How Do Propensity Score Methods Measure Up in the Presence of Measurement Error? A Monte Carlo Study

Patricia Rodríguez de Gil, Aarti P. Bellara, Rheta E. Lanehart, Reginald S. Lee, Eun Sook Kim, and Jeffrey D. Kromrey
University of South Florida

Considering that the absence of measurement error in research is a rare phenomenon and its effects can be dramatic, we examine the impact of measurement error on propensity score (PS) analysis used to minimize selection bias in behavioral and social observational studies. A Monte Carlo study was conducted to explore the effects of measurement error on the treatment effect and balance estimates in PS analysis across seven different PS conditioning methods. In general, the results indicate that even low levels of measurement error in the covariates lead to substantial bias in estimates of treatment effects and concomitant reduction in confidence interval coverage across all methods of conditioning on the PS.

KEYWORDS measurement error; Monte Carlo simulation; propensity score methods

Identifying causal relationships in social science continues to be a challenge. Causal relationships exist between two variables when the following hold true: (a) the cause precedes the effect, (b) the cause is related to the effect, and (c) no plausible alternative explanations for the effect exist other than the cause (Shadish, Cook, & Campbell, 2002). Causal relationships could be identified using a counterfactual model, which is the difference between what did happen after an individual received a treatment versus what would have happened if the same individual did not receive the treatment (Campbell & Stanley, 1963; Holland, 1986; Rubin, 2010; Shadish et al., 2002). Causality could be precisely estimated if a unit was assigned to the treatment condition and the control condition at the same time in the same context; however, in most social science settings, units cannot be assigned only to one condition. This impossibility of observing both treatment and control outcomes for each individual is referred to as the “Fundamental Problem of Causal Inference” (Holland, 1986, p. 947; Rubin, 2010).

A framework recognized in the literature as Rubin’s Causal Model (or RCM) is used to describe the statistical properties related to causal inference, specifically the concept of potential outcomes (Rubin, 2010; West & Thoemmes, 2010). In RCM, treatment effects are determined by comparing the potential outcomes that would have been observed for an individual under different conditions. These outcomes are considered “potential” as each individual cannot be observed under various conditions simultaneously. RCM combats the fundamental problem of causal inference using a statistical solution to estimate the average treatment effect (ATE) based on the expected value of the difference in outcomes, or a counterfactual model. This solution to the fundamental problem allows causal inferences to be drawn using outcome measures observed from different individuals (Holland, 1986). Since outcomes for all treatment conditions cannot be observed for all units, the RCM operates under two key assumptions: the strongly ignorable treatment assignment assumption and the stable unit treatment value assumption or SUTVA.

The strongly ignorable treatment assignment assumption refers to the mechanism or process used to assign individuals to conditions and requires the assignment to condition be independent and not associated with the outcome or other factors. When this assumption has been met, inferences to the population can be made by samples comprised of units receiving only one condition. SUTVA asserts that the
outcomes from two individuals, irrespective of treatment assignment, are independent from one another.

**PURPOSE**

Propensity score methods (PSM) have proven effective in reducing selection bias in nonexperimental (observational) studies (Rosenbaum & Rubin, 1983). Across various disciplines, methodological studies have examined the conditions under which different PSM are able to adequately estimate treatment effects (e.g. Austin, 2009; Austin, Grootendorst, Normand, & Anderson, 2007; Michalopolous, Bloom, & Hill, 2004; Rubin, 2001; Steiner, Cook, Shadish, & Clark, 2010). However, the majority of these studies have not considered the impact that measurement error may have on such estimates.

Measurement error is a common phenomenon in many fields of study (Cochran, 1968). Buonaccorsi (2011) stated, “in some sense, all statistical problems involve measurement error” (p. 1). Both continuous and categorical variables are often measured with error, and whether random or systematic, measurement error can seriously impact the estimation of treatment effects and alter their interpretation (Fuller, 1987). Although the study of measurement error has grown in recent years (Fuller, 1987; Buonaccorsi, 2011), the examination of the impact of measurement error on propensity score analysis has been scarce. In this paper, we present results from a simulation study examining the extent to which measurement error impacts the estimation of treatment effects when applying PSM to minimize selection bias in observational studies. It should be noted that this paper does not purport to provide correction of bias due to measurement error in performing PSM. However, some suggestions on models for measurement error will be provided in the discussion section.

**PROPENSITY SCORE METHODS**

Introduced by Rosenbaum and Rubin (1983) the propensity score is a statistic used to reduce bias in observational studies. It mimics the balance between treatment and control groups that occur through random assignment. All things being equal, when estimating treatment effects, randomly assigning units to treatment and control conditions is preferred. The rationale behind this procedure is that it will, on average, yield probabilistically similar groups, and any differences can be attributed to the absence or presence of a particular treatment. Results from studies in which units are not randomly assigned to treatment and control groups, however, should be interpreted cautiously because any observed differences may be attributed to various unknown, unmeasured baseline differences between groups, which are now not probabilistically similar (Shadish et al., 2002).

A propensity score (PS) is an estimate of an individual’s probability for being assigned to the treatment group conditional on observed covariates, \( p(z = 1|X) \) where \( z \) denotes treatment assignment and \( X \) represents a set of observed covariates. Variables thought to be related to such assignment are included in the estimation model. The closer an individual’s PS is to 1, the stronger the prediction for being in the treatment group, conditioned upon the observed covariates modeled. Conversely, the closer a PS is to 0, the stronger the prediction for being in the comparison group, conditioned upon the observed covariates modeled. When units from the treatment and control group (\( i \) and \( k \), respectively) have the same or similar propensity score, it is assumed that the probability of being assigned to the treatment group is the same for each of these individual units, conditional upon the observed covariates.

\[
p(z_i = 1|X) = p(z_k = 1|X)
\]  

When there is no overlap in PSs between the groups, it is believed—and hoped—that unobserved covariate(s) are accounting for the difference in groups (Stuart, 2010). The review of the procedures of PS analysis is not included here because many resources are available for interested readers (e.g., Caliendo & Kopeinig, 2008; D’Agostino, 1998; Thoemmes & Kim, 2011).

