Covariate Measurement Error in Propensity Score Analysis: An Empirical Investigation of Impacts on Treatment Effect Estimates

Jeffrey D. Kromrey[1], Patricia Rodriguez de Gil[1], Eun Sook Kim[1], Aarti Bellara[1], Rheta Lanehart[1], Tyler Hicks[1], Reginald Lee[1]
[1]University of South Florida, 4202 E. Fowler Avenue, Tampa, FL 33620

Abstract

Monte Carlo methods were used to investigate the impact of covariate measurement error on the efficacy of propensity score (PS) methods. Seven factors were crossed in the simulation design: number of covariates (3, 9, 15, 30), population treatment effect (0, .2, .5, .8), covariate relationship to treatment (.1, .2, .4), covariate relationship to outcome (.1, .2, .4), correlation among covariates (0, .2, .5), sample size (50, 100, 250, 500, 1000), and covariate reliability (.4, .6, .8, 1.0). Each sample (5000 replications) was analyzed using seven PS methods (matching with and without a caliper, ignoring covariates, ANCOVA, PS as a covariate, stratification, and PS weighting). Outcome measures included treatment effect bias, RMSE, 95% CI coverage and width. Results indicate that even low levels of measurement error lead to substantial statistical bias in treatment effect estimates and reduction in CI coverage. Such effects were evident across conditioning methods and effects increased with greater amounts of measurement error, larger numbers of covariates, and greater strength of relationship between the covariates and both the treatment assignment and the outcome variable

Key Words: observational studies, simulation, bias

1. Measurement Error in Propensity Score Analysis

Measurement error is a common phenomenon in many fields of study. Buonaccorsi (2010) stated, "in some sense, all statistical problems involve measurement error" (p. 1). Both continuous and categorical variables are often mismeasured, and whether random or systematic, measurement error can seriously impact the estimation of treatment effects and mislead their interpretation (Fuller, 1987). Although the study of measurement error has grown in recent years (Buonaccorsi, 2010; Fuller, 1987;), the examination of the impact of measurement error on propensity score analysis has been scarce.

Steiner, Cook, and Shadish (2011) conducted a simulation study in which different degrees of unreliability were induced in the covariates used to create the propensity score. Baseline data were obtained from previous studies that had demonstrated the quality of covariates for reducing bias when there were 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, Cook, Shadish, & Clark, 2010, respectively). By assuming initial reliability of covariates to be 1.00 and manipulating the amount of error variance added to the observed covariates, the researchers decreased the reliability of

each covariate (ρ_{xx} = .9, .8, .7, .6, .5) and calculated the impact of such measurement error on the bias reduction available through PS analysis.

2. The Simulation Study

The focus of the present simulation study was to determine the impact of covariate measurement error on propensity score analysis. A factorial design with eight completely crossed factors included the 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, and .5), sample size (50, 100, 250, 500, and 1000), covariate reliability (.4, .6, .8, and 1.0), and 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). 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 covariables). For each condition, 5000 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 ± .014 (Robey & Barcikowski, 1992). This complete crossed factorial design provided a total of 60,480 conditions. All simulations were conducted using SAS/IML.

3. Results

The results of this simulation study were analyzed by computing η^2 to estimate the proportion of variability in each of the outcomes (i.e., statistical bias, RMSE, CI coverage and width, and Type I error rate). Patterns of the mean values of the outcomes associated with the factors were further analyzed.

3.1 Common Support

The overall distributions of common support for sample size, reliability of covariate measurement, number of covariates, population regression weights for the propensity, population regression weights for the dependent variable, and correlation among covariates are described in Table 1. The distributions of common support for the trimmed samples are described in Table 2.

The range of common support between treatment and control groups increased as the sample size increased from 50 to 250. There was a decrease in support range for N= 500 and a slight increase as the sample size reached 1000. The non-support 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/non-support ranges occurred when N = 1000. As the reliability of the covariates increased, the range of the common support increased. However, the non-support range also increased as the reliability increased. These data suggest that an increase in reliability does not demonstrably impact the range of common support between propensity score groups.

Table 1: Distribution of Common Support Statistics

Parameter	*Non-Support Cases*	*Non-Support Range*	*Support Range*
N			
50	0.44	0.42	0.41
100	0.21	0.20	0.58
250	0.08	0.10	0.61
500	0.05	0.08	0.60
1000	0.03	0.07	0.61
ρ_{xx}			
.40	0.15	0.17	0.54
.60	0.16	0.17	0.56
.80	0.17	0.18	0.57
1.0	0.18	0.18	0.58
N of Covariates			
3	0.04	0.08	0.36
9	0.08	0.13	0.59
15	0.15	0.16	0.68
30	0.38	0.33	0.62
β_{xc}			
0.025	0.12	0.15	0.48
0.050	0.15	0.17	0.56
0.100	0.22	0.20	0.64
β_{yc}			
0.025	0.16	0.18	0.56
0.050	0.16	0.18	0.56
0.100	0.16	0.18	0.56
ρ_{12}			
.00	0.13	0.16	0.51
.20	0.16	0.18	0.57
.50	0.20	0.19	0.60

Note. ρ_{xx} = reliability, β_{xc} = strength of relationship between covariates and treatment assignment, = strength of relationship between covariates and the outcome, ρ_{12} = strength of relationship between covariates.

