The Influence of Covariate Measurement Error on Treatment Effect Estimates and Numeric Balance Diagnostics Following Several Common Methods of Propensity Score Matching: A Simulation Study

Heather D. Harris
James Madison University

Follow this and additional works at: https://commons.lib.jmu.edu/diss201019
Part of the Quantitative Psychology Commons

Recommended Citation
https://commons.lib.jmu.edu/diss201019/173

This Dissertation is brought to you for free and open access by the The Graduate School at JMU Scholarly Commons. It has been accepted for inclusion in Dissertations by an authorized administrator of JMU Scholarly Commons. For more information, please contact dc_admin@jmu.edu.
The Influence of Covariate Measurement Error on Treatment Effect Estimates and Numeric Balance Diagnostics Following Several Common Methods of Propensity Score Matching: A Simulation Study

Heather D. Harris

A dissertation submitted to the Graduate Faculty of JAMES MADISON UNIVERSITY

In Partial Fulfillment of the Requirements for the degree of Doctor of Philosophy

Department of Graduate Psychology

May 2018

FACULTY COMMITTEE:
Committee Chair: S. Jeanne Horst
Committee Members/Readers:
Allison J. Ames
Christine E. DeMars
Dena A. Pastor
Dedication

This dissertation, all of my time working on it, and whatever contribution it makes to the literature on propensity score analysis is dedicated to the one and only Casey Patrick Driscoll. Casey, you’re the very best friend and comrade I could ever hope for – thank you for being in this with me for over twelve years now. In the words of Fatboy Slim:

“We’ve come a long, long way together. Through the hard times and the good. I have to celebrate you baby. I have to praise you like I sho-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-u-d.”
Acknowledgements

First and foremost, I would like to thank my advisor, Dr. Jeanne Horst, for all her time and feedback (and patience) as I completed my dissertation. Jeanne, I am incredibly grateful to have had the opportunity to work with you this past five years. We’ve learned so much together – and I have learned so much from you – as we’ve dug ever deeper into learning propensity score analyses. Your thoughtful guidance made my dissertation project possible, and I could not have done it without you. I’m also grateful to my committee members, Dr. Allison Ames, Dr. Christine DeMars, and Dr. Dena Pastor, for their thought-provoking questions, suggested revisions, and for lending me their expertise. Christine, I am incredibly appreciative for all of your help and guidance throughout this process. It’s been a humbling experience working with you – an opportunity for which I’m incredibly grateful.

Next, I would like to thank my tribe: the folks who have supported me through thick and thin. First and foremost, I would like to thank Casey Patrick Driscoll. You are hands-down the very best part of my life. Casey, you have been my biggest source of inspiration, strength, and laughter, not only during grad school and working on my dissertation, but throughout the last twelve years. Thank you for teaching me the value of irreverence and that it’s possible (necessary, even) to find humor in all things. A lot of life has happened in the past decade, and I’m really proud of how far we’ve come. It’s an honor to be in this life with you, and I can’t imagine my life without you. I can’t wait to see what’s next for us.

Pamela Harris and Scott Harris (Mom and Dad), thank you for everything. Thank you for overcoming so many obstacles in life and working so hard to make sure your kids
had opportunities you never had. I see you two – what you’ve done together – and I think you’re amazing people. I’m so incredibly proud to have you as parents. As a kid, I was able to try just about any sport, instrument, language, or hobby. Naturally, I didn’t appreciate it as much as I should have at the time, but in hindsight, I’m incredibly grateful. And your tenacity has thankfully rubbed off on me.

Benjamin Harris and Karla Harris, thank you for cheering me on over the last five years. You’ve both been so incredibly understanding of my busy schedule, yet always ready to catch up whenever I’d have time to call. Ben, I cannot thank you enough for your unconditional love and support. You’re such an intelligent and caring human being, and of all the things I’m proud of in this life, getting to be your sister is easily at the top of the list. Jocelyn Jean Harris, although you’re still too little to understand, I want you to know that getting to meet you for the first time was such a bright light at the end of what was easily my toughest semester. Thank you for being in this world.

Jennifer Haas and Greg Haas, I’m really happy for you both and all of the things you’ve accomplished over the past few years. You both have worked really hard to build up the life you want, and I know that has taken a lot of dedication. I look forward to learning more about your future goals, and I hope there are many opportunities for us to learn from one another.

I am also incredibly grateful for the graduate students and faculty I’ve had the honor of working with over the past five years at JMU. Elisabeth Spratto, thanks for being my comrade throughout this graduate school experience over the past five years. We’ve weathered the trenches together, and during both the high points and low points, I’m really grateful to have had you as a friend. I’m incredibly proud of you. Dr. Monica
Erbacher Smith, thank you for being such an uplifting and encouraging force in my life. I’m so lucky to count you as a friend. Whenever I give a presentation, I try to channel my “inner Monica,” which helps me to focus on my audience and explain why my research results are important. I also feel incredibly lucky to have had time – in courses or otherwise – with Madison Holzman, Thai Ong, and Aaron Myers (and so many other really talented students). I respect the heck out of you all and I expect great things from you. Dr. Deborah Bandalos, thank you for always being so supportive and letting me spend quality time with your pups. Sparky and Zoomie, thank you for being excellent writing companions as I sequestered myself this summer to write my dissertation literature review.

I would also like to thank my friends, colleagues, and coworkers for being tremendous sources of support throughout my graduate school experience. Debbie Campbell, thank you for being such a brilliant light in my life and one of my dearest friends. No matter what comes my way, I always feel better knowing you’re in this world with me. Jerrod Steinmetz, thank you for being a constant friend and always cheering me on. No matter how much time passes, it feels like we can pick up right where we left off. Thank you both for reminding me throughout the past five years that there was a whole world of experiences and opportunities “on the outside.” To my colleagues and coworkers (past and present), including Dr. Alan Socha, Dr. David Paulson, Dr. Ellen Julian, Sarah Pelter, Bill Seifarth, Dr. Yu Meng, Shea Fyffe, Mike Hughes, Joeseph Costa, and Will Williams, thank you all for your encouragement throughout this process. I am so proud to be on such an indominable team of fellow nerds. Alan, I am incredibly
grateful to you in particular for your role in coordinating my summer internship at
Inteleos.

Finally, I would like to thank some of the influential teachers I’ve been lucky
enough to have along the way. Dianne Rasmussen, thank you for your support,
encouragement, and willingness to endorse my high school grant-writing efforts. You
were an incredibly important role model to me during my formative years. Barb Taves,
thank you for noticing that I was bored in class giving me alternative work. I don’t know
if I would stand out in your memory now, but that experience – and how much you cared
about your students – certainly stands out in mine. Last, but certainly not least, I owe a
tremendous thanks to Dr. April Bleske-Rechek. April, thank you for letting me join your
research lab and single-handedly igniting my curiosity and passion for conducting
research. You were a stellar advisor and academic role model, and you helped me
transform from an undergraduate fresh off academic suspension to a competitive graduate
school candidate in a few short years. I would not be where I am today without you.
# Table of Contents

Dedication .......................................................................................... ii

Acknowledgements ........................................................................ iii

Table of Contents ............................................................................ vii

List of Tables .................................................................................. xi

List of Figures ................................................................................ xv

Abstract ......................................................................................... xvii

**Chapter 1: Introduction** ................................................................ 1

  Background .................................................................................. 3

  Propensity Score Analyses ......................................................... 3

  Current Study ............................................................................. 6

**Chapter 2: Review of the Literature** ............................................ 10

  Internal Validity .......................................................................... 10

  Rubin’s Causal Model ................................................................. 12

  Propensity Scores ...................................................................... 14

  Underlying Assumptions ............................................................ 16

    Strong ignorability assumption ............................................. 16

    Sufficient overlap assumption .............................................. 17

    Stable unit of treatment assumption .................................... 18

  Propensity Score Matching ....................................................... 19

    Selection of covariates ......................................................... 19

    Sensitivity analysis ............................................................. 22

    Propensity score estimation .............................................. 23

    Logistic regression ........................................................... 23
Mahalanobis distance .......................................................... 25
Propensity score matching .................................................. 26
Nearest neighbor matching .................................................. 26
Nearest neighbor matching with caliper ............................... 28
Optimal matching ............................................................ 29
Mahalanobis matching ....................................................... 30
Additional considerations .................................................. 32
Diagnosing the quality of matches ....................................... 34
Numeric diagnosis of matches .............................................. 34
Visual diagnosis of matches ................................................ 37
Representativeness of samples ............................................. 39
Estimating effect of treatment ............................................ 40
Average treatment effect on the treated ............................... 40
Average treatment effect .................................................... 41
Average treatment effect on controls ................................. 42
Benefits and Drawbacks of Propensity Score Matching .......... 43
Benefits of using propensity score analyses ........................ 43
Drawbacks of using propensity score analyses ...................... 45
Measurement Error and Propensity Score Estimation .......... 47
Current study ................................................................. 54
Chapter 3: Method .......................................................... 56
Simulation of Data ............................................................ 56
Validation Data Sets ......................................................... 62
Propensity Score Matching .................................................. 62
Treatment Effect Estimation .............................................. 63
Criteria for Evaluating Research Questions ......................... 64
  Group balance ............................................................... 64
  Bias .............................................................................. 66
  Standard error................................................................. 67
  Root mean squared error ................................................. 67
  Confidence interval coverage .......................................... 68
  Summary ........................................................................ 68
Chapter 4: Results .................................................................. 69
  Validation Data Sets .......................................................... 69
  Simulation Study Results ................................................... 70
  Research Question 1 .......................................................... 73
  Research Question 2 .......................................................... 76
  Research Question 3 .......................................................... 82
  Research Question 4 .......................................................... 84
Chapter 5: Discussion .............................................................. 87
  Limitations and Future Research ....................................... 97
  Implications for Applied Research .................................... 101
  Conclusion ....................................................................... 104
Appendix A: Syntax for Same Measurement Error Conditions ..........151
Appendix B: Syntax for Different Measurement Error Conditions ........172
Appendix C: Calculations of Simulated Data ............................ 194
Appendix D: Validation Data Set Screening ......................................................... 195
Appendix E: SPSS Syntax for Analyses ............................................................ 213
References ............................................................................................................ 215
# List of Tables

Table 1: Correlations Among Covariates Measured Without Error ..........105

Table 2: Simulated Conditions and Measurement Error Levels on X1 and X2 .................................................................106

Table 3: Criteria Evaluated to Answer Each Research Question ............107

Table 4: Simulated Correlations Between Simulated Error-free and Error-prone Covariates....................................................108

Table 5: Descriptive Information for Simulated Validation Data Sets for Same Measurement Error Conditions........................................109

Table 6: Descriptive Information for Simulated Validation Data Sets for Different Measurement Error Conditions.................................110

Table 7: Descriptive Information for Simulated Variables by Condition for the Same Measurement Error Conditions .................................111

Table 8: Descriptive Information for Simulated Variables by Condition for the Different Measurement Error Conditions.............................112

Table 9: Average Matched Sample Size by Matching Method and Measurement Error Condition .................................................................113

Table 10: Descriptive Information for Simulated Variables by Condition for the Same Measurement Error Conditions After Nearest Neighbor Matching Using the True and Naïve Propensity Scores.................................................................114

Table 11: Descriptive Information for Simulated Variables by Condition for the Same Measurement Error Conditions After Nearest
Neighbor Matching with a Caliper Using the True and Naïve Propensity Scores

Table 12: Descriptive Information for Simulated Variables by Condition for the Same Measurement Error Conditions After Optimal Matching Using the True and Naïve Propensity Scores

Table 13: Descriptive Information for Simulated Variables by Condition for the Same Measurement Error Conditions After Mahalanobis Distance Matching Using the True and Naïve Propensity Scores

Table 14: Descriptive Information for Simulated Variables by Condition for the Different Measurement Error Conditions After Nearest Neighbor Matching Using the True and Naïve Propensity Scores

Table 15: Descriptive Information for Simulated Variables by Condition for the Different Measurement Error Conditions After Nearest Neighbor Matching with a Caliper Using the True and Naïve Propensity Scores

Table 16: Descriptive Information for Simulated Variables by Condition for the Different Measurement Error Conditions After Optimal Matching Using the True and Naïve Propensity Scores

Table 17: Descriptive Information for Simulated Variables by Condition for the Different Measurement Error Conditions After
Mahalanobis Distance Matching Using the True and Naïve Propensity Scores ................................................................. 121

Table 18: Numeric Diagnostic Information for Same Measurement Error Conditions After Nearest Neighbor Matching Using the True and Naïve Propensity Scores ......................................................... 122

Table 19: Numeric Diagnostic Information for Same Measurement Error Conditions After Nearest Neighbor Matching with Caliper Using the True and Naïve Propensity Scores ............................................. 123

Table 20: Numeric Diagnostic Information for Same Measurement Error Conditions After Optimal Matching Using the True and Naïve Propensity Scores ................................................................. 124

Table 21: Numeric Diagnostic Information for Same Measurement Error Conditions After Mahalanobis Distance Matching Using the True and Naïve Propensity Scores ................................................................. 125

Table 22: Numeric Diagnostic Information for Different Measurement Error Conditions After Nearest Neighbor Matching Using the True and Naïve Propensity Scores ................................................................. 126

Table 23: Numeric Diagnostic Information for Different Measurement Error Conditions After Nearest Neighbor Matching with Caliper Using the True and Naïve Propensity Scores ................................. 127

Table 24: Numeric Diagnostic Information for Different Measurement Error Conditions After Optimal Matching Using the True and Naïve Propensity Scores ................................................................. 128
Table 25: Numeric Diagnostic Information for Different Measurement Error Conditions After Mahalanobis Distance Matching Using the True and Naïve Propensity Scores .........................................................129

Table 26: Treatment Effect Estimate and Estimated Bias in Treatment Effect for Same Measurement Error Conditions ..........................130

Table 27: Treatment Effect Estimate and Estimated Bias in Treatment Effect for Different Measurement Error Conditions .........................131

Table 28: Results from One Four-Way and Four Three-way Mixed ANOVAs ........................................................................................................132

Table 29: Treatment Effect Estimate 95% Confidence Interval Coverage For the Same Measurement Error Conditions ..........................133

Table 30: Treatment Effect Estimate 95% Confidence Interval Coverage For the Different Measurement Error Conditions .........................134
List of Figures

Figure 1: Example of an absolute standardized mean difference plot produced by the MatchIt package in R .................................................................135

Figure 2: Example of the Q-Q plots produced by the MatchIt package in R ...............................................................................................136

Figure 3: Example of boxplots used to evaluate the distribution of two groups on continuous covariates .........................................................137

Figure 4: Example of density plots used to evaluate the distribution of treatment and comparison groups on matched covariates created via ggplots2 ................................................................................138

Figure 5: Example of a jitter plot produced by the MatchIt package in R .........................................................................................................139

Figure 6: Illustration of the jitter plots when visually diagnosing matches with incrementally larger samples ...........................................140

Figure 7: Example of the histogram plots produced by the MatchIt package in R ..........................................................................................141

Figure 8: Conceptual diagram of latent and observed simulated variables ..............................................................................................142

Figure 9: Plots of the average percent of bias reduction in L1 and X1 for the same measurement error conditions .......................................143

Figure 10: Plots of the average percent of bias reduction in L1 and X1 for the different measurement error conditions ..................................144

Figure 11: Plots of the standardized mean difference for L1 and X1 for the same measurement error conditions ......................................145
Figure 12: Plots of the standardized mean difference for L1 and X1 for the different measurement error conditions........................146

Figure 13: Plots of the estimated treatment effect and bias in the estimated treatment effect for the same measurement error conditions.................................................................147

Figure 14: Plots of the estimated treatment effect and bias in the estimated treatment effect for the different measurement error conditions.................................................................148

Figure 15: Plots of the percent of times the estimated treatment effect confidence interval excluded the true treatment effect for the same measurement error conditions.................................................................149

Figure 16: Plots of the percent of times the estimated treatment effect confidence interval excluded the true treatment effect for the same measurement error conditions.................................................................150
Abstract

In applied intervention studies, researchers frequently aim to make inferences about the impact of a treatment program on participants. However, applied researchers are often faced with threats to the internal validity of their studies, or the extent to which changes in participants’ outcomes can be attributed to the intervention. When researchers are unable to randomly assign study participants to treatment conditions, changes in the intervention outcome might be confounded with systematic differences in participants’ baseline characteristics. Propensity score matching is one technique that allows researchers to account for threats to the internal validity of a study. Specifically, using propensity score matching methods, researchers construct a qualitatively-similar comparison group based on participants’ characteristics at baseline (i.e., covariates).

In addition to threats to the internal validity of a study, measurement error is a reality with which many applied researchers must contend. However, research on the impact of covariate score measurement error on the quality of matches and the accuracy of treatment effect estimates is sparse in the propensity score matching literature. Consequently, the purpose of the current study was to evaluate how different levels and types of measurement error impacted the quality of propensity score matched groups and the accuracy of treatment effect estimates.

A simulation study was conducted to manipulate both the levels of measurement error (e.g., 10% versus 60% unreliability) and the types of measurement error (e.g., treatment and comparison group scores measured with the same level of reliability versus different levels of reliability). Four common propensity score matching methods were then used to create comparison groups, including nearest neighbor matching, nearest neighbor matching with a 0.2 caliper, optimal matching, and Mahalanobis distance.
matching. Numeric diagnostic information and the accuracy of treatment effect estimates were then evaluated. When unreliable covariates were included in the model, the final propensity score matched groups appeared balanced on the unreliable covariates. However, propensity score matching was not able to appropriately account for the full influence of the covariates on treatment effect estimates. That is, as the level of measurement error increased, the estimated treatment effect also increased, resulting in a higher estimated treatment effect than the simulated treatment effect.
CHAPTER 1

Introduction

Although randomized control trials are typically thought of as the gold standard for conducting research (Shadish, Cook, & Campbell, 2002), many researchers are unable to employ such designs when evaluating the impact of educational programs or interventions. Specifically, evaluating the effectiveness of educational interventions via randomized control trial design is particularly difficult, because withholding programming from a subset of students is typically viewed as infeasible or unethical (Nolen & Vander Putten, 2007). Consequently, educational researchers interested in estimating the impact of an educational program must contend with unaddressed threats to the internal validity of their study (i.e., the extent to which changes in an outcome are attributable to the educational intervention).

One notable threat to the internal validity of inferences drawn from a study is self-selection bias. Self-selection bias occurs when qualitatively unique individuals either opt to participate in or are assigned to participate in a treatment or intervention (Shadish et al., 2002). For example, in educational research, secondary students who are highly socially motivated might be likely to campaign for positions in their school’s student government. Moreover, program coordinators may evaluate student outcomes (e.g., whether or not the student goes on to pursue a postsecondary education) after students complete the student government program. However, the program evaluator must then disentangle the influence of social motivation from the impact of the student government program on the outcome. That is, in this example, student motivation is a confounding variable, or a variable that covaries with both students’ participation in the intervention and the intervention outcome (Shadish et al., 2002). If evaluators fail to account for levels
of social motivation at baseline, they may draw inaccurate and inappropriate inferences about the impact of the student government program – potentially overestimating the efficacy of the program.

In randomized control studies, self-selection bias is avoided via the randomization process: individuals are randomly assigned to either participate in the intervention (i.e., “treatment”) or a control group (Shadish et al., 2002). Theoretically, randomization enables all individual characteristics (e.g., student motivation, interest, or ability) to only randomly differ across the intervention and control groups (Shadish et al., 2002). That is, because students have a 0.5 probability of being assigned to either condition (assuming group sizes are equal), all other student characteristics also have a 0.5 probability of being assigned to either group as well (Shadish et al., 2002). Consequently, the two groups of students are expected to vary only randomly from one another on both observed and unobserved characteristics (Austin, 2011a; Kaplan, 2016; Rubin, 1976).

Propensity score analyses have become a popular method for accounting for confounding variables related to self-selected participation (Kim & Steiner, 2015; Pearl, 2010). Propensity score matching, in particular, is a commonly-employed method for evaluating the impact of educational interventions in applied educational research (e.g., Cham, Hughes, West, & Im, 2015; Kainz & Pan, 2014; Leow, Wen, & Korfmacher, 2015). However, despite the recent popularity of propensity score techniques, research on the impact of measurement error on the accuracy of propensity score estimates has been sparse (Rudolph & Stuart, 2016). Moreover, recommendations on the use of propensity score analyses are based upon the presumption that characteristics related to self-selected participation in programs or interventions are measured without error (Guo & Fraser,
The purpose of the current study was to evaluate the impact of varying types (e.g., measurement error that is similar across treatment groups versus measurement error that is systematically differential by group) and degree of measurement error (e.g., 10% versus 30% measurement error or unreliability) on a researcher’s ability to account for self-selection bias when estimating the impact of an educational program.

**Background**

Propensity score matching is one way researchers attempt to emulate randomization by creating a qualitatively similar comparison group balanced on individuals’ propensities for treatment (Austin, 2011a; Guo & Fraser, 2014; Shadish, Clark, & Steiner, 2008; Stuart, 2010; Stuart & Rubin, 2008a). That is, if participants’ propensities for opting into treatment can be estimated, researchers can create a qualitatively similar comparison group that has a similar average propensity for treatment. For example, students who elect to participate in the student government program might have a high probability of deciding to participate based on their observed characteristics (e.g., social motivation). Consequently, creating a comparison group of students with the same probability of participating ensures the two groups are qualitatively similar (e.g., both groups composed of equally-motivated students). Thus, techniques such as propensity score matching provide researchers one way of mitigating bias associated with self-selected participation (Austin, 2011a; Guo & Fraser, 2014).

**Propensity Score Analyses**

Propensity scores are the predicted probability of participating in an educational intervention or program, given a set of observed variables (i.e., covariates) related to self-selected participation (Austin, 2010, 2011; Luellen, Shadish, & Clark, 2005; Stuart, 2010,
Consequently, propensity scores can be thought of as the probability of self-selected assignment to an educational program or intervention. One common method of estimating propensity scores is via logistic regression. Propensity scores are estimated for both program participants and nonparticipants and indicate the probability of participating in the treatment or intervention, conditional upon the covariates included in the model (Austin, 2011a; Stuart, 2010, Stuart & Rubin, 2008a). When propensity score matching is employed, a comparison group is created that is balanced with the treatment group on propensity for treatment (Stuart, 2010; Stuart & Rubin, 2008a). Consequently, in the context of educational research, creating matched treatment-comparison groups enables researchers to disentangle the effect of an educational program from factors related to self-selected participation (Stuart, 2010; Stuart & Rubin, 2008a). However, for propensity score matching to provide accurate treatment effect estimates, certain underlying assumptions must be met.

The strong ignorability assumption is perhaps the most important assumption underlying propensity score analyses (Shadish, 2013). The strong ignorability assumption relates to whether researchers are able to account for all important confounding variables – or covariates – when estimating propensity for treatment (Guo & Fraser, 2014; Rosenbaum & Rubin, 1983a; Stuart, 2010). Specifically, assignment to treatment is considered ignorable if the probability of participating in the intervention does not systematically vary by group after balancing groups on the set of researcher-defined covariates (Guo & Fraser, 2014; Stuart & Rubin, 2008a). Thus, if a researcher is able to create balanced treatment-comparison groups, the covariates will vary only randomly across groups, similar to randomized control trials (Stuart, 2010; Stuart & Rubin, 2008a).
The validity of treatment effect estimates depends in large part on whether all confounding variables are accounted for (Brookhart et al., 2006; Guo & Fraser, 2014; Steiner, Shadish, Cook, & Clark, 2010; Stuart, 2010; Stuart & Rubin, 2008a). That is, because researchers can only create balanced groups on the variables used to estimate propensity scores, the inclusion or exclusion of important covariates directly affects the accuracy of treatment effect estimates. Guidance is mixed in the literature, with some authors suggesting researchers rely on theory to determine which covariates should be included in the propensity score estimation model (Austin, 2011a; Stuart, 2010; Stuart & Rubin, 2008a). However, simply including all relevant covariates may not be sufficient to meet the strong ignorability assumption if covariate scores are unreliably measured.

Measurement error can artificially attenuate estimates of the relationship between two variables and lead to biased estimates of association (Meyer, 2010). Because researchers aim to balance on participants’ true propensities for program participation, covariate measurement error is problematic, as estimated propensity scores will not accurately reflect participants’ true propensities for treatment (Guo & Fraser, 2014; Rudolph & Stuart, 2016). That is, the inclusion of unreliable covariates may mean that a researcher is not able to appropriately balance treatment-comparison groups based on participants’ underlying propensities for treatment. Thus, researchers may not be able to fully mitigate the influence of self-selection bias on treatment effect estimates (Rudolph & Stuart, 2016).

In order to sufficiently meet the strong ignorability assumption, it is important the covariates included in the propensity score estimation model reliably represent the constructs for which researchers are trying to account. The inclusion of unreliable
covariate scores can therefore be thought of as a form of model misspecification (Rudolph & Stuart, 2016). That is, as covariate measurement error increases, measured covariates become naïve approximations of the true confounders or students’ actual reasons for participating in the program (Guo & Fraser, 2014; Rudolph & Stuart, 2016).

Despite frequent discussion of general measurement error in the literature, only a handful of studies have discussed the impact of measurement error on treatment effect estimation following propensity score analyses (e.g., Millimet, 2011; Rudolph & Stuart, 2016; Steiner et al., 2011). Moreover, only one study has evaluated the influence of different types of measurement error on the performance of propensity score analyses – albeit under limited conditions (Rudolph & Stuart, 2016). Because covariate measurement error has implications for how well researchers can account for self-selection, further research is warranted.

**Current Study**

The current study illustrated the impact varying levels of measurement error had on the accuracy of the inferences drawn following propensity score matching. To evaluate the impact of varying degrees and types of covariate measurement error on the accuracy of inferences obtained via propensity score analyses, a simulation study was conducted. Simulation studies allow researchers to evaluate the performance of methodological and quantitative approaches in educational research under known – or simulated – conditions (Feinberg & Rubright, 2016). Moreover, simulation studies allow researchers to answer questions that cannot be answered analytically or by data collected from a naturally-occurring population (Hallgren, 2014). Consequently, research questions regarding the performance of techniques, such as propensity score matching, are best
answered if the researcher simulates realistic conditions (Burton, Altman, Royston, & Holder, 2006). The current simulation study evaluated the performance of several propensity score matching techniques under specific measurement error conditions. The research questions answered in the current study were four-fold:

**Research Question 1:** How do differing levels of covariate score measurement error (e.g., 10%, 20%, 30% unreliability) affect the quality of matches created when using common propensity score matching techniques (e.g., nearest neighbor matching within a caliper distance, optimal matching, Mahalanobis distance matching)? The first research question pertained specifically to the quality of matches after using common matching techniques (Austin, 2011a). Following best practices in the propensity score matching literature, the quality of matches was diagnosed both numerically and visually (Austin, 2011a; Guo & Fraser, 2014; Caliendo & Kopeinig, 2008; Pattanayak, 2015; Stuart, 2010; Stuart & Rubin, 2008a). Specifically, the numeric diagnosis of matches involved evaluation of the following: the standardized mean difference between propensity score matched treatment-comparison groups on key covariates (Austin, 2011a; Stuart, 2010), the variance ratio between groups’ propensity score distributions (Rubin, 2001), and the percent in bias reduction from before matching to after matching (again, both univariately and multivariately; Pan & Bai, 2015). The visual diagnosis of matches involved the evaluation of visual aids (e.g., jitter graphs; Ho, Imai, King, & Stuart, 2011) to diagnose the covariate balance between groups for one replication of each simulated condition.

**Research Question 2:** How do differing levels of measurement error (e.g., 10%, 20%, 30% unreliability) affect the accuracy of the estimated treatment effect
when using common propensity score analyses (e.g., nearest neighbor matching within a caliper distance, optimal matching, and Mahalanobis distance matching)?

To evaluate how well propensity score matching techniques performed when covariate scores were measured with differing levels of measurement error, several indices were evaluated. Specifically, bias – or the average amount by which the estimated treatment effect differs from the simulated treatment effect – was evaluated across conditions (Feinberg & Rubright, 2016). The standard error of the treatment effect estimate, which conveys the average deviation of parameter estimates across simulations, was also evaluated (Feinberg & Rubright, 2016). The root mean squared error was evaluated to provide an index of the average deviation of treatment effect estimates from the simulated treatment effect across simulations. Finally, 95% confidence interval coverage for the estimated treatment effect was evaluated across all simulated replications to evaluate how often the true (simulated) treatment effect fell within the bounds of the confidence interval. Note that Research Questions 1 and 2 focused on the amount of measurement error (e.g., 10% versus 30% unreliability). Research Questions 3 and 4, on the other hand, focused on evaluating how type of measurement error affects the quality of matches and inferences gleaned when employing propensity score analyses.

**Research Question 3: How do different types of covariate score measurement error (i.e., measurement error that is systematically similar across groups versus measurement error that is differential by group) affect the quality of matches created when using common propensity score matching techniques (e.g., nearest neighbor matching within a caliper distance, optimal matching, Mahalanobis distance matching)?** This research question focused on how the type of simulated
measurement error impacted the quality of matches, again diagnosed both numerically and visually. Specifically, the quality of matches and number of final retained matches were compared across types of measurement error to evaluate whether the quality of matches differed based on the type of measurement error simulated in each condition.

Research Question 4: How do different types of covariate score measurement error (i.e., measurement error that is systematically similar across groups versus measurement error that is differential by group) affect the accuracy of the estimated treatment effect when using common propensity score analyses (e.g., nearest neighbor matching within a caliper distance, optimal matching, and Mahalanobis distance matching)? To evaluate the final research question, the estimated treatment effect was again compared to the simulated treatment effect to determine the extent to which estimated treatment effects deviated from the simulated (true) treatment effect. Indices used to answer this research question again included bias, standard error, the root mean squared error, and 95% confidence interval coverage around the estimated treatment effect. Then, the conditions in which the type of measurement error was manipulated were compared to determine whether different types of measurement error resulted in differential treatment effect estimates.
In the research methodology literature, randomized control trials are typically considered the gold standard for drawing causal inferences (Shadish et al., 2002). That is, when researchers employ designs in which individuals are randomly assigned to a treatment condition, it is possible to obtain treatment effect estimates that are unaffected by self-selection bias. Moreover, when employing random assignment in tandem with tight control of confounds, researchers are also able to draw causal inferences regarding the relationship between the treatment and the outcome of interest (Shadish et al., 2002).

Educational settings, however, often do not lend themselves to the implementation of randomized control trials. Specifically, due to logistical or ethical considerations, researchers often find themselves in quasi-experimental situations: situations in which students cannot be randomly assigned to educational programs or interventions (Pearl, 2010; Shadish, et al., 2002). Consequently, applied educational researchers must be cognizant of threats to the internal validity of their studies. If researchers are interested in obtaining a “causal estimate” (i.e., an estimate of program impact on participants; Dehejia & Wahba, 2002, p. 151), they must first evaluate the extent to which threats to the internal validity of their study can be mitigated.

**Internal Validity**

In the context of quasi-experimental educational research, threats to the internal validity of a study should be thoroughly evaluated prior to drawing inferences (Dehejia & Wahba, 2002). Threats to the internal validity of a study include selection bias (i.e., students self-selecting or being assigned into programs), regression artifacts (e.g.,
selection of students on the high or low end of a continuum to participate in a program), attrition of students over time (e.g., students low in ability or interest dropping out of courses over time), or history and maturation effects (Shadish et al., 2002). Of the numerous potential threats to internal validity, students’ self-selected participation in educational programs may be especially problematic.

When control over treatment assignment is lacking, assignment is endogenous (Murnane & Willett, 2011). That is, unlike exogenous assignment, where individuals are assigned to conditions according to researcher specifications or protocols, students with endogenous assignment opt to either participate or not participate in an educational program on their own accord (Murnane & Willett, 2011; Rubin, 1974). This self-selected participation into programming may be problematic because the same individual differences that drive participation may make student participants qualitatively unique from students who opt not to participate in the intervention. Moreover, differences in baseline student characteristics (i.e., before treatment) might also be associated with the intended program outcome.

If differences at baseline relate both to students’ decisions to participate in an educational program and the program outcome, these student characteristics are confounding variables. That is, observed treatment/comparison group differences in an outcome might be due to the effect of treatment and also baseline differences between groups related to self-selected participation (Shadish et al., 2002; Stuart & Rubin, 2008a). For example, students who have high levels of interest in math and science may be more likely to take advanced math and science courses in high school than students uninterested in math and science. Moreover, students who have high levels of interest in
the content of advanced math and science courses might also be more engaged in the
classroom and do better on course outcomes than students who did not opt into the
courses. If factors related to program outcomes of interest (e.g., final math ability) vary
systematically with students’ reasons for participating in the program, then the impact of
the program on students is confounded with students’ incoming characteristics (e.g.,
interest in math and science; Shadish et al., 2002). Consequently, individual differences
associated with students’ decisions to participate in an educational program may threaten
the validity of inferences researchers make regarding the effectiveness of the program.

Rubin’s Causal Model

Rubin’s causal model is a mathematical and conceptual framework formalizing
the requisite conditions under which causal relationships may be established (Holland,
1986). In his seminal paper, Rubin (1974) articulated the specific conditions that must be
satisfied to establish causation. Consequently, Rubin’s causal model provides a useful
heuristic for isolating causal relationships and identifying threats to internal validity
(Holland, 1986). Thus, this framework may be useful to educational researchers aiming
to identify threats to the internal validity of a study. Moreover, it provides the framework
upon which propensity score matching is based. Since its inception, researchers have
frequently referred back to Rubin’s seminal work when attempting to draw causal
inferences in quasi-experimental research situations (e.g., Holland, 1986; Pearl, 1995,
2010; Rubin, 2001).

The potential outcomes framework is central to understanding Rubin’s Causal
Model (Rubin, 1974). That is, if researchers want to know the effect of a treatment or
intervention, they must first have an idea of the counterfactual or what the outcome
would have been had the treated not received treatment (Rosenbaum & Rubin, 1983b; Shadish et al., 2002). However, because researchers are not able to observe both outcomes for any given individual (i.e., after receiving treatment and not receiving treatment), they must approximate the counterfactual – typically through the use of a comparison group.

Because it is impossible to know the counterfactual at the individual level, the counterfactual is approximated in randomized control trials via a randomly-assigned comparison group. That is, all individuals included in the study have a 50/50 chance of being assigned to either the treatment group or the comparison group. Because assignment is exogenous – or controlled by the researcher – the probability of participation is known and therefore controlled as a specification of the study. Given a large enough sample size to offset variations due to sampling error, random assignment ensures that all other incoming baseline characteristics between the two groups are equivalent (Austin, 2011a; Kaplan, 2016; Shadish et al., 2002; Rubin, 1976). That is, because baseline characteristics also have a 50/50 chance of being assigned to either group, random assignment will offset these differences so the groups vary only randomly from one another. Consequently, the only element that varies systematically between groups is whether or not the members of that group participated in the treatment condition.

Rubin’s causal model has become foundational for a multitude of reasons. For example, deliberating over the potential outcomes helps researchers identify issues related to not only propensity for treatment, but also the expected magnitude of the treatment effect based on individual differences related to self-selected participation.
Over the past thirty years, the study of causal inference has evolved into a well-established field of estimating causal effects when exogenous assignment is not possible (Pearl, 2010).

Propensity score matching is the most popular method for dealing with issues of confoundedness related to self-selected participation in quasi-experimental research situations (Kim & Steiner, 2015; Pearl, 2010). Specifically, propensity score analyses allow researchers to account for confounding variables related to self-selection (Austin, 2011a; Caliendo & Kopeinig, 2008; Stuart, 2010; Stuart & Rubin, 2008a). When the assumptions underlying propensity score analyses are met, researchers are able to accurately estimate the effectiveness of an educational program or intervention (Rosenbaum & Rubin, 1983a; Rubin, 2004a; 2004b; Stuart & Rubin, 2008a).

**Propensity Scores**

Propensity scores are the predicted probability of selecting into a treatment or intervention given a set of observed covariates related to self-selected participation (Austin, 2010, 2011; Caliendo & Kopeinig, 2008; Luellen et al., 2005; Stuart, 2010, Rosenbaum & Rubin, 1983a; Rosenbaum & Rubin, 1985; Rubin, 2004a; 2004b; Stuart, 2010; Stuart & Rubin, 2008a). Propensity for treatment is defined via the following formula:

\[ e(x) = pr(T_i = 1|x) \]  

where the propensity score \( e(x) \) for a given case \( i \) equals the probability of receiving treatment \( T \), given the covariates included in the model \( x \) (Rosenbaum & Rubin, 1983a). Propensity scores are commonly estimated using logistic regression and indicate both participants’ and nonparticipants’ probability of participating in an intervention.
(Austin, 2011a; Caliendo & Kopeinig, 2008; Luellen et al., 2005; Stuart, 2010, Stuart & Rubin, 2008a). For example, variables thought to be associated with students’ decisions to participate in a program are entered as predictors in the logistic regression model to predict program participation (Rosenbaum & Rubin, 1983a; Stuart, 2010; Stuart & Rubin, 2008a). Creating a matched comparison group that is balanced with program participants on the propensity score ensures covariates included in the logistic regression model vary only randomly between groups (Austin, 2011a; Rosenbaum & Rubin, 1983a). Thus, matching participants to nonparticipants on their propensity scores allows researchers to isolate program effects from factors related to self-selection (Stuart, 2010; Stuart & Rubin, 2008a).

The logic underlying propensity score matching is similar to that underlying randomized control trials: to know the effect of a treatment, a researcher must also have an idea of the counterfactual (Stuart & Rubin, 2008a). That is, a researcher must have an idea of what the outcome would have been had the treated not received treatment (Holland, 1986). Because it is not possible to observe both outcomes simultaneously for program participants, researchers attempt to approximate the counterfactual by creating a qualitatively similar comparison group with a similar propensity for treatment (Steiner, Cook, & Shadish, 2011; Steiner, Shadish, Cook, & Clark, 2010; Stuart, 2010). However, there are several underlying assumptions that must be met when employing propensity score analyses to ensure treatment effect estimates are unbiased and trustworthy (Randolph, Falbe, Manuel, & Balloun, 2014).
Underlying Assumptions

Assumptions underlying propensity score analyses directly relate to threats to the validity of inferences researchers are able to make about the effectiveness of a treatment program or intervention (Guo & Fraser, 2014). Thus, before implementing propensity score analyses (e.g., propensity score matching), it is important to first understand how the assumptions underlying the technique relate to Rubin’s causal model and the potential outcomes framework. Specifically, there are three assumptions that underlie propensity score analyses: the strong ignorability assumption, the sufficient overlap (or common support) assumption, and the stable unit of treatment assumption (Guo & Fraser, 2014; Rosenbaum & Rubin, 1983b).