**MEASUREMENT ERROR IN PROPENSITY SCORE ANALYSIS**

Logistic regression is the most extensively used statistical procedure to estimate propensity scores (Shadish & Steiner, 2010). In the propensity score model, \( y \) is a binary indicator of the probability of being in the treatment group as a function of \( x \) factors or predictors. However, the measurement of these baseline predictors or covariates “is always subject to error” (Crocker & Algina, 1986, p. 6), and as a result, if the measurement error is substantial, it could bias the estimated probabilities and affect their ability to balance a two-group analysis, consequently yielding misleading inferences when estimating treatment effects.

Although research on the effects of random measurement errors in regression analysis has a fairly long history (see Pedhazur, 1997, for a brief review) and the effects of measurement errors on the validity of regression analysis can be severe (Cochran, 1968), Jencks et al. (1972) suggested that “the most frequent approach to measurement error is indifference” (p. 330). The apparent neglect of this assumption in applied research may be related to a lack of its treatment in the technical literature in social science (for a notable exception, see Huynh, 2006).

Carroll, Ruppert, Stefanski, and Crainiceanu (2006) describe measurement error in predictors as presenting a “triple whammy” for data analysis: it results in biased parameter estimates, reduced statistical power, and a masking of the
functional form of relationships (making detection of non-linearity and nonadditivity difficult). In the simplest multiple regression case of two regressors with OLS modeling, the fundamental problem of measurement error in regressors is evident in the following equation from Cochran (1968):

$$\beta_1^* = \frac{\beta_1 \rho_{11} (1 - \rho_{12}^2 \rho_{22}) + \beta_2 \rho_{21} \rho_{11} (1 - \rho_{22})}{1 - \rho_{12}^2 \rho_{11} \rho_{22}}$$

where $\beta_1^*$ is the regression coefficient for $X_1$ measured with error, $\rho_{ii} =$ the reliability coefficient for $X_i$, $\beta_i =$ the regression coefficient that would be obtained if $X_i$ were measured without error, $\rho_{ij} =$ the correlation between the two predictors, and $\beta_{2.1} =$ the regression coefficient of $X_2$ on $X_1$. Note that $\beta_1^* = \beta_1$ only if $\rho_{ii} = 1.00$, and that even if $\rho_{ii} = 1.00$ (with $\rho_{22} < 1.00$), the regression coefficient for $X_1$ will be biased. Cochran (1968) summarizes the situation with more than two regressors, stating the “effects of measurement error on individual regression coefficients in a multiple linear regression are complicated” (p. 655).

Raykov (2012) demonstrated that propensity score analysis with fallible covariates (i.e., observed covariates with measurement error) does not lead to an unbiased estimate of the ATE when treatment assignment is related to the true scores. When one or more covariates are contaminated with measurement error, Equation (1) does not hold because treatment assignment is not independent of the true scores of fallible covariates even after those fallible covariates are controlled for. That is, when individuals from treatment and control groups have the same propensity score estimated with fallible covariates, the probability of being assigned to the treatment group may not be the same conditional upon those observed covariates. Thus, a propensity score estimated with fallible covariates is neither necessarily a balancing score nor a true propensity score. To handle covariates with measurement error, Raykov proposed a modified PSM, a latent variable modeling approach in which the estimated true scores (i.e., factor scores from the latent variable model) are used to estimate propensity scores in logistic regression, which will be discussed later.

Steiner, Cook, and Shadish (2011) conducted a simulation study in which a constant or fixed treatment effect (ATE) was assumed and different degrees of unreliability were induced in the covariates used to create the propensity score, and investigated the influence of this measurement error on bias reduction. Considering that bias reduction is predicated upon covariates effectiveness, “true” baseline data were obtained from previous studies that already demonstrated the quality of covariates for reducing bias when there was a comprehensive number of covariates (23) and when a small subset of covariates (8) were shown to remove selection bias (Shadish, Clark, & Steiner, 2008; and Steiner et al., 2010, respectively). A total of 2,000 replications were conducted, by assuming initial reliability of covariates to be 1 (i.e., no error added to the covariate data), and manipulating the amount of unreliability on the observed covariates, systematically decreasing the reliability of each covariate ($\rho_{ii} = .9, .8, .7, .6, .5$). Including all covariates as main effects in the PS model, both the PS and outcome model were estimated at each replication, using stratification based on PS quintiles, linear regression including PS logits, PS weighting using weighted least squares, and linear regression with original covariates but without any PS adjustments. In the multivariable case, bias in treatment effect was recorded to estimate the attenuation rates in bias reduction (i.e., average bias across 2,000 replications). However, research on using multiple unreliable covariates (two or more) is not only complex (Steiner et al., 2011) but also scarce. In addition, because it is not possible to add or separate error from observations’ true-value, to date, this has yet to be examined through simulation research. The impact of unreliability of covariates on the performance of PSM has not been fully investigated under a variety of simulation conditions. Thus, the focus of the present simulation study was to determine the impact of covariate measurement error on propensity score analysis.