Table 2: Distribution of Trimmed Propensity Scores

Variable	*Mean*	*Standard Deviation*	*Minimum*	*Maximum*
Non-Support Cases	0.16	0.25	0.004	1.00
Non-Support Range	0.18	0.21	0.02	1.00
Support Range	0.56	0.23	0.00	0.97

The increase of the number of covariates included in the simulation model resulted in an increase in both support and non-support ranges up to 15 covariates, with support range increasing in greater increments. However, after 15 covariates, the support range decreased and the non-support range more than doubled. As the relationship between covariates and treatment assignment increased, the support range increased in a linear fashion. The non-support range also increased, but in smaller increments. The optimum combination of the support/non-support ranges occurred when the weight parameter was

0.1. There was no change in either the support range or non-support range as the relationship between the covariates and the dependent variable increased, 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 non-support range as the covariate correlation increased.

3.2 Balance

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). Regarding PS conditioning methods, stratification showed the most serious *un*balance (on average 0.59 and 0.44 for binary and continuous covariates, respectively, as presented in Table 3) followed by ignoring the covariates and matching without using a caliper. On the other hand, matching with caliper and weighting consistently showed excellent balance irrespective of design factors. PS ANCOVA achieved near zero balance when small sample size conditions were excluded.

Table 3. *Mean Balance of Covariates After Conditioning*

Conditioning method	*Binary (SD)*	*Continuous (SD)*
Caliper Match	-0.01 (1.11)	0.00 (0.02)
No Caliper Match	0.27 (0.92)	0.23 (0.22)
Ignore	0.26 (0.69)	0.24 (0.41)
PS ANCOVA	-0.00 (0.01)	-0.00 (0.00)
Weighting	-0.03 (0.67)	0.01 (0.47)
Stratification	0.59 (0.28)	0.44 (1.89)

Note. For PS ANCOVA, two sample size conditions ($N = 50$ and $N = 100$) were excluded due to extreme outlying values.

The number of covariates and sample size emerged as important factors associated with variability in balance (see Table 4) although the effects in general appear not substantial (that is, $\eta^2 < .06$). Across conditioning methods, the number of covariates was negatively associated with balance: as the number of covariates increases, the covariates are less balanced overall. Saliently, PS ANCOVA was greatly impacted by the number of covariates, but the correlation 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 outcome model resulted in extreme unbalance in the covariates. When 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 was improved in larger sample sizes.

3.3 Statistical Bias

Bias in this study was computed as the difference between the estimated treatment effect and the corresponding population parameter established in the simulation. The overall distributions of estimated bias by conditioning methods are presented in Figure 1. It should be noted that the bias estimates over 20 observed in PS ANCOVA were excluded.

Table 4. Mean Balance by Covariate Type, Sample Size, and Number of Covariates (k)

