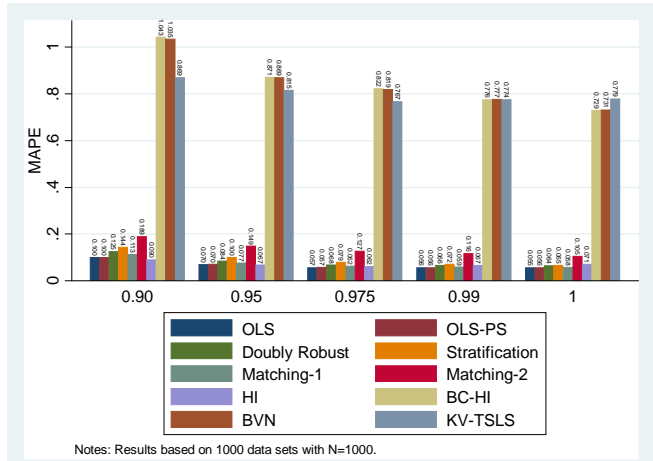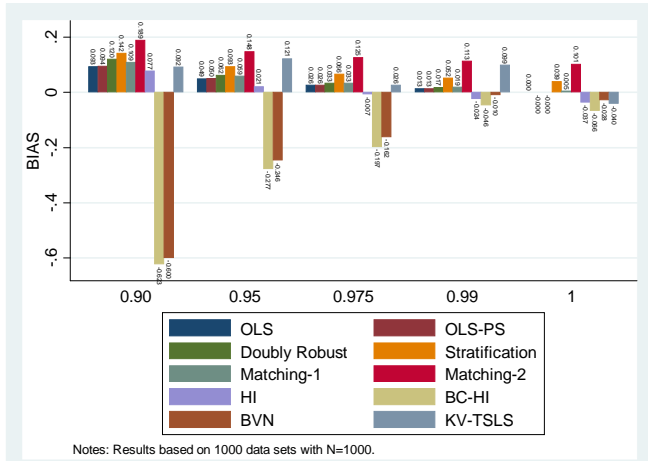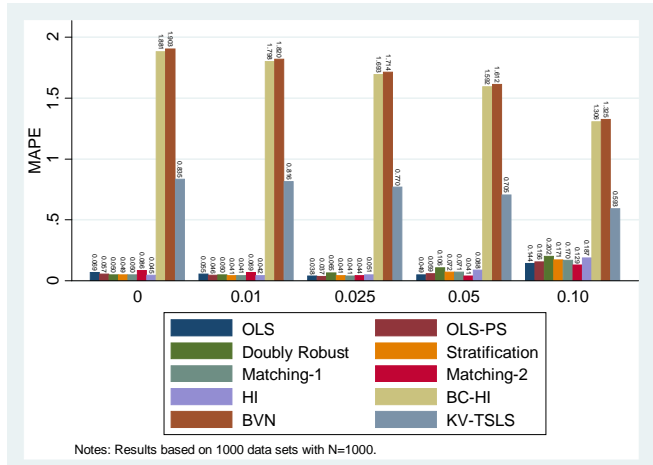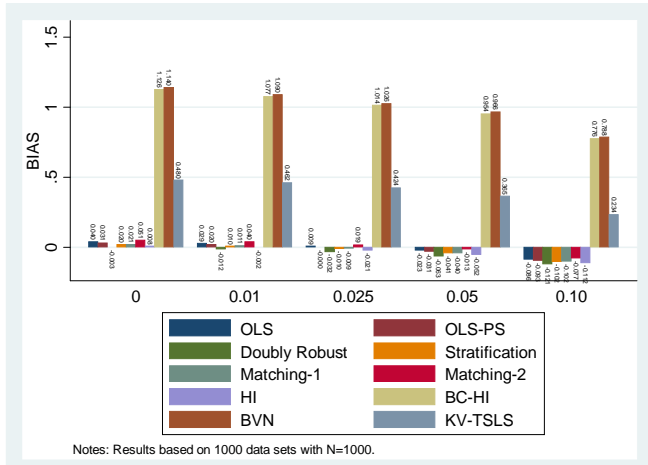
Figure 3. Classical Measurement Error in Covariates.