METHOD

Data Sources and Analysis

This research was a Monte Carlo study in which random samples were generated under known and controlled conditions from normal population distributions. A factorial mixed design with completely crossed factors included seven between-subjects factors—number of covariates (3, 9, 15, and 30), population treatment effect (0, .2, .5, and .8), covariate relationship to treatment (mean partial regression weight of 0.025, 0.050, and 0.100), covariate relationship to outcome (mean partial regression weight of 0.025, 0.050, and 0.100), correlation among the covariates (0, .2, .5), sample size (50, 100, 250, 500, and 1,000), and covariate reliability (.4, .6, .8, and 1.0) and one within-subjects factor—propensity score conditioning method (matching without a caliper, matching using $.25 SD$ of the PS as a caliper, ignoring the covariates, ANCOVA, PS as a covariate, stratification, and PS weighting using the inverse probability of treatment weight). Each condition simulated presented a continuous outcome variable, a binary treatment indicator, and both continuous and binary covariates (with a 2:1 ratio of continuous to binary variables). For each condition, 5,000 samples were simulated. The use of 5,000 estimates provided adequate precision for the investigation of the sampling behavior of point and interval estimates of the model coefficients. That is, 5,000 samples provided a maximum 95% confidence interval width around an observed proportion that is $\pm .014$ (Robey & Barcikowski, 1992). This completely crossed factorial design ($4 \times 4 \times 3 \times 3 \times 3 \times 5 \times 4 \times 7$) provided a total of 60,480 conditions.

In each simulated sample, measurement error was induced, the propensity score was estimated using logistic
regression, and the PS conditioning methods were applied to estimate the treatment effect. The ability of the propensity score conditioning methods to balance the sample data between the control and treatment groups was evaluated for each PS method. The PS distributions of the control and treatment groups were also evaluated, looking at the areas of overlap. Trimming was performed and the PS and PS methods were recalculated to investigate the impact of trimming based on common support.

Outcome measures associated with treatment effect estimates included statistical bias, RMSE, Type I error control, and 95% confidence interval coverage and width. Outcome measures associated with balance estimates included standardized mean differences for the observed covariates before and after conditioning, as well as variance ratios. In addition to estimates of balance and treatment effects, descriptive statistics were analyzed regarding the trimming of the region of common support.

RESULTS

Results were analyzed by computing $\eta^2$ values to estimate the proportion of variability in each of the outcomes (e.g., statistical bias, RMSE, CI coverage and width, Type I error rate) associated with each factor in the simulation design.
The patterns of the mean values of the outcomes associated with the factors identified were subsequently analyzed.

Common Support

The distributions of common support for sample size, reliability of covariates, number of covariates, covariate relationship to treatment, covariate relationship to outcome, and correlation among covariates are presented in Figure 1. The range of common support between treatment and control groups increased as the sample size increased from 50 to 250 (from .41 to .61), but stabilized in larger sample sizes. The nonsupport range (i.e., the range of non-overlap) almost equaled the support range at \( N = 50 \), and dramatically decreased as sample size increased. The optimum combination of the support/nonsupport ranges occurred when \( N = 1,000 \) (Figure 1a). Similarly, as sample size increased from 50 to 1,000, the proportion of nonsupport cases decreased from .44 to .03.

As the reliability of the covariates increased, the range of the common support increased very slightly, and so did the nonsupport range (Figure 1b), resulting in almost horizontal parallel lines. These data suggest that an increase in reliability does not demonstrably impact the range of common support between propensity score groups.

The increase of the number of covariates included in the simulation model resulted in an increase in both support and nonsupport ranges up to 15 covariates, with support range increasing in greater increments. However, after 15 covariates, the support range decreased and the nonsupport range more than doubled (Figure 1c). The proportion of nonsupport cases became larger as the number of covariates increased.

As the relationship between covariates and treatment assignment increased, the support range increased in a linear fashion (Figure 1d). The nonsupport range and proportion of nonsupport cases also increased. There was no change in either the support range or nonsupport range as the relationship between the covariates and the dependent variable increased, resulting in parallel lines (Figure 1e), indicating that regression weights for the dependent variable have no impact on the support range of propensity score groups. Higher correlation among covariates provided a greater support range accompanied by a slight increase in nonsupport range as the covariate correlation increased (Figure 1f).

Balance

In this study, balance was estimated using the standardized mean difference for continuous covariates and the log odds ratio for binary covariates. Thus, in both types of covariates, a value of zero indicated a perfect balance between groups in comparison (e.g., treatment and control). It should be noted that ANCOVA was not included in the evaluation of balance.

Regarding PS conditioning methods, stratification showed the most serious unbalance (on average 0.59 and 0.44 for binary and continuous covariates, respectively, as presented in Table 1) followed by ignoring the covariates and matching without using a caliper. On the other hand, matching with a caliper and weighting consistently showed excellent balance irrespective of design factors. PS ANCOVA achieved near zero balance when small sample size conditions were excluded, which will be discussed in the following section.

The balance of PS ANCOVA was greatly impacted by the number of covariates, but the negative relation between balance and the number of covariates was moderated by sample size. That is, when sample size was small (\( N = 50 \) and \( N = 100 \)) and the number of covariates was large (\( k = 15 \) and \( k = 30 \)), conditioning using the estimated PS as a covariate in the model resulted in extreme unbalance in the covariates.

As the sample size was 250 or larger, the negative effect of the number of covariates on balance was not observed. The effect of sample size was also apparent in stratification. Overall, the balance from stratification was poor, but improved in larger sample sizes. Trimming the samples for regions of non-overlap may improve or reduce the degree of balance depending on the conditioning method, sample size, and type of covariates. When stratification was employed for small samples, covariates became less balanced after trimming. Weighting also showed more unbalance after trimming. In PS ANCOVA, trimming removed the extreme outlying cases leading to substantially improved balance in continuous covariates (when \( N = 50 \), mean balance before trimming = –426.28; mean balance after trimming = 0.01). On the contrary, trimming worsened the balance of binary covariates under the same condition. In other conditioning methods, the effects of trimming appear negligible.

\[ ^{2} s \] denotes supplemental materials, available online. See Tables 3s, 4s, 5s for the mean balance by covariate type and reliability, by covariate type and the correlation among the covariates, and by trimming and sample size, respectively.
Statistical Bias in Estimates of Treatment Effect

Bias in this study was computed as the difference between the estimated treatment effect and the corresponding population parameter established in the simulation. The results of bias estimates under different simulation factors are presented first. Based on the evaluation of eta-squared for all main effects and first-order interactions, the effects of simulation factors on bias for each conditioning method are followed.