			Conditioning method					
Type of covariate	*N*	*k*	*Caliper Match*	*No Caliper Match*	*Ignore*	*PS ANCOVA*	*Weight*	*Stratification*
Binary	50	3	0.00	0.09	0.11	-0.02	-0.02	0.86
		9	0.00	0.22	0.22	-0.04	-0.06	1.03
		15	-0.08	0.18	0.14	-4.37E+06	-0.13	1.04
		30	0.10	0.47	0.16	1.68E+16	0.17	0.96
	100	3	0.00	0.10	0.12	-0.01	-0.01	0.71
		9	0.01	0.26	0.26	0.00	-0.04	0.82
		15	0.01	0.31	0.32	-0.01	-0.05	0.86
		30	-0.34	0.09	0.08	-7.10E+07	-0.23	0.91
	250	3	0.00	0.11	0.12	0.00	0.00	0.47
		9	0.01	0.29	0.29	0.00	-0.02	0.50
		15	0.01	0.36	0.36	0.00	-0.03	0.53
		30	0.02	0.43	0.43	-0.01	-0.05	0.61
	500	3	0.00	0.11	0.13	0.00	0.00	0.34
		9	0.01	0.30	0.30	0.00	-0.01	0.35
		15	0.01	0.38	0.38	0.00	-0.02	0.37
		30	0.02	0.47	0.47	0.00	-0.04	0.43
	1000	3	0.00	0.12	0.13	0.00	0.00	0.24
		9	0.01	0.31	0.31	0.00	-0.01	0.24
		15	0.01	0.39	0.39	0.00	-0.02	0.26
		30	0.02	0.49	0.49	0.00	-0.03	0.31
Continuous	50	2	0.00	0.12	0.14	0.00	-0.03	0.60
		6	0.00	0.16	0.17	-0.01	-0.03	0.72
		10	0.00	0.19	0.19	-0.77	0.01	1.50
		20	-0.01	0.23	0.18	-913.06	0.42	1.90
	100	2	0.00	0.13	0.16	0.00	-0.02	0.37
		6	0.00	0.19	0.20	0.00	-0.03	0.40
		10	0.01	0.24	0.25	0.00	-0.03	0.44
		20	0.00	0.25	0.26	-0.99	0.05	0.80
	250	2	0.00	0.15	0.18	0.00	-0.01	0.22
		6	0.00	0.22	0.23	0.00	-0.02	0.23
		10	0.01	0.28	0.28	0.00	-0.02	0.25
		20	0.01	0.32	0.33	-0.01	-0.02	0.31
	500	2	0.00	0.16	0.18	0.00	-0.01	0.15
		6	0.01	0.24	0.24	0.00	-0.01	0.16
		10	0.01	0.30	0.30	0.00	-0.02	0.17
		20	0.01	0.35	0.35	0.00	-0.02	0.21
	1000	2	0.00	0.17	0.19	0.00	0.00	0.11
		6	0.01	0.24	0.24	0.00	-0.01	0.11
		10	0.01	0.31	0.31	0.00	-0.01	0.12
		20	0.01	0.37	0.37	0.00	-0.01	0.15

The mean bias estimates by conditioning method, number of covariates, and reliability are also reported in Table 5. As previously reported for balance results, PS ANCOVA yielded exceptionally large bias when sample size was small and the number of covariates was large. When noted, these extreme conditions were excluded in the results

of PS ANCOVA. Overall, the estimates of no caliper matching and ignoring the covariates were considerably biased across all simulation conditions (0.08 ~ 1.47). We also observed substantial bias in other conditioning methods except when the number of covariates was small and their reliability was 1.0.

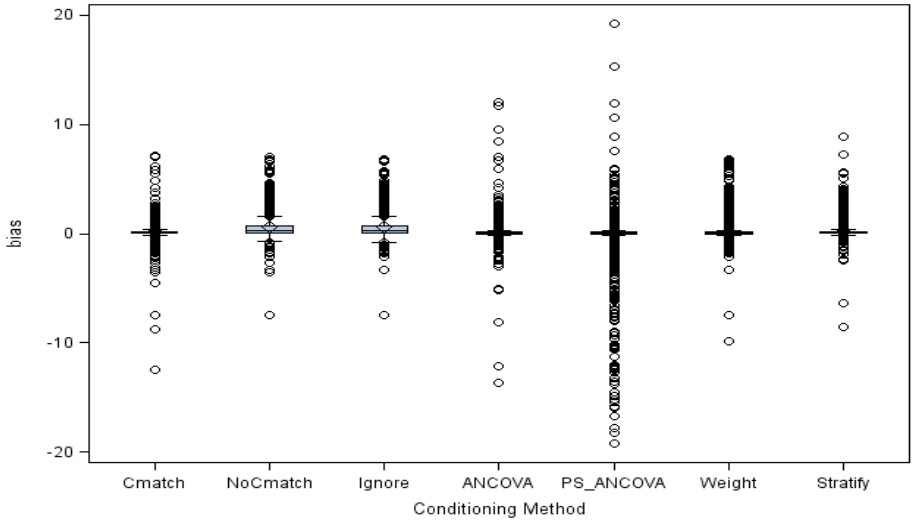

Figure 1: Distributions of estimated bias across all simulation conditions.