Notes: Numbers on horizontal axis refer to the reliability ratio for each of the covariates in the model. See Figure 1 for additional details.

31
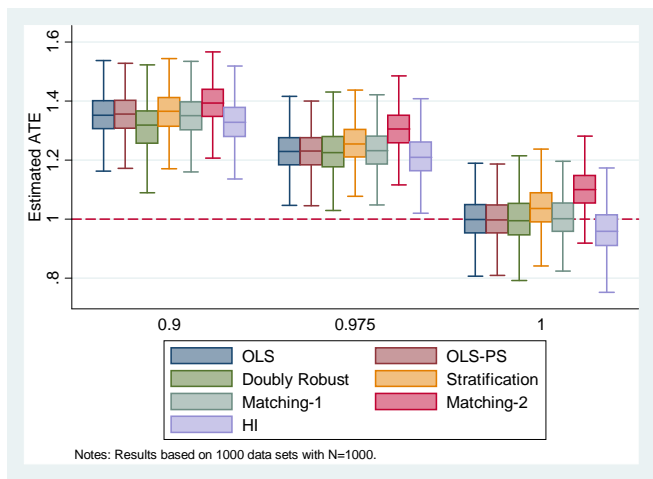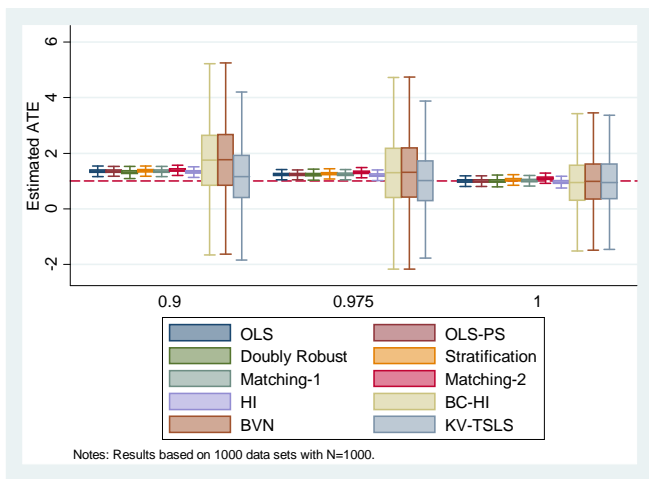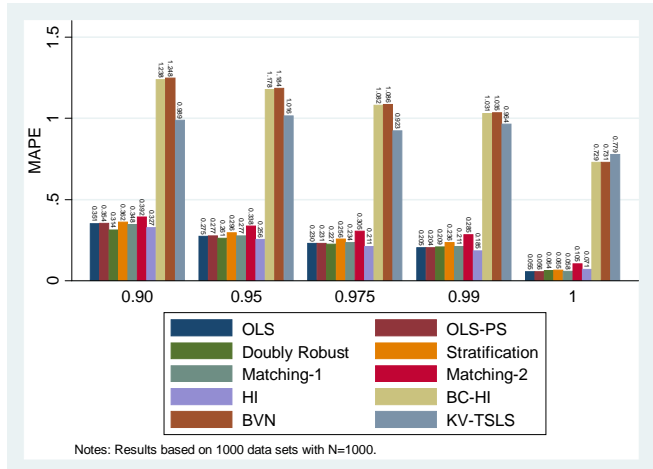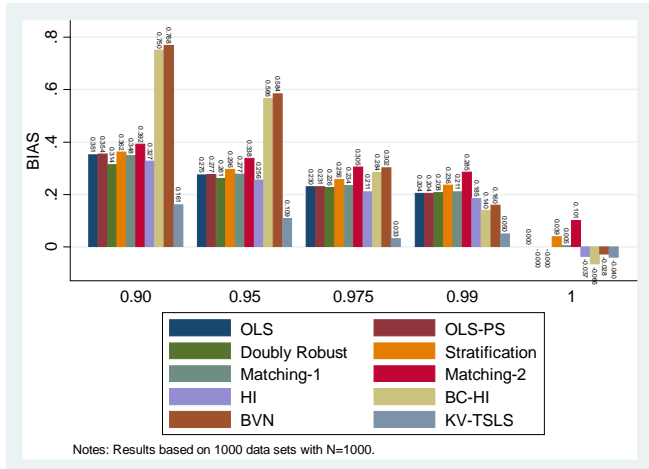
Figure 4. Nonclassical Measurement Error in Covariates.