The mean bias estimates by conditioning method, number of covariates, and reliability are reported in Figure 2 (Table 6s for the corresponding table). Overall, the estimates of no caliper matching and ignoring the covariates were considerably biased across all simulation conditions (e.g., from 0.08 to 1.47 as k increased from 3 to 30) and these two poorly performing methods are not included in Figure 2. We also observed substantial bias in other conditioning methods except when the number of covariates was small and their reliability was 1.0. As measurement error increased so did bias regardless of how the data were conditioned. Figure 2 also showed that the number of covariates in the PS model impacted the bias (9 covariates vs. 30 covariates). As we added more covariates, the bias increased consistently across all methods by reliability level. When the number of covariates was large (i.e., 30), weighting and stratification yielded larger bias compared to caliper matching, ANCOVA, and PS ANCOVA.

For each conditioning method, eta-squared values for all main effects and first-order interactions were examined to identify major factors related to variability in the bias estimates. The major factors related to bias depended upon the conditioning methods although the number of covariates was identified as a primary factor explaining bias for many conditioning methods (Table 7s).

In the case of caliper matching, both the number of covariates and their reliability emerged as major factors related to bias ($\eta^2 = \text{about } .08$). The estimate of treatment effect was more biased as the number of covariates increased and as the covariate reliability decreased. For no caliper matching, ignoring the covariates, and PS ANCOVA, the same sets of simulation factors were related to bias: the number of covariates ($\eta^2 = .27$) followed by the relation of covariates to the outcome ($\eta^2 = .11$) and to the other covariates ($\eta^2 = .10$). The interactions between the first factor and the latter two factors were also observed. That is, when the number of covariates was large, the correlation among the covariates and the relation to the outcome increased bias more seriously. Bias results from conditioning by stratification were also impacted by the number of covariates ($\eta^2 = .16$) and their relations to the outcome ($\eta^2 = .07$) whereas reliability helps the reduction of bias ($\eta^2 = .06$). No factor examined in this study was prominent in relation to bias in ANCOVA. Finally, the number of covariates ($\eta^2 = .07$) and its interaction with trimming ($\eta^2 = .09$) mainly explained the bias of weighting. Trimming reduced the bias of weighting and the effect became larger when the number of covariates was large. For example, trimming decreased bias from 0.03 to 0.02 with 3 covariates; trimming removed bias from 0.80 to –0.04 with 30 covariates.

RMSE in Estimates of Treatment Effect

The typical difference between a single sample estimate of the treatment effect and the true treatment effect was estimated by the root mean squared error (RMSE). Overall, there is a large RMSE estimate across all methods but the estimate tends to decrease with increases in sample size. Figure 3 shows boxplot distributions of RMSE by PS method when sample size equals 50 and 1,000. As evident in the boxplots, larger samples reduce both the typical value of RMSE and the variability in RMSE across the other design factors although the impact of sample size on RMSE is not prominent in the two poorly performing methods (ignoring the covariates and matching without a caliper). The smallest RMSE estimates are observed for the small number of covariates (3 and 9), across all PS methods.

Type I Error Control in Tests of Treatment Effect

The overall distributions of Type I error rate estimates are presented in Figure 4. Notable in this figure is the great inflation of Type I error rates for matching without a caliper and for ignoring the covariates. In contrast, caliper matching and PS ANCOVA provide median Type I error rates near the nominal .05 level. However, all of the conditioning methods demonstrate a large dispersion in Type I error rates across the conditions simulated. To identify the simulation design factors associated with variability in Type I error rates, eta-squared values ($\eta^2$) were computed for all main effects and first order interactions.

In addition to the main effects for sample size and covariate reliability, the interaction between these two factors was associated with variability in Type I error rates for all conditioning methods (with $\eta^2$ ranging from .07 to .14) except matching without a caliper and ignoring the covariates (Table 9s). For all conditioning methods except matching without a caliper and ignoring the covariates, the Type I error rates increased as reliability of the covariates decreased regardless of the sample size. However, the relationship between covariate reliability and Type I error rate was stronger with larger samples. For example, with caliper matching and $N = 50$, the estimated mean Type I error rates ranged from .09 with $\rho_{xx} = .40$ to .07 with $\rho_{xx} = 1.00$. In contrast, with $N = 1,000$, the estimated mean Type I error rates ranged...
FIGURE 2  Estimated bias of treatment effect by conditioning method, covariate reliability, and the number of covariates ($k$). Population treatment effect = 0, 0.2, 0.5, and 0.8.

The first-order interaction between covariate reliability and the strength of relationship between the covariates and treatment assignment was also associated with the variability in Type I error control (with $\eta^2$ ranging from .05 to .07). The mean Type I error rate estimates associated with this interaction are presented in Table 2. These data indicate that the impact of measurement error in the covariates is greater when the relationship between the covariates and treatment assignment is stronger. For example, with caliper matching and a modest relationship between the covariates and treatment assignment ($\beta_{xc} = 0.025$), the mean Type I error rate
ranged from only .03 to .12 as measurement error increases. With a stronger relationship between the covariates and treatment assignment ($\beta_{xc} = 0.100$), the mean Type I error rate ranged from .06 to .54.