Table 5: Mean Bias of Treatment Effect by Number of Covariates and Reliability

		Conditioning method						
k	ρ_{xx}	*Caliper Match*	*No Caliper Match*	*Ignore*	*ANCOVA*	*PS ANCOVA*	*Weighting*	*Stratification*
3	.40	0.05	0.08	0.09	0.05	0.05	0.05	0.05
	.60	0.03	0.08	0.08	0.03	0.03	0.03	0.04
	.80	0.02	0.07	0.08	0.02	0.02	0.01	0.02
	1.0	0.00	0.07	0.08	0.00	0.00	-0.01	0.01
9	.40	0.15	0.30	0.30	0.15	0.14	0.13	0.16
	.60	0.09	0.29	0.30	0.09	0.09	0.07	0.10
	.80	0.05	0.28	0.29	0.04	0.04	0.02	0.06
	1.0	0.01	0.27	0.28	0.00	0.00	-0.02	0.02
15	.40	0.28	0.63	0.64	0.27	0.27	0.24	0.29
	.60	0.17	0.61	0.61	0.14	0.15	0.13	0.19
	.80	0.08	0.59	0.60	0.08	0.07	0.04	0.11
	1.0	0.02	0.57	0.58	0.00	0.00	-0.03	0.04
30	.40	0.53	1.45	1.47	0.51	0.53	0.65	0.64
	.60	0.32	1.38	1.39	0.26	0.29	0.43	0.44
	.80	0.16	1.33	1.34	0.13	0.13	0.29	0.30
	1.0	0.05	1.30	1.31	0.05	-0.01	0.17	0.15

Note. $N = 50$ and $N = 100$ were removed for PS ANCOVA.

3.4 Root Mean Square Error

The typical difference between a sample estimate of the treatment effect and the true treatment effect was estimated by the root mean squared error (RMSE). The mean RMSE estimates under the study factors over the PS Methods are presented in Table 6. As

evident in the table, extremely large mean RMSE values were observed for PS ANCOVA and the smallest sample sizes (N = 50 and N=100). Overall, there was a large RMSE estimate across all methods but the estimate tended to decrease with increases in sample size. There were larger RMSE estimates in those simulation conditions in which measurement error was not induced (ρ_{xx} = 1.0) but overall, the estimate increased as the reliability decreased. The smallest RMSE estimates were observed for the smallest number of covariates (3 and 9), across all PS methods.

Table 6: *Mean RMSE by Study Factors and PS Conditioning Methods*

	Conditioning Methods						
Design Factor	*Caliper Match*	*No Caliper Match*	*Ignore*	*ANCOVA*	*PS ANCOVA*	*Weight*	*Stratification*
N							
50	1.80	1.96	1.74	207.21	891347651.2	1.80	1.56
100	0.57	1.35	1.29	7.62	7674.69	1.00	0.49
250	0.14	1.28	1.28	0.18	0.18	0.33	0.14
500	0.10	1.37	1.37	0.08	0.08	0.20	0.11
1000	0.08	1.45	1.45	0.06	0.07	0.14	0.09
Trim							
No	0.51	1.92	1.91	0.21	0.09	0.78	0.48
Yes	0.55	1.04	0.94	83.84	0.13	0.59	0.46
ρ_{xx}							
.40	0.66	1.54	1.49	5.10	0.19	0.69	0.57
.60	0.55	1.49	1.43	95.76	0.08	0.66	0.48
.80	0.48	1.47	1.41	60.39	0.04	0.68	0.45
1.0	0.45	1.42	1.37	5.50	0.13	0.72	0.40
k							
3	0.01	0.02	0.02	0.01	0.00	0.01	0.01
9	0.06	0.20	0.20	0.05	0.02	0.05	0.06
15	0.20	0.83	0.82	153.29	0.06	0.25	0.23
30	1.89	4.91	4.71	12.41	0.36	2.46	1.61
β_{xc}							
0.025	0.42	0.83	0.78	2.09	0.03	0.38	0.35
0.050	0.56	1.53	1.47	42.44	0.07	0.67	0.49
0.100	0.63	2.08	2.03	81.21	0.23	1.02	0.59
β_{yc}							
0.025	0.09	0.22	0.21	1.12	0.01	0.11	0.08
0.050	0.32	0.86	0.82	78.43	0.12	0.40	0.27
0.100	1.19	3.36	3.24	45.49	0.20	1.55	1.07
ρ_{12}							
0	0.12	0.18	0.17	5.39	0.03	0.10	0.11
0.2	0.52	1.20	1.15	38.08	0.09	0.55	0.43
0.5	0.97	3.06	2.97	82.06	0.21	1.42	0.89

Note. Estimates are based on 5,000 samples of each condition. ρ_{xx} = reliability, β_{xc} = strength of relationship between covariates and treatment assignment, = strength of relationship between covariates and the outcome, ρ_{12} = strength of relationship between covariates.

3.5 Type I Error Control

The overall distributions of Type I error rate estimates are presented in Figure 2. 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 provided median Type I error rates near the nominal .05 level. However, all of the conditioning methods demonstrated a large dispersion in Type I error rates across the conditions simulated.

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 except matching without a caliper and ignoring the covariates. The marginal mean Type I error rates by these design factors are presented in Table 7. 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 $N = 1000$, the estimated mean Type I error rates ranged from .66 with $\rho_{xx} = .40$ to .04 with $\rho_{xx} = 1.00$. Analogous results were evident for the other conditioning methods.