Strong ignorability assumption. The strong ignorability assumption, also referred to as the no unmeasured confounders assumption (Austin, 2011a), relates to whether a researcher appropriately accounts for all confounding variables when balancing on the propensity score (Guo & Fraser, 2014; Rosenbaum & Rubin, 1983a; Stuart, 2010). Specifically, assignment to treatment is considered ignorable if the probability of receiving treatment does not systematically vary by group after matching (Stuart & Rubin, 2008a). Moreover, satisfying the strong ignorability assumption is the premise on which propensity score theorems and techniques are based (Rosenbaum & Rubin, 1983b).

\[
\{y_1, y_0\} \perp x/T
\]  

That is, given the set of covariates \(x\) the measured outcome \(y_i\) for either treatment \(1\) or control \(0\) is independent of assignment to treatment \(T\); adapted from Rosenbaum & Rubin, 1983a). If a researcher is able to account for all important covariates, then
balancing on the propensity score (e.g., creating balanced treatment-comparison matched groups) will result in the measured covariates varying only randomly across groups, thereby emulating randomized control trials (Stuart, 2010; Stuart & Rubin, 2008a). For example, if treatment participants on average have a 0.7 probability of selecting into treatment given their observed characteristics at baseline, they can be matched – or balanced – with a comparison group who also have an average propensity for treatment of 0.7. Thus, despite the probability of assignment to treatment being above 0.5 (as is the case in randomized control trials), the two groups when matched are qualitatively equivalent to one another. Consequently, assignment to treatment can be considered ignorable because probability for treatment does not vary systematically with treatment assignment.

However, in observational studies, there may be important unmeasured covariates a researcher is not aware of or not able to account for. If important confounding variables are not taken into account, then the probability for treatment still systematically varies with treatment assignment and the strong ignorability assumption is not satisfied. Consequently, despite being a core underlying assumption of propensity score matching, researchers never know in practice the extent to which the strong ignorability assumption is met (Austin, 2011a).

The sufficient overlap assumption. Related to the strong ignorability assumption is the sufficient overlap – or common support – assumption (Guo & Fraser, 2014). That is, propensity score matching techniques require sufficient overlap between the treatment and comparison groups’ propensity scores to satisfy the strong ignorability assumption (Caliendo & Kopeinig, 2008). If the two groups differ too greatly on their distributions of
propensity scores, it is not possible to find quality matches and create groups balanced on the set of covariates. Consequently, the strong ignorability assumption is not satisfied as selection bias still threatens the validity of treatment effect estimate inferences.

**Stable unit of treatment assumption.** Another assumption underlying propensity score analyses is the stable unit of treatment assumption (Austin, 2011a; Stuart, 2010; Stuart & Rubin, 2008a). That is, the effect of treatment on program participants should influence only individuals who participate in the treatment or intervention (Stuart, 2010; Stuart & Rubin, 2008a). For example, university students who participate in a public speaking program under the stable unit of treatment assumption may be assumed to have completed the entire program (e.g., treatment participation coded as 0/1). Moreover, the influence of the program is presumed to only influence program participants and not individuals in the comparison group. For example, if students who participate in the program share presentation tips and materials with students included in the comparison group, the treatment (i.e., the communication and presentation skills taught during the sessions) is no longer isolated to only program participants. Consequently, treatment diffusion may have occurred and the estimated impact of the program may no longer accurately reflect the true effect of the program on students.

Similar to other statistical techniques, the extent to which assumptions are met directly relate to how well the techniques perform in estimating accurate treatment effect estimates (Austin, 2011a; Guo & Fraser, 2014; Rosenbaum & Rubin, 1983a). Moreover, the extent to which the underlying assumptions are met depends, in large part, on decisions made by the researcher. Because the inclusion of all key confounding
covariates is a foundational assumption underlying propensity score analyses, it is important to keep the strong ignorability assumption in mind when selecting covariates.

**Propensity Score Matching**

**Selection of covariates.** Because subsequent steps of many propensity score techniques rely on sufficing the strong ignorability assumption, the selection of covariates is a foundational first step (Austin, 2011a; Harris & Horst, 2016; Stuart, 2010; Stuart & Rubin, 2008a). The inclusion or exclusion of important covariates can influence the validity of inferences about the treatment effect. Thus, researchers should consider covariates that are theoretically related to self-selected participation in the treatment or intervention (Brookhart et al., 2006; Steiner et al., 2010; Stuart, 2010). Researchers can rely on previous research or trends and relevant literature to determine which covariates are important to include when estimating propensity scores (Austin, 2011a; Stuart, 2010; Stuart & Rubin, 2008a). For example, researchers may know from previous studies or substantive research in their field that math self-efficacy is related to whether or not high school students decide to participate in an afterschool mathletes program. Moreover, based on previous research published in the field, math self-efficacy may also be related to an outcome of interest to the program (e.g., senior-level math test performance). Because math self-efficacy relates to both students’ decisions to participate in the mathletes program and the program outcome, it is likely an important covariate to include in the study. Thus, educational researchers employing propensity score techniques should evaluate relevant student characteristics to determine which covariates to include in the model.
Although theoretically-relevant covariates should be included in the model, all important covariates may not be identified in the literature. That is, in addition to theoretically-relevant covariates, researchers may also wish to empirically evaluate whether additional covariates should be included in the model. Previous simulation studies evaluating the performance of different types of covariates suggested that true confounders (i.e., covariates related to both treatment selection and the outcome) perform best at reducing bias (Austin, Grootendorst, & Anderson, 2007; Harris, Horst, & DeMars, 2018; Kelcey, 2011). Moreover, covariates related to the treatment but unrelated to the outcome might inhibit researchers from effectively creating balanced propensity score matched groups (Harris et al., 2018). Thus, one option would be to empirically screen data using proxies for the outcome to identify important confounding variables (Kelcey, 2011).

When employing propensity score matching techniques, it is important to also evaluate whether it is practical or feasible to include all important covariates in propensity score estimation. It may be impossible to include all important covariates in the model if they are unreported or unmeasured (Dehejia & Wahba, 1999, 2002). For example, if students with highly engaged parents tend to have a higher probability of participating in a college preparatory program than students with parents who are less engaged, parental engagement would be an important covariate. However, unless surveys were sent home for parents to complete, it is probable researchers would not have data on levels of parental engagement. Thus, researchers would be unable to account for parental engagement as a covariate in the model. Moreover, if parental engagement is an important covariate but not included in the model, researchers conducting the study
violate the strong ignorability assumption. The exclusion of parental engagement as a covariate may result in inaccurate estimates of the effect of the program on students. In sum, researchers should be mindful to collect important covariate measures for both treatment participants and nonparticipants when planning research studies.

In addition to the feasibility of collecting data on important covariates, the inherent stability of covariates included in the model should also be evaluated (Harris & Horst, 2016). For example, individual differences such as personality characteristics may be relatively stable across time, whereas other student characteristics such as emotion or reactivity may be unstable across time. Thus, creating balanced treatment-comparison matched groups on stable traits may accurately account for the enduring influence of confounding variables, whereas less stable covariates might not be effective at reducing bias related to self-selected participation.

In addition to the stability of covariates over time, how reliably covariate scores are measured should be evaluated (Steiner et al., 2011). Specifically, unreliable covariate scores can lead to unstable estimates of the treatment effect and fail to appropriately account for bias resulting from self-selected participation into the treatment or intervention (Steiner et al., 2011). Applied studies frequently fail to account for the measurement properties of covariates (Shadish, 2013). Although the inclusion of relatively unreliable measures of covariates is better than excluding true covariates from the model, reliable measures are desirable (Steiner et al., 2011). Because covariate measurement error has direct implications for the extent to which self-selection bias can be accurately accounted for, the effect of measurement error on treatment effect estimates is an important consideration.
**Sensitivity analysis.** Sensitivity analysis is one way to evaluate the influence unmeasured confounders might have on treatment effect estimates (Rosenbaum & Rubin, 1983b). Several variations of sensitivity analyses were created to estimate the attenuated treatment effect given unobserved confounds (e.g., the use of a non-parametric Mantel and Haenszel test statistic; Caliendo & Kopeinig, 2008). However, each of the different approaches can be used to reach a common goal: to estimate how large the unmeasured bias would have to be for the confidence interval around the estimated treatment effect to include zero (Li, Shen, & Li, 2015). That is, sensitivity analysis can be used to gauge how large the impact of an unmeasured confounder would need to be for the effect of treatment to no longer be statistically significant.

Despite the usefulness of conducting sensitivity analysis in determining how robust estimated treatment effects are to violations of the strong ignorability assumption, researchers unfortunately do not commonly use sensitivity analysis in practice (Stuart, 2010). To employ sensitivity analysis, it helps to understand that when participants and nonparticipants are balanced on propensity for treatment, their odds for treatment are equal:

\[
\frac{e(x_i \mid j) / (1-e(x_i \mid j))}{e(x_i \mid k) / (1-e(x_i \mid k))} = \frac{e(x_i \mid j) / (1-e(x_i \mid j))}{e(x_i \mid j) / (1-e(x_i \mid j))} = 1
\]

where two matched cases (j treatment case and k control case with x[j] = x[k]) have the same observed probability of receiving treatment (e(x_i)) given the covariates included in the model (Guo & Fraser, 2014). To conduct sensitivity analysis, a researcher assumes that the odds ratio (\( \Gamma \)) of treatment participants receiving treatment over comparison group nonparticipants is greater than or equal to one (adapted from Guo & Fraser, 2014, p. 359):
\[
\frac{1}{\Gamma} \leq \frac{e^{(x_i[j]/1-e^{(x_i[k])}}}{e^{(x_i[k]/1-e^{(x_i[j])}}} \leq \Gamma
\]

Thus, despite the predicted odds of receiving treatment (based on the covariates) being equal, the two groups differ in their actual propensities for treatment due to the unmeasured confounders (Guo & Fraser, 2014).

**Propensity score estimation.** After evaluating and selecting covariates, the next step is to decide the most appropriate method of estimating both participants’ and nonparticipants’ propensities for treatment (e.g., Guo & Fraser, 2014; Harris & Horst, 2016). The method selected for estimating propensity scores should align with the number of levels of treatment and expected nature of the relationship between the treatment and the outcome (Guo & Fraser, 2014; Stuart, 2010). However, in practice, researchers often model the effect of treatment on the outcome as dichotomous (i.e., students either participated in treatment or they did not, coded as 0/1; e.g., Austin, 2011a; Stuart, 2010). To model a dichotomous treatment or intervention, logistic regression may be employed to estimate the probability of participating in treatment, given a set of covariates (Stuart, 2010; Stuart & Rubin, 2008a).

**Logistic regression.** Logistic regression is the most commonly employed method for estimating propensity scores when there are two levels of treatment (i.e., individuals either did or did not receive treatment; e.g., Austin, 2011a; Stuart, 2010). To estimate propensity scores via logistic regression, all relevant covariates are included as predictors in the logistic regression model predicting participation in treatment (coded as 0/1; Pan & Bai, 2015; Stuart & Rubin, 2008a). Thus, the equation for estimating propensity for treatment is as follows:
\[ \ln \left( \frac{e^{X_i}}{1 - e^{X_i}} \right) = \beta X_i \]  

where \( \beta \) is a vector of the regression coefficients, and \( X_i \) is a vector of covariates predicting participation (coded 0/1) via logistic regression (Pan & Bai, 2015). Because all treatment and comparison group members are included in the data set when predicting probability for treatment, both treatment participants and nonparticipants receive a propensity score regardless of whether or not they participated in the treatment (Austin, 2011a; Guo & Fraser, 2014; Pan & Bai, 2015; Stuart, 2010; Stuart & Rubin, 2008a).

Similar to the way in which logistic regression is implemented for inferential purposes, both categorical or continuous predictors (i.e., covariates) can be included in the model (Tabachnick & Fidell, 2013). Moreover, complex relationships among the covariates included in the model including interactions and polynomials can also be specified. However, similar to the process of selecting covariates, it is important to consult the relevant literature to determine whether additional terms or complex relationships should be specified in the model to estimate propensity scores.

Because the logistic regression model is used solely for the purpose of predicting treatment participation given the covariates included in the model, this step of the propensity score matching process is frequently referred to as a non-parametric preprocessing step (Guo & Fraser, 2014; Pan & Bai, 2015; Stuart, 2010). That is, researchers do not make inferences based upon the logistic regression model results, but rather use the model to estimate both treatment and comparison group members’ propensities for treatment. Propensity scores are then used in a separate step (e.g., for propensity score matching or weighting) to estimate the treatment effect. Thus, the model summary information for the logistic regression model is often not evaluated when
estimating propensity scores (Guo & Fraser, 2014). Moreover, models that are too predictive may be problematic. That is, models that differentiate well between treatment and comparison groups might indicate that the two groups are too different from one another to presume they are subsamples from the same population. Consequently, propensity score analyses may be inappropriate as it is not possible to meet the assumptions underlying propensity score analyses (e.g., common support).

**Mahalanobis distance.** Another matching metric is the Mahalanobis distance metric (e.g., Guo & Fraser, 2014; Rosenbaum & Rubin, 1985). Unlike propensity score estimation via logistic regression, matches created via Mahalanobis distance estimates do not rely on model-based estimates (Rubin, 1980). That is, rather than reducing propensity for treatment to a single composite score (e.g., a propensity score), the entire vector of covariate values for each individual in the treatment and control groups is used to calculate the multivariate distance between treatment participants and nonparticipants (Guo & Fraser, 2014; Rosenbaum & Rubin, 1985).

Mahalanobis distance measures are typically used to evaluate how far a given case is from a multivariate centroid (e.g., to flag outliers; Tabachnick & Fidell, 2013). However, when used for matching, the distance measures instead represent the distance between two points: a treatment group member’s vector of covariates and the vector of covariates for all possible matches from the comparison group (Guo & Fraser, 2014). The distance between the treatment participant and each individual comparison group member is calculated as follows:

\[
\text{Distance} = (X_1 - X_2)S^{-1}(X_1 - X_2)^T
\]  

(5)
where $X_1$ and $X_2$ are the vector of covariate scores for a treatment group member (1) and a comparison group member (2), and $S$ is the full sample of comparison group members’ covariance matrix (Guo & Fraser, 2014, p. 146).

Although Mahalanobis distance measures are analogous to propensity scores estimated via logistic regression in several ways, Mahalanobis distance measures are multivariate distances rather than weighted composites (Guo & Fraser, 2014). Consequently, covariates included in Mahalanobis distance matching are not weighted based on their relative contribution to predicting group membership (Guo & Fraser, 2014). That is, in Mahalanobis distance estimation, covariates contribute relatively equally to the distance metric, regardless of whether or not a given covariate is related to self-selected participation in the treatment program.

**Propensity score matching.** When propensity scores are employed to create qualitatively similar matched treatment-comparison groups, a researcher must select the algorithm by which propensity score matches will be created. There are several matching algorithms that may be used (e.g., nearest neighbor matching, nearest neighbor matching with caliper, and optimal matching; Austin, 2009b; 2011b; 2013). Consequently, researchers should be cognizant of the process by which matches are created via each algorithm. That is, the process by which matches are created not only differs across matching algorithms, they can also produce different matched treatment-comparison samples. Thus, the use of different algorithms may result in different conclusions about the effect of treatment on participants (Austin, 2013; Jacovides, Foelber, & Horst, 2016).

**Nearest neighbor matching.** Nearest neighbor matching is the most commonly used matching algorithm in the propensity score matching literature (Stuart, 2010).
Nearest neighbor matching employs a greedy algorithm (Austin, 2013). That is, the algorithm sequentially moves through the list of participants and matches each participant to the nonparticipant with the closest propensity for treatment. The algorithm sequentially iterates through the entire set of treatment participants until each participant has a nonparticipant match. Typically, the treatment cases are ordered in a descending fashion so that cases with the highest propensity for treatment are matched first (Ho et al., 2011). Although it is the default to sort descending in some software packages (e.g., Ho et al., 2013), it is not necessary, in terms of the algorithm. The algorithm for the nearest neighbor matching process is as follows:

\[ d(i,j) = \min_j \{ |e(X_i) - e(X_j)| \} \] (6)

where matches are made between two cases (case \( i \) and case \( j \)) that are the minimum distance between estimated propensity scores for each case (\( e(X_i) \); Pan & Bai, 2015, p. 7).

One drawback to implementing nearest neighbor matching is that no stipulations are placed on the distance between matches (Austin, 2013). That is, matches are created with the closest nonparticipant; there is no evaluation of the quality of that match before assigning it to the participant (Austin, 2013). Thus, participants with high propensities for treatment (e.g., propensity scores of 0.9) might be matched to nonparticipants with low propensities for treatments (e.g., propensity scores of 0.4) if the closest match has been previously assigned. Moreover, in this example, the common support and strong ignorability assumptions are not satisfied as there are not quality matches available and treatment assignment is not ignorable. To mitigate the issue and generate higher quality
matches than those obtained via the nearest neighbor matching algorithm, researchers may decide to set a caliper distance.

**Nearest neighbor matching with caliper.** Nearest neighbor matching with a caliper distance is an extension of nearest neighbor matching. The caliper distance is a researcher-specified distance in standard deviations units on the logit of the propensity score within which matches are considered acceptable (Austin, 2011a, 2011b; Stuart, 2010). That is, the nearest neighbor greedy algorithm is still employed; however, matches are only created when they fall within a designated distance of the treatment participant’s propensity score (Austin, 2011a, 2011b; Stuart, 2010). Consequently, treatment participants for whom no nonparticipants’ propensity scores fall within range are unmatched and excluded from the final matched treatment-comparison group. The algorithm for nearest neighbor matching within a designated caliper distance is as follows:

\[
d(i,j) = \min_{j} \{ |e(X_i) - e(X_j)| < b \}
\]

where matches are again made between two cases (treatment case \(i\) and control case \(j\)) that are the minimum distance between estimated propensity scores for each case \(e(X_i)\) as long as the distance between matches falls below the researcher-specified caliper width \(b\); Pan & Bai, 2015, p. 7). Suggested caliper distances range from 0.20 (Austin, 2011a, 2011b; Stuart, 2010) to 0.25 (Rosenbaum & Rubin, 1985) standard deviations on the logit of the propensity score. Previous simulation studies have championed nearest neighbor matching with a caliper width as a top performer (e.g., 0.2 standard deviations on the logit of the propensity score) over other matching methods (Austin, 2010; 2011b; 2013).
Although the use of a caliper distance provides a measure for vetting treatment-comparison matches, it does not provide a holistic evaluation of matches. That is, the quality of matches is only evaluated on a match-by-match basis. The overall quality of matches (i.e., across all matched pairs) is not evaluated to ensure the absolute distance across all matched pairs is at a minimum. Consequently, researchers may instead consider employing optimal matching to improve the overall quality of matches.

**Optimal matching.** Optimal matching, unlike nearest neighbor matching, is not a sequential algorithm used to create matched treatment-comparison pairs. Rather, optimal matching is conducted via a mathematical optimization process (Rosenbaum, 1989). That is, when optimal matching is employed, all possible combinations of treatment-comparison matched pairs are evaluated, and the final solution contains the optimal set of final matched pairs (Rosenbaum, 1989). To find the optimal solution of matched treatment-comparison pairs, a linear programming optimization process is employed (Rosenbaum, 1989).

Linear programming processes are frequently used to produce an optimal solution given a set of constraints (Feiring, 1986; Ignizio, 1985). In the case of optimal matching, the constraints in the linear programming problem include the number of treatment cases and the number of comparison pool cases to be matched to each treatment case (Rosenbaum, 1989). The constraints imposed on the linear programming problem create a multidimensional solution space containing all possible combinations of pairs (Feiring, 1986; Ignizio, 1985). Each point, or edge, of the solution space is a possible maximum or minimum because these areas represent the furthest points from the orient (or absolute zero) in Euclidean distance (Feiring, 1986; Ignizio, 1985). Thus, the edges are the only
areas of the solution space that need to be evaluated. After constructing the solution space, all possible combinations of matched pairs – or solutions – are evaluated by a search algorithm until the solution that minimizes the distance between all matched pairs is discovered (Feiring, 1986; Ignizio, 1985; Rosenbaum, 1989). Consequently, the solution found via optimal matching minimizes the total absolute distance among all matched treatment-comparison pairs (Austin, 2009b; 2013; Gu & Rosenbaum, 1993; Hansen & Klopfer, 2006; Rosenbaum, 1989).

Optimal matching provides closer matches than nearest neighbor matching and a lower average distance between propensity scores for treatment-comparison matched pairs (Gu & Rosenbaum, 1993). Although the average absolute distance between matched pairs is lower, optimal matching may not result in better quality matches than the nearest neighbor matching methods. That is, despite improvements in the overall quality of matches, optimal matching does not tend to outperform the other methods in producing quality matches and unbiased treatment effect estimates (Austin, 2011b). Consequently, researchers may opt to simply employ nearest neighbor matching.

**Mahalanobis matching.** As mentioned previously, unlike the other forms of matching, Mahalanobis distance matching is performed on a multivariate distance measure and thus does not rely on a single estimate of overall propensity for treatment. Rather, Mahalanobis distance matching involves calculating the multivariate distance from each participant to all nonparticipants (Guo & Fraser, 2014). Moreover, similar to propensity score matching within a caliper distance, researchers employing Mahalanobis distance matching may also specify a caliper distance via the following algorithm:

\[ d(i,j) = \min_j \{ D_{ij} < b \} \]  

(8)
where $D(ij)$ is the multivariate distance between $i$ and $j$ units on the vector of covariates that fall within the designated distance ($b$) in multivariate space (Pan & Bai, 2015, p. 7). Thus, similar to nearest neighbor matching within a caliper, the closest possible match will be designated unless it is otherwise already matched or it falls outside of the designated caliper width. Although the default in nearest neighbor matching is to sort treatment participants in descending order on their propensity for treatment, there is no such ordering on the Mahalanobis metric (Pan & Bai, 2015). Consequently, treatment participants are typically randomly ordered, and then the matching algorithm moves sequentially through the random ordered treatment participants to find the nearest match for each participant on the vector of covariates (Pan & Bai, 2015).

Because propensity scores are useful in weighting covariates according to how predictive they are of treatment participation, some authors recommend including propensity scores estimated via logistic regression when Mahalanobis distance matching (Rosenbaum & Rubin, 1985). Specifically, researchers may opt to include propensity scores as another value in each individual’s vector of covariates in the computation of Mahalanobis distance scores (Guo & Fraser, 2014).

Mahalanobis matching is best employed with large samples and when there is ample common support (Guo & Fraser, 2014; Kaplan, 2016). That is, because all covariates are equally prioritized when employing Mahalanobis distance matching, it can work well in creating equivalent matches across a set of covariates. However, it might be difficult to find close matches with large sets of covariates and small samples, or when common support is inadequate (Guo & Fraser, 2014). That is, as the number of covariates increases, so does the average distance, and it becomes more difficult to find quality
matches across the entire set of covariates (Guo & Fraser, 2014). Thus, with a large number of covariates and small samples, researchers may opt to use propensity scores estimated via logistic regression as it prioritizes creating quality matches on important covariates.

**Additional considerations.** In addition to the different types of matching procedures, there are also other considerations to keep in mind when employing propensity score matching techniques. Specifically, it is important to remain cognizant of the nature of the data and whether there is sufficient common support (Austin, 2011a; Claiendo & Kopeinig, 2008; Stuart, 2010). A lack of common support might make it difficult to find quality matches, and inhibit the creation of balanced treatment-comparison matched groups. Moreover, if there is not sufficient common support, employing nearest neighbor matching with a caliper distance may result in dropping participants from the final matched data set for whom there were no quality matches. If unmatched participants are at the high end of the propensity score distribution, it may be that qualitatively unique cases are excluded from the final matched groups. Finally, a lack of common support between the treatment and comparison groups might indicate that the two groups are too qualitatively distinct from one another to employ matching techniques. That is, the treatment and comparison groups may actually represent two distinct populations.

When creating matched treatment-comparison groups, an important decision is the number of comparison group members matched to each treatment group member. Researchers can opt to use a one-to-one matching ratio or a one-to-many matching ratio (Claiendo & Kopeinig, 2008; Stuart, 2010; Stuart & Rubin, 2008b). When one-to-one
matching is employed, only one comparison group member is matched to each treatment participant. In contrast, when one-to-many matching is employed, researchers can specify a disproportionate number of comparison group members to match to every treatment participant (e.g., three comparison members matched to every one treatment participant; Stuart, 2010). Matching multiple comparison group members to each treatment participant can result in more powerful between-group statistical tests; however, researchers should be mindful of the issues that may arise when conducting between-group analyses with disproportionate group sizes (Austin, 2011a).

In addition to selecting the matching ratio is the decision of whether to match with or without replacement (Guo & Fraser, 2014; Stuart, 2010; Stuart & Rubin, 2008a). Matching with replacement means matched nonparticipants remain in the pool of potential matches after being matched in a previous iteration (Guo & Fraser, 2014). That is, when matching with replacement, it is possible for the same comparison group member to be selected and matched to multiple treatment participants. Matching without replacement, however, does not allow nonparticipants to be matched to treatment participants multiple times (Guo & Fraser, 2014). That is, once a nonparticipant is matched to a treatment participant, the nonparticipant is removed from the pool of possible comparison group matches.

When comparison group members are matched with replacement, there are issues that can arise in outcome analyses. Because the comparison group is not composed of independent observations, there is a need to appropriately model the dependencies in observations in outcome analyses (Austin, 2011a; Caliendo & Kopeinig, 2008; Guo &
Fraser, 2014). Consequently, matching without replacement may be most appealing, as outcome analyses are straightforward.

When conducting propensity score matching, outcomes data should be merged only after matches have been made so as to not introduce researcher bias (Stuart & Rubin, 2008a). Because there are numerous ways to create matched treatment-comparison groups, it is common to try several methods to achieve balance (Pan & Bai, 2015). Thus, if outcomes data are available, it may be tempting to evaluate how the estimated treatment effect varies using each method. After matches are created, the next step in the propensity score matching process is to evaluate covariate balance and diagnose the quality of matches.

**Diagnosing the quality of matches.** Because the goal of propensity score matching techniques is to balance groups on the propensity score, it is important to evaluate whether the matched treatment and comparison groups are indeed balanced after matching (Guo & Fraser, 2014; Pattanayak, 2015; Stuart, 2010; Stuart & Rubin, 2008a). Typically, the quality of matches is diagnosed via numeric and visual diagnostic information (Bai, 2015; Guo & Fraser, 2014; Pattanayak, 2015; Stuart & Rubin, 2008a; Stuart, 2010). Moreover, balance between the treatment and comparison matched groups is evaluated both on the multivariate composite (e.g., on the propensity score) and on individual covariates included in the model (Austin, 2010, 2011a, 2014; Guo & Fraser, 2014; Pattanayak, 2015; Stuart, 2010; Stuart & Rubin, 2008a).

**Numeric diagnosis of matches.** There are several methods for numerically diagnosing the quality of matches, including univariate and multivariate approaches (Austin, 2011a; 2013; Guo & Fraser, 2014; Stuart, 2010; Stuart & Rubin, 2008a).
Moreover, numeric approaches are used to evaluate both absolute and relative differences between treatment-comparison matched groups. That is, the absolute difference between matched groups can be evaluated via the standardized mean difference for individual covariates (Austin, 2013). Alternately, the relative difference in balance between groups compares group balance after matching to initial group differences (e.g., percent bias reduction (PBR) to evaluate individual covariates; Pan & Bai, 2015).

*Standardized mean difference.* Univariately, individual covariate balance between groups can be evaluated via the standardized mean difference (Austin, 2013). Equivalent to calculating Cohen’s $d$ effect sizes, the equation for calculating the standardized mean difference is as follows:

$$d = \frac{(\bar{x}_{\text{treatment}} - \bar{x}_{\text{control}})}{\sqrt{\frac{\sigma^2_{\text{treatment}} + \sigma^2_{\text{control}}}{2}}}$$

(9)

where $d$ is the standardized mean difference, $\bar{x}$ is the mean of an individual covariate (for either the treatment group or the control group) and $\sigma^2$ is the variance for each group on that covariate (Austin, 2013). Typically, a standardized mean difference of less than 0.2 (Stuart, 2010) or 0.1 (Austin, 2011a) is considered sufficient covariate balance between groups.

Similarly, balance between treatment-comparison matched groups on individual dichotomous covariates can be evaluated using the following formula:

$$d = \frac{(\hat{p}_{\text{treatment}} - \hat{p}_{\text{control}})}{\sqrt{\frac{\hat{p}_{\text{treatment}}(1-\hat{p}_{\text{treatment}}) + \hat{p}_{\text{control}}(1-\hat{p}_{\text{control}})}{2}}}$$

(10)

where $\hat{p}$ equals the proportion of either treatment or control individuals coded one on that covariate (Austin, 2013).
**Percent bias reduction.** To evaluate the improvement in individual covariate balance after matching, the percent bias reduction is frequently evaluated (Bai, 2011; Pan & Bai, 2015). The equation for calculating the percent bias reduction is as follows:

\[
PBR = \frac{(M_1(X_i) - M_0(X_i))_{\text{before matching}} - (M_1(X_i) - M_0(X_i))_{\text{after matching}}}{(M_1(X_i) - M_0(X_i))_{\text{before matching}}} \times 100\%
\]

where \( M \) is the mean of the treatment group (1) or the comparison group (0) for each covariate \((X_i)\) included in the logistic regression model (Pan & Bai, 2015). Typically, an 80% improvement in balance from before matching (i.e., via the entire unmatched sample) to after matching is considered sufficient balance for individual covariates (Pan & Bai, 2015).

Multivariately, the groups can also be compared on their distributions on the propensity score. Similar to evaluating balance on individual covariates, the propensity score balance of the matched treatment-comparison groups can be evaluated via the standardized mean difference (Austin, 2009a; 2011a; 2013; Guo & Fraser, 2014; Stuart, 2010). Moreover, researchers may also calculate the percent bias reduction (Pan & Bai, 2015). In addition to evaluating mean differences, the variance ratio is used to compare the spread of propensity scores between matched treatment-comparison groups.

**Variance ratio.** The variance ratio is simply the variance of the treatment group’s propensity scores divided by the variance of the control group’s propensity scores (Rubin, 2001). Thus, the variance ratio can be calculated as follows:

\[
VR = \frac{\sigma^2_{\text{treatment group}}}{\sigma^2_{\text{control group}}}
\]

where \( \sigma^2_{\text{treatment group}} \) is the variance of the treatment group and \( \sigma^2_{\text{control group}} \) is the variance of the control group (Austin, 2013). The variance ratio is a straightforward and useful way to evaluate whether the propensity score variability is roughly equal between
matched groups. It is recommended that the variance ratio be close to one, indicating the variability in propensity scores for both groups is roughly the same (Rubin, 2001).

**Visual diagnosis of matches.** In addition to numerically evaluating the quality of matches, matches can also be diagnosed using several types of visual aids. For example, absolute standardized mean difference plots, quintile-quintile (Q-Q) plots, boxplots, density plots, jitter plots, and histograms may all be used to evaluate the quality of matches univariately and multivariately. Moreover, visual aids are available via several propensity score matching packages in R (e.g., via the MatchIt package in R; Ho et al., 2011).

**Absolute standardized mean difference plots.** Univariately, absolute standardized mean difference plots can be used to evaluate individual covariate balance between groups (see Figure 1; Pattanayak, 2015). The plots display the standardized mean difference between treatment participants and nonparticipants both before matching (left side) and after matching (right side). Typically, a line is included in the plot demarcating a benchmark (e.g., 0.25) under which the quality of matches is deemed acceptable (Ho et al., 2007).

**Q-Q plots.** Q-Q plots are visual aids used to compare the distribution of the individual covariate scores of the treatment group to that of the control group across points in the density distribution (or quintiles; Stuart, 2010). To interpret Q-Q plots, researchers visually evaluate whether the dots after matching tend to fall along the center line (see Figure 2). The dots represent the density of treatment to comparison units at each point in the distribution. Thus, Q-Q plots in which dots fall perfectly along the
center line indicate that the treatment-comparison matched groups have symmetrical density distributions on that respective covariate (Ho et al., 2011).

**Boxplots.** Similar to Q-Q plots, boxplots indicate the distribution of individual covariate scores between groups (see Figure 3). Two boxplots – or box and whisker diagrams – display the distribution of covariate scores for treatment participants and nonparticipants before and after matching. The middle line in the box represents the median, and the box represents the first and third quartile for that respective group on the covariate. The “whiskers,” or protruding lines, display the total range of scores for each group (minimum value and maximum value). Thus, boxplots provide unique visual information regarding other measures of central tendency that are not provided overtly via the Q-Q plots.

**Density plots.** Density plots provide another way for researchers to visually check whether the distribution of each group on individual covariates is roughly equivalent (Harris & Horst, 2016; Pattanayak, 2015). For example, when plotting treatment participant and nonparticipant covariate scores using semitransparent density plots (see Figure 4), differences in the two groups’ distributions may be more apparent than in Q-Q plots. Thus, density plots provide yet another means by which researchers can further diagnose the univariate balance between treatment and comparison matched groups.

**Jitter plots.** Similar to numeric diagnosis of balance, the quality of matches may also be evaluated via visual aids. Jitter plots are one useful tool for evaluating the quality of matches by treatment group. Jitter plots allow visual evaluation of whether the distribution of propensity scores appear roughly equivalent for both groups (Stuart, 2010). Moreover, jitter plots may be used to evaluate whether there is sufficient overlap
between the two groups’ propensity scores. Specifically, in the MatchIt package in R, propensity scores for treatment participants and nonparticipants to whom they are matched are plotted across the center two rows in the plot (see Figure 5; Ho et al., 2011). Conversely, unmatched treatment participants (e.g., if using a caliper) fall across the top row of the plot. Finally, the bottom row indicates the propensity scores for nonparticipants who were not matched to treatment participants. Note that the jittering of case points makes it easy to distinguish the density of treatment participants or nonparticipants across levels of propensity score when the sample size is small. However, jitter plots may become unwieldy with large sample sizes as each group instead appears as a thick black line (see Figure 6). Thus, histograms provide an appealing alternative to jitter plots with large samples.

Histories. Although commonly used as visual aids for a multitude of statistical purposes, histograms are particularly helpful to diagnose the quality of propensity score matches with large samples. Specifically, histograms can convey differences in the proportion of treatment participants or nonparticipants at each level of the propensity score. That is, researchers can evaluate whether there is the same proportion of treatment or comparison group members at each level of propensity score after creating matching (see Figure 7). Thus, with large samples, the distributions between groups are still easily compared via histograms even when jitter plots appear indistinguishable.

Representativeness of samples. After propensity score matching is employed, it is important to evaluate the representativeness of the final matched groups (Harris & Horst, 2016; Jacovidis et al., 2016). Some forms of propensity score matching, such as nearest neighbor matching with a caliper, exclude treatment participants for whom quality
matches cannot be created. Thus, researchers may run the risk of excluding qualitatively unique treatment participants from the final matched data set. Consequently, the final matched data set may no longer be representative of the original sample of treatment participants. Loss of qualitatively unique participants is problematic if the goal is to make inferences regarding the impact of the treatment on the original treated sample. Thus, researchers should be cognizant of compositional changes in their final matched treatment group. Moreover, researchers may consider selecting a matching process that retains the entire sample of treatment participants when possible.

**Estimating Effect of Treatment.** After completing the propensity score matching process, the next step is estimation of the treatment effect. Several treatment effect estimands may be obtained via propensity score analyses. Because propensity score techniques tend to be used with large data sets in intervention research, the generalization of treatment effect estimates to known populations is common (Austin, 2011a; Guo & Fraser, 2014; Stuart, 2010). Specifically, three types of treatment effects are commonly discussed: the average treatment effect on the treated (ATT), the average treatment effect (ATE), and the estimated average treatment effect on the controls (ATC; Austin, 2011a; Caliendo & Kopeinig, 2008; Guo & Fraser, 2014; Stuart, 2010). To determine which treatment effect estimate is of interest, researchers may evaluate the research question being investigated to determine the population to which inferences will be made.

**Average Treatment Effect on the Treated.** The treatment effect estimand for the average treatment effect on the treated (ATT) – also referred to as weighting by the odds – provides an estimate of the impact of a given treatment for the individuals who selected to participate (Austin, 2011a; Guo & Fraser, 2014; Caliendo & Kopeinig, 2008). That is,
as the name suggests, the ATT treatment effect estimand is used to make inferences about the effectiveness of a program for the population of individuals who self-selected into treatment (Guo & Fraser, 2014). Weights for estimating the estimated treatment effect on the treated are calculated as follows:

$$ATT = T_i + (1 - T_i) \frac{e(x_i)}{1 - e(x_i)}$$  \hspace{1cm} (13)

where $T$ is an indicator variable denoting whether each case ($i$) was treated (coded 0/1), and $e(x_i)$ is the estimated propensity for treatment for each case conditional on the set of covariates (adapted from Guo & Fraser, 2014). Note, in the above formula, that participants are weighted by their odds of participating in the treatment. Propensity score matching techniques are common methods of estimating ATT (Guo & Fraser, 2014; Stuart, 2010; Stuart & Rubin, 2008a).

**Average Treatment Effect.** Unlike ATT, the average treatment effect estimand (ATE) relates to how the treatment effect would generalize back to a population at large (e.g., the entire population including both treatment participants and nonparticipants; Austin, 2011a; Guo & Fraser, 2014; Caliendo & Kopeinig, 2008). Weights to estimate the average treatment effect can be calculated as follows:

$$ATE = \frac{T_i}{e(x_i)} + \frac{1 - T_i}{1 - e(x_i)}$$  \hspace{1cm} (14)

where $T$ is an indicator variable denoting whether each case ($i$) was treated (coded 0/1), and $e(x_i)$ is the estimated propensity for treatment for each case conditional on the set of covariates (adapted from Guo & Fraser, 2014). After calculating the ATE weights, the average treatment effect is typically estimated via techniques such as full matching or weighting, methods not employed in the current paper (Stuart, 2010). Moreover, the average treatment effect estimand may be useful in evaluating the expected outcomes for
administering treatment to a larger proportion of a given population than the proportion that received treatment (Guo & Fraser, 2014). Thus, ATE may be a useful estimand when decisions are being made about the costs and benefits of expanding treatment services.