The first-order interaction between sample size and the strength of relationship between the covariates and treatment assignment ($\beta_{xc}$) was also associated with variability in Type I error control (with $\eta^2$ ranging from .05 to .06). The impact of $\beta_{xc}$ on Type I error rates is greater with larger samples (Table 10s). With the caliper matching, the mean Type I error rates ranged from only .04 to .12 with increases in $\beta_{xc}$ when $N = 50$. In contrast, with $N = 1,000$, the mean Type I error rates ranged from .13 to .54 with increases in $\beta_{xc}$.

Finally, the number of covariates ($k$) was associated with variability in Type I error rates (with $\eta^2$ ranging from .12 to .22 for some of the conditioning methods). For all conditioning methods, larger numbers of covariates were associated with higher Type I error rates (Table 11s). This effect is attributable to the greater degree of confounding of the treatment with larger numbers of covariates.

In addition to the examination of mean Type I error rates, the Type I error control of the conditioning methods was evaluated using Bradley’s (1978) liberal criterion of robustness.

| $\beta_{xc}$ | $\rho_{xx}$ | Caliper Match | No Caliper Match | Ignore | ANCOVA | Propensity Score ANCOVA | Weighting | Stratification |
|-------------|------------|---------------|------------------|--------|--------|--------------------------|-----------|---------------|
| .025        | .40        | .12           | .40              | .41    | .17    | .13                      | .14       | .18           |
| .60         | .07        | .40           | .40              | .10    | .06    | .03                      | .06       | .11           |
| .80         | .04        | .39           | .40              | .06    | .03    | .02                      | .06       | .07           |
| 1.00        | .03        | .40           | .40              | .05    | .02    | .05                      | .05       | .05           |
| .050        | .40        | .31           | .60              | .37    | .33    | .34                      | .34       | .39           |
| .60         | .17        | .60           | .62              | .20    | .17    | .20                      | .20       | .25           |
| .80         | .08        | .60           | .61              | .10    | .07    | .13                      | .13       | .13           |
| 1.00        | .04        | .59           | .60              | .05    | .03    | .11                      | .11       | .07           |
| .100        | .40        | .54           | .77              | .59    | .57    | .57                      | .60       | .62           |
| .60         | .37        | .76           | .78              | .40    | .37    | .43                      | .46       | .46           |
| .80         | .17        | .75           | .76              | .18    | .16    | .27                      | .26       | .26           |
| 1.00        | .06        | .73           | .74              | .06    | .05    | .24                      | .24       | .09           |

Note: The mean Type I error rate meeting the Bradley’s liberal criterion (i.e., between .025 and .075) is in bold.
This criterion indicates that Type I error control is considered acceptable if the actual Type I error rate is within the bounds of $\alpha_{\text{nominal}} \pm 0.5 \cdot \alpha_{\text{nominal}}$. At the nominal $\alpha = .05$, these bounds are 0.25 and 0.75. This view of Type I error control supports the results of the mean Type I error rate analyses presented above. For example, the impact of covariate measurement reliability is striking (Table 12s). With caliper matching, adequate Type I error control is maintained in 74% of the conditions when covariates are measured without error, but in only 57% of the conditions when $\rho_{xc} = .80$, and in only 24% of the conditions when $\rho_{xc} = .40$. Similar degradation of Type I error control is evident with stratification and the ANCOVA approaches. Regardless of the conditioning method used, larger samples result in smaller proportions of conditions with adequate Type I error control, as do larger numbers of covariates. Increasing the strength of the relationship between the covariates and treatment assignment ($\beta_{xc}$) resulted in smaller proportions of conditions with adequate control, but the strength of the relationship between the covariates and the outcome variable ($\beta_{yc}$) was not related to the proportions of conditions. Among the conditioning methods, caliper matching, ANCOVA, and PS ANCOVA generally showed better control of Type I error rates followed by weighting and stratification. Although trimming the samples for common support had little effect on the proportions of conditions with adequate Type I error control, the direction of the effect differed among the conditioning methods. For example, the use of weighting provided adequate control in only 21% of the conditions before trimming, but in 29% of the conditions after trimming. In contrast, stratification provided adequate control in 33% of the conditions before trimming and only 29% of the conditions after trimming.

Confidence Interval Coverage

The distributions of confidence interval coverage estimates across all conditions simulated were first examined (Figure 3s). Evident are the overall poor interval coverage for matching without a caliper and ignoring the covariates. Use of any conditioning method besides matching without a caliper resulted in smaller proportions of conditions with adequate control. The strength of the relationship between the covariates and treatment assignment ($\beta_{xc}$) resulted in smaller proportions of conditions with adequate control, but the strength of the relationship between the covariates and the outcome variable ($\beta_{yc}$) was not related to the proportions of conditions. Among the conditioning methods, caliper matching, ANCOVA, and PS ANCOVA generally showed better control of Type I error rates followed by weighting and stratification. Although trimming the samples for common support had little effect on the proportions of conditions with adequate Type I error control, the direction of the effect differed among the conditioning methods. For example, the use of weighting provided adequate control in only 21% of the conditions before trimming, but in 29% of the conditions after trimming. In contrast, stratification provided adequate control in 33% of the conditions before trimming and only 29% of the conditions after trimming.

The mean interval coverage estimates by conditioning method and covariate reliability when $N = 100$ and $N = 1,000$ are presented in Figure 5 (see Table 13s for all sample size conditions and for all conditioning methods). These data indicate that the impact of covariate measurement error is much greater with larger samples than with smaller samples. For example, with $N = 50$, the mean interval coverage for caliper matching dropped from .92 to only .90 as covariate reliability dropped from 1.0 to .40. In contrast, with $N = 1,000$, the mean coverage estimate dropped from .95 to .40. Note that even with covariate reliability of .80, the interval coverage is substantially reduced for all methods of conditioning (with the exceptions of matching with no caliper and ignoring the covariates—methods that evidence very poor coverage regardless of the reliability of the covariates). As was noted in Type I error control, the dramatic impact of covariate measurement error with large samples is the result of bias in the point estimate of the treatment effect combined with smaller standard errors obtained with larger samples. The smaller standard errors lead to narrower confidence intervals with a concomitant reduction in interval coverage.