Notes: Measurement errors for each of the covariates are correlated with each other and treatment assignment. See Figure 1 for additional details.
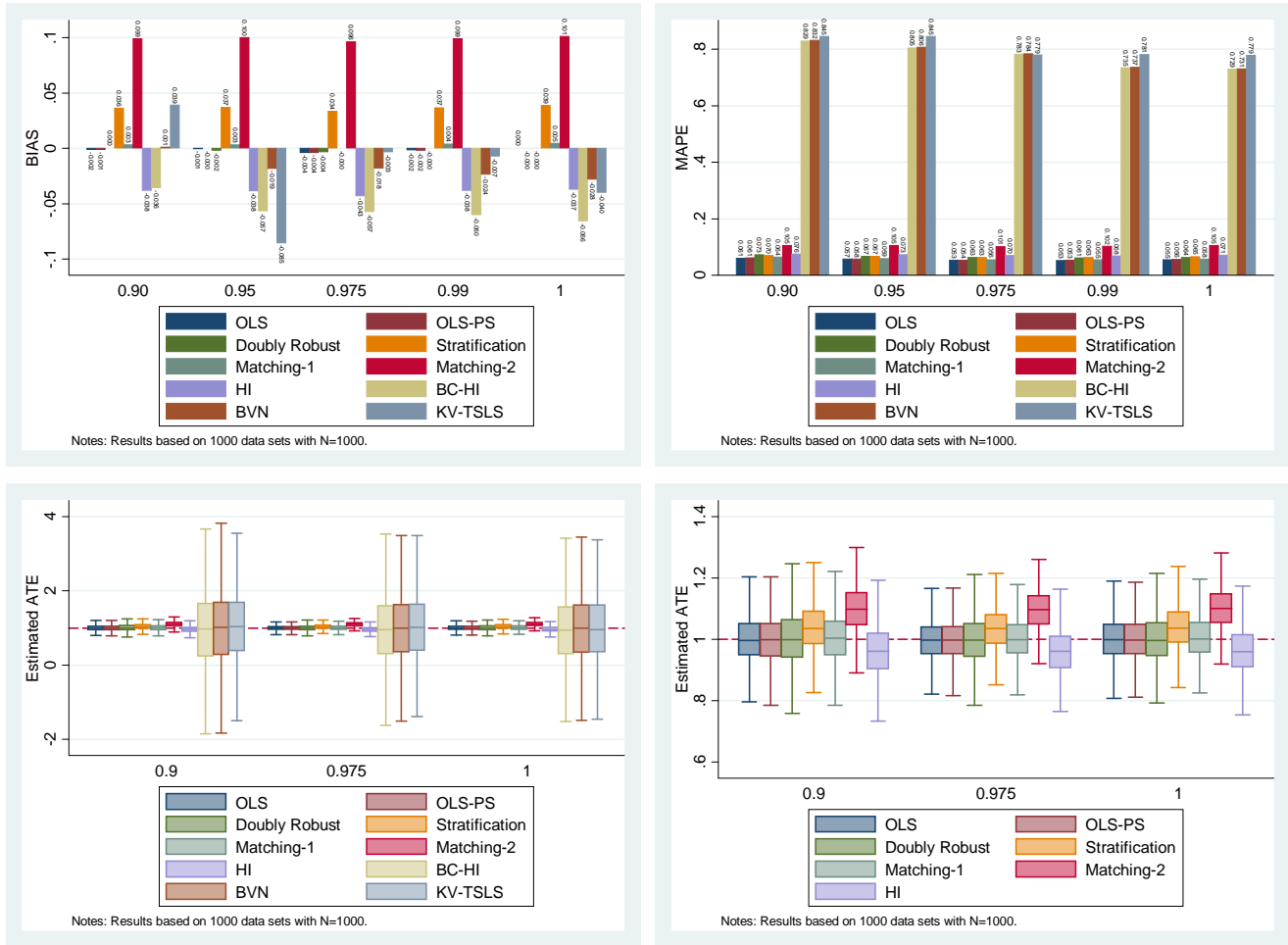
Figure 5. Classical Measurement Error in the Outcome.