**Average Treatment Effect on Controls.** In addition to estimating the treatment effect for a population overall (ATE) and on only treatment participants (ATT), researchers can estimate how the effect of treatment would generalize to only a population of untreated individuals (Austin, 2011a). This treatment effect estimand is often referred to as the average effect of treatment on the controls (ATC; Austin, 2011a). The equation to estimate ATC weights is as follows:

\[
ATC = \frac{T_i(1-e(x)_i)}{e(x)_i} + (1 - T_i)
\]

where \( T \) is an indicator variable denoting whether each case \((i)\) was treated, and \( e(x) \) is the estimated propensity for treatment for each untreated case (adapted from Austin, 2011a). Note that to calculate ATC weights, a researcher weights the treated by the inverse of the odds, whereas the untreated cases are weighted by one. To estimate the treatment effect for a population of untreated individuals, researchers can use the calculated ATC weights and then include the weights in a regression model (Austin, 2011a).

Of interest in the current study is ATT estimated after employing propensity score matching techniques. Because propensity score matching is simply a nonparametric preprocessing step, the outcome analyses are conducted in a second, separate step (Guo & Fraser, 2014; Pan & Bai, 2015; Stuart, 2010; Stuart & Rubin, 2008a). However, there is conflicting advice in the literature regarding the best way to analyze outcomes. For example, regression adjustments using propensity scores can be conducted on outcome
data from the final matched samples (Stuart, 2010). Alternately, a researcher could simply compare mean differences between the treatment and comparison groups (Austin, 2011a). However, there are disagreements in the literature regarding whether or not the matched treatment-comparison pairs should be treated as independent observations. Authors who argue the pairs are independent recommend using simple between-subjects outcome analyses (e.g., Schafer & Kang, 2008). Yet, others have argued that matched treatment-comparison pairs are not independent from one another and should not be treated as independent observations (e.g., Austin, 2011a; Imbens, 2004). Consequently, the propensity score matching literature should be consulted to determine which approach (i.e., treating matched groups as independent or dependent) most aligns with a specific data situation (e.g., Austin, 2011a; Guo & Fraser, 2014; Imbens, 2004; Schafer & Kang, 2008; Stuart, 2010).

**Benefits and Drawbacks of Propensity Score Matching**

Propensity score matching techniques are intuitive techniques for modeling the counterfactual (Shadish, 2013; Shadish & Cook, 2009). That is, because many researchers are used to simple between-groups comparisons, propensity score matching techniques may seem straightforward (Guo & Fraser, 2014). However, as is the case with any methodology, propensity score matching techniques also have several notable limitations. Thus, researchers should keep in mind both the benefits and drawbacks of using propensity score matching techniques when selecting the most appropriate technique to employ with their data.

**Benefits of using propensity score analyses.** Propensity score matching techniques offer several benefits over traditional regression approaches. For example, in a
multiple regression framework, covariates may be used to control for differences between
groups at baseline (e.g., academic ability pretest scores collected before an educational
intervention; Tabachnick & Fidell, 2013). However, when conducting multiple
regression, researchers assume 1) that it is appropriate to use the variable as a covariate in
the same model as the outcome and 2) that the relationship between group membership
and the outcome is fixed or systematic across levels of the covariate (Rosenbaum &
Rubin, 1983b). In propensity score analyses, the relationship of the covariates and
selection into treatment is captured in a separate model (i.e., the model used to estimate
propensity scores; Guo & Fraser, 2014). That is, in propensity score analyses, the
mechanism by which individuals selected into treatment is modeled in a separate data
preprocessing step (Leite, 2017). Thus, propensity score analyses help researchers avoid
misspecification of the self-selection mechanism (McCaffrey, Ridgeway, & Morral,

Another benefit of propensity score matching is that results are easy to
communicate to a broad range of audiences. That is, because simple between-groups
comparisons are easily understood by many audiences, the results from propensity score
matching analyses may be more intuitive and straightforward than traditional multiple
regression techniques. Thus, techniques like propensity score matching may be appealing
when communicating results to stakeholders with a range of statistical expertise
(McCaffrey et al., 2013).

Propensity score techniques also provide a means by which researchers can easily
and overtly check the equivalence of groups at baseline (Harder, Stuart, & Anthony,
2010; Shadish, 2013). Because the goal of propensity score matching is to evaluate and
create balanced groups, researchers may employ these methods to evaluate the equivalence of groups in other research settings. That is, researchers can use propensity score balance diagnostic information to simply evaluate group differences on a set of covariates, even if propensity score matching is not employed.

Finally, and perhaps most importantly, propensity score matching provides researchers with a statistical way of emulating randomized control trials, possibly foreshadowing results that could be obtained via future randomized control trials (D’Agostino, 1998; Shadish, 2013; Steyer, Gabler, von Davier, & Nochtigall, 2000). Thus, propensity score matching provides a useful framework for researchers who conduct both quasi-experimental and experimental research (Shadish, Clark, & Steiner, 2008; Shadish & Cook, 2009). Moreover, propensity score matching techniques are especially useful for researchers who wish to answer research questions regarding the estimated treatment effect under conditions similar to randomized control trials.

**Drawbacks of using propensity score analyses.** The primary limitation of propensity score analyses is that the extent to which researchers meet the strong ignorability assumption is never known (Shadish, 2013). That is, in practice, researchers are not able to adjust for hidden selection bias, let alone test for it in a statistical manner (Rubin, 1997). Thus, propensity score matching techniques cannot account for unobserved or unmeasured covariates. As noted in the literature, “…the strong ignorability assumption seems to be strongly ignored by most users of propensity score analysis” (Shadish, 2013, p. 134).

Another limitation of propensity score analyses is that researchers are not able to treat true confounding variables (i.e., variables related to both self-selected participation
and the outcome) differently from variables related only to self-selected participation (Shadish, 2013). Specifically, when logistic regression is employed, covariates in the model are weighted according to how well they predict participation in the treatment, not whether they also relate to the outcome (Guo & Fraser, 2014; Rubin, 1997). Because covariates unrelated to the outcome might be weighted more heavily in the logistic regression model, variables that are not true confounders may carry more weight when creating treatment-comparison matches. Thus, variables strongly related to only selection into treatment may be weighted more heavily than true confounding variables in the logistic regression model. This is problematic if true confounding variables are not appropriately balanced, as it may result in inaccurate treatment effect estimates (Harris et al., 2018). Consequently, researchers may wish to rely on theory or empirically identify true confounding variables when selecting covariates to include in the model to estimate propensity scores (Kelcey, 2011).

The distinction between rigorous research design and accounting for threats to internal validity in hindsight is an important one. That is, propensity score matching techniques cannot make up for poor research design (Shadish, 2013). If the goal is to establish causation between two elements in a study, it is most appropriate to conduct a randomized control trial. Unfortunately, applied researchers are not always in the position to do so. Consequently, propensity score matching provides an appealing alternative.

In order to satisfy the strong ignorability assumption, propensity score analyses also necessitate researchers have data on all important covariates and the outcome for both the treatment group and a comparison group (Shadish, 2013; Zhao, 2004). Moreover, there must also be common support between the treatment participants’ and
nonparticipants’ propensity scores if a researcher aims to create quality matched groups (Shadish, 2013; Zhao, 2004). Because of the data requirements necessary for conducting propensity score analyses, propensity score matching techniques do not lend themselves to situations in which covariate data are sparse for both treatment participants and nonparticipants.

A final drawback to propensity score matching techniques is that, although seemingly intuitive (McCaffrey et al., 2004; Shadish, 2013), the process does not appear to be fully understood by applied researchers. That is, in the applied propensity score matching literature, researchers tend to use whatever covariate measures are convenient or available rather than intentionally planning and collecting covariate data for their study (Shadish, 2013). However, several of the drawbacks mentioned above can be mitigated by being familiar with propensity score matching best practices and basing covariate selection on previous research.

**Measurement Error and Propensity Score Estimation**

When propensity score analyses are used appropriately, they can be a powerful tool for modeling the counterfactual and obtaining causal estimates. However, one area in which more research is needed is on the performance of propensity score analyses when covariates are measured with error. Similar to other areas of research in the social sciences, measurement error is an unfortunate reality with which applied educational researchers must often contend (Shadish et al., 2002). Specifically, in educational and psychological research, measurement error can artificially attenuate – or bias – estimates of the relationship between two variables (Meyer, 2010). In the context of propensity score analyses, the inclusion of unreliable covariates means that researchers are unable to
accurately estimate students’ *true* propensities for treatment. Thus, if propensity for treatment is a *naïve* estimate (i.e., not an accurate representation of students’ actual propensity for treatment), balancing on propensity scores may not properly mitigate group differences associated with self-selected participation (Rudolph & Stuart, 2016).

The inclusion of unreliable scores on important covariates can therefore be thought of as a form of model misspecification. That is, as scores for important covariates included in the model become less reliable representations of the true constructs, measured covariates become proxies for the true confounders (Rudolph & Stuart, 2016). Consequently, to sufficiently meet the strong ignorability assumption, covariates included in the logistic regression model should be valid and reliable representations of the latent constructs driving student participation.

In simple bivariate analyses, the relationship between two variables measured with error will predictably be attenuated (Cohen, Cohen, West, & Aiken, 2003; Pedhauzer, 1997). That is, because the shared variance between the two variables is decreased due to decreased measurement precision, the relationship between the two variables is underestimated. However, when multiple error-prone predictor variables are included in a model, the regression coefficients will be unsystematically biased (i.e., either attenuated or augmented) and the standard errors will be biased upward (Cohen et al., 2003; Pedhauzer, 1997). Moreover, semipartial correlations are also biased when predictors are measured with error (Liu, 1988; Pedhauzer, 1997). Finally, the R squared estimate - or the total amount of variance in the outcome explained by predictors - decreases when predictors are measured with error (Pedhauzer, 1997).
Similar to multiple regression models, measurement error is problematic when employing logistic regression as well. With one predictor (e.g., covariate), the regression coefficient estimate will be attenuated (i.e., biased downward). However, with more than one correlated covariate, the influence of measurement error is once again more complex (Cohen et al., 2003; Pedhauzer, 1997).

Intercorrelations among covariates included in the propensity score estimation model make it difficult to predict how individual coefficients are impacted by covariate measurement error. For example, if a researcher includes scores from two personality subscales and two motivation subscales as covariates in a propensity score estimation model, scores from all four scales are likely imperfect representations of the latent constructs, and each of the four scales is likely related (to some extent) to the other three scales included in the model. Moreover, it would be difficult to predict in what way measurement error would impact each of the four regression coefficients estimated via the propensity score estimation model. However, without accounting for measurement error, the regression coefficient estimates are biased (similar to linear regression models; Clark, 1982). Consequently, the estimated relative contribution of covariates included in the propensity score estimation model might vary depending on the level of measurement error and the intercorrelations among covariates in the model.

Measurement error in the logistic regression model predictor variables (e.g., covariates) also affects parameter estimates at the omnibus level. For example, the presence of measurement error decreases the proportion of null deviance accounted for by covariates included in the logistic regression model (Osborne, 2014). As one may expect, statistical significance tests via a logistic regression model that includes error-
prone covariates are also affected by the presence of measurement error. Specifically, omnibus statistical significance tests are underpowered when error-prone covariates are included in the model, and efficiency consequently decreases (Spiegelman, Rosner, & Logan, 2000).

Significance tests for logistic regression coefficients also decrease in power as the level of measurement error increases (Stefanski & Carroll, 1985). Although decreases in power at both the omnibus and individual regression coefficient level are problematic for hypothesis testing, statistical significance is not of consequence when conducting propensity score analysis. That is, the logistic regression model is simply used for preprocessing the data and estimating propensity scores, or the predicted probability of self-selecting into the treatment condition.

When covariates are measured with error, the estimated predicted probability – or propensity – for treatment is also impacted. That is, the inclusion of error-prone covariates results in biased predicted probabilities (Stefanski & Carroll, 1985). However, the attenuation of propensity scores is bidirectional, as it affects extreme scores at both ends of the distribution (Stefanski & Carroll, 1985). That is, low probability cases are estimated to be higher than they truly are, and high probability cases are estimated to be lower than they truly are (Clark, 1982; Michalik & Tripathi, 1980; Stefanski & Carroll, 1985). This squeezing of the extremes towards the center is due to the estimated intercept being asymptotically biased (Clark, 1982; Michalik & Tripathi, 1980; Stefanski & Carroll, 1985). That is, the bias is towards the null, and the resulting model makes it appear as though the error-prone covariates are less able to differentiate between groups (e.g., treatment versus comparison group). Moreover, as the amount of measurement
error increases, so does the omnibus bidirectional attenuation toward the null (Clark, 1982).

When conducting propensity score matching, this bidirectional attenuation may be especially problematic for extreme cases on the high end of the distribution of propensity scores. Specifically, individuals with high propensity scores typically belong to the treatment group and are therefore included in the final matched datasets. Thus, attenuated scores at the high end of the distribution may result in cases with the highest propensity for treatment being matched to cases with slightly lower propensities for treatment. Consequently, when researchers evaluate the quality of matched treatment-comparison groups, it may appear between-group balance was achieved. However, due to the inclusion of error-prone covariates in the model, groups may simply appear balanced because cases with the highest true propensity for treatment were underestimated, and therefore seemed well matched.

Because of the influence measurement error has on parameter estimation, recommendations in the literature typically involve explicitly modeling measurement error rather than neglecting it (Rabe-Hesketh, Pickles, & Skrondal, 2003; Rosner, Spiegelman, & Willett, 1990; Stefanski & Carroll, 1985). However, best practices for handling logistic regression measurement error have not been widely implemented in the propensity score analysis domain. That is, researchers employing propensity score analyses still frequently fail to account for measurement error in the logistic regression model (Pan & Bai, 2015; Stuart, 2010).

One approach for dealing with error-prone covariates is to use estimation procedures that take asymptotic bias into account. Such estimation procedures are able to
take into account the impact of error-prone covariates and have been recommended in the literature (e.g., Spiegelman et al., 2000). For example, the results of a Monte Carlo study of nonparametric maximum likelihood estimation (NPMLE) found that unbiased parameter estimates were recovered even when error-prone covariates were included in the model (Rabe-Hesketh et al., 2003). However, one drawback is that unique estimation methods typically require large sample sizes, which may make employing these approaches infeasible in practice (Spiegelman et al., 2000).

In addition to employing superior estimation methods, researchers can also use generalized linear latent mixed models to account for measurement error when employing logistic regression (Rabe-Hesketh et al., 2003). That is, propensity for treatment can be modeled at the latent level, which allows researchers to account for the influence of measurement error. For example, the four covariates mentioned previously (i.e., two personality subscales and two motivation subscales) can be included as predictors in a latent generalized linear model to estimate propensity for belonging to the treatment group.

Although the use of generalized linear latent mixed models has been suggested in the logistic regression methodology literature, other within-person measurement error corrections for logistic regression models have also been explored (e.g., Rosner et al., 1992). However, both latent class and longitudinal statistical approaches for correcting measurement error are of little help when conducting single time point propensity score matching.

Despite frequent discussion of issues related to measurement error in the logistic regression literature, only a handful of studies in the propensity score matching literature
have discussed the impact systematic measurement error has on estimating treatment effects via propensity score analyses (e.g., Millimet, 2011; Rudolph & Stuart, 2016; Steiner et al., 2011). One such simulation study evaluated the effect of increasing levels of covariate measurement error on accurate treatment effect estimation (Steiner et al., 2011). As levels of covariate measurement error increased, balancing groups using the covariates was less likely to reduce bias (Steiner et al., 2011). Similarly, the results of another simulation study suggested that the true treatment effect could only be recovered if covariates were measured without error or if measurement error was included in the propensity score estimation model (Millimet, 2011). Consequently, treatment effect estimates became increasingly biased when there were high levels of covariate measurement error (Millimet, 2011).

In addition to measurement error inhibiting a researcher’s ability to balance on participants’ true propensities for treatment, measurement error might vary differentially by group, which adds an additional level of complexity. When covariate measurement error is differential by treatment status, correcting for measurement error becomes an involved process. In the Bayesian framework, a two-step approach for specifying conditional hierarchical models for dealing with measurement error has been suggested (Hong, Rudolph, & Stuart, 2016). However, these approaches presume a level of familiarity with Bayesian inference and require researchers to specify prior distributions that align with the distribution of measurement error (Hong et al., 2016). Consequently, this approach may be difficult to employ in practice, as the researcher must already have an idea of the distributional tendencies of the covariate score measurement error.
In the frequentist framework, models incorporating both measurement error and missing at random data have been proposed; however, the proposed functional weighting methods operate under the assumption that the researcher is able to specify the distribution of measurement error (McCaffrey, Lockwood, & Setodji, 2013). Additional work on estimating the measurement error distributions for both continuous and categorical variables has been conducted (e.g., determining the measurement error distribution either empirically or theoretically; Lockwood & McCaffrey, 2016). These methods include first estimating propensity scores using generalized linear models (e.g., logistic regression), and then correcting for measurement error using integral equations for either the observed or assumed measurement error distribution (Lockwood & McCaffrey, 2016). However, despite an uptick in the research evaluating measurement error in the context of propensity score analyses, strong surrogacy – or that the unreliable measure accurately approximates the latent construct in the model – is still for the most part assumed (Lockwood & McCaffrey, 2016). Moreover, the impact of measurement error on researchers’ conclusions about the quality of treatment-comparison matched pairs has not been discussed in the literature.

Current study. Because measurement error has not been studied heavily in the context of propensity score analysis (Rudolph & Stuart, 2016), additional research is warranted. However, in applied propensity score matching studies, a researcher never knows the true treatment effect. Therefore, a simulation study is deemed the most appropriate technique for the current study because the true values are known (i.e., the simulated treatment effect) and can be used to evaluate how well the simulated values are recovered (Burton et al., 2006; Hallgren, 2014). Specifically, the current simulation study
evaluates the numeric diagnostic quality of matches and the extent to which treatment effects are accurately recovered when unreliable covariates are included in the propensity score estimation model. Thus, the research questions for the current study were four-fold:

1. How do differing levels of measurement error (e.g., 10%, 20%, 30%) affect the quality of matches created when using common propensity score matching techniques (e.g., nearest neighbor matching within a caliper distance, optimal matching, Mahalanobis matching)?

2. How do differing levels of measurement error (e.g., 10%, 20%, 30%) affect the accuracy of the estimated treatment effect when using common propensity score analyses (e.g., nearest neighbor matching within a caliper distance, optimal matching, and Mahalanobis matching)?

3. How do different types of measurement error (i.e., measurement error that is similar across groups versus measurement error that is differential by group) affect the quality of matches created when using common propensity score matching techniques (e.g., nearest neighbor matching within a caliper distance, optimal matching, Mahalanobis matching)?

4. How do different types of measurement error (i.e., measurement error that is similar across groups versus measurement error that is differential by group) affect the accuracy of the estimated treatment effect when using common propensity score analyses (e.g., nearest neighbor matching within a caliper distance, optimal matching, and Mahalanobis matching)?
CHAPTER 3

Method

The goal of the current study was to evaluate both the quality of treatment-comparison group balance (Research Questions 1 and 3) and the accuracy of treatment effect estimates (Research Questions 2 and 4) via a propensity score matching simulation study. Specifically, the aim of the study was to empirically evaluate the performance of different propensity score matching techniques as the level of covariate measurement error (i.e., 10% measurement error to 60% measurement error) and the type of covariate measurement error (i.e., measurement error that is similar across groups versus measurement error that is systematically differential by group) were manipulated across conditions.

Simulation of Data

The current study was conducted using R freeware program version 3.4.1 (R Core Team, 2017). Data were generated for 1,000 simulees across 1,000 replications using the mvtnorm package (Genz et al., 2015). The code used for simulating measurement error that is the same between groups and for measurement error differing between the two groups may be found in Appendices A and B, respectively. The treatment-comparison group size ratio was set to 1:4 (treatment simulees to comparison simulees) and varied slightly across replications due to the probabilistic nature of assignment. The 1:4 ratio was selected because it is recommended in the propensity score matching literature that the comparison pool be at least three times the size of the treatment group (Bai, 2015). Moreover, the sample size of 1,000 simulees (200 treatment to 800 control) was selected because intervention research simulation studies frequently mirror sample sizes typical of
applied studies in the field (e.g., Tumlison, Sass, & Cano, 2014). Figure 8 displays the conceptual relationship among variables in the simulated data set (adapted from Harris et al., 2018), where T is the simulated treatment condition, Y is the outcome variable, and X1-X5 are covariates. Note the arrows going from L1 to X1 and from L2 to X2: these arrows indicate that L1 and L2 are error-free latent traits driving both self-selection into the program (W), performance on the outcome (Y), and also causing responses on the measured, error-prone covariates (X1 and X2). Thus, X1 and X2 are observed covariates measured with error. Finally, X3-X5 are error-free covariates that relate to both self-selected participation in the treatment (W) and the outcome (Y). The performance of two sets of covariates in the context of PSM were evaluated: a) a covariate set including X1, X2 (i.e., the two covariates measured with error) and X3-X5 (i.e., three covariates measured without error), versus b) a covariate set including L1, L2 (i.e., substituting the error-free versions of X1 and X2 in the covariate set), and X3-X5. Henceforth, propensity scores estimated using the error-prone covariate (i.e., X1-X5) set are referred to as “naïve propensity scores,” and propensity scores estimated using the error-free covariate set (i.e., L1, L2, and X3-X5) are referred to as “true propensity scores.” However, it is important to note that, although the term “true propensity score” is used in the current study, propensity scores estimated using the error-free covariate set were simply an estimate given the error-free set of covariates.

The data for the current study were generated in a series of steps. First, five error-free covariates were simulated. Second, the simulated covariates were used to assign each simulee a probability of treatment. Third, simulees were probabilistically assigned to either the treatment or comparison group. Fourth, the outcome measure was simulated via
a linear function of the covariates. Finally, measurement error was added to two of the simulated covariates (L1 and L2) to create the error-prone covariates via the addition of a random error term (see Appendices A and B for simulation code).

To simulate the covariates, a correlation matrix was first defined for the latent (or error-free) covariates (i.e., L1, L2, and X3-X5). The correlations among covariates were set to between 0.4 and 0.5 (see Table 1) to mimic typical relationships (disattenuated for measurement error) observed among non-cognitive and dispositional measures in the social sciences (Lester, Inman, & Bishop, 2014). The five error-free covariates were then each set to correlate 0.4 with a simulated latent propensity for selecting into treatment, which mirrors the moderate relationship between covariates and treatment used in previous studies (e.g., Austin et al., 2007).

Similar to past propensity score analysis simulation studies (e.g., Austin 2009b; 2010), simulee propensity for treatment was assigned probabilistically via a function of the set of covariates via the below equation:

\[
P(T=1) = \Phi (-0.94 + 0.11L1 + 0.15L2 + 0.14X3 + 0.13X4 + 0.16X5)
\]

(16)

where the probability of treatment (P(T=1)) is calculated via a normal probability density function of the error-free covariates. Appendix C includes the mathematical equations for how the coefficients in Equation 16 were computed. Note that the coefficients in the above equation were calculated based on the correlations between each covariate and latent propensity for treatment (simulated at 0.4 for all covariates) and the intercorrelations among the simulated covariates (Table 1). Simulees were then assigned to the treatment group via probabilistic assignment. That is, for each case, a random draw via a random number generator was taken (ranging from 0-1), and cases for which the
simulated propensity value exceeded the random draw value were assigned to the
treatment group. Consequently, across all 1,000 simulees, there was a higher probability
that simulees with high propensity scores were assigned to the treatment group than to
the comparison group.

To simulate X1 and X2 (i.e., covariates measured with error), an error term was
added to the values of L1 and L2. Specifically, two types of measurement error were
imposed to create the two observed covariates (X1 and X2): measurement error that was
similar across groups (referred to hereafter as the “same measurement error conditions”),
and measurement error that was systematically differential by group (referred to hereafter
as the “different measurement error conditions”). The X1 and X2 error terms in the same
measurement error conditions was simulated to have a mean of zero and a standard
deviation that varied across the six simulated conditions so that the percent of
measurement error associated with X1 and X2 increased from ten percent (in Condition
1) to sixty percent (in Condition 6). The standard deviation for the simulated
measurement error for each condition was set via the below formula:

\[ \text{sd}_c = \sqrt{(1 - r) \times (1/r)} \]  \hspace{1cm} (17)

where \( r \) equals the simulated reliability level for each condition. Note that the above
equation includes the calculation for the standard error of measurement (with a standard
deviation of 1, the standard deviation drops out of the equation) divided by the simulated
reliability level. Multiplying the calculated standard error of measurement by the inverse
of the simulated reliability was mathematically necessary because the variance for the
simulated variable increased as additional error variance was added. Conversely, the
above equation can be thought of as the square root of the error variance (i.e., 1-\( r \)) added
via the random error term out of the total variance (via multiplying by 1/r). Thus, this method of imposing measurement error ensured that the resulting proportion of error variance in the final error-prone covariates accurately reflected the intended level of unreliability in scores in the same measurement error conditions. As reliability decreased across simulated conditions, the error variance increased (see left half of Table 2). The different measurement error conditions were similarly simulated; however, the measurement error in each condition varied systematically by group (right half of Table 2). That is, in the different measurement error conditions, the percent of measurement error imposed on X1 and X2 only varied across conditions for the control group and not for the treatment group. Measurement error varied across conditions for the control group paralleled applied research situations in which the control group was administered a different – perhaps condensed – version of a non-cognitive form. For example, the Ten-Item Personality Inventory (TIPI) was developed as a condensed measure of the 44-item Big-Five Inventory, and the five subscale scores have coefficient alpha reliability estimates hovering around 0.4 (Gosling, Rentfrow, & Swann, 2003). Although the condensed TIPI measure has lower reliability than the long form (Gosling et al., 2003), using a condensed measure may allow researchers to gather covariate data from a larger number of possible matches. Thus, the different measurement error conditions mirrored research situations ranging from all respondents completing the full form of covariate measures (i.e., simulated reliability levels of 0.8 or 0.9) to the treatment group responding to the full form and the comparison group responding to a considerably condensed form (i.e., comparison group simulated reliability level of 0.4). Consequently, there were twelve simulated conditions included in the current study: two types of simulated
measurement error (same measurement error and different measurement error) crossed with six levels of measurement error (ranging from 10% unreliability to 60% unreliability).

After the five error-free covariates (i.e., L1, L2, and X3-X5) and the two error-prone covariates (i.e., X1 and X2) were simulated, they were then transformed to have a mean of 25 and a standard deviation of five. The outcome variable (Y) was then simulated as a function of the five transformed error-free covariates (L1, L2, X3, X4, and X5) and an additional error term (v, with a simulated mean of zero and standard deviation of 20) via the equation below:

\[ Y = 100 + 10T + 2L1 + 2L2 + 2X3 + 2X4 + 2X5 + v \quad (18) \]

The effect of treatment on the outcome for program participants (as indicated by the unstandardized path coefficient for the grouping variable, T) was simulated to be 10 points higher than the comparison group (or \( d = 0.27 \)), controlling for the covariates in the model. A small effect was simulated to test how well the true treatment effect could be recovered. Moreover, the unstandardized path coefficients for L1, L2, and X3 – X5 were simulated to be two, which is a two-unit increase in Y for every one-unit increase in the respective covariate, controlling for the other variables in the model. Note that the error-prone naïve covariates (X1 and X2) were not included in the outcome model (Equation 17). The naïve covariates were not included in the model because the simulated “true” latent variables (L1 and L2), though unobserved, were what determined both self-selection into the treatment and performance on the outcome measure.
**Validation Data Sets**

After the data were simulated, a validation data set was evaluated from one replication of each of the twelve simulated conditions (Burton et al., 2006; Hallgren, 2014). Descriptive statistics for each of the simulated variables, the amount of imposed measurement error, the number of simulees, and the relationship among variables were evaluated to ensure the data were correctly simulated. Moreover, descriptive statistic information for one replication of each of the twelve conditions was reported in the final results of the study. After the data were simulated, propensity score matching was conducted for each of the 1,000 replications and across the twelve simulated conditions using both the true and naïve propensity scores.

**Propensity Score Matching**

The MatchIt package in R (Ho, Imai, King, & Stuart, 2011) was used to conduct propensity score matching for each of the 1,000 replications across the twelve conditions of the simulation study. Because only two grouping levels were simulated (i.e., simulees were either in the treatment group or the comparison group), logistic regression was employed to estimate propensity scores used for nearest neighbor matching, nearest neighbor matching with a 0.2 caliper width, and optimal matching (Austin, 2009; Ho et al., 2011; Stuart, 2010). Thus, propensity scores indicated the probability of selecting into the treatment condition given the covariates in the model. Two sets of covariates (to estimate “true” and “naïve” propensity scores) were included in the model for each of the four matching methods: one including the error-free, latent covariates (i.e., L1 and L2) to estimate true propensity scores and the other including the two error-prone covariates (i.e., X1 and X2) to estimate naïve propensity scores. Simulated cases were matched
using both the error-free and the error-prone covariates across all levels of measurement error to compare the performance of matching methods using the same simulated cases within each replication. Note that X3-X5 were error-free covariates and included in both covariate sets.

Mahalanobis distance matching was also conducted via the MatchIt package in R using the same sets of covariates as the other matching methods (Ho et al., 2011). However, for Mahalanobis matching, both sets of covariates (to estimate true and naïve propensity scores) also included propensity scores estimated via logistic regression (Rosenbaum & Rubin, 1985). Thus, Mahalanobis distance matching included a vector of five covariates and an estimated propensity score (six values total) to create matches. All matches across the four matching techniques and two types of propensity scores (true versus naïve) were created using a one-to-one matching ratio. That is, one control simulee was matched to every treatment simulee. Moreover, matches were created without replacement: once control simulees were matched to treatment simulees, they were not matched to other treatment simulees.

**Treatment Effect Estimation**

Once propensity score matches were creating using each of the matching methods, the treatment effect was estimated using data from the final matched groups. When propensity score matched groups are not exactly matched on covariates (e.g., via exact matching), the two groups may be treated as independent in outcome analyses (Austin, 2011a; Ho et al., 2007). Consequently, the estimated treatment effect was calculated as the mean difference between the matched treatment and comparison groups.
The mean difference was then used to evaluate whether the simulated treatment effect \( T = 10 \) was accurately recovered.

**Criteria for Evaluating Research Questions**

After saving out relevant information from each replication of a simulation study, parameter estimates or statistics can be evaluated across simulated conditions. Specific criteria include the bias of estimated parameters, the accuracy of statistics or inferences made under certain conditions, and the coverage of statistics under specific circumstances. The performance of test statistics and parameter recovery can be evaluated by calculating the bias, standard error, the root mean squared error values, and confidence interval coverage. Because each measure provides unique information about the performance of the quantitative methods employed, it is important to select measures that best align with the research question (Feinberg & Rubright, 2016). To answer the four research questions posed in the current study, numeric diagnostic information evaluating group balance, the bias of treatment effect estimates, the standard error, the root mean squared error, and confidence interval coverage were evaluated.

**Group balance.** After creating matches, the quality of matches is typically diagnosed both visually and numerically (Bai, 2015; Caliendo & Kopeinig, 2008; Guo & Fraser, 2014; Harder et al., 2010; Stuart, 2010; Stuart & Rubin, 2008a). Visual diagnosis includes both univariate and multivariate aids (Ho et al., 2011). For example, Q-Q plots aid in the evaluating the distribution of groups on individual covariates, whereas jitter plots and histograms allow multivariate visual comparison of the distribution of propensity scores between groups (Ho et al., 2011; Stuart, 2010). Jitter plots for each of the validation data sets were saved and reported in the final results. However, the large
number of simulated replications in the current study made the visual evaluation of matches impractical across all 1,000 replications. Consequently, the quality of matches within each condition of the current simulation study were evaluated numerically and not visually.

Univariate and multivariate numeric approaches to diagnosing the quality of matches are recommended in the literature. One method of evaluating the quality of matches univariately is to calculate the standardized mean difference between groups on individual covariates after matching (Stuart, 2010; Stuart & Rubin, 2008a). Matched groups are considered balanced across the set of covariates if the standardized mean difference on each individual covariate falls below a specified benchmark of either 0.10 (Ho et al., 2007) or 0.2 (Austin, 2011a). Similarly, the percent of bias reduction (PBR) was evaluated. An 80% reduction in bias from before matching to after matching is viewed as sufficient in the PSM literature (Pan & Bai, 2015).

Multivariately, the quality of matches can be evaluated by calculating the standardized mean difference on the propensity scores. Similar to evaluating the balance on individual covariates, the standardized mean difference between groups on the propensity score should be low (e.g., below 0.2), indicating the two groups are balanced (Austin, 2011a; Ho et al., 2007; Stuart, 2010). Moreover, the ratio of the variance of the matched participant group divided by the variance of the comparison group should be close to one, indicating the two groups have similar distributions of propensity scores (Rubin, 2001).

In the current study, numeric diagnostic information including the univariate PBR, standardized mean difference, and the propensity score variance ratio were saved
following each of 1,000 replications for each of the twelve simulation conditions. Moreover, these numeric diagnostic results were saved following the creation of matches using each of the four matching methods included in the current study (nearest neighbor matching, nearest neighbor matching with a 0.2 caliper width, optimal matching, and Mahalanobis distance matching) using both the true and naïve propensity scores. Balance across replications was then evaluated according to the benchmarks suggested in the literature (e.g., Austin, 2011a; Ho et al., 2007; Stuart, 2010) to answer Research Question 1 and 3. To answer Research Question 2 and 4, the accuracy of the estimated treatment effect was evaluated via the bias, standard error, root mean squared error, and confidence interval coverage as criteria.

**Bias.** Bias is the degree to which estimated values differ from the true values simulated in the population for each condition. Because the idiosyncrasies of each simulated data set will deviate slightly from the values in the population, if unbiased, these deviations will cancel out. Thus, values near zero indicate that the simulated parameters were recovered well, whereas nonzero values indicate the degree to which estimated values deviate from the true values. Bias is calculated as follows:

\[
Bias = \frac{\sum_{i=1}^{n}(\hat{\theta}_i - \theta_{true})}{n}
\]

(19)

where \(n\) is the total number of simulated replications, \(\theta_{true}\) is the parameter as simulated, and \(\hat{\theta}_i\) is the parameter estimated in each replication (\(i\)) of the simulation study. Of particular interest in the current study was the degree to which true (simulated) treatment effects were recovered via propensity score matching using both the true and naïve propensity scores. Thus, the bias – or extent to which the mean estimated treatment effect deviated from the true treatment effect – was evaluated in the current study.
**Standard error.** The standard error (SE) indicates the extent to which the estimated values deviate from the average estimated value across replications. Thus, the smaller the standard error, the more precisely the values in the simulated population were recovered. The standard error is the standard deviation of the parameter estimates, and it is calculated as follows:

\[
SE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{\theta}_i - \bar{\theta})^2}{n-1}}
\]  

(20)

where \( n \) is the total number of simulated replications, \( \hat{\theta}_i \) is the parameter estimated in each replication (\( i \)) of the simulation study, and \( \bar{\theta} \) refers to the arithmetic mean of that estimated parameter. Note that the standard error in simulation studies is the deviation from the estimated parameter across replications and not an indicator of how much the estimated values deviate from the true value. The standard error of the estimated treatment effect was evaluated in the current study for each of the four propensity score matching technique using both the true and naïve propensity scores across the twelve simulated conditions.

**Root mean squared error.** The root mean squared error (RMSE) can be conceptualized as a combination of both bias and the standard error in Equation 20. Thus, when bias across replications averages to zero, the RMSE value will equal the standard error. The calculation for RMSE is as follows:

\[
RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{\theta}_i - \theta_{true})^2}{n-1}}
\]  

(21)

where \( n \) is the total number of simulated replications, \( \theta_{true} \) is the parameter as simulated, and \( \hat{\theta}_i \) is the parameter estimated in each replication (\( i \)) of the simulation study.
**Confidence interval coverage.** The confidence interval coverage indicates the proportion of times the true parameter falls within a confidence interval upon repeated sampling. The treatment effect confidence interval was calculated for each replication using the analytical formulas (i.e., taking the square root of the weighted sum of the within-group variance divided by the square root of the within-group sample size, and then multiplying by 1.96). Confidence interval coverage was evaluated to determine the proportion of times the true treatment effect fell within the constructed confidence interval. Because a 95% confidence interval was used, the treatment effect was expected to fall outside of the interval approximately 5% of the time (2.5% below the lower bound and 2.5% above the upper bound of the confidence interval) if the estimates were unbiased and the analytical standard errors were accurate. Thus, if the proportion of times the treatment effect fell outside the confidence interval deviated from the expected rate of 5%, the confidence interval was not an accurate representation of the estimated treatment effect upon repeated sampling.

**Summary.** In summary, Table 3 displays the evaluation criteria that were employed to answer each of the four research questions posed in the current study. Note that the numeric diagnostic evaluation of balance between treatment-comparison matched groups was used to answer Research Questions 1 and 3, whereas the estimated treatment effect evaluation criteria were used to answer Research Questions 2 and 4.
CHAPTER 4

Results

Prior to conducting the simulation study, validation data sets were created to ensure data were simulated correctly for each of the twelve conditions. To ensure the error-prone covariates were simulated correctly, the correlations between the error-free covariates (i.e., L1 and L2) and the error-prone covariates (i.e., X1 and X2) were evaluated. Note in Table 4 that the correlation between the error-free and error-prone covariates equals roughly the square root of the simulated reliability level for each condition. The correlation between the error-free and error-prone covariates equals the square root of reliability because the correlation between measures equals the proportion of true score variance squared. Descriptive information for each of the simulated validation data sets is displayed in Table 5 (same measurement error conditions) and Table 6 (different measurement error conditions). Note that the treatment group and control group have different average covariate scores across all simulated covariates and across all twelve conditions. Moreover, note that the group averages differ for each of the simulated covariates, indicating the data were simulated to represent realistic data situations in which propensity score matching is warranted. That is, prior to matching, the groups differ on average on important covariates related to self-selected participation in the treatment group.