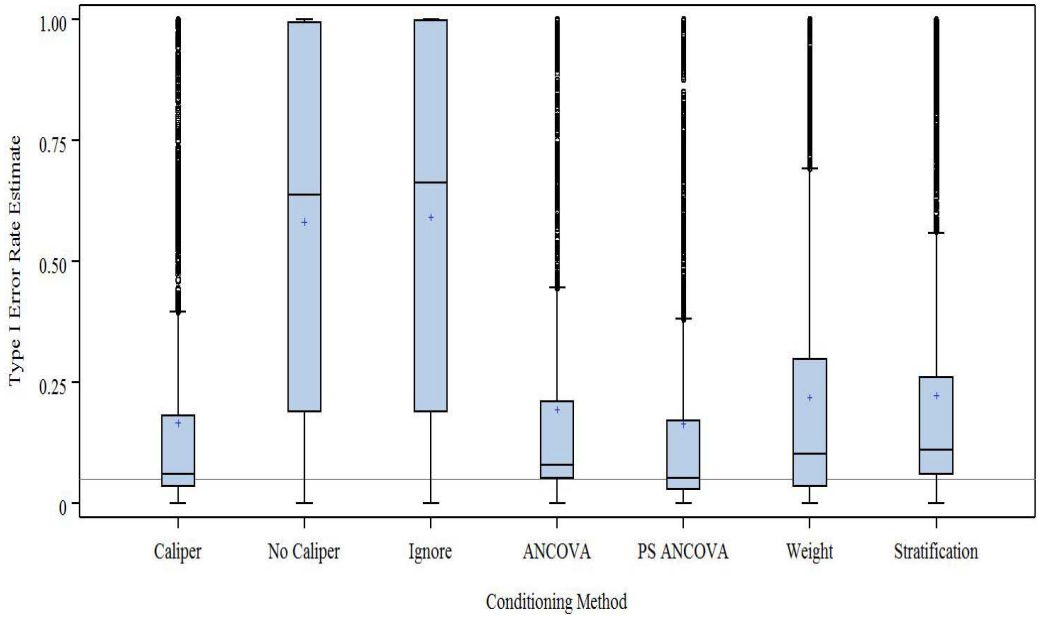

Figure 2: Distributions of estimated Type I error rates across all conditions.

The first-order interaction between covariate reliability and the strength of relationship between the covariates and treatment assignment was also associated with variability in Type I error control. The mean Type I error rate estimates associated with this interaction are presented in Table 8. As with the interaction results presented in Table 7, these data indicated that the impact of measurement error in the covariates was greater when the relationship between the covariates and treatment assignment was 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.

Table 7: *Mean Type I Error Rates by Sample Size and Covariate Reliability (ρ_{xx})*

		Analysis Method						
N	ρ_{xx}	*Caliper Match*	*No Caliper Match*	*Ignore*	*ANCOVA*	*PS ANCOVA*	*Weight*	*Strati-fication*
50	.40	.09	.22	.25	.11	.08	.21	.16
	.60	.08	.22	.24	.09	.06	.20	.14
	.80	.07	.21	.23	.08	.05	.19	.12
	1.00	.07	.21	.22	.07	.05	.19	.12
100	.40	.11	.37	.41	.17	.12	.22	.17
	.60	.07	.35	.38	.10	.07	.17	.11
	.80	.05	.34	.35	.06	.04	.15	.07
	1.00	.04	.33	.34	.05	.04	.15	.05
250	.40	.28	.66	.68	.36	.31	.31	.36
	.60	.14	.65	.67	.19	.15	.18	.20
	.80	.06	.64	.65	.09	.05	.12	.09
	1.00	.03	.63	.64	.05	.03	.12	.04
500	.40	.47	.80	.81	.53	.50	.45	.56
	.60	.26	.80	.80	.31	.28	.25	.36
	.80	.11	.80	.80	.13	.10	.13	.17
	1.00	.04	.80	.80	.05	.03	.11	.05
1000	.40	.66	.89	.89	.70	.68	.62	.73
	.60	.44	.90	.90	.48	.45	.37	.56
	.80	.19	.90	.89	.20	.18	.17	.31
	1.00	.04	.90	.89	.05	.04	.11	.08

Note. Estimates are based on 5,000 samples of each condition.