Notes: Numbers on horizontal axis refer to the reliability ratio for the outcome. See Figure 1 for additional details.
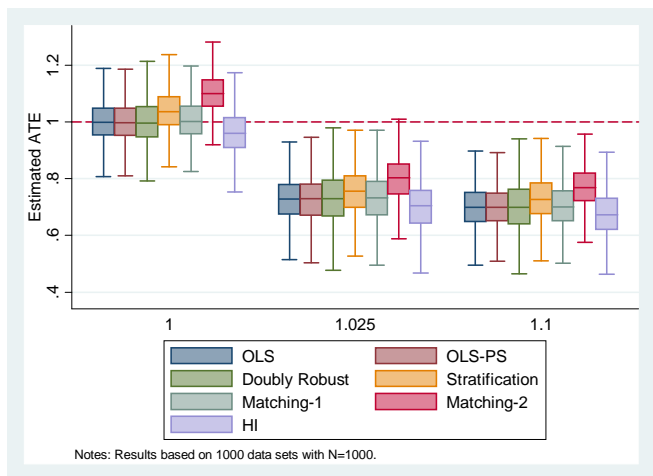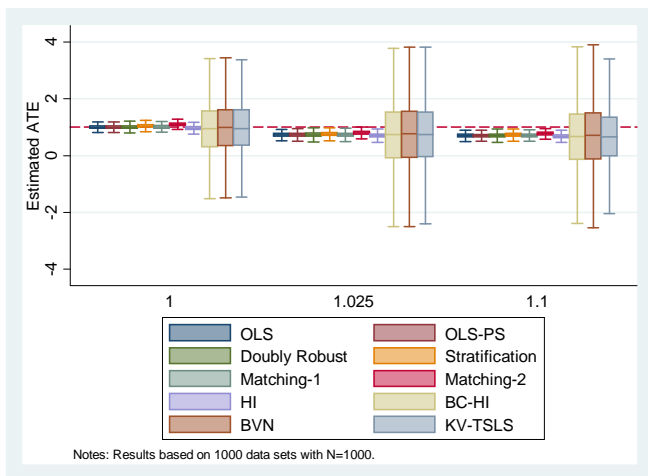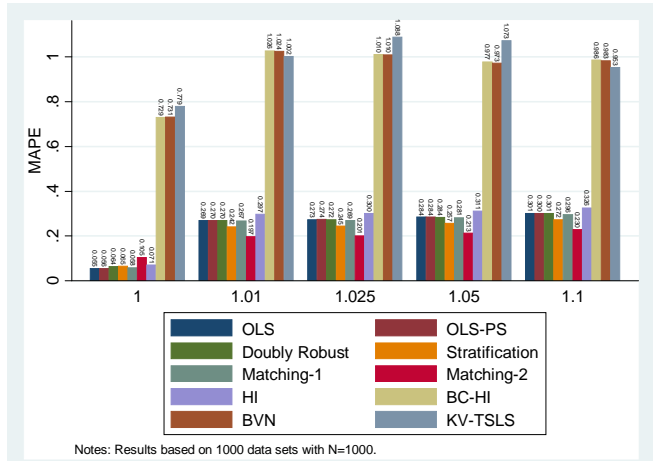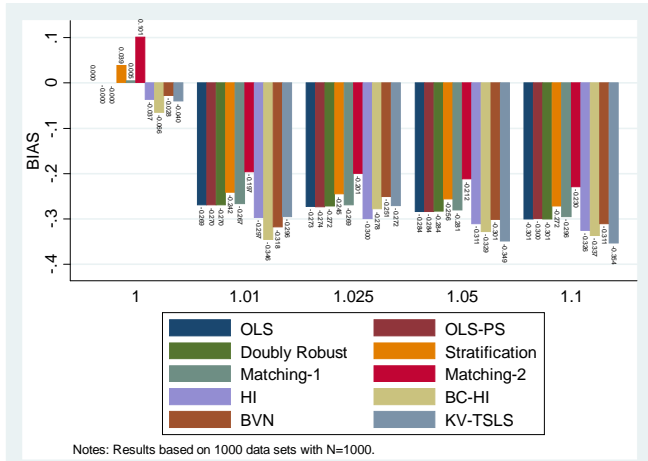