Validation Data Sets

Figures D1a through D12a in Appendix D display the correlations among simulated variables, histograms, and scatterplots for each of the twelve validation data sets. Note that the correlations between the error-free covariates (L1 and L2) and the
error-prone covariates (X1 and X2) equal approximately the square root of the set reliability level in the systematic measurement error conditions (see Table 4 for a reference guide to the square root of the reliability across conditions). Moreover, the correlations between the error-free covariates (L1 and L2) and the error-prone covariates (X1 and X2) also equal the square root of the reliability level set for the treatment group and control group in the different measurement error conditions. Recall that in the different measurement error conditions, the treatment group reliability on X1 and X2 was simulated to equal 0.8, and the comparison group reliability level decreased incrementally by 0.1 in reliability from 0.9 in Condition 1 to 0.4 in Condition 6. Finally, in Appendix D, Figures D1b through D12b display jitter graphs of the distribution of propensity score matched groups for each condition using the true propensity scores (left side) and naïve propensity scores (right side). The jitter graphs were created after conducting nearest neighbor matching to ensure the process worked properly with the simulated data. Nearest neighbor matching was selected for creating the jitter graphs because all of the treatment group members were retained, and the plots were straightforward for screening the simulated data sets. Note that the two groups overlap in their distributions of both true and naïve propensity scores, indicating sufficient common support to conduct propensity score matching.

**Simulation Study Results**

Tables 7 and 8 display the average means, standard deviations, and the standard errors for the mean and standard deviations for all simulated variables across the six conditions for both same measurement error condition (Table 7) and the different measurement error condition (Table 8). Notice that the standard error for both the means
and standard deviations in both Table 7 and Table 8 are higher for the treatment group than for the comparison group. Because the treatment group averages around one fourth the number of simulated cases as the comparison group, we expect the standard error to be higher than the comparison group due to sampling error. Also notice that the standard deviation for the simulated error-prone covariates (i.e., X1 and X2) increases as the simulated measurement error increases. Recall that in Equation 17, the standard error of measurement is multiplied by one divided by reliability to ensure the correct proportion of variance is error variance. Consequently, the standard deviation for the simulated error-prone covariates was expected to increase as the simulated levels of unreliability increased.

Table 9 displays the average number of matched treatment-comparison cases by condition and matching method for the same measurement error conditions (top of table) and the different measurement error conditions (bottom of table). Notice that for nearest neighbor matching, optimal matching, and Mahalanobis distance matching, the number of matches equals the total number of simulated treatment cases across replication in each condition. Because nearest neighbor matching, optimal matching, and Mahalanobis distance matching retain all treatment cases (i.e., no treatment cases are excluded from the final treatment-comparison matched samples), the average number of cases matched across these matching methods is the same. However, when nearest neighbor matching with a caliper width of 0.2 standard deviations on the logit of the propensity score was employed, the average number of cases dropped from 200 matched treatment cases to approximately 175 matched treatment cases, regardless of the measurement error condition or the type of simulated measurement error.
Tables 10 through 13 display the average means and standard deviations for all simulated variables after creating matches using each of the four matching methods (nearest neighbor matching, nearest neighbor matching using a 0.2 caliper, optimal matching, and Mahalanobis distance matching, respectively) for the same measurement error conditions. The top half of the tables display values for matches created using true propensity scores (i.e., using the error-free covariate set of L1, L2, and X3-X5). Conversely, the bottom half of the tables display values for matches created using naïve propensity scores (i.e., cases matched using the error-prone covariates X1 and X2 in addition to X3-X5). Similarly, Tables 14 through 17 display the average means and standard deviations for all simulated variables after creating matching using each of the four matching methods (nearest neighbor matching, nearest neighbor matching using a 0.2 caliper, optimal matching, and Mahalanobis distance matching, respectively) for the different measurement error conditions. The top half of the tables again display values for matches created using true propensity scores (i.e., using the error-free covariate set of L1, L2, and X3-X5), and the bottom half of the tables display values for matches created using naïve propensity scores (i.e., using the error-prone covariates X1 and X2 in addition to X3-X5). Notice that the mean differences between the matched treatment and comparison groups on the simulated covariates in Tables 10 through 17 (i.e., across all matching methods and the two types of measurement error) are more similar than in the pre-matched data sets in Tables 7 and 8. Because the propensity score matching process produces qualitatively similar groups, the treatment and comparison groups were expected to be similar after matching. Also notice that the standard deviation of the propensity scores is lower in the high measurement error conditions than in the low
measurement error conditions when the naïve propensity score was used to create matches (bottom of Tables 10-17). This pattern illustrates a bidirectional attenuation of propensity scores towards the null. That is, high propensity scores were estimated to be lower than they truly were, and low propensity scores were estimated to be higher than they truly were, leading to a decrease in propensity score variance. The results of the four research questions posed in the current study follow and are organized by research question.

**Research Question 1: How do differing levels of covariate score measurement error (e.g., 10%, 20%, 30% unreliability) affect the quality of matches created when using common propensity score matching techniques (e.g., nearest neighbor matching within a caliper distance, optimal matching, Mahalanobis distance matching)?**

The first research question focused on the quality of matches after using common matching techniques in the presence of six levels of measurement error. Following best practices in the propensity score matching literature, the quality of matches was diagnosed numerically by calculating the percent in bias reduction (Pan & Bai, 2015), the standardized mean difference between propensity score matched treatment-comparison groups (Austin, 2011a; Stuart, 2010), and the variance ratio between groups’ propensity score distributions (Rubin, 2001).

Tables 18 through 21 display the percent in bias reduction (PBR), standardized mean difference ($d$), and variance ratio (Var Ratio) for each of the matching methods (nearest neighbor matching, nearest neighbor matching using a 0.2 caliper, optimal matching, and Mahalanobis distance matching, respectively) for the same measurement error conditions. Similarly, Tables 22 through 25 display the percent in bias reduction
standardized mean difference \((d)\), and variance ratio \((\text{Var} \ \text{Ratio})\) for each of the matching methods (nearest neighbor matching, nearest neighbor matching using a 0.2 caliper, optimal matching, and Mahalanobis distance matching, respectively) for the different measurement error conditions. The top of Tables 18-25 display diagnostic information by matching method and across conditions for when the true propensity score was used to create matches (i.e., the covariate set including error-free covariates \(L1, L2, \) and \(X3-X5\)), whereas the bottom of Tables 18-25 display the same information for when the naïve propensity score was used (i.e., the covariate set including error-prone covariates \(X1, X2, \) and \(X3-X5\)).

The percent in bias reduction (displayed as a proportion in the tables) met or exceeded the 80% improvement (or 0.8) recommended in the literature (Pan & Bai, 2015) for most of the twelve simulated conditions. Figures 9 and 10 display the average percent in bias reduction for \(L1\) (dark grey bars) and \(X1\) (light grey bars) by condition for each matching method for the same measurement error conditions (Figure 9) and for the different measurement error conditions (Figure 10). Notice the identical pattern was present for both the same measurement error and the different measurement error conditions. That is, as the amount of measurement error increased (i.e., as unreliability goes from 10% in Condition 1 to 60% in Condition 6), matches created using the naïve propensity score did not appropriately balance the true covariate (i.e., \(L1\)). Moreover, in Condition 6, the true covariate (\(L1\)) just met the 80% improvement in bias reduction (indicated via the red dotted line) when the naïve propensity score was used to create matches. However, the error-prone covariate (\(X1\)) appeared well balanced across all six conditions when the true or naïve propensity scores were used to create matches. Also
note that, of the four matching methods, Mahalanobis distance matching performed the worst at balancing groups on covariates, regardless of the type of propensity score used. Although not illustrated via figures (see Tables 18-25), the same pattern was present when comparing the error-free covariate L2 to the error-prone covariate X2. Moreover, X3-X5 were well-balanced across all conditions, regardless of whether the true or naïve propensity score was used to create matches (see Tables 18-25). This finding makes sense given only error-free permutations of X3-X5 were included in both the error-free and error-prone covariate set.

Figures 11 and 12 display the average standardized mean difference for L1 (dark grey bars) and X1 (light grey bars) across the four matching methods for the same measurement error conditions (Figure 11) and for the different measurement error conditions (Figure 12). Notice that the identical pattern was again present for both the same measurement error conditions and the different measurement error conditions. As the amount of measurement error increased across conditions, so did the average standardized mean difference between groups on the true covariate (i.e., L1) when the error-prone covariate (X1) was used to estimate naïve propensity scores. Moreover, if the 0.1 benchmark was used to diagnose balance (indicated via the bottom red dotted line; Austin, 2011a), L1 would be considered unbalanced in several measurement error conditions. Specifically, treatment-comparison matched groups were imbalanced using the 0.1 cutoff in the same measurement error conditions after nearest neighbor matching with a 0.2 caliper in Condition 6 (0.4 reliability), in Conditions 5 and 6 (0.5 and 0.4 reliability, respectively) after using nearest neighbor matching or optimal matching, and in all six conditions after using Mahalanobis matching. Moreover, the same pattern of
exceeding the 0.1 benchmark for each matching method was found for the different measurement error conditions (i.e., when measurement error differed systematically by treatment group). However, X1 again appeared to be well balanced across each of the conditions, regardless of whether the true or naïve propensity score was used for matching.

Again, note that, of the four matching methods, Mahalanobis distance matching performed the worst at balancing groups both when true and naïve propensity scores were used to create matches. Unlike the patterns found by condition for the percent in bias reduction and the standardized mean difference by condition, no such pattern was found for the variance ratio (see right side of Tables 18-25). That is, with the exception of Mahalanobis distance matching, which performed slightly worse than the other matching methods, the variance ratio was close to the benchmark of one (Rubin, 2001) across conditions for each matching method and type of measurement error.

Research Question 2: How do differing levels of measurement error (e.g., 10%, 20%, 30% unreliability) affect the accuracy of the estimated treatment effect when using common propensity score analyses (e.g., nearest neighbor matching within a caliper distance, optimal matching, and Mahalanobis distance matching)?

To evaluate how well propensity score matching techniques perform when covariate scores are measured with differing levels of measurement error, several indices were evaluated. Specifically, bias – or the average amount by which the estimated treatment effect differs from the simulated treatment effect – was first examined across the six measurement error conditions. Tables 26 and 27 display the average mean difference and bias across replications between matched treatment-comparison groups on
the outcome variable (Y) by condition (i.e., Condition 1 to Condition 6). Note that Table 26 displays the values for the same measurement error conditions and Table 27 displays the values for the different measurement error conditions. Moreover, the top half of each table displays these values for matches created using the true propensity scores (i.e., matched using the error-free covariate set including L1, L2, and X3-X5), whereas the bottom half of the table displays the values for matches created using the naïve propensity scores (i.e., matched using the error-prone covariate set including X1, X2, and X3-X5). Recall the true treatment effect was simulated to be 10 points (see Equation 17). Consequently, the bias values displayed in Tables 26 and 27 were calculated via Equation 19 by subtracting 10 from the estimated treatment effect (i.e., the mean difference between groups), and then averaging across 1,000 replications. Tables 26 and 27 also display the standard error (SE) of the treatment effect estimate, which conveys the average deviation of parameter estimates across simulations, and the root mean squared error (RMSE), which provides an index of the average deviation of treatment effect estimates from the simulated (true) treatment effect across simulations. Based on the results presented in Tables 26 and 27, there was no apparent pattern in the SE across levels of measurement error. However, RMSE did increase as the amount of bias in the estimated treatment effect increased.

To evaluate whether the amount of treatment effect bias significantly varied by simulated condition, a series of five ANOVAs were conducted (see Appendix E for SASS syntax). Note that for each of the five ANOVAs conducted in the current study, there were two levels of propensity scores (i.e., either true or naïve propensity scores), which were treated as a within-subjects factor. The type of propensity score employed
was treated as a within-subjects factor because the same simulated cases within each replication were matched twice: once using the true propensity score and once using the naïve propensity score.

Table 28 displays the results from one four-way 2x2x6x4 mixed ANOVA (top of Table 28), which included two levels of propensity scores (true versus naïve; treated as a within-subjects variable), two types of measurement error (same versus different measurement error), and six levels of measurement error (Condition 1 through Condition 6, ranging from 0.9 reliability to 0.4 reliability), and four levels of matching methods (nearest neighbor matching, nearest neighbor matching with a 0.2 caliper, optimal matching, and Mahalanobis distance matching). Consequently, the 2x2x6x4 ANOVA was conducted to evaluate whether there was a four-way interaction between the type of measurement error, type of propensity score used to create matches, level of measurement error, and the matching method employed.

The bottom of Table 28 also displays the results from four three-way 2x2x6 mixed ANOVAs (true versus naïve propensity scores, same versus different measurement error, and six levels of measurement error). Although unnecessary, given the non-significant four-way interaction, four 2x2x6 mixed ANOVAs were conducted following the 2x2x6x4 mixed ANOVA to separately evaluate the results from each of the four matching methods employed in the current study.

The four-way interaction between type of measurement error, type of propensity score used, the level of measurement error, and the type of matching method employed was not statistically or practically significant ($F(18, 47952)=1.17, p = 0.280, \eta^2 = 0$). And, although statistically significant, none of the remaining interactions were practically
significant, with the exception of the type of propensity score by level of measurement error interaction ($\eta^2 = 0.06$). This is not surprising, given treatment effect estimates tended to increase as the level of measurement error increased – irrespective of matching method employed – when matches were created using the naïve propensity score.

The results for the four three-way ANOVAs are also presented in Table 28. Note the three-way interaction among type of measurement error (same versus different), type of propensity score used (i.e., true versus naïve), the level of measurement error (i.e., Condition 1 through Condition 6) was not statistically significant for the nearest neighbor matching, nearest neighbor matching with a caliper, or optimal matching methods. The three-way interaction was, however, statistically significant for the bias in the estimated treatment effect when Mahalanobis distance matching was employed ($F(5,11988)=3.26$, $p = 0.006$, $\eta^2 = 0$). However, this three-way interaction was not practically significant, as the effect did not explain a substantial amount of variance in bias in the estimated treatment effect.

The two-way interaction between type of propensity score used to create matches (either true or naïve) and the level of measurement error (Condition 1 through Condition 6) for bias in the estimated treatment effect was statistically and practically significant (see Table 28). The percent of variance in the estimated treatment effect bias explained by the interaction between type of propensity score (true versus naïve) and level of measurement error ranged from 5% (when nearest neighbor with a 0.2 caliper was employed) to 7% (when optimal matching was employed). This interaction indicated that, as the level of measurement error increased, the amount of bias in the estimated treatment
effect increased when matching on naïve propensity scores and not when matching on true propensity scores.

Figures 13 and 14 display the average treatment effect and bias in the average treatment effect across the six measurement error conditions for matches created using the true propensity scores (dark grey bars) and the naïve propensity scores (light grey bars). Figure 13 displays the trend for the same measurement error conditions, whereas Figure 14 displays the trend for the different measurement error conditions. Notice that, for both types of measurement error (same versus different), the use of the true propensity score resulted in overestimating the treatment effect when three of the four matching methods were used (i.e., nearest neighbor matching, optimal matching, and Mahalanobis distance matching). Use of the naïve propensity score increased the amount of treatment effect overestimation when all four matching methods were employed. Moreover, as the amount of unreliability increased (i.e., from Condition 1 to Condition 6), the amount of overestimation in the treatment effect also increased. Also note that Mahalanobis distance matching resulted in the most biased treatment effect estimates out of the four matching methods, regardless of whether matches were created using the true or naïve propensity scores. Conversely, nearest neighbor matching within a 0.2 caliper width resulted in the least biased treatment effect estimates out of the four matching methods, regardless of whether matches were created using the true or naïve propensity scores.

Tables 29 and 30 display the percentage of times the true (simulated) treatment effect was excluded from the 95% confidence interval upper or lower bound. Table 29 displays the confidence interval coverage information for the same measurement error
conditions, and Table 30 displays the confidence interval coverage information for the different measurement error conditions. Note that, in both Tables 29 and 30, the true treatment effect never fell above the upper bound of the estimated treatment effect confidence interval. However, the true treatment effect did fall below the lower bound, indicating the treatment effect was consistently overestimated. Note that, when the true propensity score was used to create matches (top panel of Tables 29 and 30), the lower bound of the confidence interval excluded the true treatment effect at a lower rate than expected (i.e., less than 2.5% of the time) for three of the four matching methods.

The high rate of 95% confidence interval coverage might be due, in part, to the simulated level of common support and similarity among final matched pairs. That is, because four potential matches were simulated for every treatment case, it may have resulted in dependencies among high quality matched treatment-comparison pairs. Although previous research supports treating matched treatment-comparison groups as independent in outcome analyses (e.g., Austin, 2011a; Ho et al., 2007), other research suggests that it is important to account for the dependencies among matched pairs when covariates explain a moderate to high proportion of variance in the outcome (Austin, 2009c). However, there are no guidelines in the literature regarding what level of relationship between the covariates and the outcome constitutes treating the final matched groups as independent or dependent.

Once again, Mahalanobis distance matching performed the worst of the four matching methods when the true propensity score was used to create matches. When the naïve propensity score was used to create matches, the percentage of times the lower bound of the treatment effect estimate confidence interval excluded the true treatment
effect increased. Notably, the percentage of times the lower bound of the confidence interval excluded the true treatment effect increased as the level of measurement error increased (e.g., from 10% unreliability in Condition 1 to 60% unreliability in Condition 6). Moreover, Mahalanobis distance matching again performed the worst of the four matching methods, with the lower bound of the confidence interval excluding the true treatment effect 21-25% of the time in the highest measurement error conditions (i.e., Condition 6 with 60% unreliability, or a simulated reliability level of 0.4). The standard error for the treatment effect in each condition did not systematically vary with the level of simulated measurement error. Consequently, changes in the level of confidence interval coverage are attributed to bias in the estimated treatment effect. Figures 15 and 16 display these trends in the estimated treatment effect confidence interval coverage for the same measurement error conditions (Figure 15) and the different measurement error conditions (Figure 16).

Research Question 3: How do different types of covariate score measurement error (i.e., measurement error that is systematically similar across groups versus measurement error that is differential by group) affect the quality of matches created when using common propensity score matching techniques (e.g., nearest neighbor matching within a caliper distance, optimal matching, Mahalanobis distance matching)?

The third research question focused on whether the type of simulated measurement error impacted the quality of matches, again diagnosed numerically. Tables 18 through 21 display the percent in bias reduction (PBR), standardized mean difference ($d$), and variance ratio (Var Ratio) for each of the matching methods for the same
measurement error conditions. Tables 22 through 25 display the percent in bias reduction, standardized mean difference, and variance ratio for each of the matching methods for the different measurement error conditions. The difference in quality of matches appeared miniscule between the two types of measurement error. For example, the difference in quality of matches on L1 between nearest neighbor matching conditions (comparing Table 18 to Table 22) for Condition 6 indicated a less than one percent difference in the percent bias reduction (PBR) for the two types of measurement error (same versus different). Moreover, the standardized mean difference ($d$) also appears similar between the two types of measurement error across matching conditions. Finally, although the variance ratio was near one across all of the matching conditions and between the two types of measurement error, it was slightly higher for the conditions in which there was different simulated measurement error between the treatment and control group.

The trends in quality of matches between measurement conditions are again displayed in Figures 9-12. Figure 9 displays the average percent in bias reduction for L1 (dark grey bars) and X1 (light grey bars) by condition for each matching method in the same measurement error conditions. Figure 10 displays the average percent in bias reduction for L1 and X1 by condition for each matching method in the different measurement error conditions. Figures 11 and 12 display the standardized mean difference for L1 (dark grey bars) and X1 (light grey bars) for the same measurement error conditions (Figure 11) and for the different measurement error conditions (Figure 12). Notice that, in addition to the pattern of quality of matches being equivalent across conditions (as discussed in Research Question 1), there was also no noticeable change in the average quality of matches for the true covariate (L1; dark grey bars) and the error-
prone covariate (X1; light grey bars) based on the type of measurement error across matching methods. That is, the same level of quality matches was present for both the same measurement error and the different measurement error conditions. As the amount of measurement error increased (i.e., as unreliability goes from 10% in condition 1 to 60% in condition 6), matches created using the naïve propensity performed similarly poor at reducing bias for the true covariate (i.e., L1) for both types of simulated measurement error. Again, although not illustrated via figures, the same pattern was present when comparing the error-free covariate L2 to the error-prone covariate X2 (see Tables 18-25). Moreover, X3-X5 were again well-balanced for both types of measurement error, regardless of whether the true or naïve propensity score was used to create matches (see Tables 18-25).

**Research Question 4: How do different types of covariate score measurement error (i.e., measurement error that is systematically similar across groups versus measurement error that is differential by group) affect the accuracy of the estimated treatment effect when using common propensity score analyses (e.g., nearest neighbor matching within a caliper distance, optimal matching, and Mahalanobis distance matching)?**

To evaluate the fourth research question, the estimated treatment effect was again compared to the true (simulated) treatment effect to determine whether treatment effect estimates deviate from the simulated values across conditions. To evaluate whether the amount of treatment effect bias statistically varied by type of measurement error, the interaction between condition (i.e., level of measurement error; Condition 1 through Condition 6) and type of measurement error (i.e., same versus different measurement
error) was examined via the four 2x2x6 mixed ANOVAs. Again, the ANOVAs included two types of measurement error (same measurement error versus different measurement error) two types of propensity scores (true versus naïve propensity scores) and the six measurement error conditions. The results are displayed in Table 28. The interaction between the type of simulated measurement error (same versus different) and type of propensity score (true versus naïve) was not statistically significant for two of the four matching methods (nearest neighbor matching within a 0.2 caliper width and Mahalanobis distance matching). The interaction was statistically significant for nearest neighbor matching and optimal matching. However, the $\eta^2$ value indicated that essentially no variance in treatment effect bias was accounted for by the interaction between type of measurement error (same versus different) and type of propensity score used to create matches (true versus naïve).

Figures 13 and 14 again display the average treatment effect and bias in estimated treatment effect across the six measurement error conditions for matches created using the true propensity score and the naïve propensity score. Figure 13 displays the trend for the same measurement error conditions, whereas Figure 14 displays the trend for the different measurement error conditions. Notice that average treatment effect and level of bias were fairly equivalent between the two types of measurement error (comparing Figure 13 to Figure 14).

Tables 29 and 30 display the confidence interval coverage for the same measurement error conditions (Table 29) and the different measurement error conditions (Table 30). Note again that the upper bound of the treatment effect estimate confidence interval never excluded the true treatment effect in either type of simulated measurement error.
error. However, note that, when the naïve propensity score was used to create matches, the percentage of times the lower bound excluded the true treatment effect increased for each of the four matching methods. Note also that the percentage of times the lower bound of the treatment effect confidence interval excluded the true treatment effect increased most when Mahalanobis distance matching was employed. Moreover, the percentage of times the lower bound of the confidence interval excluded the true treatment effect following Mahalanobis distance matching differed between types of measurement error (i.e., same versus different). That is, the percentage was higher in the different measurement error conditions than in the same measurement error conditions (see bottom of Figure 15 versus bottom of Figure 16).
CHAPTER 5

Discussion

The influence of covariate measurement error on the performance of numeric balance diagnostics and treatment effect estimates was evaluated in the current study following propensity score matching. Several data scenarios were simulated to emulate covariate measurement error situations applied researchers may encounter in practice. Specifically, two types of measurement error were simulated: measurement error that was the same across both treatment and comparison groups (i.e., the “same” measurement error conditions), and measurement error that was systematically different across the treatment and comparison groups (i.e., the “different” measurement error conditions).

In total, twelve measurement error conditions were simulated (1,000 replications each) in the current study: two types of measurement error (same versus different measurement error) crossed with six levels of simulated measurement error. For the same measurement error conditions, the levels of simulated measurement error ranged from 0.1 unreliability (i.e., 0.9 reliability) in the Condition 1 to 0.6 unreliability (i.e., 0.4 reliability) in Condition 6. For the different measurement error conditions, the levels of simulated measurement error again ranged from 0.1 unreliability in Condition 1 to 0.6 unreliability in Condition 6. However, in the different measurement error conditions, the level of measurement error was manipulated across conditions only for the control group, and the level of measurement error for the treatment group remained the same (0.2 unreliability; conversely, 0.8 reliability).

Following the simulation of data for each of the twelve measurement error conditions, four common matching methods were employed: nearest neighbor matching,
nearest neighbor matching within a 0.2 caliper width, optimal matching, and Mahalanobis distance matching. Matching was conducted twice for each type of matching method: once using error-free covariates (i.e., balancing groups on their “true” propensity scores) and once using error-prone covariates (i.e., balancing groups on their “naïve” propensity scores).

Four research questions were investigated in the current study. The first research question pertained to the quality of matches after using common matching techniques (Austin, 2011a): How do differing levels of covariate measurement error (e.g., 10%, 20%, 30% unreliability) affect the quality of matches created when using common propensity score matching techniques (e.g., nearest neighbor matching within a caliper distance, optimal matching, Mahalanobis matching)? The quality of matches was diagnosed numerically via benchmarks in the propensity score matching literature (e.g., Austin, 2011a; Guo & Fraser, 2014; Caliendo & Kopeinig, 2008; Pattanayak, 2015; Stuart, 2010; Stuart & Rubin, 2008a). The numeric diagnosis of matches involved evaluation of the following: the standardized mean difference between propensity score matched treatment-comparison groups (Austin, 2011a; Stuart, 2010), the variance ratio between groups’ propensity score distributions (Rubin, 2001), and the percent in bias reduction from before matching to after matching (Pan & Bai, 2015).

Overall, when groups were matched using the true propensity scores, the final matched treatment-comparison groups were adequately balanced on the true confounding variables (i.e., L1 and L2). However, when groups were matched using the naïve propensity score, the final matched treatment-comparison groups were adequately balanced only on the error-prone covariates (i.e., naïve covariates, X1 and X2). That is,
the final naïve propensity score matched treatment-comparison groups appeared to be well-balanced on the error-prone covariates after evaluating the percent bias reduction, standardized mean differences, and variance ratios. However, the error-free covariates (i.e., true covariates, L1 and L2) were not well balanced – particularly in the high measurement error conditions - unless the error-free covariates were used to create matches. For example, the percent bias reduction (PBR) for the true covariates (L1 and L2) fell below the 80% benchmark recommended in the literature (Pan & Bai, 2015) when reliability fell below 0.7 and Mahalanobis distance matches were created using the naïve propensity scores. This lack of balance in the true covariates as the proportion of measurement error increased is likely due to the unaccounted for influence of measurement error. That is, as the level of measurement error increased, the error-prone covariates (X1 and X2) became worse proxies for the true (latent) covariates, and led to an inability to fully account for self-selection bias.

The second research question focused on the recovery of the simulated treatment effect following propensity score matching: How do differing levels of measurement error (e.g., 10%, 20%, 30% unreliability) affect the accuracy of the estimated treatment effect when using common propensity score analyses (e.g., nearest neighbor matching within a caliper distance, optimal matching, and Mahalanobis matching)? To evaluate how well propensity score matching techniques performed when covariate scores were measured with differing levels of measurement error, several indices were evaluated. Bias, or the average amount by which the estimated treatment effect differs from the simulated treatment effect, was evaluated across conditions. The standard error of the treatment effect estimate - which conveyed the average deviation of parameter estimates
across replications - was also evaluated. Additionally, the root mean squared error provided an index of the average deviation of treatment effect estimates from the simulated treatment effect across simulations. Finally, 95% confidence interval coverage of the estimated treatment effect was evaluated in the current study.

Across the four matching methods, both the estimated treatment effect and bias in the estimated treatment effect increased as the level of measurement error increased (e.g., from 10% to 60% unreliability; alternatively, from 0.9 to 0.4 reliability). The 95% confidence interval included the true treatment effect more than 95% of the time in the low measurement error conditions for three of the matching methods (nearest neighbor matching, nearest neighbor matching with a 0.2 caliper, and optimal matching). This high rate of confidence interval coverage when balancing on the true propensity scores may be due, in part, to the high quality of treatment-comparison group matches. That is, one reason for the high rate of coverage might be because the groups were well-matched on the true propensity scores (e.g., PBR of 1.0 for nearest neighbor matching with a 0.2 caliper). Consequently, there may have been dependencies among well-matched treatment-comparison pairs (Austin, 2009c).

However, when the naïve propensity score was used to create matches, the percent of times the true treatment effect fell below the lower bound of the confidence interval increased slightly (e.g., from 0% to 4% with nearest neighbor matching). Moreover, when Mahalanobis distance matching was used, the 95% confidence interval excluded the true treatment in over 20% of the replications. The substantial lack of coverage indicated that, because the true treatment effect fell below the lower bound of
the confidence interval, that the treatment effect tended to be overestimated across replications.

Recall from Research Question 1 that the final matched treatment-comparison groups appeared balanced on the covariates included in the estimation of the naïve propensity score (i.e., the covariate set including X1 and X2). However, the final matched groups were *imbalanced* on the simulated true covariates (i.e., L1 and L2). Consequently, the simulated bias associated with self-selected participation into treatment was not fully mitigated through the creation of naïve propensity score matched treatment-comparison groups.

Recall also that one assumption underlying propensity score analysis is the strong ignorability assumption (Rosenbaum & Rubin, 1983a). The strong ignorability assumption states that, for treatment assignment to be considered ignorable, all important covariates must be adequately balanced between groups. However, when measurement error is present, the treatment-comparison matched groups were not sufficiently balanced on the true covariates when matches were created using the naïve, error-prone covariates. Specifically, because high probability cases were estimated to be lower than they truly were and low probability cases were estimated to be higher than they truly were, matches were not created based on each cases “true” propensity for treatment. Consequently, the strong ignorability assumption was not met, and the error-prone covariates became worse proxies for the true (latent) covariates as the amount of measurement error increased. Thus, the groups were not adequately balanced on the true covariates that accounted for self-selection bias, resulting in an overestimation of the treatment effect (i.e., the effect of treatment was still conflated with self-selection bias). These findings mirror research
conducted by Steiner et al. (2011) and Millimet (2011), who also found that self-selection bias was not appropriately mitigated via propensity score analysis when error-prone covariates were included in the model.

Research Questions 1 and 2 focused on the effects of changes in the amount of measurement error on the quality of matches and accuracy of inferences when employing propensity score matching. Research Questions 3 and 4, on the other hand, focused on evaluating how type of measurement error affected the quality of matches and inferences gleaned when employing propensity score matching. Specifically, the third research question focused on whether the quality of matches – again diagnosed numerically – differed between the two types of simulated measurement error: How do different types of covariate score measurement error (i.e., measurement error that is systematically similar across groups versus measurement error that is differential by group) affect the quality of matches created when using common propensity score matching techniques (e.g., nearest neighbor matching within a caliper distance, optimal matching, Mahalanobis matching)?

Similar to the first research question, the numeric diagnosis of matches involved evaluation of the following: the percent in bias reduction from before matching to after matching (Pan & Bai, 2015), the standardized mean difference between propensity score matched treatment-comparison groups (Austin, 2011a; Stuart, 2010), and the variance ratio between groups’ propensity score distributions (Rubin, 2001).

After evaluating the PBR, standardized mean difference, and the variance ratio, the final matched treatment-comparison groups appeared similarly balanced on both the true and naïve propensity scores across the two types of simulated measurement error (i.e., same versus different measurement error). The quality of naïve covariate balance
(i.e., X1 and X2) for the treatment-comparison groups was similar regardless of the type of simulated measurement error (i.e., same versus different). Moreover, when groups were matched on the naïve propensity scores, the true covariates (i.e., L1 and L2) were equally imbalanced across the two types of simulated measurement error conditions. The final matched treatment-comparison groups again also appeared equally well-balanced after evaluating the variance ratio when groups were matched using both the true and naïve propensity scores.

Between the two types of measurement error (same versus different), the variance ratio was slightly higher – indicating a lower quality of matches – when measurement error was simulated to differ between the treatment and comparison groups. However, the difference in quality of matches based upon evaluation of the variance ratio makes sense given the way in which the data were simulated. Recall that the error variance added to the simulated covariates was multiplied by one divided by a set reliability level (see Equation 17). Consequently, we would expect the two groups to differ in their propensity score variance if they differed in their level of simulated measurement error. Specifically, as the amount of simulated error variance increased for error prone covariates for the comparison group, so did the overall variance of their propensity scores.

In applied situations, we might find the same pattern (i.e., variance ratio higher than one in different measurement error conditions). That is, if the comparison group covariate scores are more variant in the presence of measurement error, the comparsion group would also have high variance in their propensity score distribution. Consequently, as the matched treatment-comparison groups differ in their propensity score variance, the variance ratio deviates further from one.
Finally, the fourth research question focused on whether the type of covariate measurement error impacted the recovery of treatment effect estimates: *How do different types of covariate score measurement error (i.e., measurement error that is systematically similar across groups versus measurement error that is differential by group) affect the accuracy of the estimated treatment effect when using common propensity score analyses (e.g., nearest neighbor matching within a caliper distance, optimal matching, and Mahalanobis matching)?* Similar to how the second research question was evaluated, bias, the standard error of the treatment effect estimate, the root mean squared error, and 95% confidence interval coverage of the estimated treatment effect were evaluated.

The amount of bias in the estimated treatment effect did not systematically vary according to the type of simulated measurement error across any of the four matching methods. That is, across the four matching methods, both the estimated treatment effect and bias in the estimated treatment effect were approximately the same across each of the six corresponding measurement error conditions (e.g., comparing Condition 4 with the same measurement error to Condition 4 with different measurement error). Moreover, the similarity in bias and estimated treatment effects across the two measurement error conditions held when both the true and naïve propensity scores were used for matching.

The results of the fourth research question (treatment effect estimates by type of measurement error) are perhaps the most surprising, particularly because the highest measurement error condition for the comparison group (Condition 6) included a reliability level of 0.4. One might expect that the treatment effect would be recovered better in the different measurement error conditions than the same measurement error
conditions because the treatment group had a consistently high simulated reliability level. However, recall that previous research found extreme probabilities to be most impacted by measurement error (Stefanski & Carroll, 1985). Specifically, in the presence of measurement error, low probability cases were estimated to be higher than they truly were, and high probability cases were estimated to be lower than they truly were (Clark, 1982; Michalik & Tripathi, 1980; Stefanski & Carroll, 1985). In the context of propensity score matching, this attenuation in predicted probabilities at the extremes is most likely to impact the high end of the treatment group (i.e., those with high predicted probabilities) and the low end of the control group (i.e., those with low predicted probabilities). Moreover, cases with a low propensity for treatment (i.e., those with low predicted probabilities) may have had an increased likelihood of being selected and matched to treatment cases as the level of measurement error increased. Consequently, it is likely that the presence of measurement error for the comparison group resulted in treatment-comparison matched pairs that were imbalanced on the true covariates (i.e., L1 and L2).

Across the four matching methods employed in the current study, Mahalanobis distance matching systematically performed worse than the other matching methods in both attainment of balance and recovery of the treatment effect. Specifically, the final Mahalanobis distance matched treatment-comparison groups were less balanced than matches created via the other matching methods. Moreover, Mahalanobis distance matching also performed worse than the other matching methods at recovering the simulated treatment effect, regardless of whether the true or naïve propensity scores were used to balance groups. This finding was surprising because Mahalanobis distance matching was recently championed by one well-known researcher as superior to other
propensity score matching methods (King & Nielsen, 2016). Recall that, unlike matching methods that rely on logistic regression, Mahalanobis distance matching does not weight covariates according to how well they predict participation in treatment (coded 0/1). That is, comparison group cases are simply matched to the treatment cases they are closest to in multivariate space on the vector of covariates. Recall also that the covariates were not equally weighted in the equation predicting latent propensity for treatment (see Equation 16). Consequently, in the current simulation study, Mahalanobis distance matching weighted the covariates equally; resulting in imbalanced matched treatment-comparison groups.

Nearest neighbor matching within a 0.2 caliper width performed the best of all the matching methods in both attainment of balance and recovery of the treatment effect. Specifically, it performed well across the six levels of simulated measurement error, the two types of measurement error, and when using either the true or naïve propensity scores to balance groups. Numeric balance diagnostics indicated groups matched via nearest neighbor matching within a 0.2 caliper width were also the best matched among the four matching methods. For example, the percent bias reduction approached 100% in several of the simulated conditions. Moreover, nearest neighbor matching within a 0.2 caliper width also performed best at recovering the true (simulated) treatment effect. In fact, the highest measurement error condition (Condition 6) for nearest neighbor matching recovered the true treatment effect better than the lowest measurement error condition (Condition 1) for Mahalanobis distance matching.

Although the number of treatment-comparison matches retained in the final data set was not specifically a research question in the current study, nearest neighbor
matching using a caliper width was the only method that did not retain all treatment cases. Recall that, when nearest neighbor matching with a caliper distance is employed, treatment cases for whom there are not quality comparison group matches are excluded from the final matched data set. In the current study, the inclusion of a 0.2 caliper width resulted in approximately 25 treatment cases being excluded from the final matched data sets.

In practice, researchers should weigh the benefit of improvement in balance against the cost of excluding part of the original sample in the final matched data set. Specifically, if the goal is to make inferences regarding the effectiveness of treatment back to the original sample of treatment participants, then researchers should evaluate whether the final matched group is still representative of the original sample (Harris & Horst, 2016; Jacovidis et al., 2016).

**Limitations and Future Research**

There are several notable limitations to the current study. For example, it was not feasible to simulate all possible propensity score matching research scenarios in which researchers might encounter measurement error. Moreover, one difficulty inherent in simulation studies is that the conditions are contrived - by design - and it is difficult to fully emulate reality. Specifically, the types of simulated measurement error, the types of covariates included in the model, the propensity score analysis methods, treatment and comparison group sizes, and the treatment/comparison group ratios were all held constant.

In the current study, only two types of measurement error were simulated: measurement error simulated to be the same level between the two groups and
measurement error that was systematically different between groups. However, it is plausible that researchers may encounter other types of measurement error in applied settings, such as measurement error that varies systematically across levels of the covariates (e.g., due to mediation effects). For example, reading ability might be a source of construct-irrelevant variance (or error) that varies across levels of an academic achievement covariate. However, reading ability may also relate to the academic program treatment outcome. Consequently, both the level of measurement error and the magnitude of the bias in the treatment effect might vary systematically across levels of propensity for treatment. Future simulation studies can investigate the effect of different types of measurement error on the accuracy of estimated treatment effects and increase the number of error-prone covariates.