Table 8: *Mean Type I Error Rates by Strength of Relationship between Treatment Assignment and Covariates (β_{xc}) and Covariate Reliability (ρ_{xx})*

		Analysis Method						
β_{xc}	ρ_{xx}	*Caliper Match*	*No Caliper Match*	*Ignore*	*ANCOVA*	*PS ANCOVA*	*Weight*	*Strati-fication*
0.025	.40	.12	.40	.41	.17	.13	.14	.18
	.60	.07	.40	.40	.10	.06	.08	.11
	.80	.04	.39	.40	.06	.03	.06	.07
	1.00	.03	.40	.40	.05	.02	.05	.05
0.050	.40	.31	.60	.62	.37	.33	.34	.39
	.60	.17	.60	.62	.20	.17	.20	.25
	.80	.08	.60	.61	.10	.07	.13	.13
	1.00	.04	.59	.60	.05	.03	.11	.07
0.100	.40	.54	.77	.79	.59	.57	.60	.62
	.60	.37	.76	.78	.40	.37	.43	.46
	.80	.17	.75	.76	.18	.16	.27	.26
	1.00	.06	.73	.74	.06	.05	.24	.09

Note. Estimates are based on 5,000 samples of each condition.

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. This criterion indicates that Type I error control is considered acceptable if the actual Type I error rate is within the bounds of $\alpha_{nominal} \pm 0.5\ \alpha_{nominal}$. At the nominal α = .05, these bounds are .025 and .075. The proportions of conditions that provided acceptable Type I error control according to Bradley's criterion are presented in Table 9. 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 strikingly observed in Table 9. With caliper matching, adequate Type I error control was maintained in 74% of the conditions when covariates were measured without error, but in only 57% of the conditions when ρ_{xx} = .80, and in only 24% of the conditions when ρ_{xx} = .40. Similar degradation of Type I error control was evident with stratification and the ANCOVA approaches. Regardless of the conditioning method used, larger samples resulted in smaller proportions of conditions with adequate Type I error control, as did larger numbers of covariates. Increasing the strength of relationship between the covariates and treatment assignment (β_{xc}) resulted in smaller proportions of conditions with adequate control, but the strength of relationship between the covariates and the outcome variable (β_{yc}) was not related to the proportions of conditions. An interesting result of the Bradley criterion analysis was that trimming the samples for common support had little effect on the proportions of conditions with adequate Type I error control and 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.

3.6 Confidence Interval Coverage

The distributions of confidence interval coverage estimates across all conditions simulated are presented in Figure 3. Evident in this figure were the overall poor interval coverage for matching without a caliper and ignoring the covariables. Use of any conditioning method besides matching without a caliper resulted in substantial improvement in the coverage of confidence intervals, but the overall coverage was best with caliper matching, PS ANCOVA, and ANCOVA with the original covariates. However, the distributions presented in Figure 3 showed notable dispersion in coverage estimates regardless of the conditioning method employed.

Three research design factors were substantially associated with variability in interval coverage: sample size, covariate reliability, and strength of relationship between the covariates and treatment assignment. However, the interaction between reliability and sample size and the interaction between reliability and strength of relationship were both considerable.

The mean interval coverage estimates by conditioning method, sample size, and covariate reliability are presented in Table 10. These data indicated that the impact of covariate measurement error was 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 = 1000, the mean coverage estimate dropped from .95 to .40. Note that even with covariate reliability of .80, the interval coverage was substantially reduced for all methods of conditioning

(with the exceptions of matching with no caliper and ignoring the covariates – methods that evidenced very poor coverage regardless of the reliability of the covariates).

Table 9: *Proportion of Conditions with Adequate Type I Error Control (Bradley Criterion) by Research Design Factors*

	Analysis Method						
Design Factor	*Caliper Match*	*No Caliper Match*	*Ignore*	*ANCOVA*	*PS ANCOVA*	*Weight*	*Strati-fication*
N							
50	.64	.33	.30	.67	.42	.31	.20
100	.60	.11	.11	.60	.43	.33	.53
250	.48	.00	.00	.46	.43	.24	.41
500	.40	.00	.00	.36	.39	.22	.28
1000	.31	.00	.00	.29	.33	.15	.15
Trim							
No	.49	.03	.03	.50	.39	.21	.33
Yes	.48	.14	.14	.45	.41	.29	.29
ρ_{xx}							
.40	.24	.07	.06	.12	.24	.17	.07
.60	.42	.09	.08	.26	.38	.33	.19
.80	.57	.09	.09	.55	.46	.31	.37
1.00	.71	.10	.10	.96	.52	.20	.63
k							
3	.65	.15	.17	.58	.53	.40	.38
9	.49	.09	.08	.49	.36	.34	.32
15	.48	.07	.07	.44	.36	.19	.28
30	.33	.03	.01	.38	.34	.08	.26
β_{xc}							
0.025	.61	.17	.16	.67	.43	.36	.52
0.050	.51	.07	.07	.45	.42	.27	.29
0.100	.34	.01	.02	.31	.34	.13	.13
β_{yc}							
0.025	.49	.09	.08	.47	.40	.25	.31
0.050	.48	.08	.08	.47	.41	.25	.31
0.100	.49	.09	.08	.48	.40	.25	.32
ρ_{12}							
.00	.46	.12	.10	.44	.37	.26	.31
.20	.49	.08	.08	.47	.40	.26	.31
.50	.51	.06	.07	.52	.43	.23	.32

Note. Estimates are based on 5,000 samples of each condition. k = number of covariates, ρ_{xx} = reliability, β_{xc} = strength of relationship between covariates and treatment assignment, β_{yc} = strength of relationship between covariates and the outcome, ρ_{12} = strength of relationship between covariates.