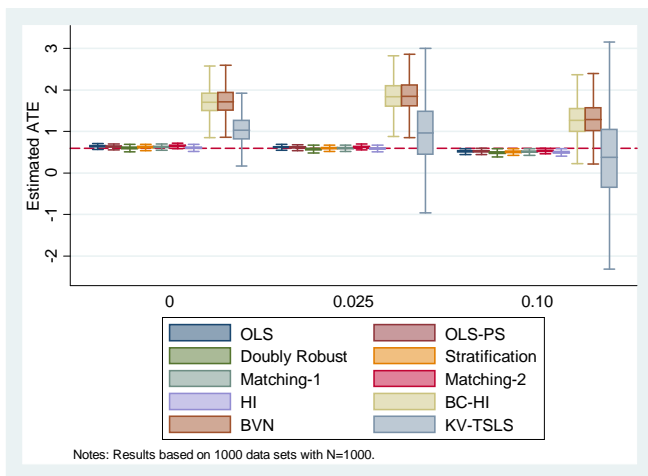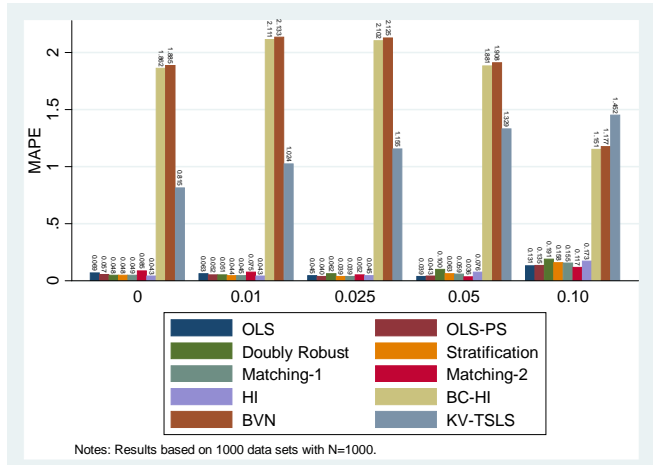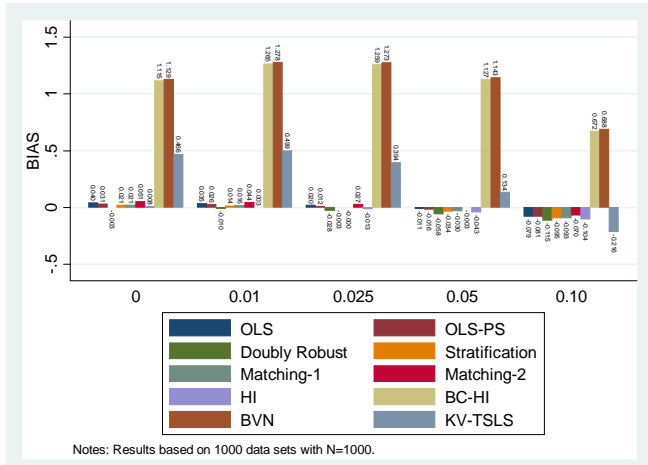
Figure 6. Mean-Reverting Measurement Error in the Outcome.

Notes: Measurement error in the outcome is negatively correlated with the true value. See Figure 1 for additional details.