The levels of measurement error were not manipulated in the current study for the treatment group for the different measurement error conditions. Consequently, it is unclear how changes in the level of measurement error for the treatment group in the different measurement error conditions might impact the quality of matches and accuracy of treatment effect estimates. Based on the current results, it is plausible that the quality of matches would decrease, and estimated treatment effects would increase in bias as the level of measurement error increases. However, additional conditions in which the level of measurement error varied for the treatment group would need to be simulated to investigate.

Both the number and types of covariates used to estimate propensity scores were held constant in the current study. Specifically, only five covariates were included; however, researchers conducting applied propensity score matching studies frequently
include large sets of covariates with the goal of meeting the strong ignorability assumption (Pan & Bai, 2015). Moreover, only continuous covariates were included to estimate propensity scores, whereas in applied studies, researchers typically include both continuous and categorical variables (e.g., gender or ethnicity; Guo & Fraser, 2014; Pan & Bai, 2015). Consequently, future research studies should include conditions with differing – and larger – covariate sets that include both continuous and categorical variables.

The level of correlations among the covariates was also held constant across conditions; however, it is likely that researchers in practice would encounter situations in which the covariates are more or less related to one another. The correlations among covariates in the current study mirrored latent correlations among attitudinal and dispositional inventory scores encountered in the social sciences. However, it is plausible that subsets of the covariates could vary systematically in the strength and pattern of the relationships. For example, several measures of motivation may be collected as covariates for a propensity score matching study. Moreover, the motivation measures might be more related to one another than to the other covariates used to predict treatment participation (e.g., extraversion or agreeableness). Consequently, future simulation studies should include a variety of patterns of relationships among simulated covariates.

Only four matching methods – albeit popular methods – were included in the current study. For example, the current study employed nearest neighbor matching at only one caliper width (i.e., 0.2 standard deviations on the logit of the propensity score). However, other caliper widths might perform better in terms of the number of treatment
cases retained than the one employed in the current study when covariates are measured with error. In practice, researchers typically employ several methods and then evaluate balance; selecting the method that produces the best balance between the treatment-comparison matched groups (Guo & Fraser, 2014). Future studies might include different caliper widths, exact matching, and other matching algorithms (e.g., genetic matching). In the future, research might also include exact matching simulated treatment-comparison cases on important covariates (e.g., age or gender), and evaluate potential interactions between the types of matching methods and simulated covariate sets.

Finally, the number of simulated cases and the ratio of treatment group cases to comparison group cases were held constant across conditions. The ratio was held constant in order to isolate the effects that changes in the simulated measurement error had on the quality of matches and the estimated treatment effect. In the future, researchers should consider manipulating the ratio of simulated treatment cases to comparison group cases. Moreover, researchers might also consider manipulating the region of common support, or the proportion of quality comparison group matches that overlap in propensity score distribution with the treatment group.

Research on the influence of measurement error on the quality of matches and the accuracy of treatment effect estimates is still sparse in the propensity score matching literature (Rudolph & Stuart, 2016). Consequently, additional studies are needed to expand upon the research conducted in the current study. Although the current study focused solely on the performance of several propensity score matching techniques, many other propensity score methods exist. For example, researchers interested in estimating a treatment effect for an entire population may use the estimated propensity scores as
weights rather than to create matches (e.g., via inverse probability of treatment weighting). Similarly, generalized boosted models (GBM) have increased in popularity over the past ten years; however, additional research is needed to evaluate the influence measurement error has on the accuracy of GBM estimated treatment effects. Because GBMs employ statistical learning algorithms to model the complex relationships between covariates and self-selected participation in treatment, it is possible these methods will perform better than others at mitigating specific types of measurement error (e.g., measurement error that varies systematically across levels of a covariate). Conversely, it is possible that GBM estimates may perform worse than other propensity score analyses if they capitalize on chance idiosyncrasies in the data resulting from measurement error.

Researchers in the future might also evaluate the performance of different methods of estimating propensity for treatment. For example, propensity scores estimated at the latent level (e.g., using structural equation modeling or SEM; Guo & Fraser, 2014) might allow researchers to better account for measurement error than logistic regression models. If propensity scores are estimated via SEM, measurement error could be accounted for prior to creating matches. However, in order to estimate propensity for treatment at the latent level, researchers would require access to item-level data and it may necessitate a larger sample size.

**Implications for Applied Research**

As sure as researchers are to encounter selection bias, applied researchers will encounter covariate measurement error. The current study illustrated the impact varying levels of measurement error had on the accuracy of the inferences drawn following propensity score matching. Specifically, although final matched treatment-comparison
groups appeared to be balanced on covariates included in the model, treatment effect estimates were systematically biased upward. Although the treatment effect in the current study was biased upward, researchers in practice would obtain biased estimates aligning with whichever direction the confounding variables conflated outcome estimates. In the current study, this bias in treatment effect estimates occurred because the true confounding factors associated with self-selected participation in treatment were not fully mitigated in the presence of measurement error. Consequently, researchers should be mindful that, in the presence of measurement error, achieving visual and numeric balance between matched groups does not necessarily signify that the strong ignorability assumption has been met. That is, regardless of achieving balance on error-prone covariates, the treatment effects would still remain biased and conflated with the factors driving self-selection bias.

One way in which researchers can improve the accuracy of treatment effect estimates when using propensity score matching is to administer and include measures that result in reliable covariates scores. Applied researchers familiar with the 0.7 reliability benchmark implied by Nunnally (1978) may continue using this benchmark as a guideline when selecting covariates. However, researchers should also keep in mind that, for self-selection bias to be completely mitigated via propensity score matching, covariate scores must be perfectly reliable.

Applied researchers may opt to administer the long form of a measure if administering the long form results in more reliable covariate scores than a short measure. Researchers may also consider modeling propensity for treatment at the latent level to mitigate the influence of measurement error – particularly if researchers have
access to item-level covariate data. For example, to estimate propensity for treatment, a researcher could employ a full structural model to predict participation given the latent covariates (Guo & Fraser, 2014). However, researchers should keep in mind that estimating propensity for treatment using structural equation modeling would likely require a large sample. Moreover, as the number of covariates increases, so does the requisite size of the sample to yield stable parameter estimates.

The present study provided an illustration of how increased levels of measurement error can lead to an inability to adequately balance matched treatment-comparison groups. Moreover, the inability to sufficiently balance groups using error-prone covariates resulted in biased treatment effect estimates. However, in the present study, the accuracy of treatment effect estimates did not substantially vary depending on whether both groups’ covariate scores were measured with high reliability (e.g., same versus different measurement error conditions).

Researchers conducting large-scale studies should also weigh the benefits of using the long forms of a measure – possibly to collect data on reliable covariate scores - against the possible cost of a decrease in comparison group survey response rates. For example, potential comparison group members may opt not to respond to surveys collecting data on covariates if they perceive the surveys to be too long and tedious. If the creation of a quality comparison group depends, in part, on whether respondents are willing to complete a survey, researchers may need to weigh the potential increase in score reliability against a possible increase in the number of potential matches.
Conclusion

In social sciences research, measurement error is an inconvenient reality with which applied researchers must contend. Additional research is needed to evaluate the impact of measurement error when employing other methods of matching (e.g., exact matching). However, the current study illustrated several simulated scenarios researchers might encounter in practice. Similar to other statistical methods, measurement error impedes a researcher’s ability to draw accurate inferences about the true effect of an intervention on treatment participants. Consequently, it behooves applied researchers to understand the ways in which measurement error impacts the accuracy of treatment effect estimates. However, researchers aware of the implications of measurement error on applied propensity score matching studies can remain cognizant and adjust their study designs accordingly.
Table 1  
*Correlations Among Covariates Measured Without Error*

<table>
<thead>
<tr>
<th>Covariates</th>
<th>L1</th>
<th>L2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L2</td>
<td>0.58</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X3</td>
<td>0.42</td>
<td>0.56</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X4</td>
<td>0.49</td>
<td>0.44</td>
<td>0.59</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>X5</td>
<td>0.60</td>
<td>0.41</td>
<td>0.44</td>
<td>0.51</td>
<td>1.00</td>
</tr>
</tbody>
</table>

*Note.* As measurement error is imposed on covariates L1 and L2, the correlations between these covariates and other covariates in the model is attenuated.
Table 2  
**Simulated Conditions and Measurement Error Levels on X1 and X2**

<table>
<thead>
<tr>
<th></th>
<th>CTT Measurement Error</th>
<th>Differential by Group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Treatment &amp; Control</td>
<td>Treatment</td>
</tr>
<tr>
<td></td>
<td>% Error</td>
<td>Reliability</td>
</tr>
<tr>
<td>Condition 1</td>
<td>10</td>
<td>0.90</td>
</tr>
<tr>
<td>Condition 2</td>
<td>20</td>
<td>0.80</td>
</tr>
<tr>
<td>Condition 3</td>
<td>30</td>
<td>0.70</td>
</tr>
<tr>
<td>Condition 4</td>
<td>40</td>
<td>0.60</td>
</tr>
<tr>
<td>Condition 5</td>
<td>50</td>
<td>0.50</td>
</tr>
<tr>
<td>Condition 6</td>
<td>60</td>
<td>0.40</td>
</tr>
</tbody>
</table>

*Note.* For each of the six conditions, nearest neighbor matching, nearest neighbor matching with a 0.2 caliper, optimal matching, and Mahalanobis matching were conducted.
<table>
<thead>
<tr>
<th>Criteria Evaluated to Answer Each Research Question</th>
<th>RQ1</th>
<th>RQ2</th>
<th>RQ3</th>
<th>RQ4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Diagnosing Matches</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percent Bias Reduction</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardized Mean Difference</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variance Ratio</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td><strong>Estimated Treatment Effect</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bias</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Standard Error</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Root Mean Squared Error</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Confidence Interval Coverage</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>
Table 4

*Simulated Correlations Between Simulated Error-free and Error-prone Covariates*

<table>
<thead>
<tr>
<th></th>
<th>Same Measurement Error</th>
<th>Different Measurement Error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Treatment and Control Groups</td>
<td>Treatment Group</td>
</tr>
<tr>
<td>Reliability</td>
<td>Square Root of Reliability</td>
<td>L1_X1</td>
</tr>
<tr>
<td>0.90</td>
<td>0.95</td>
<td>0.95</td>
</tr>
<tr>
<td>0.80</td>
<td>0.89</td>
<td>0.90</td>
</tr>
<tr>
<td>0.70</td>
<td>0.84</td>
<td>0.84</td>
</tr>
<tr>
<td>0.60</td>
<td>0.77</td>
<td>0.76</td>
</tr>
<tr>
<td>0.50</td>
<td>0.71</td>
<td>0.69</td>
</tr>
<tr>
<td>0.40</td>
<td>0.63</td>
<td>0.66</td>
</tr>
</tbody>
</table>

*Note.* L1_X1 indicates the correlation between the first error-free simulated covariate (L1) and the corresponding error-prone covariate (X1). L2_X2 indicates the correlation between the second error-free simulated covariate (L2) and the corresponding error-prone covariate (X2).
<table>
<thead>
<tr>
<th>Condition</th>
<th>Treatment</th>
<th>Control</th>
<th>All</th>
<th>True Prop</th>
<th>Naïve Prop</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cond 1</td>
<td>188</td>
<td>812</td>
<td>1000</td>
<td>0.28</td>
<td>0.29</td>
</tr>
<tr>
<td>X1</td>
<td>26.98</td>
<td>26.09</td>
<td>24.63</td>
<td>4.68</td>
<td>4.92</td>
</tr>
<tr>
<td>X2</td>
<td>27.66</td>
<td>24.26</td>
<td>24.90</td>
<td>4.72</td>
<td>4.86</td>
</tr>
<tr>
<td>X3</td>
<td>4.72</td>
<td>5.20</td>
<td>5.01</td>
<td>4.87</td>
<td>4.86</td>
</tr>
<tr>
<td>X4</td>
<td>27.70</td>
<td>24.11</td>
<td>24.64</td>
<td>4.91</td>
<td>5.20</td>
</tr>
<tr>
<td>X5</td>
<td>27.73</td>
<td>24.18</td>
<td>24.84</td>
<td>4.78</td>
<td>5.19</td>
</tr>
<tr>
<td>Y</td>
<td>27.25</td>
<td>24.43</td>
<td>25.05</td>
<td>4.75</td>
<td>5.19</td>
</tr>
<tr>
<td>Cond 2</td>
<td>210</td>
<td>799</td>
<td>1000</td>
<td>0.31</td>
<td>0.31</td>
</tr>
<tr>
<td>X1</td>
<td>27.74</td>
<td>24.37</td>
<td>25.11</td>
<td>4.48</td>
<td>4.89</td>
</tr>
<tr>
<td>X2</td>
<td>27.74</td>
<td>24.35</td>
<td>25.04</td>
<td>4.68</td>
<td>4.89</td>
</tr>
<tr>
<td>X3</td>
<td>5.36</td>
<td>5.55</td>
<td>5.03</td>
<td>5.54</td>
<td>5.55</td>
</tr>
<tr>
<td>X4</td>
<td>28.13</td>
<td>24.40</td>
<td>25.04</td>
<td>4.71</td>
<td>5.01</td>
</tr>
<tr>
<td>X5</td>
<td>27.54</td>
<td>24.39</td>
<td>25.15</td>
<td>4.72</td>
<td>5.03</td>
</tr>
<tr>
<td>Cond 3</td>
<td>200</td>
<td>791</td>
<td>1000</td>
<td>0.31</td>
<td>0.31</td>
</tr>
<tr>
<td>X1</td>
<td>27.96</td>
<td>24.43</td>
<td>25.16</td>
<td>4.96</td>
<td>5.04</td>
</tr>
<tr>
<td>X2</td>
<td>28.15</td>
<td>24.47</td>
<td>25.24</td>
<td>5.07</td>
<td>5.22</td>
</tr>
<tr>
<td>X3</td>
<td>5.07</td>
<td>5.03</td>
<td>5.25</td>
<td>5.88</td>
<td>5.57</td>
</tr>
<tr>
<td>X4</td>
<td>28.02</td>
<td>24.44</td>
<td>25.24</td>
<td>4.98</td>
<td>6.13</td>
</tr>
<tr>
<td>X5</td>
<td>27.79</td>
<td>24.29</td>
<td>25.16</td>
<td>4.66</td>
<td>6.10</td>
</tr>
<tr>
<td>Cond 4</td>
<td>207</td>
<td>794</td>
<td>1000</td>
<td>0.31</td>
<td>0.31</td>
</tr>
<tr>
<td>X1</td>
<td>27.78</td>
<td>24.52</td>
<td>25.19</td>
<td>4.84</td>
<td>5.17</td>
</tr>
<tr>
<td>X2</td>
<td>28.05</td>
<td>24.84</td>
<td>25.24</td>
<td>4.80</td>
<td>5.22</td>
</tr>
<tr>
<td>X3</td>
<td>6.34</td>
<td>4.75</td>
<td>5.25</td>
<td>4.78</td>
<td>6.13</td>
</tr>
<tr>
<td>X4</td>
<td>28.51</td>
<td>24.76</td>
<td>25.24</td>
<td>4.78</td>
<td>6.10</td>
</tr>
<tr>
<td>X5</td>
<td>27.98</td>
<td>24.81</td>
<td>25.16</td>
<td>4.72</td>
<td>6.04</td>
</tr>
<tr>
<td>Cond 5</td>
<td>206</td>
<td>794</td>
<td>1000</td>
<td>0.28</td>
<td>0.28</td>
</tr>
<tr>
<td>X1</td>
<td>27.26</td>
<td>24.47</td>
<td>25.19</td>
<td>4.38</td>
<td>5.01</td>
</tr>
<tr>
<td>X2</td>
<td>27.21</td>
<td>24.29</td>
<td>25.47</td>
<td>4.66</td>
<td>5.01</td>
</tr>
<tr>
<td>X3</td>
<td>7.27</td>
<td>4.86</td>
<td>5.01</td>
<td>7.27</td>
<td>5.25</td>
</tr>
<tr>
<td>X4</td>
<td>26.93</td>
<td>24.38</td>
<td>25.31</td>
<td>6.90</td>
<td>6.40</td>
</tr>
<tr>
<td>X5</td>
<td>27.69</td>
<td>24.35</td>
<td>25.57</td>
<td>27.21</td>
<td>6.44</td>
</tr>
<tr>
<td>Cond 6</td>
<td>198</td>
<td>802</td>
<td>1000</td>
<td>0.31</td>
<td>0.31</td>
</tr>
<tr>
<td>X1</td>
<td>28.18</td>
<td>24.63</td>
<td>25.33</td>
<td>4.59</td>
<td>5.13</td>
</tr>
<tr>
<td>X2</td>
<td>28.19</td>
<td>24.46</td>
<td>25.20</td>
<td>4.94</td>
<td>5.12</td>
</tr>
<tr>
<td>X3</td>
<td>7.83</td>
<td>4.90</td>
<td>5.12</td>
<td>27.48</td>
<td>7.90</td>
</tr>
<tr>
<td>X4</td>
<td>8.03</td>
<td>24.16</td>
<td>24.95</td>
<td>28.39</td>
<td>7.90</td>
</tr>
<tr>
<td>X5</td>
<td>28.39</td>
<td>24.36</td>
<td>24.98</td>
<td>28.42</td>
<td>8.18</td>
</tr>
</tbody>
</table>

Note. M refers to the mean value and SD refers to the standard deviation. True Prop in the table refers to the true propensity score, which included the error-free covariates (L1 and L2) in the covariate set. Naïve Prop in the table refers to the naïve propensity score, which included the error-prone covariates (X1 and X2) in the covariate set. Cond indicates the corresponding measurement error condition.
Table 6

Descriptive Information for Simulated Validation Data Sets for Different Measurement Error Conditions

<table>
<thead>
<tr>
<th></th>
<th>L1</th>
<th>L2</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>Y</th>
<th>True Prop</th>
<th>Naïve Prop</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td>SD</td>
</tr>
<tr>
<td>Cond 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>189</td>
<td>27.09</td>
<td>4.97</td>
<td>27.60</td>
<td>4.62</td>
<td>27.40</td>
<td>5.53</td>
<td>27.46</td>
<td>5.15</td>
<td>27.85</td>
</tr>
<tr>
<td>Control</td>
<td>811</td>
<td>24.55</td>
<td>4.76</td>
<td>24.33</td>
<td>4.75</td>
<td>24.56</td>
<td>5.04</td>
<td>24.38</td>
<td>5.03</td>
<td>24.42</td>
</tr>
<tr>
<td>All</td>
<td>1000</td>
<td>25.03</td>
<td>4.90</td>
<td>24.95</td>
<td>4.89</td>
<td>25.09</td>
<td>5.25</td>
<td>24.96</td>
<td>5.19</td>
<td>25.07</td>
</tr>
<tr>
<td>Cond 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>197</td>
<td>26.81</td>
<td>4.97</td>
<td>26.48</td>
<td>4.90</td>
<td>27.02</td>
<td>5.32</td>
<td>26.59</td>
<td>5.30</td>
<td>27.54</td>
</tr>
<tr>
<td>Control</td>
<td>803</td>
<td>24.33</td>
<td>4.65</td>
<td>24.46</td>
<td>4.85</td>
<td>24.33</td>
<td>5.31</td>
<td>24.53</td>
<td>5.52</td>
<td>24.57</td>
</tr>
<tr>
<td>All</td>
<td>1000</td>
<td>24.82</td>
<td>4.82</td>
<td>24.86</td>
<td>4.92</td>
<td>24.86</td>
<td>5.42</td>
<td>24.94</td>
<td>5.54</td>
<td>25.16</td>
</tr>
<tr>
<td>Cond 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>204</td>
<td>27.85</td>
<td>4.77</td>
<td>27.61</td>
<td>4.72</td>
<td>27.71</td>
<td>5.08</td>
<td>27.74</td>
<td>5.09</td>
<td>27.19</td>
</tr>
<tr>
<td>Control</td>
<td>796</td>
<td>24.49</td>
<td>4.68</td>
<td>24.43</td>
<td>4.74</td>
<td>24.44</td>
<td>5.51</td>
<td>24.37</td>
<td>5.76</td>
<td>24.50</td>
</tr>
<tr>
<td>All</td>
<td>1000</td>
<td>25.18</td>
<td>4.89</td>
<td>25.08</td>
<td>4.90</td>
<td>25.11</td>
<td>5.58</td>
<td>25.05</td>
<td>5.79</td>
<td>25.05</td>
</tr>
<tr>
<td>Cond 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>213</td>
<td>27.34</td>
<td>4.49</td>
<td>27.55</td>
<td>4.45</td>
<td>27.61</td>
<td>5.04</td>
<td>27.76</td>
<td>5.33</td>
<td>27.39</td>
</tr>
<tr>
<td>Control</td>
<td>787</td>
<td>24.38</td>
<td>4.95</td>
<td>24.56</td>
<td>4.97</td>
<td>24.22</td>
<td>6.52</td>
<td>24.67</td>
<td>6.32</td>
<td>24.54</td>
</tr>
<tr>
<td>All</td>
<td>1000</td>
<td>25.01</td>
<td>5.00</td>
<td>25.20</td>
<td>5.01</td>
<td>24.95</td>
<td>6.39</td>
<td>25.33</td>
<td>6.25</td>
<td>25.15</td>
</tr>
<tr>
<td>Cond 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>222</td>
<td>27.50</td>
<td>4.37</td>
<td>26.94</td>
<td>4.64</td>
<td>27.59</td>
<td>5.13</td>
<td>27.09</td>
<td>5.27</td>
<td>26.98</td>
</tr>
<tr>
<td>Control</td>
<td>778</td>
<td>24.27</td>
<td>4.98</td>
<td>24.48</td>
<td>4.87</td>
<td>24.40</td>
<td>6.91</td>
<td>24.39</td>
<td>7.13</td>
<td>24.41</td>
</tr>
<tr>
<td>All</td>
<td>1000</td>
<td>24.99</td>
<td>5.03</td>
<td>25.03</td>
<td>4.92</td>
<td>25.11</td>
<td>6.68</td>
<td>24.99</td>
<td>6.85</td>
<td>24.98</td>
</tr>
<tr>
<td>Cond 6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>202</td>
<td>27.52</td>
<td>4.95</td>
<td>27.59</td>
<td>4.83</td>
<td>27.73</td>
<td>5.72</td>
<td>27.73</td>
<td>5.75</td>
<td>27.51</td>
</tr>
<tr>
<td>Control</td>
<td>798</td>
<td>24.39</td>
<td>5.03</td>
<td>24.32</td>
<td>4.88</td>
<td>24.68</td>
<td>7.93</td>
<td>24.20</td>
<td>7.78</td>
<td>24.43</td>
</tr>
<tr>
<td>All</td>
<td>1000</td>
<td>25.03</td>
<td>5.17</td>
<td>24.98</td>
<td>5.04</td>
<td>25.30</td>
<td>7.63</td>
<td>24.91</td>
<td>7.55</td>
<td>25.05</td>
</tr>
</tbody>
</table>

Note. M refers to the mean value and SD refers to the standard deviation. True Prop in the table refers to the true propensity score, which included the error-free covariates (L1 and L2) in the covariate set. Naïve Prop in the table refers to the naïve propensity score, which included the error-prone covariates (X1 and X2) in the covariate set. Cond indicates the corresponding measurement error condition.
Table 7
Descriptive Information for Simulated Variables by Condition for the Same Measurement Error Conditions

<table>
<thead>
<tr>
<th>Cond 1</th>
<th>Treatment</th>
<th>L1 M</th>
<th>SE</th>
<th>L2 M</th>
<th>SE</th>
<th>X1 M</th>
<th>SE</th>
<th>X2 M</th>
<th>SE</th>
<th>X3 M</th>
<th>SE</th>
<th>X4 M</th>
<th>SE</th>
<th>X5 M</th>
<th>SE</th>
<th>Y M</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Control</td>
<td>27.49</td>
<td>0.34</td>
<td>27.49</td>
<td>0.34</td>
<td>27.49</td>
<td>0.36</td>
<td>27.49</td>
<td>0.36</td>
<td>27.49</td>
<td>0.33</td>
<td>27.49</td>
<td>0.34</td>
<td>27.49</td>
<td>0.36</td>
<td>384.86</td>
<td>2.99</td>
</tr>
<tr>
<td>Cond 2</td>
<td>Treatment</td>
<td>27.50</td>
<td>0.34</td>
<td>27.50</td>
<td>0.33</td>
<td>27.50</td>
<td>0.38</td>
<td>27.50</td>
<td>0.37</td>
<td>27.49</td>
<td>0.33</td>
<td>27.49</td>
<td>0.32</td>
<td>27.50</td>
<td>0.34</td>
<td>384.92</td>
<td>2.85</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>24.38</td>
<td>0.17</td>
<td>24.38</td>
<td>0.18</td>
<td>24.37</td>
<td>0.20</td>
<td>24.37</td>
<td>0.17</td>
<td>24.38</td>
<td>0.17</td>
<td>24.38</td>
<td>0.17</td>
<td>24.38</td>
<td>0.17</td>
<td>343.76</td>
<td>1.49</td>
</tr>
<tr>
<td>Cond 3</td>
<td>Treatment</td>
<td>27.48</td>
<td>0.33</td>
<td>27.48</td>
<td>0.35</td>
<td>27.48</td>
<td>0.40</td>
<td>27.48</td>
<td>0.41</td>
<td>27.48</td>
<td>0.34</td>
<td>27.50</td>
<td>0.33</td>
<td>27.50</td>
<td>0.32</td>
<td>384.87</td>
<td>2.80</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>24.37</td>
<td>0.17</td>
<td>24.37</td>
<td>0.17</td>
<td>24.37</td>
<td>0.20</td>
<td>24.37</td>
<td>0.20</td>
<td>24.38</td>
<td>0.17</td>
<td>24.38</td>
<td>0.17</td>
<td>24.38</td>
<td>0.17</td>
<td>343.75</td>
<td>1.46</td>
</tr>
<tr>
<td>Cond 4</td>
<td>Treatment</td>
<td>27.49</td>
<td>0.34</td>
<td>27.49</td>
<td>0.33</td>
<td>27.49</td>
<td>0.45</td>
<td>27.48</td>
<td>0.44</td>
<td>27.50</td>
<td>0.33</td>
<td>27.49</td>
<td>0.34</td>
<td>27.50</td>
<td>0.34</td>
<td>385.03</td>
<td>2.84</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>24.38</td>
<td>0.18</td>
<td>24.38</td>
<td>0.23</td>
<td>24.38</td>
<td>0.22</td>
<td>24.38</td>
<td>0.17</td>
<td>24.38</td>
<td>0.17</td>
<td>24.38</td>
<td>0.17</td>
<td>24.38</td>
<td>0.17</td>
<td>343.75</td>
<td>1.49</td>
</tr>
<tr>
<td>Cond 5</td>
<td>Treatment</td>
<td>27.49</td>
<td>0.33</td>
<td>27.48</td>
<td>0.34</td>
<td>27.51</td>
<td>0.48</td>
<td>27.48</td>
<td>0.50</td>
<td>27.50</td>
<td>0.32</td>
<td>27.51</td>
<td>0.32</td>
<td>27.50</td>
<td>0.36</td>
<td>384.95</td>
<td>2.85</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>24.38</td>
<td>0.18</td>
<td>24.39</td>
<td>0.25</td>
<td>24.37</td>
<td>0.25</td>
<td>24.38</td>
<td>0.25</td>
<td>24.38</td>
<td>0.17</td>
<td>24.37</td>
<td>0.17</td>
<td>24.37</td>
<td>0.17</td>
<td>343.81</td>
<td>1.47</td>
</tr>
<tr>
<td>Cond 6</td>
<td>Treatment</td>
<td>27.49</td>
<td>0.33</td>
<td>27.48</td>
<td>0.33</td>
<td>27.53</td>
<td>0.53</td>
<td>27.48</td>
<td>0.53</td>
<td>27.49</td>
<td>0.34</td>
<td>27.49</td>
<td>0.33</td>
<td>27.50</td>
<td>0.34</td>
<td>384.91</td>
<td>2.83</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>24.38</td>
<td>0.17</td>
<td>24.38</td>
<td>0.28</td>
<td>24.37</td>
<td>0.28</td>
<td>24.39</td>
<td>0.17</td>
<td>24.38</td>
<td>0.18</td>
<td>24.38</td>
<td>0.17</td>
<td>24.38</td>
<td>0.17</td>
<td>343.81</td>
<td>1.51</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>L1 SD</th>
<th>L2 SD</th>
<th>X1 SD</th>
<th>X2 SD</th>
<th>X3 SD</th>
<th>X4 SD</th>
<th>X5 SD</th>
<th>Y SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cond 1</td>
<td>Treatment</td>
<td>4.74</td>
<td>0.23</td>
<td>4.73</td>
<td>0.24</td>
<td>5.03</td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>4.85</td>
<td>0.12</td>
<td>4.86</td>
<td>0.12</td>
<td>5.13</td>
<td>0.13</td>
</tr>
<tr>
<td>Cond 2</td>
<td>Treatment</td>
<td>4.74</td>
<td>0.24</td>
<td>4.74</td>
<td>0.24</td>
<td>5.36</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>4.87</td>
<td>0.15</td>
<td>4.87</td>
<td>0.14</td>
<td>5.47</td>
<td>0.14</td>
</tr>
<tr>
<td>Cond 3</td>
<td>Treatment</td>
<td>4.74</td>
<td>0.23</td>
<td>4.73</td>
<td>0.24</td>
<td>5.75</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>4.86</td>
<td>0.12</td>
<td>4.86</td>
<td>0.12</td>
<td>5.86</td>
<td>0.15</td>
</tr>
<tr>
<td>Cond 4</td>
<td>Treatment</td>
<td>4.73</td>
<td>0.24</td>
<td>4.72</td>
<td>0.24</td>
<td>6.25</td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>4.87</td>
<td>0.13</td>
<td>4.87</td>
<td>0.12</td>
<td>6.36</td>
<td>0.16</td>
</tr>
<tr>
<td>Cond 5</td>
<td>Treatment</td>
<td>4.73</td>
<td>0.24</td>
<td>4.73</td>
<td>0.24</td>
<td>6.89</td>
<td>0.35</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>4.86</td>
<td>0.12</td>
<td>4.86</td>
<td>0.12</td>
<td>6.97</td>
<td>0.17</td>
</tr>
<tr>
<td>Cond 6</td>
<td>Treatment</td>
<td>4.75</td>
<td>0.23</td>
<td>4.74</td>
<td>0.23</td>
<td>7.75</td>
<td>0.38</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>4.87</td>
<td>0.12</td>
<td>4.86</td>
<td>0.12</td>
<td>7.82</td>
<td>0.19</td>
</tr>
</tbody>
</table>

Note: M indicates the average mean, SD indicates the average standard deviation, and SE indicates the standard error. Cond indicates the condition.
| Condition | L1  | SE  | M   | SE  | M   | SE  | M   | SE  | M   | SE  | M   | SE  | M   | SE  | M   | SE  | Y   |
|-----------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| **Cond 1** |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| Treatment | 27.49 | 0.33 | 27.49 | 0.34 | 27.49 | 0.38 | 27.48 | 0.38 | 27.50 | 0.34 | 27.49 | 0.35 | 27.50 | 0.34 | 384.94 | 2.94 |
| Control   | 24.38 | 0.18 | 24.38 | 0.17 | 24.38 | 0.19 | 24.37 | 0.19 | 24.38 | 0.18 | 24.38 | 0.17 | 24.38 | 0.17 | 343.76 | 1.51 |
| **Cond 2** |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| Treatment | 27.49 | 0.33 | 27.50 | 0.34 | 27.48 | 0.38 | 27.50 | 0.38 | 27.50 | 0.33 | 27.48 | 0.33 | 27.48 | 0.33 | 384.86 | 2.96 |
| Control   | 24.38 | 0.17 | 24.38 | 0.17 | 24.38 | 0.19 | 24.38 | 0.20 | 24.38 | 0.17 | 24.38 | 0.17 | 24.38 | 0.17 | 343.75 | 1.47 |
| **Cond 3** |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| Treatment | 27.49 | 0.33 | 27.50 | 0.35 | 27.49 | 0.38 | 27.50 | 0.39 | 27.49 | 0.34 | 27.49 | 0.34 | 27.49 | 0.33 | 384.92 | 2.93 |
| Control   | 24.37 | 0.17 | 24.37 | 0.17 | 24.37 | 0.21 | 24.37 | 0.20 | 24.37 | 0.17 | 24.37 | 0.17 | 24.37 | 0.17 | 343.70 | 1.48 |
| **Cond 4** |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| Treatment | 27.49 | 0.32 | 27.49 | 0.34 | 27.49 | 0.36 | 27.49 | 0.37 | 27.50 | 0.34 | 27.48 | 0.34 | 27.48 | 0.33 | 384.80 | 2.89 |
| Control   | 24.38 | 0.17 | 24.38 | 0.17 | 24.38 | 0.21 | 24.38 | 0.23 | 24.38 | 0.17 | 24.38 | 0.17 | 24.38 | 0.17 | 343.81 | 1.53 |
| **Cond 5** |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| Treatment | 27.48 | 0.34 | 27.47 | 0.34 | 27.48 | 0.39 | 27.48 | 0.38 | 27.48 | 0.33 | 27.49 | 0.33 | 27.49 | 0.34 | 384.82 | 2.97 |
| Control   | 24.37 | 0.17 | 24.37 | 0.17 | 24.38 | 0.25 | 24.37 | 0.25 | 24.38 | 0.17 | 24.38 | 0.17 | 24.38 | 0.17 | 343.74 | 1.47 |
| **Cond 6** |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| Treatment | 27.50 | 0.33 | 27.49 | 0.33 | 27.50 | 0.37 | 27.49 | 0.37 | 27.49 | 0.32 | 27.49 | 0.33 | 27.51 | 0.33 | 384.95 | 2.76 |
| Control   | 24.38 | 0.17 | 24.38 | 0.17 | 24.37 | 0.27 | 24.38 | 0.28 | 24.38 | 0.17 | 24.38 | 0.17 | 24.37 | 0.17 | 343.81 | 1.44 |

**Note.** M indicates the average mean, SD indicates the average standard deviation, and SE indicates the standard error. Cond indicates the condition.
### Table 9

**Average Matched Sample Size by Matching Method and Measurement Error Condition**

<table>
<thead>
<tr>
<th></th>
<th>NN Matching</th>
<th>NN w/ 0.2 Caliper</th>
<th>Optimal Matching</th>
<th>Mahal Matching</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>True</td>
<td>Naïve</td>
<td>True</td>
<td>Naïve</td>
</tr>
<tr>
<td><strong>Same ME</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cond 1</td>
<td>200.18</td>
<td>200.18</td>
<td>175.35</td>
<td>175.51</td>
</tr>
<tr>
<td>Cond 2</td>
<td>200.21</td>
<td>200.21</td>
<td>175.51</td>
<td>175.92</td>
</tr>
<tr>
<td>Cond 3</td>
<td>200.34</td>
<td>200.34</td>
<td>175.79</td>
<td>176.09</td>
</tr>
<tr>
<td>Cond 4</td>
<td>200.50</td>
<td>200.50</td>
<td>176.03</td>
<td>176.50</td>
</tr>
<tr>
<td>Cond 5</td>
<td>200.07</td>
<td>200.07</td>
<td>175.53</td>
<td>176.14</td>
</tr>
<tr>
<td>Cond 6</td>
<td>200.27</td>
<td>200.27</td>
<td>175.61</td>
<td>176.39</td>
</tr>
<tr>
<td><strong>Different ME</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cond 1</td>
<td>199.14</td>
<td>199.14</td>
<td>174.71</td>
<td>174.48</td>
</tr>
<tr>
<td>Cond 2</td>
<td>199.96</td>
<td>199.96</td>
<td>175.48</td>
<td>175.56</td>
</tr>
<tr>
<td>Cond 3</td>
<td>199.83</td>
<td>199.83</td>
<td>175.23</td>
<td>175.91</td>
</tr>
<tr>
<td>Cond 4</td>
<td>200.04</td>
<td>200.04</td>
<td>175.47</td>
<td>176.68</td>
</tr>
<tr>
<td>Cond 5</td>
<td>199.53</td>
<td>199.53</td>
<td>175.03</td>
<td>176.17</td>
</tr>
<tr>
<td>Cond 6</td>
<td>199.87</td>
<td>199.87</td>
<td>175.47</td>
<td>176.85</td>
</tr>
</tbody>
</table>

**Note.** NN Matching indicates the nearest neighbor matching conditions, NN w/ 0.2 Caliper indicates the nearest neighbor matching condition including a caliper width of 0.2 standard deviations on the logit of the propensity score, Optimal Matching indicates the optimal matching conditions, and Mahal Matching indicates the Mahalanobis distance matching conditions. True in the table refers to the true propensity score, which included the error-free covariates (L1 and L2) in the covariate set. Naïve in the table refers to the naïve propensity score, which included the error-prone covariates (X1 and X2) in the covariate set. Same ME refers to the same measurement error conditions. Different ME refers to the different measurement error conditions. Cond indicates the corresponding measurement error condition.
Table 10

**Descriptive Information for Simulated Variables by Condition for the Same Measurement Error Conditions After Nearest Neighbor Matching Using the True and Naïve Propensity Scores**