The mean coverage estimates by covariate reliability and the strength of the relationship between the covariates and treatment assignment ($\beta_{xc}$) were examined (Table 14s). As expected, these data indicate that the impact of covariate measurement error on confidence interval coverage is greater when the covariates are more strongly related to treatment assignment. With caliper matching and $\beta_{xc} = 0.025$, the coverage estimates drop from .96 to .86 as covariate reliability drops from 1.00 to .40. However, with a stronger relationship between the covariates and treatment assignment ($\beta_{xc} = 0.100$), the coverage for caliper matching drops from .94 to .52 with the same reduction in covariate reliability. As noted with the sample size interaction, with a strong relationship between the covariates and treatment assignment, a reliability value as high as .80 presents sufficient measurement error to notably reduce the average coverage probability to .86 (from .94) with caliper matching.

Confidence Interval Width

The distributions of confidence interval widths across all conditions (Figure 4s) suggest that the differences in the average widths of intervals across PS methods are relatively small and that substantial variability in widths is evident for all PS conditioning methods. The eta-squared analysis indicated that two research design factors were substantially associated with variability in interval width: sample size and number of covariates. However, the interaction between these factors was also sizable (with $\eta^2$ ranging from .08 to .17).

As expected, larger samples result in narrower confidence intervals regardless of the conditioning method or number
of covariates (Table 15s). Similarly, increasing the number of covariates results in wider intervals across the methods and the sample sizes. However, the impact of increasing the number of covariates is greater with smaller samples than with larger samples. For example, with caliper matching and small samples \((N = 50)\) the mean widths of the intervals increased from 0.73 to 11.16 as the covariates increased from 3 to 30. Conversely, with \(N = 1,000\), the mean interval widths increased from 0.12 to only 0.71 across the same range of covariates.

**DISCUSSION**

While PS methods have primarily been applied in medical research, recently there has been an increase in its use in social science research (Thoemmes & Kim, 2011). Much of social science research relies on effects estimated from nonrandomized studies. Given the inability to use random assignment in many studies, for example, in education, there has been a call for methodologists interested in education research to examine methods that approximate randomization (Schneider, Carnoy, Kilpatrick, Schmidt, & Shavelson, 2007). One of the most promising alternatives to randomization is propensity score analysis. In an effort to increase the methodological knowledge-base of propensity score analysis this study empirically investigated its performance under conditions common in behavioral and social science research.

The simulation data for common support suggest that sample size and the number of covariates exert the strongest influence on the support range of propensity score groups. The larger the sample size, the higher support range and lower nonsupport range. In addition, the inclusion of more than 15 covariates in the models examined in this simulation study reduced the support range and increased the nonsupport range.

With respect to balance, no simulation design factor makes a notable impact. On average, both continuous and binary covariates of caliper matching, weighting, and large sample PS ANCOVA were well balanced across simulation factors. Trimming in general does not help reduce imbalance. However, under certain circumstances (e.g., continuous covariates of PS ANCOVA with small sample size) trimming is recommended for balance improvement. As a side note, trimming may help the bias reduction and adequate control of Type I error for weighting.

Depending on conditioning methods, different simulation factors were associated with the bias of treatment effect estimates. Overall, the number of covariates and the reliability of the covariates emerged as major simulation design factors related to bias: more covariates and more measurement error induced more bias. The number of covariates often interacted with other simulation factors such as the correlation of covariates with the outcome and the other covariates in explaining the variability in bias estimates. When the number of covariates is small and the reliability of the covariates is high, the treatment effect estimates are generally unbiased. The smaller bias associated with the smaller number of co-
variates is not surprising. When the model is correctly specified and the other factors are held constant, the PS and the corresponding treatment effect are likely better estimated in a smaller model with a smaller number of covariates. Among the conditioning methods, no caliper matching and ignoring the covariates yielded the most biased estimates.

Covariate reliability was seen to have a profound impact on both Type I error control and confidence interval coverage for the estimation of treatment effects. This effect was more prominent with larger samples (in which standard errors are smaller) and with conditions in which the relationship between treatment assignment and the covariates was strong (in which selection bias is stronger). Especially important is that even small amounts of measurement error (i.e., $\rho_{xx} = .80$) result in notable decrements in the accuracy of inferences. In summary, the impact of covariate measurement error varies depending on the outcome criteria in the simulation. In general, common support, balance, RMSE, and confidence interval width are not affected by the unreliability of covariates, whereas such unreliability makes substantial impacts on bias, Type I error control, and confidence interval coverage.

Given the adverse impact of measurement error on the performance of PSM, methods that deal with measurement error and that can be easily incorporated into PSM are worthy of extra attention. In the extant literature, there are at least a couple of models that explicitly take into account or correct for measurement error: latent variable models and errors-in-variables logistic regression models. Raykov (2012) proposed a modified PSM replacing fallible observed covariates with the corresponding factor scores estimated from the latent variable model. To obtain estimated true scores of fallible covariates without measurement error (i.e., factor scores), each fallible covariate needs two or more indicators congeneric to the corresponding latent construct. The estimated true scores ($T$) of the fallible covariates ($W$) in addition to a set of error-free covariates ($X$) are used in a model estimating propensity scores. Subsequently, instead of the uncorrected PS with the fallible covariates $p(z = 1|X, W)$, the modified PS, $p(z = 1|X, T)$ is used in the conditioning methods of PSM.