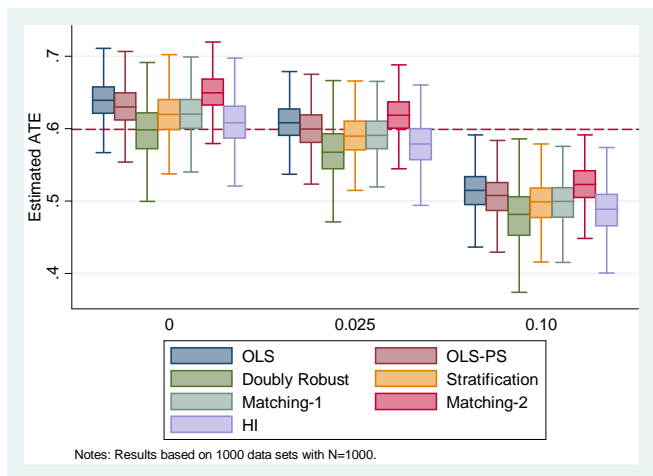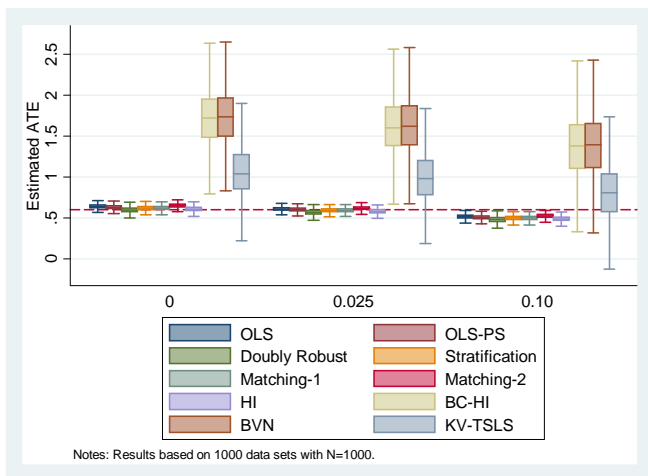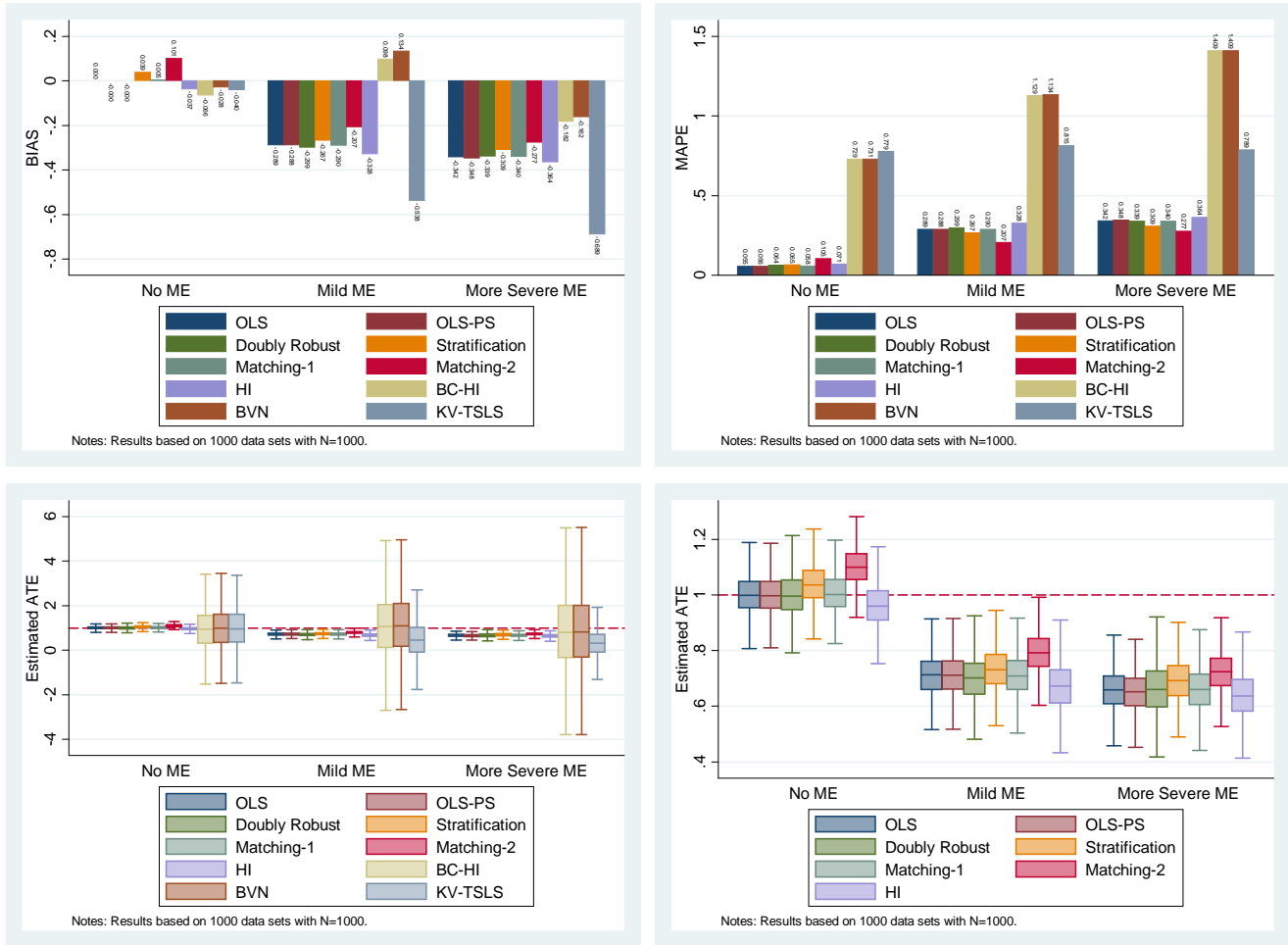
Figure 7. Random Classification Error in a Binary Outcome.

Notes: Numbers on horizontal axis refer to the misclassification rate. See Figure 1 for additional details.
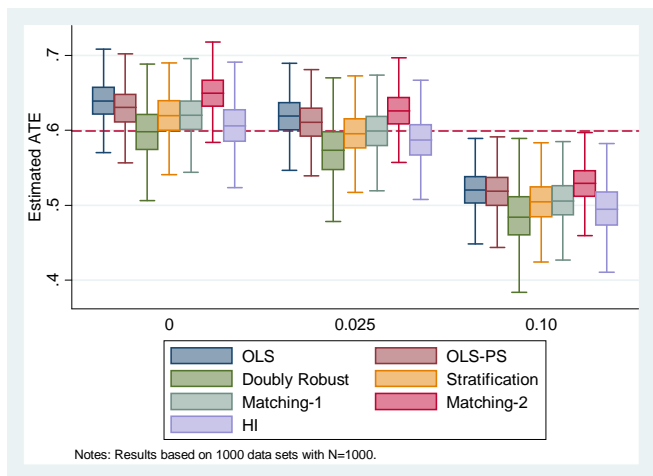
Figure 8. Correlated Classification Error in a Binary Outcome.

Notes: Classification errors are correlated with covariates in the model. See Figure 1 for additional details.

Figure 9. Simultaneous Measurement Error in Treatment Assignment, Outcome, and Covariates.

Notes: 'No ME' case entails no measurement error in any of the data. 'Mild ME' case entails a misclassification rate of 1% for treatment assignment correlated with the true values of the covariates, mean-reverting measurement error in the outcome yielding a reliability ratio of 1.01, and measurement error in the covariates yielding a reliability ratio of 0.99. 'More Severe ME' case entails a misclassification rate of 5% for treatment assignment correlated with the true values of the covariates, mean-reverting measurement error in the outcome yielding a reliability ratio of 1.05, and measurement error in the covariates yielding a reliability ratio of 0.95. See Figure 1 for additional details.