<table>
<thead>
<tr>
<th></th>
<th>True</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cond</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1</td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>L2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prop</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cond 1</td>
<td>Treatment</td>
<td>27.48</td>
<td>4.74</td>
<td>27.49</td>
<td>4.73</td>
<td>27.49</td>
<td>5.03</td>
<td>27.49</td>
<td>5.02</td>
<td>27.49</td>
</tr>
<tr>
<td>Cond 2</td>
<td>Treatment</td>
<td>27.31</td>
<td>4.58</td>
<td>27.32</td>
<td>4.58</td>
<td>27.32</td>
<td>4.87</td>
<td>27.32</td>
<td>4.87</td>
<td>27.33</td>
</tr>
<tr>
<td>Cond 3</td>
<td>Treatment</td>
<td>27.50</td>
<td>4.74</td>
<td>27.50</td>
<td>4.74</td>
<td>27.50</td>
<td>5.26</td>
<td>27.50</td>
<td>5.36</td>
<td>27.50</td>
</tr>
<tr>
<td>Cond 4</td>
<td>Treatment</td>
<td>27.33</td>
<td>4.58</td>
<td>27.33</td>
<td>4.58</td>
<td>27.33</td>
<td>5.23</td>
<td>27.33</td>
<td>5.20</td>
<td>27.33</td>
</tr>
<tr>
<td>Cond 5</td>
<td>Treatment</td>
<td>27.48</td>
<td>4.74</td>
<td>27.48</td>
<td>4.74</td>
<td>27.48</td>
<td>5.75</td>
<td>27.48</td>
<td>5.75</td>
<td>27.48</td>
</tr>
<tr>
<td>Cond 6</td>
<td>Treatment</td>
<td>27.33</td>
<td>4.58</td>
<td>27.33</td>
<td>4.58</td>
<td>27.33</td>
<td>5.63</td>
<td>27.33</td>
<td>5.64</td>
<td>27.33</td>
</tr>
<tr>
<td>Naïve</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cond 1</td>
<td>Treatment</td>
<td>27.49</td>
<td>4.73</td>
<td>27.49</td>
<td>4.73</td>
<td>27.49</td>
<td>5.03</td>
<td>27.49</td>
<td>5.02</td>
<td>27.49</td>
</tr>
<tr>
<td>Cond 2</td>
<td>Treatment</td>
<td>27.34</td>
<td>4.58</td>
<td>27.34</td>
<td>4.58</td>
<td>27.34</td>
<td>4.87</td>
<td>27.34</td>
<td>4.87</td>
<td>27.34</td>
</tr>
<tr>
<td>Cond 3</td>
<td>Treatment</td>
<td>27.50</td>
<td>4.74</td>
<td>27.50</td>
<td>4.74</td>
<td>27.50</td>
<td>5.36</td>
<td>27.50</td>
<td>5.36</td>
<td>27.50</td>
</tr>
<tr>
<td>Cond 4</td>
<td>Treatment</td>
<td>27.34</td>
<td>4.58</td>
<td>27.34</td>
<td>4.58</td>
<td>27.34</td>
<td>4.87</td>
<td>27.34</td>
<td>4.87</td>
<td>27.34</td>
</tr>
<tr>
<td>Cond 5</td>
<td>Treatment</td>
<td>27.49</td>
<td>4.73</td>
<td>27.49</td>
<td>4.73</td>
<td>27.49</td>
<td>5.03</td>
<td>27.49</td>
<td>5.02</td>
<td>27.49</td>
</tr>
<tr>
<td>Cond 6</td>
<td>Treatment</td>
<td>27.34</td>
<td>4.58</td>
<td>27.34</td>
<td>4.58</td>
<td>27.34</td>
<td>4.87</td>
<td>27.34</td>
<td>4.87</td>
<td>27.34</td>
</tr>
</tbody>
</table>

*Note. M indicates the average mean value across 1,000 replications for each condition, and SD indicates the average standard deviation value across 1,000 replications for each condition. True in the table refers to the true propensity score, which included the error-free covariates (L1 and L2) in the covariate set. Naïve in the table refers to the naïve propensity score, which included the error-prone covariates (X1 and X2) in the covariate set. Cond indicates the corresponding measurement error condition.*
Table 11  
Descriptive Information for Simulated Variables by Condition for the Same Measurement Error Conditions After Nearest Neighbor Matching with a Caliper  
Using the True and Naïve Propensity Scores

<p>| | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L1 M</td>
<td>SD</td>
<td>L2 M</td>
<td>SD</td>
<td>X1 M</td>
<td>SD</td>
<td>X2 M</td>
<td>SD</td>
<td>X3 M</td>
<td>SD</td>
<td>X4 M</td>
</tr>
<tr>
<td>Cond 1</td>
<td>Treatment</td>
<td>26.79</td>
<td>4.43</td>
<td>26.80</td>
<td>4.42</td>
<td>26.79</td>
<td>4.74</td>
<td>26.80</td>
<td>4.73</td>
<td>26.80</td>
<td>4.42</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>26.78</td>
<td>4.43</td>
<td>26.79</td>
<td>4.44</td>
<td>26.78</td>
<td>4.73</td>
<td>26.79</td>
<td>4.74</td>
<td>26.80</td>
<td>4.43</td>
</tr>
<tr>
<td>Cond 2</td>
<td>Treatment</td>
<td>26.80</td>
<td>4.43</td>
<td>26.80</td>
<td>4.43</td>
<td>26.81</td>
<td>5.08</td>
<td>26.81</td>
<td>5.09</td>
<td>26.80</td>
<td>4.43</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>26.80</td>
<td>4.44</td>
<td>26.80</td>
<td>4.43</td>
<td>26.81</td>
<td>5.10</td>
<td>26.79</td>
<td>4.44</td>
<td>26.79</td>
<td>4.43</td>
</tr>
<tr>
<td>Cond 3</td>
<td>Treatment</td>
<td>26.80</td>
<td>4.43</td>
<td>26.79</td>
<td>4.42</td>
<td>26.79</td>
<td>5.49</td>
<td>26.80</td>
<td>4.44</td>
<td>26.81</td>
<td>4.43</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>26.79</td>
<td>4.43</td>
<td>26.78</td>
<td>4.42</td>
<td>26.79</td>
<td>5.50</td>
<td>26.78</td>
<td>5.51</td>
<td>26.80</td>
<td>4.43</td>
</tr>
<tr>
<td>Cond 4</td>
<td>Treatment</td>
<td>26.80</td>
<td>4.42</td>
<td>26.81</td>
<td>4.41</td>
<td>26.80</td>
<td>6.02</td>
<td>26.80</td>
<td>6.00</td>
<td>26.81</td>
<td>4.42</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>26.79</td>
<td>4.43</td>
<td>26.79</td>
<td>4.43</td>
<td>26.79</td>
<td>6.01</td>
<td>26.79</td>
<td>6.02</td>
<td>26.80</td>
<td>4.43</td>
</tr>
<tr>
<td>Cond 5</td>
<td>Treatment</td>
<td>26.80</td>
<td>4.42</td>
<td>26.80</td>
<td>4.43</td>
<td>26.82</td>
<td>6.69</td>
<td>26.80</td>
<td>6.67</td>
<td>26.81</td>
<td>4.43</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>26.80</td>
<td>4.43</td>
<td>26.79</td>
<td>4.44</td>
<td>26.80</td>
<td>6.68</td>
<td>26.78</td>
<td>6.69</td>
<td>26.81</td>
<td>4.42</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>26.76</td>
<td>4.44</td>
<td>26.78</td>
<td>4.43</td>
<td>26.76</td>
<td>7.55</td>
<td>26.76</td>
<td>7.55</td>
<td>26.79</td>
<td>4.43</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. $M$ indicates the average mean value across 1,000 replications for each condition, and SD indicates the average standard deviation value across 1,000 replications for each condition. True in the table refers to the true propensity score, which included the error-free covariates (L1 and L2) in the covariate set. Naïve in the table refers to the naïve propensity score, which included the error-prone covariates (X1 and X2) in the covariate set. Cond indicates the corresponding measurement error condition.
<table>
<thead>
<tr>
<th>Treatment</th>
<th>L1</th>
<th>SD</th>
<th>L2</th>
<th>SD</th>
<th>X1</th>
<th>SD</th>
<th>X2</th>
<th>SD</th>
<th>X3</th>
<th>SD</th>
<th>X4</th>
<th>SD</th>
<th>X5</th>
<th>SD</th>
<th>Y</th>
<th>Prop</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cond 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>27.48</td>
<td>4.74</td>
<td>27.49</td>
<td>4.73</td>
<td>27.49</td>
<td>5.03</td>
<td>27.49</td>
<td>5.02</td>
<td>27.49</td>
<td>4.73</td>
<td>27.49</td>
<td>4.73</td>
<td>384.86</td>
<td>40.64</td>
<td>0.28</td>
<td>0.15</td>
</tr>
<tr>
<td>Control</td>
<td>27.33</td>
<td>4.59</td>
<td>27.34</td>
<td>4.58</td>
<td>27.33</td>
<td>4.89</td>
<td>27.33</td>
<td>4.88</td>
<td>27.33</td>
<td>4.58</td>
<td>27.33</td>
<td>4.58</td>
<td>373.42</td>
<td>38.86</td>
<td>0.27</td>
<td>0.13</td>
</tr>
<tr>
<td>Cond 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>27.50</td>
<td>4.74</td>
<td>27.50</td>
<td>4.74</td>
<td>27.50</td>
<td>5.36</td>
<td>27.50</td>
<td>5.36</td>
<td>27.50</td>
<td>4.74</td>
<td>27.50</td>
<td>4.74</td>
<td>384.92</td>
<td>40.69</td>
<td>0.25</td>
<td>0.12</td>
</tr>
<tr>
<td>Control</td>
<td>27.35</td>
<td>4.60</td>
<td>27.35</td>
<td>4.60</td>
<td>27.34</td>
<td>5.24</td>
<td>27.35</td>
<td>5.22</td>
<td>27.35</td>
<td>4.59</td>
<td>27.35</td>
<td>4.58</td>
<td>373.54</td>
<td>38.98</td>
<td>0.25</td>
<td>0.12</td>
</tr>
<tr>
<td>Cond 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>27.48</td>
<td>4.74</td>
<td>27.48</td>
<td>4.73</td>
<td>27.48</td>
<td>5.75</td>
<td>27.48</td>
<td>5.75</td>
<td>27.48</td>
<td>4.74</td>
<td>27.48</td>
<td>4.74</td>
<td>384.87</td>
<td>40.69</td>
<td>0.26</td>
<td>0.15</td>
</tr>
<tr>
<td>Control</td>
<td>27.34</td>
<td>4.59</td>
<td>27.35</td>
<td>4.58</td>
<td>27.35</td>
<td>5.64</td>
<td>27.35</td>
<td>5.64</td>
<td>27.35</td>
<td>4.58</td>
<td>27.35</td>
<td>4.58</td>
<td>373.45</td>
<td>39.08</td>
<td>0.26</td>
<td>0.14</td>
</tr>
<tr>
<td>Cond 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>27.49</td>
<td>4.73</td>
<td>27.49</td>
<td>4.72</td>
<td>27.49</td>
<td>6.25</td>
<td>27.49</td>
<td>6.23</td>
<td>27.49</td>
<td>4.73</td>
<td>27.49</td>
<td>4.73</td>
<td>385.03</td>
<td>40.69</td>
<td>0.28</td>
<td>0.15</td>
</tr>
<tr>
<td>Control</td>
<td>27.35</td>
<td>4.60</td>
<td>27.35</td>
<td>4.59</td>
<td>27.35</td>
<td>6.14</td>
<td>27.35</td>
<td>6.13</td>
<td>27.35</td>
<td>4.58</td>
<td>27.35</td>
<td>4.58</td>
<td>373.58</td>
<td>38.96</td>
<td>0.27</td>
<td>0.14</td>
</tr>
<tr>
<td>Cond 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>27.49</td>
<td>4.73</td>
<td>27.48</td>
<td>4.73</td>
<td>27.51</td>
<td>6.89</td>
<td>27.48</td>
<td>6.87</td>
<td>27.50</td>
<td>4.73</td>
<td>27.50</td>
<td>4.73</td>
<td>384.95</td>
<td>40.60</td>
<td>0.26</td>
<td>0.14</td>
</tr>
<tr>
<td>Control</td>
<td>27.35</td>
<td>4.59</td>
<td>27.34</td>
<td>4.60</td>
<td>27.33</td>
<td>6.78</td>
<td>27.34</td>
<td>6.79</td>
<td>27.35</td>
<td>4.58</td>
<td>27.35</td>
<td>4.58</td>
<td>373.56</td>
<td>38.97</td>
<td>0.26</td>
<td>0.13</td>
</tr>
<tr>
<td>Cond 6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>27.49</td>
<td>4.75</td>
<td>27.48</td>
<td>4.74</td>
<td>27.53</td>
<td>7.65</td>
<td>27.48</td>
<td>7.74</td>
<td>27.49</td>
<td>4.75</td>
<td>27.49</td>
<td>4.75</td>
<td>384.91</td>
<td>40.79</td>
<td>0.28</td>
<td>0.14</td>
</tr>
<tr>
<td>Control</td>
<td>27.35</td>
<td>4.59</td>
<td>27.34</td>
<td>4.58</td>
<td>27.37</td>
<td>7.04</td>
<td>27.32</td>
<td>7.65</td>
<td>27.35</td>
<td>4.60</td>
<td>27.35</td>
<td>4.60</td>
<td>373.54</td>
<td>38.99</td>
<td>0.28</td>
<td>0.13</td>
</tr>
<tr>
<td>Note:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 12: Descriptive Information for Simulated Variables by Condition for the Same Measurement Error Conditions After Optimal Matching Using the True and Naive Propensity Scores

Note: M indicates the average mean value across 1,000 replications for each condition, and SD indicates the average standard deviation value across 1,000 replications for each condition. True in the table refers to the true propensity score, which included the error-free covariates (L1 and L2) in the covariate set. Naive in the table refers to the naive propensity score, which included the error-prone covariates (X1 and X2) in the covariate set. Cond indicates the corresponding measurement error condition.
<table>
<thead>
<tr>
<th>Table 13</th>
<th>Descriptive Information for Simulated Variables by Condition for the Same Measurement Error Conditions After Mahalanobis Distance Matching Using the True and Naïve Propensity Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L1</td>
</tr>
<tr>
<td></td>
<td>M</td>
</tr>
<tr>
<td>Cond 1</td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>27.48</td>
</tr>
<tr>
<td>Control</td>
<td>27.08</td>
</tr>
<tr>
<td>Cond 2</td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>27.50</td>
</tr>
<tr>
<td>Control</td>
<td>27.09</td>
</tr>
<tr>
<td>Cond 3</td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>27.48</td>
</tr>
<tr>
<td>Control</td>
<td>27.08</td>
</tr>
<tr>
<td>Cond 4</td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>27.49</td>
</tr>
<tr>
<td>Control</td>
<td>27.10</td>
</tr>
<tr>
<td>Cond 5</td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>27.49</td>
</tr>
<tr>
<td>Control</td>
<td>27.08</td>
</tr>
<tr>
<td>Cond 6</td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>27.49</td>
</tr>
<tr>
<td>Control</td>
<td>27.09</td>
</tr>
</tbody>
</table>

| Note: M indicates the average mean value across 1,000 replications for each condition, and SD indicates the average standard deviation value across 1,000 replications for each condition. True in the table refers to the true propensity score, which included the error-free covariates (L1 and L2) in the covariate set. Naïve in the table refers to the naïve propensity score, which included the error-prone covariates (X1 and X2) in the covariate set. Cond indicates the corresponding measurement error condition. |
Table 14
Descriptive Information for Simulated Variables by Condition for the Different Measurement Error Conditions After Nearest Neighbor Matching
Using the True and Naïve Propensity Scores

<table>
<thead>
<tr>
<th></th>
<th>TRUE</th>
<th>L1</th>
<th>L2</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>Y</th>
<th>Prep</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>Cond 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>27.49</td>
<td>4.73</td>
<td>27.49</td>
<td>4.74</td>
<td>27.49</td>
<td>5.34</td>
<td>27.48</td>
<td>5.37</td>
<td>27.50</td>
<td>4.74</td>
</tr>
<tr>
<td>Control</td>
<td>27.33</td>
<td>4.58</td>
<td>27.34</td>
<td>4.58</td>
<td>27.34</td>
<td>4.87</td>
<td>27.34</td>
<td>4.87</td>
<td>27.34</td>
<td>4.58</td>
</tr>
<tr>
<td>Cond 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>27.49</td>
<td>4.74</td>
<td>27.50</td>
<td>4.74</td>
<td>27.48</td>
<td>5.37</td>
<td>27.50</td>
<td>5.35</td>
<td>27.50</td>
<td>4.73</td>
</tr>
<tr>
<td>Control</td>
<td>27.32</td>
<td>4.57</td>
<td>27.34</td>
<td>4.57</td>
<td>27.32</td>
<td>5.21</td>
<td>27.35</td>
<td>5.21</td>
<td>27.35</td>
<td>4.57</td>
</tr>
<tr>
<td>Cond 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>27.49</td>
<td>4.73</td>
<td>27.50</td>
<td>4.73</td>
<td>27.49</td>
<td>5.36</td>
<td>27.50</td>
<td>5.35</td>
<td>27.50</td>
<td>4.73</td>
</tr>
<tr>
<td>Control</td>
<td>27.34</td>
<td>4.58</td>
<td>27.34</td>
<td>4.56</td>
<td>27.33</td>
<td>5.62</td>
<td>27.34</td>
<td>5.61</td>
<td>27.33</td>
<td>4.57</td>
</tr>
<tr>
<td>Cond 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>27.49</td>
<td>4.74</td>
<td>27.49</td>
<td>4.74</td>
<td>27.49</td>
<td>5.35</td>
<td>27.49</td>
<td>5.37</td>
<td>27.50</td>
<td>4.75</td>
</tr>
<tr>
<td>Control</td>
<td>27.34</td>
<td>4.59</td>
<td>27.33</td>
<td>4.58</td>
<td>27.33</td>
<td>6.15</td>
<td>27.32</td>
<td>6.14</td>
<td>27.34</td>
<td>4.59</td>
</tr>
<tr>
<td>Cond 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>27.48</td>
<td>4.74</td>
<td>27.47</td>
<td>4.75</td>
<td>27.48</td>
<td>5.37</td>
<td>27.48</td>
<td>5.36</td>
<td>27.48</td>
<td>4.75</td>
</tr>
<tr>
<td>Control</td>
<td>27.33</td>
<td>4.60</td>
<td>27.32</td>
<td>4.58</td>
<td>27.33</td>
<td>6.78</td>
<td>27.31</td>
<td>6.78</td>
<td>27.32</td>
<td>4.58</td>
</tr>
<tr>
<td>Cond 6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>27.50</td>
<td>4.73</td>
<td>27.49</td>
<td>4.73</td>
<td>27.50</td>
<td>5.34</td>
<td>27.49</td>
<td>5.35</td>
<td>27.49</td>
<td>4.73</td>
</tr>
<tr>
<td>Control</td>
<td>27.35</td>
<td>4.58</td>
<td>27.34</td>
<td>4.58</td>
<td>27.36</td>
<td>7.64</td>
<td>27.34</td>
<td>7.64</td>
<td>27.35</td>
<td>4.60</td>
</tr>
</tbody>
</table>

Note. *M* indicates the average mean value across 1,000 replications for each condition, and *SD* indicates the average standard deviation value across 1,000 replications for each condition. True in the table refers to the true propensity score, which included the error-free covariates (L1 and L2) in the covariate set. Naïve in the table refers to the naïve propensity score, which included the error-prone covariates (X1 and X2) in the covariate set. Cond indicates the corresponding measurement error condition.
<table>
<thead>
<tr>
<th></th>
<th>L1 M</th>
<th>L1 SD</th>
<th>L2 M</th>
<th>L2 SD</th>
<th>X1 M</th>
<th>X1 SD</th>
<th>X2 M</th>
<th>X2 SD</th>
<th>X3 M</th>
<th>X3 SD</th>
<th>X4 M</th>
<th>X4 SD</th>
<th>X5 M</th>
<th>X5 SD</th>
<th>Y M</th>
<th>Y SD</th>
<th>Prop M</th>
<th>Prop SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cond 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>26.80</td>
<td>4.42</td>
<td>26.80</td>
<td>4.43</td>
<td>26.80</td>
<td>5.07</td>
<td>26.79</td>
<td>5.10</td>
<td>26.81</td>
<td>4.43</td>
<td>26.80</td>
<td>4.44</td>
<td>26.81</td>
<td>4.43</td>
<td>378.06</td>
<td>37.11</td>
<td>0.27</td>
<td>0.14</td>
</tr>
<tr>
<td>Control</td>
<td>26.79</td>
<td>4.42</td>
<td>26.80</td>
<td>4.43</td>
<td>26.79</td>
<td>4.73</td>
<td>26.79</td>
<td>4.74</td>
<td>26.79</td>
<td>4.43</td>
<td>26.79</td>
<td>4.43</td>
<td>26.80</td>
<td>4.44</td>
<td>367.78</td>
<td>37.19</td>
<td>0.27</td>
<td>0.14</td>
</tr>
<tr>
<td>Cond 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>26.80</td>
<td>4.42</td>
<td>26.81</td>
<td>4.42</td>
<td>26.79</td>
<td>5.10</td>
<td>26.80</td>
<td>5.06</td>
<td>26.81</td>
<td>4.41</td>
<td>26.79</td>
<td>4.43</td>
<td>26.79</td>
<td>4.43</td>
<td>377.98</td>
<td>36.97</td>
<td>0.21</td>
<td>0.08</td>
</tr>
<tr>
<td>Control</td>
<td>26.79</td>
<td>4.43</td>
<td>26.79</td>
<td>4.43</td>
<td>26.79</td>
<td>5.08</td>
<td>26.81</td>
<td>5.08</td>
<td>26.78</td>
<td>4.42</td>
<td>26.78</td>
<td>4.44</td>
<td>26.79</td>
<td>4.43</td>
<td>367.99</td>
<td>37.11</td>
<td>0.21</td>
<td>0.08</td>
</tr>
<tr>
<td>Cond 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>26.80</td>
<td>4.42</td>
<td>26.80</td>
<td>4.41</td>
<td>26.80</td>
<td>5.08</td>
<td>26.80</td>
<td>5.07</td>
<td>26.80</td>
<td>4.43</td>
<td>26.80</td>
<td>4.43</td>
<td>26.80</td>
<td>4.43</td>
<td>378.02</td>
<td>36.99</td>
<td>0.26</td>
<td>0.13</td>
</tr>
<tr>
<td>Cond 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>26.79</td>
<td>4.42</td>
<td>26.78</td>
<td>4.43</td>
<td>26.78</td>
<td>6.01</td>
<td>26.79</td>
<td>6.02</td>
<td>26.79</td>
<td>4.43</td>
<td>26.79</td>
<td>4.43</td>
<td>26.78</td>
<td>4.44</td>
<td>377.85</td>
<td>37.09</td>
<td>0.24</td>
<td>0.13</td>
</tr>
<tr>
<td>Control</td>
<td>26.78</td>
<td>4.42</td>
<td>26.78</td>
<td>4.43</td>
<td>26.78</td>
<td>6.01</td>
<td>26.79</td>
<td>6.02</td>
<td>26.79</td>
<td>4.43</td>
<td>26.79</td>
<td>4.43</td>
<td>26.78</td>
<td>4.44</td>
<td>367.90</td>
<td>37.17</td>
<td>0.24</td>
<td>0.13</td>
</tr>
<tr>
<td>Cond 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>26.79</td>
<td>4.42</td>
<td>26.78</td>
<td>4.44</td>
<td>26.79</td>
<td>5.09</td>
<td>26.78</td>
<td>5.09</td>
<td>26.79</td>
<td>4.43</td>
<td>26.79</td>
<td>4.43</td>
<td>26.79</td>
<td>4.43</td>
<td>377.84</td>
<td>37.12</td>
<td>0.23</td>
<td>0.10</td>
</tr>
<tr>
<td>Control</td>
<td>26.77</td>
<td>4.44</td>
<td>26.77</td>
<td>4.43</td>
<td>26.77</td>
<td>6.67</td>
<td>26.79</td>
<td>6.69</td>
<td>26.78</td>
<td>4.44</td>
<td>26.79</td>
<td>4.43</td>
<td>26.79</td>
<td>4.43</td>
<td>367.82</td>
<td>37.10</td>
<td>0.23</td>
<td>0.10</td>
</tr>
<tr>
<td>Cond 6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>26.81</td>
<td>4.41</td>
<td>26.79</td>
<td>4.44</td>
<td>26.81</td>
<td>7.51</td>
<td>26.79</td>
<td>7.57</td>
<td>26.80</td>
<td>4.44</td>
<td>26.79</td>
<td>4.43</td>
<td>26.81</td>
<td>4.42</td>
<td>368.07</td>
<td>37.03</td>
<td>0.23</td>
<td>0.12</td>
</tr>
</tbody>
</table>

### Note
- *M* indicates the average mean value across 1,000 replications for each condition, and *SD* indicates the average standard deviation value across 1,000 replications for each condition. True in the table refers to the true propensity score, which included the error-free covariates (L1 and L2) in the covariate set. Naïve in the table refers to the naïve propensity score, which included the error-prone covariates (X1 and X2) in the covariate set. Cond indicates the corresponding measurement error condition.
Table 16
Descriptive Information for Simulated Variables by Condition for the Different Measurement Error Conditions After Optimal Matching Using the True and Naïve Propensity Scores

<table>
<thead>
<tr>
<th></th>
<th>L1 M</th>
<th>L2 M</th>
<th>X1 M</th>
<th>X2 M</th>
<th>X3 M</th>
<th>X4 M</th>
<th>X5 M</th>
<th>Y M</th>
<th>Prop M</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SD</td>
<td>SD</td>
<td>SD</td>
<td>SD</td>
<td>SD</td>
<td>SD</td>
<td>SD</td>
<td>SD</td>
<td>SD</td>
</tr>
<tr>
<td>Cond 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>27.49</td>
<td>4.73</td>
<td>27.49</td>
<td>4.74</td>
<td>27.49</td>
<td>5.34</td>
<td>27.48</td>
<td>5.37</td>
<td>27.50</td>
</tr>
<tr>
<td>Control</td>
<td>27.35</td>
<td>4.59</td>
<td>27.35</td>
<td>4.59</td>
<td>27.35</td>
<td>4.88</td>
<td>27.35</td>
<td>4.88</td>
<td>27.36</td>
</tr>
<tr>
<td>Cond 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>27.49</td>
<td>4.74</td>
<td>27.50</td>
<td>4.74</td>
<td>27.48</td>
<td>5.37</td>
<td>27.50</td>
<td>5.35</td>
<td>27.50</td>
</tr>
<tr>
<td>Control</td>
<td>27.34</td>
<td>4.59</td>
<td>27.35</td>
<td>4.58</td>
<td>27.34</td>
<td>5.22</td>
<td>27.36</td>
<td>4.58</td>
<td>27.34</td>
</tr>
<tr>
<td>Cond 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>27.49</td>
<td>4.73</td>
<td>27.50</td>
<td>4.73</td>
<td>27.49</td>
<td>5.36</td>
<td>27.50</td>
<td>5.35</td>
<td>27.49</td>
</tr>
<tr>
<td>Control</td>
<td>27.35</td>
<td>4.59</td>
<td>27.35</td>
<td>4.58</td>
<td>27.33</td>
<td>5.63</td>
<td>27.35</td>
<td>5.62</td>
<td>27.34</td>
</tr>
<tr>
<td>Cond 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>27.49</td>
<td>4.74</td>
<td>27.49</td>
<td>4.74</td>
<td>27.49</td>
<td>5.35</td>
<td>27.49</td>
<td>5.37</td>
<td>27.49</td>
</tr>
<tr>
<td>Control</td>
<td>27.34</td>
<td>4.60</td>
<td>27.35</td>
<td>4.59</td>
<td>27.34</td>
<td>6.14</td>
<td>27.35</td>
<td>6.14</td>
<td>27.34</td>
</tr>
<tr>
<td>Cond 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>27.48</td>
<td>4.74</td>
<td>27.47</td>
<td>4.75</td>
<td>27.48</td>
<td>5.37</td>
<td>27.48</td>
<td>5.36</td>
<td>27.48</td>
</tr>
<tr>
<td>Control</td>
<td>27.34</td>
<td>4.60</td>
<td>27.34</td>
<td>4.59</td>
<td>27.33</td>
<td>6.79</td>
<td>27.33</td>
<td>6.79</td>
<td>27.34</td>
</tr>
<tr>
<td>Cond 6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>27.50</td>
<td>4.73</td>
<td>27.49</td>
<td>4.73</td>
<td>27.50</td>
<td>5.34</td>
<td>27.49</td>
<td>5.35</td>
<td>27.49</td>
</tr>
<tr>
<td>Control</td>
<td>27.36</td>
<td>4.59</td>
<td>27.35</td>
<td>4.59</td>
<td>27.37</td>
<td>7.63</td>
<td>27.37</td>
<td>7.63</td>
<td>27.36</td>
</tr>
</tbody>
</table>

Note: M indicates the average mean value across 1,000 replications for each condition, and SD indicates the average standard deviation value across 1,000 replications for each condition. True in the table refers to the true propensity score, which included the error-free covariates (L1 and L2) in the covariate set. Naïve in the table refers to the naïve propensity score, which included the error-prone covariates (X1 and X2) in the covariate set. Cond indicates the corresponding measurement error condition.
Table 17
Descriptive Information for Simulated Variables by Condition for the Different Measurement Error Conditions After Mahalanobis Distance Matching Using the True and Naïve Propensity Scores

<table>
<thead>
<tr>
<th></th>
<th>L1</th>
<th>L2</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>Y</th>
<th>Prop</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cond 1</td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td>M</td>
</tr>
<tr>
<td>Treatment</td>
<td>27.49</td>
<td>4.73</td>
<td>27.49</td>
<td>4.73</td>
<td>27.49</td>
<td>5.34</td>
<td>27.48</td>
<td>5.37</td>
<td>27.50</td>
</tr>
<tr>
<td>Control</td>
<td>27.08</td>
<td>4.38</td>
<td>27.09</td>
<td>4.37</td>
<td>27.08</td>
<td>4.68</td>
<td>27.10</td>
<td>4.37</td>
<td>27.09</td>
</tr>
<tr>
<td>Cond 2</td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td>M</td>
</tr>
<tr>
<td>Treatment</td>
<td>27.49</td>
<td>4.74</td>
<td>27.50</td>
<td>4.74</td>
<td>27.48</td>
<td>5.37</td>
<td>27.50</td>
<td>5.35</td>
<td>27.50</td>
</tr>
<tr>
<td>Control</td>
<td>27.08</td>
<td>4.38</td>
<td>27.09</td>
<td>4.37</td>
<td>27.09</td>
<td>5.04</td>
<td>27.09</td>
<td>4.36</td>
<td>27.08</td>
</tr>
<tr>
<td>Cond 3</td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td>M</td>
</tr>
<tr>
<td>Treatment</td>
<td>27.49</td>
<td>4.73</td>
<td>27.50</td>
<td>4.73</td>
<td>27.49</td>
<td>5.36</td>
<td>27.50</td>
<td>5.35</td>
<td>27.49</td>
</tr>
<tr>
<td>Control</td>
<td>27.09</td>
<td>4.37</td>
<td>27.09</td>
<td>4.36</td>
<td>27.08</td>
<td>5.45</td>
<td>27.10</td>
<td>5.45</td>
<td>27.09</td>
</tr>
<tr>
<td>Cond 4</td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td>M</td>
</tr>
<tr>
<td>Treatment</td>
<td>27.49</td>
<td>4.74</td>
<td>27.49</td>
<td>4.74</td>
<td>27.49</td>
<td>5.35</td>
<td>27.49</td>
<td>5.37</td>
<td>27.50</td>
</tr>
<tr>
<td>Control</td>
<td>27.08</td>
<td>4.39</td>
<td>27.09</td>
<td>4.38</td>
<td>27.08</td>
<td>5.98</td>
<td>27.09</td>
<td>5.98</td>
<td>27.09</td>
</tr>
<tr>
<td>Cond 5</td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td>M</td>
</tr>
<tr>
<td>Treatment</td>
<td>27.48</td>
<td>4.74</td>
<td>27.47</td>
<td>4.75</td>
<td>27.48</td>
<td>5.37</td>
<td>27.48</td>
<td>5.36</td>
<td>27.48</td>
</tr>
<tr>
<td>Control</td>
<td>27.08</td>
<td>4.38</td>
<td>27.07</td>
<td>4.38</td>
<td>27.05</td>
<td>6.65</td>
<td>27.05</td>
<td>6.63</td>
<td>27.06</td>
</tr>
<tr>
<td>Cond 6</td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td>M</td>
</tr>
<tr>
<td>Treatment</td>
<td>27.50</td>
<td>4.73</td>
<td>27.49</td>
<td>4.73</td>
<td>27.50</td>
<td>5.24</td>
<td>27.49</td>
<td>5.35</td>
<td>27.49</td>
</tr>
<tr>
<td>Control</td>
<td>27.10</td>
<td>4.37</td>
<td>27.08</td>
<td>4.37</td>
<td>27.08</td>
<td>7.53</td>
<td>27.09</td>
<td>7.52</td>
<td>27.09</td>
</tr>
</tbody>
</table>

Note: M indicates the average mean value across 1,000 replications for each condition, and SD indicates the average standard deviation value across 1,000 replications for each condition. True in the table refers to the true propensity score, which included the error-free covariates (L1 and L2) in the covariate set. Naïve in the table refers to the naïve propensity score, which included the error-prone covariates (X1 and X2) in the covariate set. Cond indicates the corresponding measurement error condition.
Table 18

<table>
<thead>
<tr>
<th>Cond</th>
<th>PBR</th>
<th>d</th>
<th>PBR</th>
<th>d</th>
<th>PBR</th>
<th>d</th>
<th>PBR</th>
<th>d</th>
<th>PBR</th>
<th>d</th>
<th>PBR</th>
<th>d</th>
<th>Var Ratio</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>0.94</td>
<td>0.04</td>
<td>0.95</td>
<td>0.04</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
<td>0.03</td>
<td>1.25</td>
<td>0.06</td>
</tr>
<tr>
<td>Cond 2</td>
<td>0.95</td>
<td>0.04</td>
<td>0.95</td>
<td>0.03</td>
<td>0.94</td>
<td>0.03</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
<td>0.03</td>
<td>1.10</td>
<td>0.02</td>
</tr>
<tr>
<td>Cond 3</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
<td>0.03</td>
<td>1.11</td>
<td>0.02</td>
</tr>
<tr>
<td>Cond 4</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
<td>0.02</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
<td>0.03</td>
<td>1.26</td>
<td>0.06</td>
</tr>
<tr>
<td>Cond 5</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
<td>0.03</td>
<td>0.94</td>
<td>0.03</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
<td>0.03</td>
<td>1.14</td>
<td>0.03</td>
</tr>
<tr>
<td>Cond 6</td>
<td>0.95</td>
<td>0.04</td>
<td>0.95</td>
<td>0.03</td>
<td>0.94</td>
<td>0.03</td>
<td>0.94</td>
<td>0.02</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
<td>0.03</td>
<td>1.10</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Naïve

| Cond 1| 0.92 | 0.05 | 0.91 | 0.06 | 0.95 | 0.03 | 0.95 | 0.03 | 0.95 | 0.03 | 0.95 | 0.04 | 1.23       | 0.06 |
| Cond 2| 0.90 | 0.07 | 0.88 | 0.08 | 0.95 | 0.03 | 0.95 | 0.03 | 0.95 | 0.03 | 0.95 | 0.03 | 1.12       | 0.02 |
| Cond 3| 0.87 | 0.09 | 0.84 | 0.10 | 0.95 | 0.02 | 0.96 | 0.03 | 0.96 | 0.03 | 0.96 | 0.03 | 1.12       | 0.03 |
| Cond 4| 0.85 | 0.10 | 0.81 | 0.12 | 0.96 | 0.02 | 0.96 | 0.02 | 0.95 | 0.03 | 0.96 | 0.03 | 1.20       | 0.05 |
| Cond 5| 0.82 | 0.12 | 0.78 | 0.14 | 0.95 | 0.02 | 0.95 | 0.02 | 0.95 | 0.03 | 0.96 | 0.03 | 1.14       | 0.03 |
| Cond 6| 0.80 | 0.13 | 0.76 | 0.16 | 0.95 | 0.02 | 0.96 | 0.02 | 0.96 | 0.03 | 0.96 | 0.03 | 1.10       | 0.02 |

Note. PBR indicates the average percent in bias reduction across 1,000 replications for each condition, d indicates the average standardized mean difference across 1,000 replications for each condition (control group subtracted from treatment group divided by the pooled standard deviation), and Var Ratio indicates the average variance ratio across 1,000 replications for each condition (treatment group's propensity score variance divided by the control group's propensity score variance). True in the table refers to the true propensity score, which included the error-free covariates (L1 and L2) in the covariate set. Naïve in the table refers to the naïve propensity score, which included the error-prone covariates (X1 and X2) in the covariate set. Cond indicates the corresponding measurement error condition.
Table 19  
Numeric Diagnostic Information for Same Measurement Error Conditions After Nearest Neighbor Matching with a Caliper Using the True and Naïve Propensity Scores

<table>
<thead>
<tr>
<th>Propensity Score</th>
<th>L1</th>
<th>L2</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>Var Ratio</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cond 1</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Cond 2</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Cond 3</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Cond 4</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Cond 5</td>
<td>1.00</td>
<td>0.00</td>
<td>0.99</td>
<td>0.00</td>
<td>0.99</td>
<td>0.00</td>
<td>0.99</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Cond 6</td>
<td>0.99</td>
<td>0.01</td>
<td>1.00</td>
<td>0.00</td>
<td>0.98</td>
<td>0.01</td>
<td>0.99</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Naïve</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cond 1</td>
<td>0.96</td>
<td>0.03</td>
<td>0.96</td>
<td>0.03</td>
<td>0.99</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Cond 2</td>
<td>0.94</td>
<td>0.04</td>
<td>0.92</td>
<td>0.05</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Cond 3</td>
<td>0.91</td>
<td>0.06</td>
<td>0.88</td>
<td>0.08</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Cond 4</td>
<td>0.88</td>
<td>0.08</td>
<td>0.85</td>
<td>0.10</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Cond 5</td>
<td>0.86</td>
<td>0.10</td>
<td>0.82</td>
<td>0.12</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Cond 6</td>
<td>0.83</td>
<td>0.11</td>
<td>0.79</td>
<td>0.14</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Note. PBR indicates the average percent in bias reduction across 1,000 replications for each condition, $d$ indicates the average standardized mean difference across 1,000 replications for each condition (control group subtracted from treatment group divided by the pooled standard deviation), and Var Ratio indicates the average variance ratio across 1,000 replications for each condition (treatment group's propensity score variance divided by the control group's propensity score variance). True in the table refers to the true propensity score, which included the error-free covariates (L1 and L2) in the covariate set. Naïve in the table refers to the naïve propensity score, which included the error-prone covariates (X1 and X2) in the covariate set. Cond indicates the corresponding measurement error condition.
Table 20