On the other hand, the errors-in-variables logistic regression model (Carroll et al., 2006) corrects bias in the parameter estimates and respective standard errors caused by measurement error in the covariates. For simplicity of discussion, in simple linear regression under measurement error, the regression coefficient of a predictor $x (\beta_x)$ is a biased estimate. The unbiased estimate, $\hat{\beta}_x$ can be obtained as $\rho_{xx} \beta_x$ if the reliability ($\rho_{xx}$) of the predictor or measurement error variance is available. For a general approach applicable to nonlinear models such as logistic regression, they proposed regression calibration in which the unknown true scores of fallible covariates are replaced by the regression of true scores on the observable covariates consisting of both fallible and measurement error-free covariates. The regression of true scores on the observed covariates can be done in different ways, for example, with known measurement error variance, using replicate data, or through bootstrapping.

Given the problems of measurement error in PSM, researchers are strongly encouraged to adopt a measurement error correction method when performing propensity score analysis with fallible covariates. In fact, the methods introduced above are easily applied in practice. Stata program code is available for regression calibration in the Stata website. A measure with multiple indicators is often available and any software program for latent variable modeling can provide factor scores of a measure that can be used for propensity score estimation.

AN EMPIRICAL STUDY

The data for this analysis were survey data from the High School Longitudinal Survey (HSLS:09), a nationally representative study of 9th graders in 2009 with follow-up data in 2011 that were obtained from the Institute of Education Sciences (IES), National Center for Educational Statistics (NCES). The data included surveys of students, school administrators, school counselors, and parents. These data were used as part of a funded project from the National Science Foundation focused on persistence in STEM course pathways (NSF #1139510).

The analysis sample consisted of 11,250 students (STEM = 2830; Non-STEM = 8420) enrolled in regular, charter, or magnet high schools ($n = 870$) in the United States. STEM students were identified by enrollment in rigorous 8th grade math courses with a grade of C or better and intent to enroll in rigorous 9th grade math courses or enrollment in rigorous 8th grade science courses with a grade of C or better and intent to enroll in rigorous 9th grade science courses.

Logistic regression was used to estimate the propensity to be in a STEM pathway in the 8th grade. The propensity to be STEM was predicted by the 227 covariates selected from the surveys, as well as the normalized student survey weight. The covariates included characteristics of the students (e.g., gender, race, participation in math or science clubs), the families (e.g., language spoken in the home, attendance at school meetings), and the schools (e.g., AYP status, program offerings in math and science). The standardized math score at

| Conditioning Method | Treatment Effect Estimate | 95% CI |  |
|---------------------|---------------------------|-------|---|
| Caliper match       | 5.25                      | [4.37, 6.14] |
| No caliper match    | 6.29                      | [5.45, 7.13] |
| Ignore              | 10.35                     | [9.58, 11.13] |
| ANCOVA              | 5.48                      | [4.73, 6.22] |
| PS ANCOVA           | 5.60                      | [4.75, 6.44] |
| Weighting           | 4.50                      | [3.82, 5.18] |
| Stratification      | 6.05                      | [5.16, 6.93] |
first-year follow-up was the continuous dependent variable for all conditioning methods. Because these are complex sample survey data, statistics were weighted with student survey weights and variance estimates were obtained from 200 supplied BRR weights.

Table 3 presents point and interval estimates of the treatment effects obtained from each conditioning method. Relative to the unadjusted treatment effect estimate obtained when ignoring the covariates (10.35), all conditioning methods resulted in substantially reduced estimated magnitudes of the mean difference (with adjusted treatment effect estimates ranging from 4.50 to 6.29). The widths of the 95% confidence intervals were less than 2.0 points for all methods except PS weighting. The larger interval provided by weighting resulted from the use of the products of PS weights and sample survey weights in the analysis.

Table 4 presents summary statistics for balance of the binary and continuous covariates. Although researchers typically report balance for each covariate individually in a table format, to conserve space we present summaries of the distributions of balance statistics (standardized mean differences for continuous covariates and odds ratios for binary covariates). The balance results from the applied study mirror the results of the simulation study with all the conditioning methods improving the balance between the groups for both continuous and binary covariates when compared to ignoring the covariates. As Table Y indicates, the average balance is very close to 0 (continuous covariates) or 1 (binary covariates) for each conditioning method. However the dispersions of balance statistics (for both continuous and binary covariates) are smaller for PS ANCOVA, caliper matching and stratification. Larger dispersions of balance statistics are evident for weighting and non-caliper matching, although the dispersions are notably smaller than those observed with ignoring the covariates.

CONCLUSIONS AND FUTURE RESEARCH

Much of the previous research on PS analysis has stressed the importance of covariate selection, namely the inclusion of all potential confounding variables, to remove the bias associated with the nonrandomized design. Covariate selection is indeed a critical aspect of PS analysis, but much of the previous research assumed the covariates were all measured without error, which is unrealistic in social science research. This study explored the impact of measurement error on the PS estimates and found that even low levels of measurement error had a negative impact on the accuracy and precision of the estimates. These results suggest that the psychometric quality of covariates may be as important as the inclusion of all potentially confounding covariates.

Given the impact of covariate measurement error on the accuracy and precision of treatment effect estimates, this study provides several implications for practice. First, researchers should be cognizant of the importance of psychometric qualities of covariates. In planning research, consideration of covariate reliability needs to be a part of the covariate selection process. Concomitantly, covariate reliability estimates should be reported as a standard element of research results. If reliability estimates are not available (e.g., in many secondary data analyses covariate selection is necessarily opportunistic), this absence should be noted as an important caveat in the interpretation of results. If multiple indicators are available for fallible covariates, the latent variable approach of Raykov (2012) provides a promising vehicle for unbiased estimation. Conversely, if only single indicators are available, the errors-in-variables logistic models (Carroll et al., 2006) may provide a method by which treatment effect estimates can be improved. It is important to note, however, that these errors-in-variables models have not been investigated in the context of propensity score modeling. Finally, these recently developed modeling strategies may provide an approach to sensitivity analysis by comparing treatment effect estimates to those obtained with a standard logistic or probit model.