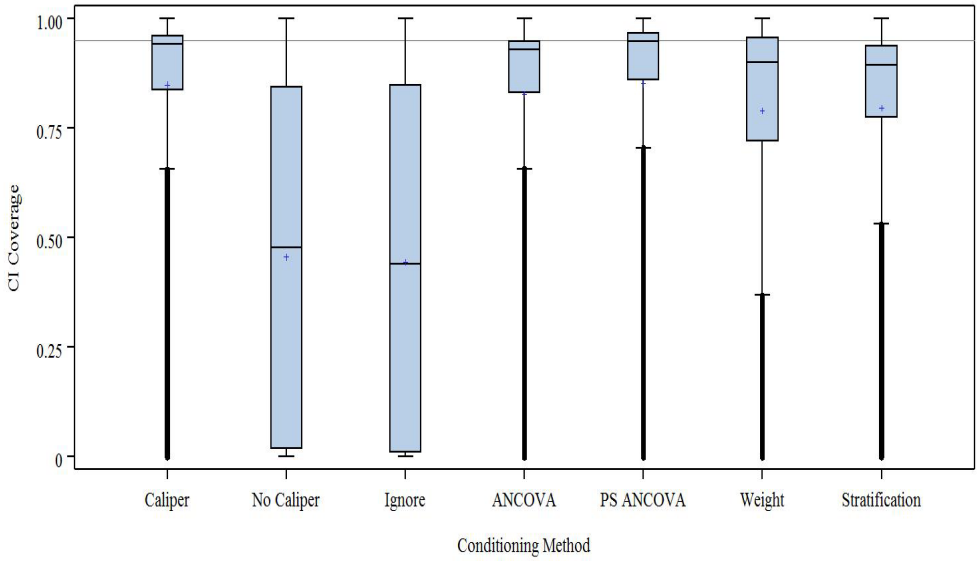

Figure 3: Distributions of estimated confidence interval coverage across all conditions.

Table 10: *Mean Confidence Interval Coverage Estimates by Sample Size and Covariate Reliability(ρ_{xx})*

		Analysis Method						
N	ρ_{xx}	*Caliper Match*	*No Caliper Match*	*Ignore*	*ANCOVA*	*PS ANCOVA*	*Weight*	*Strati-fication*
50	.40	.90	.79	.78	.89	.93	.79	.84
	.60	.91	.79	.78	.92	.94	.80	.86
	.80	.92	.80	.79	.92	.95	.81	.87
	1.00	.92	.80	.80	.93	.95	.82	.88
100	.40	.90	.66	.63	.86	.89	.79	.84
	.60	.94	.68	.66	.91	.94	.83	.90
	.80	.95	.70	.68	.94	.96	.85	.93
	1.00	.96	.71	.69	.95	.96	.85	.94
250	.40	.76	.39	.37	.70	.73	.72	.69
	.60	.88	.40	.38	.84	.87	.83	.82
	.80	.94	.42	.40	.92	.95	.88	.91
	1.00	.96	.43	.42	.95	.97	.88	.96
500	.40	.59	.25	.24	.53	.56	.60	.50
	.60	.78	.25	.24	.74	.77	.77	.69
	.80	.91	.25	.24	.89	.91	.87	.85
	1.00	.96	.25	.25	.95	.97	.89	.95
1000	.40	.40	.15	.14	.36	.38	.43	.33
	.60	.62	.14	.14	.59	.61	.66	.51
	.80	.84	.14	.15	.83	.85	.84	.74
	1.00	.95	.14	.15	.95	.96	.89	.92

Note. Estimates are based on 5,000 samples of each condition.

The mean coverage estimates by covariate reliability and the strength of relationship between the covariates and treatment assignment are presented in Table 11. As expected, these data indicated that the impact of covariate measurement error on confidence interval coverage was greater when the covariates were more strongly related to treatment assignment. With caliper matching and $\beta_{xc} = 0.025$, the coverage estimates dropped from .96 to .88 as covariate reliability drops from 1.00 to .40. However, with a stronger relationship between the covariates and the treatment assignment ($\beta_{xc} = 0.100$), the coverage for caliper matching dropped 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 presented sufficient measurement error to notably reduce the average coverage probability to .86 with caliper matching.