<table>
<thead>
<tr>
<th></th>
<th>L1</th>
<th>L2</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>Propensity Score Variance Ratio</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cond 1</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
<td>0.03</td>
<td>1.24</td>
</tr>
<tr>
<td>Cond 2</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
<td>0.03</td>
<td>0.96</td>
<td>0.03</td>
<td>1.06</td>
</tr>
<tr>
<td>Cond 3</td>
<td>0.95</td>
<td>0.03</td>
<td>0.96</td>
<td>0.03</td>
<td>0.95</td>
<td>0.02</td>
<td>0.96</td>
<td>0.03</td>
<td>1.09</td>
</tr>
<tr>
<td>Cond 4</td>
<td>0.96</td>
<td>0.03</td>
<td>0.96</td>
<td>0.03</td>
<td>0.95</td>
<td>0.03</td>
<td>0.96</td>
<td>0.03</td>
<td>1.26</td>
</tr>
<tr>
<td>Cond 5</td>
<td>0.96</td>
<td>0.03</td>
<td>0.95</td>
<td>0.03</td>
<td>0.94</td>
<td>0.03</td>
<td>0.96</td>
<td>0.03</td>
<td>1.13</td>
</tr>
<tr>
<td>Cond 6</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
<td>0.02</td>
<td>0.96</td>
<td>0.03</td>
<td>1.09</td>
</tr>
<tr>
<td>Naive</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cond 1</td>
<td>0.93</td>
<td>0.05</td>
<td>0.92</td>
<td>0.05</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
<td>0.03</td>
<td>1.22</td>
</tr>
<tr>
<td>Cond 2</td>
<td>0.90</td>
<td>0.06</td>
<td>0.88</td>
<td>0.08</td>
<td>0.96</td>
<td>0.03</td>
<td>0.96</td>
<td>0.03</td>
<td>1.11</td>
</tr>
<tr>
<td>Cond 3</td>
<td>0.87</td>
<td>0.09</td>
<td>0.85</td>
<td>0.10</td>
<td>0.96</td>
<td>0.02</td>
<td>0.96</td>
<td>0.03</td>
<td>1.11</td>
</tr>
<tr>
<td>Cond 4</td>
<td>0.85</td>
<td>0.10</td>
<td>0.82</td>
<td>0.12</td>
<td>0.96</td>
<td>0.02</td>
<td>0.96</td>
<td>0.03</td>
<td>1.19</td>
</tr>
<tr>
<td>Cond 5</td>
<td>0.83</td>
<td>0.11</td>
<td>0.79</td>
<td>0.14</td>
<td>0.96</td>
<td>0.02</td>
<td>0.96</td>
<td>0.03</td>
<td>1.14</td>
</tr>
<tr>
<td>Cond 6</td>
<td>0.81</td>
<td>0.13</td>
<td>0.76</td>
<td>0.16</td>
<td>0.96</td>
<td>0.02</td>
<td>0.96</td>
<td>0.03</td>
<td>1.09</td>
</tr>
</tbody>
</table>

**Note.** PBR indicates the average percent in bias reduction across 1,000 replications for each condition, *d* indicates the average standardized mean difference across 1,000 replications for each condition (control group subtracted from treatment group divided by the pooled standard deviation), and Var Ratio indicates the average variance ratio across 1,000 replications for each condition (treatment group's propensity score variance divided by the control group's propensity score variance). True in the table refers to the true propensity score, which included the error-free covariates (L1 and L2) in the covariate set. Naïve in the table refers to the naïve propensity score, which included the error-prone covariates (X1 and X2) in the covariate set. Cond indicates the corresponding measurement error condition.
Table 21
Numeric Diagnostic Information for Same Measurement Error Conditions After Mahalanobis Distance Matching Using the True and Naïve Propensity Scores

<table>
<thead>
<tr>
<th></th>
<th>L1</th>
<th>L2</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>Propensity Score</th>
<th>Var Ratio</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cond 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True</td>
<td>0.87</td>
<td>0.09</td>
<td>0.87</td>
<td>0.09</td>
<td>0.87</td>
<td>0.08</td>
<td>0.87</td>
<td>0.09</td>
<td>0.87</td>
<td>0.09</td>
</tr>
<tr>
<td>Cond 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True</td>
<td>0.87</td>
<td>0.09</td>
<td>0.87</td>
<td>0.09</td>
<td>0.87</td>
<td>0.08</td>
<td>0.87</td>
<td>0.09</td>
<td>0.87</td>
<td>0.09</td>
</tr>
<tr>
<td>Cond 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True</td>
<td>0.87</td>
<td>0.09</td>
<td>0.87</td>
<td>0.09</td>
<td>0.87</td>
<td>0.07</td>
<td>0.87</td>
<td>0.09</td>
<td>0.87</td>
<td>0.09</td>
</tr>
<tr>
<td>Cond 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True</td>
<td>0.87</td>
<td>0.09</td>
<td>0.87</td>
<td>0.09</td>
<td>0.87</td>
<td>0.06</td>
<td>0.87</td>
<td>0.09</td>
<td>0.87</td>
<td>0.09</td>
</tr>
<tr>
<td>Cond 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True</td>
<td>0.87</td>
<td>0.09</td>
<td>0.87</td>
<td>0.09</td>
<td>0.86</td>
<td>0.07</td>
<td>0.87</td>
<td>0.09</td>
<td>0.87</td>
<td>0.09</td>
</tr>
<tr>
<td>Cond 6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True</td>
<td>0.87</td>
<td>0.09</td>
<td>0.87</td>
<td>0.09</td>
<td>0.86</td>
<td>0.06</td>
<td>0.87</td>
<td>0.09</td>
<td>0.87</td>
<td>0.09</td>
</tr>
<tr>
<td>Naïve</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cond 1</td>
<td>0.84</td>
<td>0.11</td>
<td>0.84</td>
<td>0.11</td>
<td>0.87</td>
<td>0.08</td>
<td>0.87</td>
<td>0.09</td>
<td>0.87</td>
<td>0.09</td>
</tr>
<tr>
<td>Cond 2</td>
<td>0.82</td>
<td>0.12</td>
<td>0.80</td>
<td>0.13</td>
<td>0.87</td>
<td>0.08</td>
<td>0.87</td>
<td>0.09</td>
<td>0.87</td>
<td>0.09</td>
</tr>
<tr>
<td>Cond 3</td>
<td>0.80</td>
<td>0.14</td>
<td>0.77</td>
<td>0.15</td>
<td>0.87</td>
<td>0.07</td>
<td>0.87</td>
<td>0.09</td>
<td>0.87</td>
<td>0.09</td>
</tr>
<tr>
<td>Cond 4</td>
<td>0.78</td>
<td>0.15</td>
<td>0.75</td>
<td>0.17</td>
<td>0.87</td>
<td>0.07</td>
<td>0.88</td>
<td>0.09</td>
<td>0.87</td>
<td>0.09</td>
</tr>
<tr>
<td>Cond 5</td>
<td>0.75</td>
<td>0.17</td>
<td>0.72</td>
<td>0.19</td>
<td>0.87</td>
<td>0.06</td>
<td>0.87</td>
<td>0.09</td>
<td>0.87</td>
<td>0.09</td>
</tr>
<tr>
<td>Cond 6</td>
<td>0.73</td>
<td>0.18</td>
<td>0.69</td>
<td>0.21</td>
<td>0.87</td>
<td>0.05</td>
<td>0.87</td>
<td>0.09</td>
<td>0.87</td>
<td>0.09</td>
</tr>
</tbody>
</table>

Note. PBR indicates the average percent in bias reduction across 1,000 replications for each condition, \( d \) indicates the average standardized mean difference across 1,000 replications for each condition (control group subtracted from treatment group divided by the pooled standard deviation), and Var Ratio indicates the average variance ratio across 1,000 replications for each condition (treatment group's propensity score variance divided by the control group's propensity score variance). True in the table refers to the true propensity score, which included the error-free covariates (L1 and L2) in the covariate set. Naïve in the table refers to the naïve propensity score, which included the error-prone covariates (X1 and X2) in the covariate set. Cond indicates the corresponding measurement error condition.
Table 22

Numeric Diagnostic Information for Different Measurement Error Conditions After Nearest Neighbor Matching Using the True and Naïve Propensity Scores

<table>
<thead>
<tr>
<th></th>
<th>L1</th>
<th>L2</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>Propensity Score</th>
<th>Var Ratio</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cond 1</td>
<td>PBR</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
</tr>
<tr>
<td>Cond 2</td>
<td>PBR</td>
<td>0.95</td>
<td>0.04</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
</tr>
<tr>
<td>Cond 3</td>
<td>PBR</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
</tr>
<tr>
<td>Cond 4</td>
<td>PBR</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
</tr>
<tr>
<td>Cond 5</td>
<td>PBR</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
</tr>
<tr>
<td>Cond 6</td>
<td>PBR</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
<td>0.03</td>
<td>0.95</td>
<td>0.02</td>
<td>0.95</td>
<td>0.02</td>
<td>0.95</td>
</tr>
</tbody>
</table>

Naïve

|       | PBR | 0.92| 0.05| 0.90| 0.06| 0.95| 0.03| 0.94             | 0.04      | 0.95| 0.03| 0.95| 0.03| 1.24| 0.07|
| Cond 2| PBR | 0.90| 0.07| 0.88| 0.08| 0.95| 0.03| 0.95             | 0.03      | 0.95| 0.03| 0.95| 0.04| 1.17| 0.03|
| Cond 3| PBR | 0.87| 0.08| 0.85| 0.10| 0.96| 0.02| 0.96             | 0.02      | 0.95| 0.03| 0.95| 0.03| 1.17| 0.04|
| Cond 4| PBR | 0.85| 0.10| 0.83| 0.11| 0.96| 0.02| 0.98             | 0.01      | 0.95| 0.03| 0.94| 0.04| 1.14| 0.03|
| Cond 5| PBR | 0.83| 0.11| 0.80| 0.13| 0.98| 0.01| 0.99             | 0.01      | 0.95| 0.03| 0.95| 0.03| 1.02| 0.01|
| Cond 6| PBR | 0.81| 0.13| 0.77| 0.15| 1.00| 0.00| 1.00             | 0.00      | 0.95| 0.03| 0.95| 0.03| 1.15| 0.04|

Note. PBR indicates the average percent in bias reduction across 1,000 replications for each condition, d indicates the average standardized mean difference across 1,000 replications for each condition (control group subtracted from treatment group divided by the pooled standard deviation), and Var Ratio indicates the average variance ratio across 1,000 replications for each condition (treatment group's propensity score variance divided by the control group's propensity score variance). True in the table refers to the true propensity score, which included the error-free covariates (L1 and L2) in the covariate set. Naïve in the table refers to the naïve propensity score, which included the error-prone covariates (X1 and X2) in the covariate set. Cond indicates the corresponding measurement error condition.
Table 23

Numeric Diagnostic Information for Different Measurement Error Conditions After Nearest Neighbor Matching with a Caliper Using the True and Naïve Propensity Scores

<table>
<thead>
<tr>
<th></th>
<th>L1</th>
<th>L2</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>Propensity Score</th>
<th>Var Ratio</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>PBR</td>
<td>d</td>
<td>PBR</td>
<td>d</td>
<td>PBR</td>
<td>d</td>
<td>PBR</td>
<td>d</td>
<td>PBR</td>
<td>d</td>
</tr>
<tr>
<td>Cond 1</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>0.99</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Cond 2</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Cond 3</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Cond 4</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>0.99</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Cond 5</td>
<td>1.00</td>
<td>0.00</td>
<td>0.99</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Cond 6</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>0.99</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Naïve

| Cond 1 | 0.96   | 0.03   | 0.95| 0.04| 0.99| 0.00| 1.00| 0.00   | 1.00     | 0.00| 0.99| 0.01| 1.00| 0.00| 1.00| 0.00| 1.00| 0.00|
| Cond 2 | 0.94   | 0.04   | 0.92| 0.05| 0.99| 0.00| 1.00| 0.00   | 1.00     | 0.00| 1.00| 0.00| 1.00| 0.00| 1.00| 0.00| 1.00| 0.00|
| Cond 3 | 0.92   | 0.06   | 0.89| 0.08| 1.00| 0.00| 1.00| 0.00   | 0.99     | 0.00| 1.00| 0.00| 1.00| 0.00| 1.00| 0.00| 1.00| 0.00|
| Cond 4 | 0.89   | 0.08   | 0.86| 0.09| 1.00| 0.00| 1.00| 0.00   | 1.00     | 0.00| 1.00| 0.00| 1.00| 0.00| 1.00| 0.00| 1.00| 0.00|
| Cond 5 | 0.86   | 0.10   | 0.83| 0.11| 0.99| 0.00| 1.00| 0.00   | 1.00     | 0.00| 1.00| 0.00| 1.00| 0.00| 1.00| 0.00| 1.00| 0.00|
| Cond 6 | 0.84   | 0.11   | 0.80| 0.13| 1.00| 0.00| 1.00| 0.00   | 1.00     | 0.00| 1.00| 0.00| 1.00| 0.00| 1.00| 0.00| 1.00| 0.00|

Note. PBR indicates the average percent in bias reduction across 1,000 replications for each condition. d indicates the average standardized mean difference across 1,000 replications for each condition (control group subtracted from treatment group divided by the pooled standard deviation), and Var Ratio indicates the average variance ratio across 1,000 replications for each condition (treatment group's propensity score variance divided by the control group's propensity score variance). True in the table refers to the true propensity score, which included the error-free covariates (L1 and L2) in the covariate set. Naïve in the table refers to the naïve propensity score, which included the error-prone covariates (X1 and X2) in the covariate set. Cond indicates the corresponding measurement error condition.
Table 24

<table>
<thead>
<tr>
<th>L1</th>
<th>L2</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>Propensity Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>True PBR</td>
<td>d PBR</td>
<td>d PBR</td>
<td>d PBR</td>
<td>d PBR</td>
<td>d PBR</td>
<td>d PBR</td>
<td>d Var Ratio</td>
</tr>
<tr>
<td>Cond 1 0.95 0.03</td>
<td>0.96 0.03</td>
<td>0.95 0.03</td>
<td>0.96 0.03</td>
<td>0.95 0.03</td>
<td>0.96 0.03</td>
<td>0.96 0.03</td>
<td>1.27</td>
</tr>
<tr>
<td>Cond 2 0.95 0.03</td>
<td>0.95 0.03</td>
<td>0.95 0.03</td>
<td>0.95 0.03</td>
<td>0.95 0.03</td>
<td>0.96 0.03</td>
<td>0.96 0.03</td>
<td>1.18</td>
</tr>
<tr>
<td>Cond 3 0.95 0.03</td>
<td>0.95 0.03</td>
<td>0.95 0.03</td>
<td>0.95 0.03</td>
<td>0.95 0.03</td>
<td>0.96 0.03</td>
<td>0.96 0.03</td>
<td>1.25</td>
</tr>
<tr>
<td>Cond 4 0.95 0.03</td>
<td>0.95 0.03</td>
<td>0.96 0.02</td>
<td>0.95 0.03</td>
<td>0.96 0.03</td>
<td>0.96 0.03</td>
<td>0.96 0.03</td>
<td>1.16</td>
</tr>
<tr>
<td>Cond 5 0.95 0.03</td>
<td>0.95 0.03</td>
<td>0.95 0.03</td>
<td>0.95 0.03</td>
<td>0.95 0.03</td>
<td>0.95 0.03</td>
<td>0.95 0.03</td>
<td>1.05</td>
</tr>
<tr>
<td>Cond 6 0.96 0.03</td>
<td>0.96 0.03</td>
<td>0.96 0.02</td>
<td>0.96 0.02</td>
<td>0.96 0.03</td>
<td>0.96 0.03</td>
<td>0.96 0.03</td>
<td>1.17</td>
</tr>
</tbody>
</table>

Naïve
| Cond 1 0.93 0.05 | 0.91 0.06 | 0.95 0.03 | 0.94 0.03 | 0.96 0.03 | 0.96 0.03 | 0.96 0.03 | 1.24 | 0.07 |
| Cond 2 0.90 0.07 | 0.88 0.08 | 0.95 0.03 | 0.95 0.03 | 0.96 0.03 | 0.96 0.03 | 0.96 0.03 | 1.17 | 0.03 |
| Cond 3 0.88 0.08 | 0.85 0.10 | 0.96 0.02 | 0.96 0.02 | 0.96 0.03 | 0.96 0.03 | 0.96 0.03 | 1.17 | 0.04 |
| Cond 4 0.86 0.09 | 0.83 0.11 | 0.97 0.02 | 0.98 0.01 | 0.96 0.03 | 0.96 0.03 | 0.96 0.03 | 1.13 | 0.03 |
| Cond 5 0.84 0.11 | 0.80 0.13 | 0.99 0.01 | 1.00 0.00 | 0.96 0.03 | 0.96 0.03 | 0.95 0.03 | 1.01 | 0.00 |
| Cond 6 0.81 0.12 | 0.78 0.15 | 1.00 0.00 | 1.01 0.00 | 0.96 0.03 | 0.96 0.03 | 0.96 0.03 | 1.15 | 0.04 |

Note. PBR indicates the average percent in bias reduction across 1,000 replications for each condition, d indicates the average standardized mean difference across 1,000 replications for each condition (control group subtracted from treatment group divided by the pooled standard deviation), and Var Ratio indicates the average variance ratio across 1,000 replications for each condition (treatment group's propensity score variance divided by the control group's propensity score variance). True in the table refers to the true propensity score, which included the error-free covariates (L1 and L2) in the covariate set. Naïve in the table refers to the naïve propensity score, which included the error-prone covariates (X1 and X2) in the covariate set. Cond indicates the corresponding measurement error condition.
Table 25
Numeric Diagnostic Information for Different Measurement Error Conditions After Mahalanobis Distance Matching Using the True and Naïve Propensity Scores

<table>
<thead>
<tr>
<th></th>
<th>L1</th>
<th>L2</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>Propensity Score</th>
<th>Var Ratio</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>True</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cond 1</td>
<td>0.87</td>
<td>0.09</td>
<td>0.87</td>
<td>0.09</td>
<td>0.87</td>
<td>0.08</td>
<td>0.87</td>
<td>0.08</td>
<td>0.87</td>
<td>0.09</td>
</tr>
<tr>
<td>Cond 2</td>
<td>0.87</td>
<td>0.09</td>
<td>0.87</td>
<td>0.09</td>
<td>0.87</td>
<td>0.08</td>
<td>0.87</td>
<td>0.08</td>
<td>0.87</td>
<td>0.09</td>
</tr>
<tr>
<td>Cond 3</td>
<td>0.87</td>
<td>0.09</td>
<td>0.87</td>
<td>0.09</td>
<td>0.87</td>
<td>0.08</td>
<td>0.87</td>
<td>0.07</td>
<td>0.87</td>
<td>0.09</td>
</tr>
<tr>
<td>Cond 4</td>
<td>0.87</td>
<td>0.09</td>
<td>0.87</td>
<td>0.09</td>
<td>0.87</td>
<td>0.07</td>
<td>0.87</td>
<td>0.07</td>
<td>0.87</td>
<td>0.09</td>
</tr>
<tr>
<td>Cond 5</td>
<td>0.87</td>
<td>0.09</td>
<td>0.87</td>
<td>0.09</td>
<td>0.87</td>
<td>0.07</td>
<td>0.87</td>
<td>0.06</td>
<td>0.87</td>
<td>0.09</td>
</tr>
<tr>
<td>Cond 6</td>
<td>0.87</td>
<td>0.09</td>
<td>0.87</td>
<td>0.09</td>
<td>0.87</td>
<td>0.06</td>
<td>0.87</td>
<td>0.09</td>
<td>0.87</td>
<td>0.09</td>
</tr>
<tr>
<td><strong>Naïve</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cond 1</td>
<td>0.84</td>
<td>0.11</td>
<td>0.83</td>
<td>0.12</td>
<td>0.86</td>
<td>0.08</td>
<td>0.86</td>
<td>0.09</td>
<td>0.87</td>
<td>0.09</td>
</tr>
<tr>
<td>Cond 2</td>
<td>0.82</td>
<td>0.12</td>
<td>0.80</td>
<td>0.14</td>
<td>0.87</td>
<td>0.08</td>
<td>0.87</td>
<td>0.08</td>
<td>0.87</td>
<td>0.09</td>
</tr>
<tr>
<td>Cond 3</td>
<td>0.80</td>
<td>0.13</td>
<td>0.78</td>
<td>0.15</td>
<td>0.88</td>
<td>0.07</td>
<td>0.88</td>
<td>0.07</td>
<td>0.87</td>
<td>0.09</td>
</tr>
<tr>
<td>Cond 4</td>
<td>0.78</td>
<td>0.15</td>
<td>0.76</td>
<td>0.17</td>
<td>0.89</td>
<td>0.07</td>
<td>0.89</td>
<td>0.06</td>
<td>0.87</td>
<td>0.09</td>
</tr>
<tr>
<td>Cond 5</td>
<td>0.75</td>
<td>0.17</td>
<td>0.72</td>
<td>0.19</td>
<td>0.89</td>
<td>0.06</td>
<td>0.90</td>
<td>0.06</td>
<td>0.87</td>
<td>0.09</td>
</tr>
<tr>
<td>Cond 6</td>
<td>0.74</td>
<td>0.18</td>
<td>0.70</td>
<td>0.20</td>
<td>0.90</td>
<td>0.06</td>
<td>0.91</td>
<td>0.05</td>
<td>0.87</td>
<td>0.09</td>
</tr>
</tbody>
</table>

**Note.** PBR indicates the average percent in bias reduction across 1,000 replications for each condition, d indicates the average standardized mean difference across 1,000 replications for each condition (control group subtracted from treatment group divided by the pooled standard deviation), and Var Ratio indicates the average variance ratio across 1,000 replications for each condition (treatment group's propensity score variance divided by the control group's propensity score variance). True in the table refers to the true propensity score, which included the error-free covariates (L1 and L2) in the covariate set. Naïve in the table refers to the naïve propensity score, which included the error-prone covariates (X1 and X2) in the covariate set. Cond indicates the corresponding measurement error condition.
### Table 26
*Treatment Effect Estimate and Estimated Bias in Treatment Effect for Same Measurement Error Conditions*

<table>
<thead>
<tr>
<th>Cond</th>
<th>NN Matching</th>
<th>NN w/ 0.2 Caliper</th>
<th>Optimal Matching</th>
<th>Mahal Matching</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>True</td>
<td>Mean Diff</td>
<td>Bias</td>
<td>SE</td>
</tr>
<tr>
<td>1</td>
<td>11.52</td>
<td>1.52</td>
<td>2.33</td>
<td>2.78</td>
</tr>
<tr>
<td>2</td>
<td>11.54</td>
<td>1.54</td>
<td>2.17</td>
<td>2.66</td>
</tr>
<tr>
<td>3</td>
<td>11.51</td>
<td>1.51</td>
<td>2.22</td>
<td>2.68</td>
</tr>
<tr>
<td>4</td>
<td>11.63</td>
<td>1.63</td>
<td>2.28</td>
<td>2.80</td>
</tr>
<tr>
<td>5</td>
<td>11.47</td>
<td>1.47</td>
<td>2.14</td>
<td>2.59</td>
</tr>
<tr>
<td>6</td>
<td>11.57</td>
<td>1.57</td>
<td>2.23</td>
<td>2.73</td>
</tr>
</tbody>
</table>

Naïve

<table>
<thead>
<tr>
<th>Cond</th>
<th>Mean Diff</th>
<th>Bias</th>
<th>SE</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11.92</td>
<td>1.92</td>
<td>2.33</td>
<td>3.03</td>
</tr>
<tr>
<td>2</td>
<td>12.22</td>
<td>2.22</td>
<td>2.31</td>
<td>3.20</td>
</tr>
<tr>
<td>3</td>
<td>12.60</td>
<td>2.60</td>
<td>2.35</td>
<td>3.51</td>
</tr>
<tr>
<td>4</td>
<td>13.08</td>
<td>3.08</td>
<td>2.32</td>
<td>3.85</td>
</tr>
<tr>
<td>5</td>
<td>13.26</td>
<td>3.26</td>
<td>2.33</td>
<td>4.01</td>
</tr>
<tr>
<td>6</td>
<td>13.58</td>
<td>3.58</td>
<td>2.42</td>
<td>4.32</td>
</tr>
</tbody>
</table>

**Note.** NN Matching indicates the nearest neighbor matching conditions, NN w/ 0.2 Caliper indicates the nearest neighbor matching condition including a caliper width of 0.2 standard deviations on the logit of the propensity score, Optimal Matching indicates the optimal matching conditions, and Mahal Matching indicates the Mahalanobis distance matching conditions. Mean Diff indicates the mean difference between groups on the outcome variable (Y) averaged across 1,000 replications, Bias indicates the amount by which the estimated treatment effect (calculated as a mean difference) deviates from the simulated treatment effect of 10 on average across the 1,000 replications. SE indicates the standard error and RMSE indicates the root mean squared error. True in the table refers to the true propensity score, which included the error-free covariates (L1 and L2) in the covariate set. Naïve in the table refers to the naïve propensity score, which included the error-prone covariates (X1 and X2) in the covariate set. Cond indicates the corresponding measurement error condition.
Table 27
Treatment Effect Estimate and Estimated Bias in Treatment Effect for Different Measurement Error Conditions.

<table>
<thead>
<tr>
<th>True</th>
<th>Mean</th>
<th>Diff</th>
<th>Bias</th>
<th>SE</th>
<th>RMSE</th>
<th>Mean</th>
<th>Diff</th>
<th>Bias</th>
<th>SE</th>
<th>RMSE</th>
<th>Mean</th>
<th>Diff</th>
<th>Bias</th>
<th>SE</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN Matching</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>NN w/ 0.2 Caliper</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Optimal Matching</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cond 1</td>
<td>11.60</td>
<td>1.60</td>
<td>2.26</td>
<td>2.77</td>
<td>10.27</td>
<td>0.27</td>
<td>2.25</td>
<td>2.27</td>
<td>11.48</td>
<td>1.48</td>
<td>2.31</td>
<td>2.74</td>
<td>14.10</td>
<td>4.10</td>
<td>2.34</td>
</tr>
<tr>
<td>Cond 2</td>
<td>11.54</td>
<td>1.54</td>
<td>2.41</td>
<td>2.86</td>
<td>9.99</td>
<td>-0.01</td>
<td>2.26</td>
<td>2.26</td>
<td>11.40</td>
<td>1.40</td>
<td>2.38</td>
<td>2.77</td>
<td>14.02</td>
<td>4.02</td>
<td>2.36</td>
</tr>
<tr>
<td>Cond 3</td>
<td>11.66</td>
<td>1.66</td>
<td>2.23</td>
<td>2.78</td>
<td>10.14</td>
<td>0.14</td>
<td>2.16</td>
<td>2.16</td>
<td>11.52</td>
<td>1.52</td>
<td>2.23</td>
<td>2.70</td>
<td>14.10</td>
<td>4.10</td>
<td>2.27</td>
</tr>
<tr>
<td>Cond 4</td>
<td>11.48</td>
<td>1.48</td>
<td>2.23</td>
<td>2.68</td>
<td>9.95</td>
<td>-0.05</td>
<td>2.23</td>
<td>2.23</td>
<td>11.32</td>
<td>1.32</td>
<td>2.25</td>
<td>2.61</td>
<td>13.94</td>
<td>3.94</td>
<td>2.23</td>
</tr>
<tr>
<td>Cond 5</td>
<td>11.56</td>
<td>1.56</td>
<td>2.28</td>
<td>2.76</td>
<td>10.02</td>
<td>0.02</td>
<td>2.19</td>
<td>2.19</td>
<td>11.43</td>
<td>1.43</td>
<td>2.27</td>
<td>2.68</td>
<td>14.02</td>
<td>4.02</td>
<td>2.34</td>
</tr>
<tr>
<td>Cond 6</td>
<td>11.51</td>
<td>1.51</td>
<td>2.20</td>
<td>2.67</td>
<td>9.98</td>
<td>-0.02</td>
<td>2.16</td>
<td>2.16</td>
<td>11.31</td>
<td>1.31</td>
<td>2.30</td>
<td>2.64</td>
<td>14.00</td>
<td>4.00</td>
<td>2.24</td>
</tr>
<tr>
<td>Naïve</td>
<td>11.98</td>
<td>1.98</td>
<td>2.36</td>
<td>3.08</td>
<td>10.70</td>
<td>0.70</td>
<td>2.30</td>
<td>2.41</td>
<td>11.86</td>
<td>1.86</td>
<td>2.36</td>
<td>3.00</td>
<td>14.61</td>
<td>4.61</td>
<td>2.42</td>
</tr>
<tr>
<td>Cond 1</td>
<td>12.25</td>
<td>2.25</td>
<td>2.38</td>
<td>3.28</td>
<td>10.91</td>
<td>0.91</td>
<td>2.36</td>
<td>2.53</td>
<td>12.09</td>
<td>2.09</td>
<td>2.40</td>
<td>3.19</td>
<td>14.77</td>
<td>4.77</td>
<td>2.48</td>
</tr>
<tr>
<td>Cond 2</td>
<td>12.71</td>
<td>2.71</td>
<td>2.30</td>
<td>3.55</td>
<td>11.32</td>
<td>1.32</td>
<td>2.23</td>
<td>2.59</td>
<td>12.55</td>
<td>2.55</td>
<td>2.34</td>
<td>3.46</td>
<td>15.03</td>
<td>5.03</td>
<td>2.29</td>
</tr>
<tr>
<td>Cond 3</td>
<td>12.74</td>
<td>2.74</td>
<td>2.31</td>
<td>3.58</td>
<td>11.52</td>
<td>1.52</td>
<td>2.32</td>
<td>2.78</td>
<td>12.62</td>
<td>2.62</td>
<td>2.36</td>
<td>3.53</td>
<td>15.17</td>
<td>5.17</td>
<td>2.39</td>
</tr>
<tr>
<td>Cond 4</td>
<td>13.22</td>
<td>3.22</td>
<td>2.41</td>
<td>4.03</td>
<td>11.89</td>
<td>1.89</td>
<td>2.39</td>
<td>3.05</td>
<td>13.11</td>
<td>3.11</td>
<td>2.37</td>
<td>3.91</td>
<td>15.69</td>
<td>5.69</td>
<td>2.37</td>
</tr>
<tr>
<td>Cond 5</td>
<td>13.42</td>
<td>3.42</td>
<td>2.44</td>
<td>4.20</td>
<td>12.18</td>
<td>2.18</td>
<td>2.45</td>
<td>3.28</td>
<td>13.24</td>
<td>3.24</td>
<td>2.41</td>
<td>4.04</td>
<td>16.04</td>
<td>6.04</td>
<td>2.33</td>
</tr>
<tr>
<td>Cond 6</td>
<td>11.60</td>
<td>1.60</td>
<td>2.26</td>
<td>2.77</td>
<td>10.27</td>
<td>0.27</td>
<td>2.25</td>
<td>2.27</td>
<td>11.48</td>
<td>1.48</td>
<td>2.31</td>
<td>2.74</td>
<td>14.10</td>
<td>4.10</td>
<td>2.34</td>
</tr>
</tbody>
</table>

*Note.* NN Matching indicates the nearest neighbor matching conditions, NN w/ 0.2 Caliper indicates the nearest neighbor matching condition including a caliper width of 0.2 standard deviations on the logit of the propensity score, Optimal Matching indicates the optimal matching conditions, and Mahal Matching indicates the Mahalanobis distance matching conditions. Mean Diff indicates the mean difference between groups on the outcome variable (Y) averaged across 1,000 replications, Bias indicates the amount by which the estimated treatment effect (calculated as a mean difference) deviates from the simulated treatment effect of 10 on average across the 1,000 replications. SE indicates the standard error and RMSE indicates the root mean squared error. True in the table refers to the true propensity score, which included the error-free covariates (L1 and L2) in the covariate set. Naïve in the table refers to the naïve propensity score, which included the error-prone covariates (X1 and X2) in the covariate set. Cond indicates the corresponding measurement error condition.
Table 28

Results from One Four-Way and Four Three-way Mixed ANOVAs

<table>
<thead>
<tr>
<th></th>
<th>Four-way</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SS</td>
<td>df</td>
<td>MS</td>
<td>F</td>
<td>p</td>
<td>η²</td>
</tr>
<tr>
<td>Prop</td>
<td>36847.27</td>
<td>1</td>
<td>36847.2</td>
<td>22277.92</td>
<td>0.000</td>
<td>0.30</td>
</tr>
<tr>
<td>Prop*ME</td>
<td>7.21</td>
<td>1</td>
<td>7.21</td>
<td>4.36</td>
<td>0.037</td>
<td>0.00</td>
</tr>
<tr>
<td>Prop*Condition</td>
<td>7423.30</td>
<td>5</td>
<td>1484.66</td>
<td>897.63</td>
<td>0.000</td>
<td>0.06</td>
</tr>
<tr>
<td>Prop*Method</td>
<td>153.86</td>
<td>3</td>
<td>51.29</td>
<td>31.01</td>
<td>0.000</td>
<td>0.00</td>
</tr>
<tr>
<td>Prop<em>ME</em>Condition</td>
<td>24.82</td>
<td>5</td>
<td>4.96</td>
<td>3.00</td>
<td>0.010</td>
<td>0.00</td>
</tr>
<tr>
<td>Prop<em>Condition</em>Method</td>
<td>45.08</td>
<td>15</td>
<td>3.01</td>
<td>1.82</td>
<td>0.027</td>
<td>0.00</td>
</tr>
<tr>
<td>Prop<em>ME</em>Condition*Method</td>
<td>34.70</td>
<td>18</td>
<td>1.93</td>
<td>1.17</td>
<td>0.280</td>
<td>0.00</td>
</tr>
<tr>
<td>Error</td>
<td>79311.71</td>
<td>47952</td>
<td>1.65</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NN Matching

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SS</td>
<td>df</td>
<td>MS</td>
<td>F</td>
<td>p</td>
<td>η²</td>
</tr>
<tr>
<td>Prop</td>
<td>8653.09</td>
<td>1</td>
<td>8653.09</td>
<td>5855.35</td>
<td>0.000</td>
<td>0.31</td>
</tr>
<tr>
<td>Prop*ME</td>
<td>9.24</td>
<td>1</td>
<td>9.24</td>
<td>6.25</td>
<td>0.012</td>
<td>0.00</td>
</tr>
<tr>
<td>Prop*Condition</td>
<td>1795.76</td>
<td>5</td>
<td>359.15</td>
<td>243.03</td>
<td>0.000</td>
<td>0.06</td>
</tr>
<tr>
<td>Prop<em>ME</em>Condition</td>
<td>8.65</td>
<td>5</td>
<td>1.73</td>
<td>1.17</td>
<td>0.321</td>
<td>0.00</td>
</tr>
<tr>
<td>Error</td>
<td>17715.97</td>
<td>11988</td>
<td>1.48</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NN w/ 0.2 Caliper

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SS</td>
<td>df</td>
<td>MS</td>
<td>F</td>
<td>p</td>
<td>η²</td>
</tr>
<tr>
<td>Prop</td>
<td>11360.77</td>
<td>1</td>
<td>11360.77</td>
<td>5039.29</td>
<td>0.000</td>
<td>0.28</td>
</tr>
<tr>
<td>Prop*ME</td>
<td>0.96</td>
<td>1</td>
<td>0.96</td>
<td>0.42</td>
<td>0.515</td>
<td>0.00</td>
</tr>
<tr>
<td>Prop*Condition</td>
<td>2220.04</td>
<td>5</td>
<td>444.01</td>
<td>196.95</td>
<td>0.000</td>
<td>0.05</td>
</tr>
<tr>
<td>Prop<em>ME</em>Condition</td>
<td>3.81</td>
<td>5</td>
<td>0.76</td>
<td>0.34</td>
<td>0.890</td>
<td>0.00</td>
</tr>
<tr>
<td>Error</td>
<td>27026.18</td>
<td>11988</td>
<td>2.25</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Optimal Matching

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SS</td>
<td>df</td>
<td>MS</td>
<td>F</td>
<td>p</td>
<td>η²</td>
</tr>
<tr>
<td>Prop</td>
<td>8730.12</td>
<td>1</td>
<td>8730.12</td>
<td>6563.74</td>
<td>0.000</td>
<td>0.33</td>
</tr>
<tr>
<td>Prop*ME</td>
<td>7.86</td>
<td>1</td>
<td>7.86</td>
<td>5.91</td>
<td>0.015</td>
<td>0.00</td>
</tr>
<tr>
<td>Prop*Condition</td>
<td>1805.01</td>
<td>5</td>
<td>361.00</td>
<td>271.42</td>
<td>0.000</td>
<td>0.07</td>
</tr>
<tr>
<td>Prop<em>ME</em>Condition</td>
<td>8.82</td>
<td>5</td>
<td>1.76</td>
<td>1.33</td>
<td>0.250</td>
<td>0.00</td>
</tr>
<tr>
<td>Error</td>
<td>15944.68</td>
<td>11988</td>
<td>1.33</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Mahal Matching

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SS</td>
<td>df</td>
<td>MS</td>
<td>F</td>
<td>p</td>
<td>η²</td>
</tr>
<tr>
<td>Prop</td>
<td>8257.15</td>
<td>1</td>
<td>8257.15</td>
<td>5314.76</td>
<td>0.000</td>
<td>0.29</td>
</tr>
<tr>
<td>Prop*ME</td>
<td>2.10</td>
<td>1</td>
<td>2.10</td>
<td>1.35</td>
<td>0.245</td>
<td>0.00</td>
</tr>
<tr>
<td>Prop*Condition</td>
<td>1647.57</td>
<td>5</td>
<td>329.51</td>
<td>212.09</td>
<td>0.000</td>
<td>0.06</td>
</tr>
<tr>
<td>Prop<em>ME</em>Condition</td>
<td>25.29</td>
<td>5</td>
<td>5.06</td>
<td>3.26</td>
<td>0.006</td>
<td>0.00</td>
</tr>
<tr>
<td>Error</td>
<td>18624.88</td>
<td>11988</td>
<td>1.55</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. Four-way indicates the results for the 2x2x6x4 ANOVA testing the four-way interaction. NN Matching indicates the results of the nearest neighbor matching conditions 2x2x6 ANOVA, NN w/ 0.2 Caliper indicates the results of the nearest neighbor matching with a caliper width of 0.2 standard deviations on the logit of the propensity score 2x2x6 ANOVA, Optimal Matching indicates the results of the optimal matching conditions 2x2x6 ANOVA, and Mahal Matching indicates the results of the Mahalanobis distance matching conditions 2x2x6 ANOVA. Prop indicates the effect of type of propensity score (true versus naïve), ME indicates the effect of the type of imposed measurement error, Condition indicates the effect of the six levels of simulated measurement error, and Method indicates the effect of the type of matching method employed.
### Table 29