This current study used simulation methods that maximized the information of the covariate set for the PS estimation models. Future research should begin to look at the impact of both measurement error and misspecification of the PS model on the treatment effect and balance estimates, an intersected phenomenon likely to occur in applied research settings. In addition, errors-in-variables logistic models (e.g., Carroll et al., 2006) and a latent variable modeling approach (Raykov, 2012) may provide improved estimation of propensity scores in the presence of measurement error. Empirical inquiry into such methods is a worthy avenue for further research.

SUPPLEMENTAL MATERIAL

Supplemental data for this article can be accessed on the publisher’s website.

ARTICLE INFORMATION

Conflict of Interest Disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors
reported any financial or other conflicts of interest in relation to the work described.

**Ethical Principles:** The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

**Funding:** This work was supported by Grant 1139510 from the National Science Foundation.

**Role of the Funders/Sponsors:** None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

**Acknowledgments:** The authors would like to thank the anonymous reviewers for their comments on prior versions of this manuscript. The ideas and opinions expressed herein are those of the authors alone, and endorsement by the authors’ institution or the National Science Foundation is not intended and should not be inferred. The order of authorship is arbitrary and does not necessarily indicate authors’ contribution levels.

**REFERENCES**

Austin, P. C. (2009). Some methods of propensity score matching had superior performance to others: Results of an empirical investigation and Monte Carlo simulations. *Biometrical Journal, 51*(1), 171–184. doi:10.1002/bimj.200810488

Austin, P. C., Grootendorst, P., Normand, S. L., & Anderson, G. M. (2007). Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: A Monte Carlo study. *Statistics in Medicine, 26*, 754–768. doi:10.1002/sim.2618

Buonaccorsi, J. P. (2011). *Measurement error*. Boca Raton, FL: Chapman & Hall.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*, 144–152. doi:10.1111/j.2044-8317.1978.tb00581.x

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys, 22*(1), 31–72. doi:10.1111/j.1467-6419.2007.00527.x

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Boston, MA: Houghton Mifflin.

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: A modern perspective* (2nd ed.). Boca Raton, FL: Chapman and Hall.

Cochran, W. G. (1968). Errors of measurement in statistics. *Technometrics, 10*, 637–666. doi:10.2307/1267450

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont, CA: Wadsworth Group.

D’Agostino, R. B. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine, 17*(19), 2265–2281. doi:10.1002/(SICI)1097-0258(19981015)17:19<2265::AID-SIM918>3.0.CO;2-B

Fuller, W. A. (1987). *Measurement Error Models*. New York: Wiley.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*, 945–960. doi:10.2307/2289064

Huynh, C. (2006, April). Estimation and diagnostics for errors-in-variables mixed-effects models. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Jencks, C., Smith, M., Ackland, H., Bane, M. J., Cohen, D., Gintis, H., Heyns, B., & Michelson, S. (1972). *Inequality: A reassessment of the effect of family and schooling in America*. New York, NY: Basic Books.

Michalopoulos, C., Bloom, H. S., & Hill, C. J. (2004). Can propensity score methods match the findings from a random assignment evaluation of mandatory welfare to work programs? *The Review of Econometrics and Statistics, 86*(1), 156–179. http://www.jstor.org/stable/3211665

Pedhazur, E. J. (1997). *Multiple regression in behavioral research* (3rd ed.). Fort Worth, Texas: Harcourt Brace.

Raykov, T. (2012). Propensity score analysis with fallible covariates: A note on a latent variable modeling approach. *Educational and Psychological Measurement, 72*(5), 715–733. doi:10.1177/0013164412440999

Robey, R. R., & Barckiovski, R. S. (1992). Type I error and the number of iterations in Monte Carlo studies of robustness. *British Journal of Mathematical and Statistical Psychology, 45*, 283–288. doi:10.1111/j.2044-8317.1992.tb00993.x

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41–45. doi:10.2307/2335942

Rubin, D. B. (2010). On the limitations of comparative effectiveness research. *Statistics in Medicine, 29*(19), 1991–1995. doi:10.1002/sim.3960

Rubin, D. B. (2001). Using propensity score to help design observational studies: Application to the Tobacco litigation. *Health Services and Outcomes Research Methodology, 2*, 169–188. doi:10.1023/A:1020363010465

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, R. J. (2007). Estimating causal effects using experimental and observational designs (report from the Governing Board of the American Educational Research Association Grants Program). Washington, DC: American Educational Research Association.

Shadish, W. R., & Steiner, P. M. (2010). A primer on propensity score analysis. *Newborn and Infant Nursing Reviews, 10*(1), 19–26. doi:10.1053/j.nain.2009.12.010

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can non-randomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association, 103*(484), 1334–1344. doi:10.1198/01621450800000733

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Belmont, CA: Wadsworth, Cengage Learning.

Steiner, P. M., Cook, T. D., & Shadish, W. R. (2011). On the importance of reliable covariate measurement in selection bias adjustments using propensity scores. *Journal of Educational and Behavioral Statistics, 36*(2), 213–236. doi:10.3102/1076998610375835

Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods, 15*(3), 250–267. doi:10.1037/a0018719

Stuart, E. A. (2010). Matching methods for causal inference: A review and look forward. *Statistical Science, 25*(1), 1–21. doi:10.1214/09-STS313

Thoemmes, F., & Kim, E. S. (2011). A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research, 46*(1), 90–118. doi:10.1080/00273171.2011.540475

West, S. G., & Thoemmes, F. (2010). Campbell’s and Rubin’s perspectives on causal inference. *Psychological Methods, 15*(1), 18–37. doi:10.1037/a0015917