Table 11: *Mean Confidence Interval Coverage Estimates by Strength of Relationship between Covariates (β_{xc}) and Treatment Assignment and Covariate Reliability (ρ_{xx})*

		Analysis Method						
β_{xc}	ρ_{xx}	*Caliper Match*	*No Caliper Match*	*Ignore*	*ANCOVA*	*PS ANCOVA*	*Weight*	*Strati-fication*
0.025	.40	.88	.63	.62	.85	.88	.87	.83
	.60	.93	.63	.62	.91	.94	.92	.89
	.80	.95	.63	.63	.94	.97	.94	.93
	1.00	.96	.63	.63	.94	.97	.95	.95
0.050	.40	.72	.44	.42	.68	.71	.69	.65
	.60	.85	.44	.43	.82	.85	.81	.78
	.80	.92	.44	.44	.91	.94	.87	.88
	1.00	.95	.45	.45	.95	.96	.88	.93
0.100	.40	.52	.28	.26	.47	.50	.45	.44
	.60	.69	.29	.27	.66	.68	.61	.60
	.80	.86	.30	.29	.86	.87	.74	.78
	1.00	.94	.32	.30	.95	.95	.76	.91

Note. Estimates are based on 5,000 samples of each condition.

3.7 Confidence Interval Width

The distributions of confidence interval widths across all conditions in the simulation suggested that the differences in the average widths of intervals across PS methods were relatively small and that substantial variability in widths was 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.

The mean interval widths by conditioning method, sample size, and number of covariates are presented in Table 12. As expected, larger samples resulted in narrower confidence intervals regardless of the conditioning method or number of covariates. Similarly, increasing the number of covariates resulted in wider intervals across the methods and the sample sizes. However, the impact of increasing the number of covariates was 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 $N = 1000$, the mean interval widths increased from 0.12 to only 0.71 across the same range of covariates.

Table 12: Mean Confidence Interval Width by Sample Size and Number of Covariates

		Analysis Method						
N	*k*	*Caliper Match*	*No Caliper Match*	*Ignore*	*ANCOVA*	*PS ANCOVA*	*Weight*	*Strati-fication*
50	3	0.73	0.64	0.61	0.54	0.60	0.60	0.56
	9	1.54	1.25	1.13	1.05	1.22	1.10	1.14
	15	3.29	2.43	2.00	2.24	6.12	1.86	2.20
	30	11.16	6.35	5.90	5.79	18407.81	5.18	8.06
100	3	0.45	0.42	0.42	0.37	0.40	0.41	0.39
	9	0.88	0.77	0.74	0.66	0.75	0.74	0.74
	15	1.48	1.22	1.14	1.08	1.25	1.13	1.21
	30	5.06	3.44	2.96	2.95	9.65	2.73	3.35
250	3	0.26	0.25	0.26	0.23	0.24	0.26	0.24
	9	0.49	0.46	0.46	0.40	0.44	0.46	0.44
	15	0.79	0.70	0.69	0.63	0.70	0.70	0.69
	30	1.66	1.33	1.24	1.27	1.46	1.23	1.42
500	3	0.18	0.17	0.18	0.16	0.17	0.18	0.17
	9	0.33	0.32	0.32	0.28	0.30	0.32	0.30
	15	0.52	0.48	0.48	0.44	0.47	0.49	0.47
	30	1.06	0.89	0.84	0.85	0.94	0.85	0.92
1000	3	0.12	0.12	0.13	0.11	0.12	0.13	0.12
	9	0.22	0.22	0.23	0.20	0.20	0.23	0.21
	15	0.36	0.34	0.34	0.30	0.32	0.35	0.32
	30	0.71	0.61	0.58	0.59	0.63	0.60	0.62

Note. Estimates are based on 5,000 samples of each condition.

4. Conclusions

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, especially educational research, relies on effects estimated from nonrandomized studies. Given the inability to use random assignment in many studies 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 educational research.

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. 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 were smaller) and with conditions in which the relationship between treatment assignment and the covariates was strong (in which selection bias was stronger). Especially important is that even small amounts of measurement error (i.e., ρ_{xx} = .80) resulted in notable decrements in the accuracy of inferences.

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. 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., Carrol et al., 2006) 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.

References

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31,* 144-152.

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

Carrol, 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.

Fuller, W. A. (1987). *Measurement Error Models.* Wiley, & Sons, Inc.

Robey, R. R. & Barcikowski, 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.

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., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random to nonrandom assignment. *Journal of the American Statistical Association, 103,* 1334-1343.

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.

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*, 250-267.

Thoemmes, F., & Kim, E.S., (2011). A Systematic Review of Propensity Score Methods in the Social Sciences', *Multivariate Behavioral Research*, *46* (1), 90 -118.