*Treatment Effect Estimate 95% Confidence Interval Coverage for the Same Measurement Error Conditions*

<table>
<thead>
<tr>
<th>Cond</th>
<th>True Lower Bound</th>
<th>True Upper Bound</th>
<th>NN Matching Lower Bound</th>
<th>NN Matching Upper Bound</th>
<th>NN w/ 0.2 Caliper Lower Bound</th>
<th>NN w/ 0.2 Caliper Upper Bound</th>
<th>Optimal Matching Lower Bound</th>
<th>Optimal Matching Upper Bound</th>
<th>Mahal Matching Lower Bound</th>
<th>Mahal Matching Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cond 1</td>
<td>1%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>1%</td>
<td>0%</td>
<td>6%</td>
<td>0%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cond 2</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>5%</td>
<td>0%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cond 3</td>
<td>1%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>6%</td>
<td>0%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cond 4</td>
<td>1%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>6%</td>
<td>0%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cond 5</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>5%</td>
<td>0%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cond 6</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>5%</td>
<td>0%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Naïve**

| Cond 1 | 1% | 0% | 0% | 0% | 1% | 0% | 7% | 0% |
| Cond 2 | 1% | 0% | 0% | 0% | 1% | 0% | 11% | 0% |
| Cond 3 | 1% | 0% | 0% | 0% | 2% | 0% | 14% | 0% |
| Cond 4 | 3% | 0% | 1% | 0% | 2% | 0% | 17% | 0% |
| Cond 5 | 3% | 0% | 1% | 0% | 2% | 0% | 20% | 0% |
| Cond 6 | 4% | 0% | 1% | 0% | 4% | 0% | 21% | 0% |

*Note.* The values in the above table indicate the percentage of times the true treatment effect was excluded from the 95% confidence interval. NN Matching indicates the nearest neighbor matching conditions, NN w/ 0.2 Caliper indicates the nearest neighbor matching condition including a caliper width of 0.2 standard deviations on the logit of the propensity score, Optimal Matching indicates the optimal matching conditions, and Mahal Matching indicates the Mahalanobis distance matching conditions. Lower Bound indicates the percent of times the lower bound of the confidence interval excluded the true treatment effect of 10 across 1,000 replications. Upper Bound indicates the percent of times the upper bound of the confidence interval excluded the true treatment effect of 10 across 1,000 replications. True in the table refers to the true propensity score, which included the error-free covariates (L1 and L2) in the covariate set. Naïve in the table refers to the naïve propensity score, which included the error-prone covariates (X1 and X2) in the covariate set. Cond indicates the corresponding measurement error condition.
Table 30
*Treatment Effect Estimate 95% Confidence Interval Coverage for the Different Measurement Error Conditions*

<table>
<thead>
<tr>
<th>Condition</th>
<th>True Lower Bound</th>
<th>True Upper Bound</th>
<th>NN Matching Lower Bound</th>
<th>NN Matching Upper Bound</th>
<th>NN w/ 0.2 Caliper Lower Bound</th>
<th>NN w/ 0.2 Caliper Upper Bound</th>
<th>Optimal Matching Lower Bound</th>
<th>Optimal Matching Upper Bound</th>
<th>Mahal Matching Lower Bound</th>
<th>Mahal Matching Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cond 1</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>7%</td>
<td>0%</td>
</tr>
<tr>
<td>Cond 2</td>
<td>1%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>7%</td>
<td>0%</td>
</tr>
<tr>
<td>Cond 3</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>6%</td>
<td>0%</td>
</tr>
<tr>
<td>Cond 4</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>5%</td>
<td>0%</td>
</tr>
<tr>
<td>Cond 5</td>
<td>1%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>7%</td>
<td>0%</td>
</tr>
<tr>
<td>Cond 6</td>
<td>1%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>1%</td>
<td>0%</td>
<td>6%</td>
<td>0%</td>
</tr>
<tr>
<td><strong>Naïve</strong></td>
<td><strong>Conditon</strong> 1</td>
<td><strong>1%</strong></td>
<td><strong>0%</strong></td>
<td><strong>0%</strong></td>
<td><strong>0%</strong></td>
<td><strong>0%</strong></td>
<td><strong>0%</strong></td>
<td><strong>0%</strong></td>
<td><strong>9%</strong></td>
<td><strong>0%</strong></td>
</tr>
<tr>
<td>Cond 2</td>
<td><strong>1%</strong></td>
<td><strong>0%</strong></td>
<td><strong>0%</strong></td>
<td><strong>0%</strong></td>
<td><strong>1%</strong></td>
<td><strong>0%</strong></td>
<td><strong>0%</strong></td>
<td><strong>12%</strong></td>
<td><strong>0%</strong></td>
<td></td>
</tr>
<tr>
<td>Cond 3</td>
<td><strong>1%</strong></td>
<td><strong>0%</strong></td>
<td><strong>0%</strong></td>
<td><strong>0%</strong></td>
<td><strong>1%</strong></td>
<td><strong>0%</strong></td>
<td><strong>12%</strong></td>
<td><strong>0%</strong></td>
<td><strong>13%</strong></td>
<td><strong>0%</strong></td>
</tr>
<tr>
<td>Cond 4</td>
<td><strong>2%</strong></td>
<td><strong>0%</strong></td>
<td><strong>1%</strong></td>
<td><strong>0%</strong></td>
<td><strong>1%</strong></td>
<td><strong>0%</strong></td>
<td><strong>1%</strong></td>
<td><strong>0%</strong></td>
<td><strong>18%</strong></td>
<td><strong>0%</strong></td>
</tr>
<tr>
<td>Cond 5</td>
<td><strong>3%</strong></td>
<td><strong>0%</strong></td>
<td><strong>1%</strong></td>
<td><strong>0%</strong></td>
<td><strong>3%</strong></td>
<td><strong>0%</strong></td>
<td><strong>4%</strong></td>
<td><strong>0%</strong></td>
<td><strong>25%</strong></td>
<td><strong>0%</strong></td>
</tr>
<tr>
<td>Cond 6</td>
<td><strong>5%</strong></td>
<td><strong>0%</strong></td>
<td><strong>2%</strong></td>
<td><strong>0%</strong></td>
<td><strong>4%</strong></td>
<td><strong>0%</strong></td>
<td><strong>4%</strong></td>
<td><strong>0%</strong></td>
<td><strong>25%</strong></td>
<td><strong>0%</strong></td>
</tr>
</tbody>
</table>

*Note.* The values in the above table indicate the percentage of times the true treatment effect was excluded from the 95% confidence interval. NN Matching indicates the nearest neighbor matching conditions, NN w/ 0.2 Caliper indicates the nearest neighbor matching condition including a caliper width of 0.2 standard deviations on the logit of the propensity score, Optimal Matching indicates the optimal matching conditions, and Mahal Matching indicates the Mahalanobis distance matching conditions. Lower Bound indicates the percent of times the lower bound of the confidence interval excluded the true treatment effect of 10 across 1,000 replications. Upper Bound indicates the percent of times the upper bound of the confidence interval excluded the true treatment effect of 10 across 1,000 replications. True in the table refers to the true propensity score, which included the error-free covariates (L1 and L2) in the covariate set. Naïve in the table refers to the naïve propensity score, which included the error-prone covariates (X1 and X2) in the covariate set. Cond indicates the corresponding measurement error condition.
Figure 1. Example of an absolute standardized mean difference plot produced by the MatchIt package in R (Ho et al., 2013).
Figure 2. Example of the Q-Q plots produced by the MatchIt package in R (Ho et al., 2013).
Figure 3. Example of boxplots used to evaluate the distribution of two groups on continuous covariates.
Figure 4. Example of density plots used to evaluate the distribution of treatment and comparison groups on matched covariates created via ggplot2 (Wickham, 2009).
Figure 5. Example of a jitter plot produced by the MatchIt package in R (Ho et al., 2013).
Figure 6. Illustration of the jitter plots when visually diagnosing matches with incrementally larger samples (sample size increases from the plot on the left to the plot on the right).
Figure 7. Example of the histogram plots produced by the MatchIt package in R (Ho et al., 2013).
Figure 8. Conceptual diagram of latent and observed simulated variables where L1 and L2 are latent variables related to self-selection into treatment (T) and the outcome (Y), X1 and X2 are observed error-prone measures of L1 and L2, X3-X5 are observed error-free covariates, and V is a random disturbance term with a mean of zero and a standard deviation of one. Note the grey lines indicate correlations or indirect paths among the simulated covariates and black lines indicate the direct paths between simulated variables.
Figure 9. Plots of the average percent of bias reduction in L1 (dark grey) and X1 (light grey) for the same measurement error conditions. Note that the red dashed line indicates the 80% reduction in bias benchmark recommended in the literature (Pan & Bai, 2015).
Figure 10. Plots of the average percent of bias reduction in L₁ (dark grey) and X₁ (light grey) for the different measurement error conditions. Note that the red dashed line indicates the 80% reduction in bias benchmark recommended in the literature (Pan & Bai, 2015).
Figure 11. Plots of the standardized mean difference for L1 (dark grey) and X1 (light grey) for the same measurement error conditions. Note that the bottom red dashed line indicates the 0.10 benchmark (Ho et al., 2007) and the top red dashed line indicates the 0.2 benchmark (Austin, 2011a) recommended in the literature.
Figure 12. Plots of the standardized mean difference for L1 (dark grey) and X1 (light grey) for the different measurement error conditions. Note that the bottom red dashed line indicates the 0.10 benchmark (Ho et al., 2007) and the top red dashed line indicates the 0.2 benchmark (Austin, 2011a) recommended in the literature.
Figure 13. Plots of the estimated treatment effect (left panels) and bias in the estimated treatment effect (right panels) created using true propensity scores (dark grey bars) and naïve propensity scores (light grey bars) for the same measurement error conditions. Note the red dotted line indicates the true (simulated) treatment effect.
Figure 14. Plots of the estimated treatment effect (left panels) and bias in the estimated treatment effect (right panels) created using true propensity scores (dark grey bars) and naïve propensity scores (light grey bars) for the different measurement error conditions. Note the red dotted line indicates the true (simulated) treatment effect.
Figure 15. Plots of the percent of times the estimated treatment effect confidence interval upper (light grey – value of zero across all conditions) and lower bound (dark grey) excluded the true treatment effect for the same measurement error conditions.
Figure 16. Plots of the percent of times the estimated treatment effect confidence interval upper (light grey – value of zero across all conditions) and lower bound (dark grey) excluded the true treatment effect for the different measurement error conditions.
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# ~~~~~~~~~~~~~~       Dissertation Code
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# ~~~~~~~~~~~~~~              Heather Harris
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

## Quick note: Because I adapted this code from Christine and Jeanne’s code
## I left (most) of the previous comments as one "#" so that they can be distinguished
## from
## my comments.

## Follow-up, because the script has been edited numerous times, this may no longer be
## the case.

## Before starting, I set my working directory to my desktop R folder
## because I plan on saving out a final PDF file of my plot(s) there.
setwd("~heatherdawnharris/Desktop/Validation Sets")
## In this step, I double-check that it worked
getwd()
install.packages("mvtnorm")
require(mvtnorm)
library(mvtnorm)
install.packages("optmatch")
require(optmatch)

## In the following steps, I compare the propensity scores created above with those
## calculated by
## the MatchIt package.
## I would first install the package then require it (if I didn’t already have it)
install.packages("MatchIt")
require(MatchIt)
require(psych)

## In this step, I create the for loop for simulated data. First, I create places for the
## values I want to save out.

#Numeric diagnostics for PSM WITH COVARIATES
PropMeanTreatW <- rep(NA, 1000)
PropMeanContrW <- rep(NA, 1000)
PropVarTreatW <- rep(NA, 1000)
PropVarContrW <- rep(NA, 1000)

#Numeric diagnostics for PSM WITHOUT COVARIATES
PropMeanTreatWO <- rep(NA, 1000)
PropMeanContrWO <- rep(NA, 1000)
PropVarTreatWO <- rep(NA, 1000)
PropVarContrWO <- rep(NA, 1000)

#All variables BEFORE matching
pAvgX1Treat <- rep(NA, 1000)
pAvgX2Treat <- rep(NA, 1000)
pAvgL1Treat <- rep(NA, 1000)
pAvgL2Treat <- rep(NA, 1000)
pAvgX3Treat <- rep(NA, 1000)
pAvgX4Treat <- rep(NA, 1000)
pAvgX5Treat <- rep(NA, 1000)
pAvgY1Treat <- rep(NA, 1000)
pAvgX1Cont <- rep(NA, 1000)
pAvgX2Cont <- rep(NA, 1000)
pAvgL1Cont <- rep(NA, 1000)
pAvgL2Cont <- rep(NA, 1000)
pAvgX3Cont <- rep(NA, 1000)
# All variables AFTER matching True Prop NN

\[
\begin{align*}
\text{pAvgX4Cont} & \leftarrow \text{rep}(\text{NA}, 1000) \\
\text{pAvgX5Cont} & \leftarrow \text{rep}(\text{NA}, 1000) \\
\text{pAvgYCont} & \leftarrow \text{rep}(\text{NA}, 1000) \\
\text{pSDX1Treat} & \leftarrow \text{rep}(\text{NA}, 1000) \\
\text{pSDX2Treat} & \leftarrow \text{rep}(\text{NA}, 1000) \\
\text{pSDX1Treat} & \leftarrow \text{rep}(\text{NA}, 1000) \\
\text{pSDX2Treat} & \leftarrow \text{rep}(\text{NA}, 1000) \\
\text{pSDX3Treat} & \leftarrow \text{rep}(\text{NA}, 1000) \\
\text{pSDX4Treat} & \leftarrow \text{rep}(\text{NA}, 1000) \\
\text{pSDX5Treat} & \leftarrow \text{rep}(\text{NA}, 1000) \\
\text{pSDYTreat} & \leftarrow \text{rep}(\text{NA}, 1000) \\
\text{pSDX1Cont} & \leftarrow \text{rep}(\text{NA}, 1000) \\
\text{pSDX2Cont} & \leftarrow \text{rep}(\text{NA}, 1000) \\
\text{pSDL1Treat} & \leftarrow \text{rep}(\text{NA}, 1000) \\
\text{pSDL2Cont} & \leftarrow \text{rep}(\text{NA}, 1000) \\
\text{pSDL3Cont} & \leftarrow \text{rep}(\text{NA}, 1000) \\
\text{pSDL4Cont} & \leftarrow \text{rep}(\text{NA}, 1000) \\
\text{pSDL5Cont} & \leftarrow \text{rep}(\text{NA}, 1000) \\
\text{pCorX1.X2} & \leftarrow \text{rep}(\text{NA}, 1000) \\
\text{pCorX1.X3} & \leftarrow \text{rep}(\text{NA}, 1000) \\
\text{pCorX1.X4} & \leftarrow \text{rep}(\text{NA}, 1000) \\
\text{pCorX1.X5} & \leftarrow \text{rep}(\text{NA}, 1000) \\
\text{pCorX2.X3} & \leftarrow \text{rep}(\text{NA}, 1000) \\
\text{pCorX2.X4} & \leftarrow \text{rep}(\text{NA}, 1000) \\
\text{pCorX2.X5} & \leftarrow \text{rep}(\text{NA}, 1000) \\
\text{pCorX3.X4} & \leftarrow \text{rep}(\text{NA}, 1000) \\
\text{pCorX3.X5} & \leftarrow \text{rep}(\text{NA}, 1000) \\
\text{pCorX4.X5} & \leftarrow \text{rep}(\text{NA}, 1000) \\
\text{pCorY.X1} & \leftarrow \text{rep}(\text{NA}, 1000) \\
\text{pCorY.X2} & \leftarrow \text{rep}(\text{NA}, 1000) \\
\text{pCorY.X3} & \leftarrow \text{rep}(\text{NA}, 1000) \\
\text{pCorY.X4} & \leftarrow \text{rep}(\text{NA}, 1000) \\
\text{pCorY.X5} & \leftarrow \text{rep}(\text{NA}, 1000) \\
\text{pPropMeanTT} & \leftarrow \text{rep}(\text{NA}, 1000) \\
\text{pPropMeanCT} & \leftarrow \text{rep}(\text{NA}, 1000) \\
\text{pPropSDTT} & \leftarrow \text{rep}(\text{NA}, 1000) \\
\text{pPropSDCT} & \leftarrow \text{rep}(\text{NA}, 1000) \\
\text{pPropMeanTN} & \leftarrow \text{rep}(\text{NA}, 1000) \\
\text{pPropMeanCN} & \leftarrow \text{rep}(\text{NA}, 1000) \\
\text{pPropSDTN} & \leftarrow \text{rep}(\text{NA}, 1000) \\
\text{pPropSDCN} & \leftarrow \text{rep}(\text{NA}, 1000) \\
\text{pCorL1X1} & \leftarrow \text{rep}(\text{NA}, 1000) \\
\text{pCorL2X2} & \leftarrow \text{rep}(\text{NA}, 1000)
\end{align*}
\]
#All variables AFTER matching Naive Prop NN
NNTavgX1Cont <- rep(NA, 1000)
NNTavgX2Cont <- rep(NA, 1000)
NNTavgL1Cont <- rep(NA, 1000)
NNTavgL2Cont <- rep(NA, 1000)
NNTavgX3Cont <- rep(NA, 1000)
NNTavgX4Cont <- rep(NA, 1000)
NNTavgX5Cont <- rep(NA, 1000)
NNTmatchedN <- rep(NA, 1000)
NNTCohensD <- rep(NA, 1000)
NNTPropMeanT <- rep(NA, 1000)
NNTPropMeanC <- rep(NA, 1000)
NNTPropVarT <- rep(NA, 1000)
NNTPropVarC <- rep(NA, 1000)

NNTCorY.X1 <- rep(NA, 1000)
NNTCorY.X2 <- rep(NA, 1000)
NNTCorY.X3 <- rep(NA, 1000)
NNTCorY.X4 <- rep(NA, 1000)
NNTCorY.X5 <- rep(NA, 1000)

NNTCorX1.X2 <- rep(NA, 1000)
NNTCorX1.X3 <- rep(NA, 1000)
NNTCorX1.X4 <- rep(NA, 1000)
NNTCorX1.X5 <- rep(NA, 1000)
NNTCorX2.X3 <- rep(NA, 1000)
NNTCorX2.X4 <- rep(NA, 1000)
NNTCorX2.X5 <- rep(NA, 1000)
NNTCorX3.X4 <- rep(NA, 1000)
NNTCorX3.X5 <- rep(NA, 1000)
NNTCorX4.X5 <- rep(NA, 1000)

NNTCorY.X1Treat <- rep(NA, 1000)
NNTCorX1.X2Treat <- rep(NA, 1000)
NNTCorX1.X3Treat <- rep(NA, 1000)
NNTCorX1.X4Treat <- rep(NA, 1000)
NNTCorX1.X5Treat <- rep(NA, 1000)
NNTCorX2.X3Treat <- rep(NA, 1000)
NNTCorX2.X4Treat <- rep(NA, 1000)
NNTCorX2.X5Treat <- rep(NA, 1000)
NNTCorX3.X4Treat <- rep(NA, 1000)
NNTCorX3.X5Treat <- rep(NA, 1000)
NNTCorX4.X5Treat <- rep(NA, 1000)

NNTsdYCont <- rep(NA, 1000)
NNTsdX1Cont <- rep(NA, 1000)
NNTsdX2Cont <- rep(NA, 1000)
NNTsdL1Cont <- rep(NA, 1000)
NNTsdL2Cont <- rep(NA, 1000)
NNTsdX3Cont <- rep(NA, 1000)
NNTsdX4Cont <- rep(NA, 1000)
NNTsdX5Cont <- rep(NA, 1000)
NNTsdYX1Cont <- rep(NA, 1000)
NNTsdYX2Cont <- rep(NA, 1000)
NNTsdYX3Cont <- rep(NA, 1000)
NNTsdYX4Cont <- rep(NA, 1000)
NNTsdYX5Cont <- rep(NA, 1000)
NNTsdL1X1Cont <- rep(NA, 1000)
NNTsdL2X1Cont <- rep(NA, 1000)
NNTsdL3X1Cont <- rep(NA, 1000)
NNTsdL4X1Cont <- rep(NA, 1000)
NNTsdL5X1Cont <- rep(NA, 1000)
NNTsdL2X2Cont <- rep(NA, 1000)
NNTsdL3X2Cont <- rep(NA, 1000)
NNTsdL4X2Cont <- rep(NA, 1000)
NNTsdL5X2Cont <- rep(NA, 1000)
NNTsdL1X3Cont <- rep(NA, 1000)
NNTsdL2X3Cont <- rep(NA, 1000)
NNTsdL3X3Cont <- rep(NA, 1000)
NNTsdL4X3Cont <- rep(NA, 1000)
NNTsdL5X3Cont <- rep(NA, 1000)
NNTsdL2X4Cont <- rep(NA, 1000)
NNTsdL3X4Cont <- rep(NA, 1000)
NNTsdL4X4Cont <- rep(NA, 1000)
NNTsdL5X4Cont <- rep(NA, 1000)
NNTsdL1X5Cont <- rep(NA, 1000)
NNTsdL2X5Cont <- rep(NA, 1000)
NNTsdL3X5Cont <- rep(NA, 1000)
NNTsdL4X5Cont <- rep(NA, 1000)
NNTsdL5X5Cont <- rep(NA, 1000)
#All variables AFTER matching True Prop NN w/ 0.2 Caliper
NNCTAvgX1Treat <- rep(NA, 1000)
NNCTAvgX2Treat <- rep(NA, 1000)
NNCTAvgX3Treat <- rep(NA, 1000)
NNCTAvgX4Treat <- rep(NA, 1000)
NNCTAvgX5Treat <- rep(NA, 1000)
NNCTAvgY1Treat <- rep(NA, 1000)
NNCTAvgY2Treat <- rep(NA, 1000)
NNCTAvgY3Treat <- rep(NA, 1000)
NNCTAvgY4Treat <- rep(NA, 1000)
NNCTAvgY5Treat <- rep(NA, 1000)
NNCTAvgX1Cont <- rep(NA, 1000)
NNCTAvgX2Cont <- rep(NA, 1000)
NNCTAvgX3Cont <- rep(NA, 1000)
NNCTAvgX4Cont <- rep(NA, 1000)
NNCTAvgX5Cont <- rep(NA, 1000)
NNCTAvgY1Cont <- rep(NA, 1000)
NNCTAvgY2Cont <- rep(NA, 1000)
NNCTAvgY3Cont <- rep(NA, 1000)
NNCTAvgY4Cont <- rep(NA, 1000)
NNCTAvgY5Cont <- rep(NA, 1000)
NNCTAvgX1Treat <- rep(NA, 1000)
NNCTAvgX2Treat <- rep(NA, 1000)
NNCTAvgX3Treat <- rep(NA, 1000)
NNCTAvgX4Treat <- rep(NA, 1000)
NNCTAvgX5Treat <- rep(NA, 1000)
NNCTAvgX1Cont <- rep(NA, 1000)
NNCTAvgX2Cont <- rep(NA, 1000)
NNCTAvgX3Cont <- rep(NA, 1000)
NNCTAvgX4Cont <- rep(NA, 1000)
NNCTAvgX5Cont <- rep(NA, 1000)
NNCTAvgY1Treat <- rep(NA, 1000)
NNCTAvgY2Treat <- rep(NA, 1000)
NNCTAvgY3Treat <- rep(NA, 1000)
NNCTAvgY4Treat <- rep(NA, 1000)
NNCTAvgY5Treat <- rep(NA, 1000)
NNCTAvgX1Cont <- rep(NA, 1000)
NNCTAvgX2Cont <- rep(NA, 1000)
NNCTAvgX3Cont <- rep(NA, 1000)
NNCTAvgX4Cont <- rep(NA, 1000)
NNCTAvgX5Cont <- rep(NA, 1000)
# All variables AFTER matching True Prop Optimal

```r
optTAvgX1Treat <- rep(NA, 1000)
optTAvgX2Treat <- rep(NA, 1000)
optTAvgX3Treat <- rep(NA, 1000)
optTAvgX4Treat <- rep(NA, 1000)
optTAvgX5Treat <- rep(NA, 1000)
optTAvgYTreat <- rep(NA, 1000)
optTAvgX1Cont <- rep(NA, 1000)
optTAvgX2Cont <- rep(NA, 1000)
optTAvgX3Cont <- rep(NA, 1000)
optTAvgX4Cont <- rep(NA, 1000)
optTAvgX5Cont <- rep(NA, 1000)
optTAvgCont <- rep(NA, 1000)
```

# All variables AFTER matching Naive Prop NN w/ 0.2 Caliper

```r
NNCNPropVarC <- rep(NA, 1000)
NNCNPropVarT <- rep(NA, 1000)
NNCNPropMeanC <- rep(NA, 1000)
NNCNPropMeanT <- rep(NA, 1000)
NNCNMatchedN <- rep(NA, 1000)
NNCNcohensDW <- rep(NA, 1000)
NNCNCorY.X1 <- rep(NA, 1000)
NNCNCorY.X2 <- rep(NA, 1000)
NNCNCorY.X3 <- rep(NA, 1000)
NNCNCorY.X4 <- rep(NA, 1000)
NNCNCorY.X5 <- rep(NA, 1000)
```

# All variables AFTER matching True Prop Optimal

```r
optTAvgX1Treat <- rep(NA, 1000)
optTAvgX2Treat <- rep(NA, 1000)
optTAvgX3Treat <- rep(NA, 1000)
optTAvgX4Treat <- rep(NA, 1000)
optTAvgX5Treat <- rep(NA, 1000)
optTAvgYTreat <- rep(NA, 1000)
optTAvgX1Cont <- rep(NA, 1000)
optTAvgX2Cont <- rep(NA, 1000)
```

# All variables AFTER matching True Prop Optimal

```r
# All variables AFTER matching True Prop Optimal
```
# All variables AFTER matching Naive Prop Optimal

optNAvgX1Cont <- rep(NA, 1000)
optNAvgX2Cont <- rep(NA, 1000)
optNAvgX3Cont <- rep(NA, 1000)
optNAvgX4Cont <- rep(NA, 1000)
optNAvgX5Cont <- rep(NA, 1000)
optTSdx1Treat <- rep(NA, 1000)
optTSdx2Treat <- rep(NA, 1000)
optTSdl1Treat <- rep(NA, 1000)
optTSdl2Treat <- rep(NA, 1000)
optTSdx3Treat <- rep(NA, 1000)
optTSdx4Treat <- rep(NA, 1000)
optTSdx5Treat <- rep(NA, 1000)
optTSdyTreat <- rep(NA, 1000)
optTSdx1Cont <- rep(NA, 1000)
optTSdx2Cont <- rep(NA, 1000)
optTSdl1Cont <- rep(NA, 1000)
optTSdl2Cont <- rep(NA, 1000)
optTSdx3Cont <- rep(NA, 1000)
optTSdx4Cont <- rep(NA, 1000)
optTSdx5Cont <- rep(NA, 1000)
optTSdyCont <- rep(NA, 1000)
optTcorX1.X2 <- rep(NA, 1000)
optTcorX1.X3 <- rep(NA, 1000)
optTcorX1.X4 <- rep(NA, 1000)
optTcorX1.X5 <- rep(NA, 1000)
optTcorX2.X3 <- rep(NA, 1000)
optTcorX2.X4 <- rep(NA, 1000)
optTcorX2.X5 <- rep(NA, 1000)
optTcorX3.X4 <- rep(NA, 1000)
optTcorX3.X5 <- rep(NA, 1000)
optTcorX4.X5 <- rep(NA, 1000)
optTcorY.X1 <- rep(NA, 1000)
optTcorY.X2 <- rep(NA, 1000)
optTcorY.X3 <- rep(NA, 1000)
optTcorY.X4 <- rep(NA, 1000)
optTcorY.X5 <- rep(NA, 1000)
optTmatchedN <- rep(NA, 1000)
optTcohenSDW <- rep(NA, 1000)
optTpropMeanT <- rep(NA, 1000)
optTpropMeanC <- rep(NA, 1000)
optTpropVarT <- rep(NA, 1000)
optTpropVarC <- rep(NA, 1000)

# All variables AFTER matching Naive Prop Optimal
<table>
<thead>
<tr>
<th>Variable</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>optNSDL1Cont</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>optNSDL2Cont</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>optNSDL3Cont</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>optNSDL4Cont</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>optNSDL5Cont</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>optNSDYCont</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>optNCorX1.X2</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>optNCorX1.X3</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>optNCorX1.X4</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>optNCorX1.X5</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>optNCorX2.X3</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>optNCorX2.X4</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>optNCorX2.X5</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>optNCorX3.X4</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>optNCorX3.X5</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>optNCorX4.X5</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>optNCorY.X1</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>optNCorY.X2</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>optNCorY.X3</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>optNCorY.X4</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>optNCorY.X5</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>optNMatchedN</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>optNCohensDW</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>optNPropMeanT</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>optNPropMeanC</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>optNPropVarT</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>optNPropVarC</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>MahalTavgX1Treat</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>MahalTavgX2Treat</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>MahalTavgL1Treat</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>MahalTavgL2Treat</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>MahalTavgX3Treat</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>MahalTavgX4Treat</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>MahalTavgX5Treat</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>MahalTavgYTreat</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>MahalTavgX1Cont</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>MahalTavgX2Cont</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>MahalTavgL1Cont</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>MahalTavgL2Cont</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>MahalTavgX3Cont</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>MahalTavgX4Cont</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>MahalTavgX5Cont</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>MahalTavgYCont</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>MahalTSDX1Treat</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>MahalTSDX2Treat</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>MahalTSDL1Treat</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>MahalTSDL2Treat</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>MahalTSDX1Cont</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>MahalTSDX2Cont</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>MahalTSDL1Cont</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>MahalTSDL2Cont</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>MahalTSDX3Cont</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>MahalTSDX4Cont</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>MahalTSDX5Cont</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>MahalTSDYCont</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>MahalTCorX1.X2</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>MahalTCorX1.X3</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>MahalTCorX1.X4</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>MahalTCorX1.X5</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>MahalTCorX2.X3</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>MahalTCorX2.X4</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>MahalTCorX2.X5</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>MahalTCorX3.X4</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>MahalTCorX3.X5</td>
<td>rep(NA, 1000)</td>
</tr>
</tbody>
</table>
MahalTCorX4.X5 <- rep(NA, 1000)
MahalTCorY.X1 <- rep(NA, 1000)
MahalTCorY.X2 <- rep(NA, 1000)
MahalTCorY.X3 <- rep(NA, 1000)
MahalTCorY.X4 <- rep(NA, 1000)
MahalTMatchedN <- rep(NA, 1000)
MahalTCohensDW <- rep(NA, 1000)
MahalTPropMeanT <- rep(NA, 1000)
MahalTPropMeanC <- rep(NA, 1000)
MahalTPropVarT <- rep(NA, 1000)
MahalTPropVarC <- rep(NA, 1000)

# All variables AFTER matching Naive Prop Mahal
MahalNAvgX1Treat <- rep(NA, 1000)
MahalNAvgX2Treat <- rep(NA, 1000)
MahalNAvgL1Treat <- rep(NA, 1000)
MahalNAvgL2Treat <- rep(NA, 1000)
MahalNAvgX3Treat <- rep(NA, 1000)
MahalNAvgX4Treat <- rep(NA, 1000)
MahalNAvgX5Treat <- rep(NA, 1000)
MahalNAvgYTreat <- rep(NA, 1000)
MahalNAvgX1Cont <- rep(NA, 1000)
MahalNAvgX2Cont <- rep(NA, 1000)
MahalNAvgL1Cont <- rep(NA, 1000)
MahalNAvgL2Cont <- rep(NA, 1000)
MahalNAvgX3Cont <- rep(NA, 1000)
MahalNAvgX4Cont <- rep(NA, 1000)
MahalNAvgX5Cont <- rep(NA, 1000)
MahalNAvgYC <- rep(NA, 1000)
MahalNSDX1Treat <- rep(NA, 1000)
MahalNSDX2Treat <- rep(NA, 1000)
MahalNSDL1Treat <- rep(NA, 1000)
MahalNSDL2Treat <- rep(NA, 1000)
MahalNSDX3Treat <- rep(NA, 1000)
MahalNSDX4Treat <- rep(NA, 1000)
MahalNSDX5Treat <- rep(NA, 1000)
MahalNSDYTreat <- rep(NA, 1000)
MahalNSDYCont <- rep(NA, 1000)
MahalNSDX1Cont <- rep(NA, 1000)
MahalNSDX2Cont <- rep(NA, 1000)
MahalNSDL1Cont <- rep(NA, 1000)
MahalNSDL2Cont <- rep(NA, 1000)
MahalNSDX3Cont <- rep(NA, 1000)
MahalNSDX4Cont <- rep(NA, 1000)
MahalNSDX5Cont <- rep(NA, 1000)
MahalNCorX1.X2 <- rep(NA, 1000)
MahalNCorX1.X3 <- rep(NA, 1000)
MahalNCorX1.X4 <- rep(NA, 1000)
MahalNCorX1.X5 <- rep(NA, 1000)
MahalNCorX2.X3 <- rep(NA, 1000)
MahalNCorX2.X4 <- rep(NA, 1000)
MahalNCorX2.X5 <- rep(NA, 1000)
MahalNCorX3.X4 <- rep(NA, 1000)
MahalNCorX3.X5 <- rep(NA, 1000)
MahalNCorX4.X5 <- rep(NA, 1000)
MahalNCorY.X1 <- rep(NA, 1000)
MahalNCorY.X2 <- rep(NA, 1000)
MahalNCorY.X3 <- rep(NA, 1000)
MahalNCorY.X4 <- rep(NA, 1000)
MahalNCorY.X5 <- rep(NA, 1000)
MahalNMatchedN <- rep(NA, 1000)
MahalNCohensDW <- rep(NA, 1000)
MahalNPropMeanT <- rep(NA, 1000)
MahalNPropMeanC <- rep(NA, 1000)
MahalNPropVarT <- rep(NA, 1000)
MahalNPropVarC <- rep(NA, 1000)
## In case I want to figure out how long it will take. :)  
```
str(mSDYCont)  
proc.time()  
now<-proc.time()  
proc.time()<-now
```

```
set.seed(21)
for(i in 1:1000){
    # 1. Simulating Examinees
    egetP=qnorm(1-TreatP) # threshold
    corrX=matrix(c(1,.58,.42,.49,.60,  
                    .58,1,.56,.44,.41,  
                    .42,.56,1,.59,.44,  
                    .49,.44,.59,1,.51,  
                    .60,.41,.44,.51,1),5,5)
    corrXlp=c(.4,.4,.4,.4,.4)
    Pcoef=solve(corrX) %*% corrXlp  #solve means inverse
    varExpP=sum(Pcoef %*% t(Pcoef)*corrX)
    library(mvtnorm)
    library(psych)
    # simulate covariates
    Nexaminee=1000
    Nrep=1000
    X=rmvnorm(Nexaminee, rep(0,5), corrX, method="chol")
    # the portion of latent propensity accounted for by the covariates
    noErr=as.vector(X %*% Pcoef)
    #varErr
    #var(noErr) # should be close to varExpP
    Rsq=varExpP/(1+varExpP) # because the error variance of a probit is 1
    # probability that examinee is above cutoff
    #Rsq
    dist=noErr-mycut/sqrt(1-Rsq)
    #head(dist)
    #mean(dist)
    #mean(randraw)
    trueprop=dnorm(dist)
    #head(trueprop)
    #mean(trueprop)
    # randomly assign each person Nrep times
    randraw=matrix(runif(Nexaminee*Nrep),nrow=Nexaminee,ncol=Nrep)
    #randraw
    #trueprop
    group=ifelse(trueprop>randraw,1,0)
    meangrp=rowMeans(group)
```
#group
# using first replication, check if P coef is estimated well
dataZ= data.frame(X, group[,1])
#str(dataZ)
#describe(temp1)
#head(temp1)
#describeBy(temp1, temp1$group)
#describe(dataZ)

# Saving out the error-free variables
L1<- dataZ$X1
L2<- dataZ$X2

# Creating the SD for the error term to create measurement error.

#ErrorSD<- sqrt((0.1)*(1/0.9)) #Condition 1
#ErrorSD<- sqrt((0.2)*(1/0.8)) #Condition 2
#ErrorSD<- sqrt((0.3)*(1/0.7)) #Condition 3
#ErrorSD<- sqrt((0.4)*(1/0.6)) #Condition 4
#ErrorSD<- sqrt((0.5)*(1/0.5)) #Condition 5
ErrorSD<- sqrt((0.6)*((1/0.4))) #Condition 6

# ERROR FOR X1
v1<- rnorm(Nexaminee, mean=0, sd=ErrorSD)
X1<- dataZ$X1+v1
# cor(L1, X1)

# ERROR FOR X2
v2<- rnorm(Nexaminee, mean=0, sd=ErrorSD)
X2<- dataZ$X2+v2
# cor(L2, X2)

## Now transforming the simulated variables from a mean of 0 and SD of 1 to a mean of 25 and an SD of 5.
L1<- L1*5+25
L2<- L2*5+25
X1<- X1*5+25
X2<- X2*5+25
X3<- dataZ$X3*5+25
X4<- dataZ$X4*5+25
X5<- dataZ$X5*5+25

group<- group[,1]
finaldata<- cbind(L1, L2, X1, X2, X3, X4, X5, group)
finaldata<- as.data.frame(finaldata)
Pb<- glm(formula= group ~ L1+L2+X3+X4+X5, data=finaldata, family=binomial)
finaldata$TRUEprop<- predict(Pb, type="response")

#describe(finaldata)
#pairs.panels(finaldata)

## In the below steps, I set the model specifications where Y is the outcome variable with an intercept of 100, and a function of X1, X2, X3, X4, and X5, with a treatment effect of 10.
Yv<- rnorm(Nexaminee, mean=0, sd=20)

Y<- 100 + 10*group +2*finaldata$X1 + 2*finaldata$X2 + 2*finaldata$X3 + 2*finaldata$X4 + 2*finaldata$X5 + Yv
mean(Y)
finaldata$Y<- Y

## Now adding naive propensity scores using the error-prone covariates
Pb2<- glm(formula=group ~ X1+X2+X3+X4+X5, data=finaldata, family=binomial)
finaldata$NAIVEprop<- predict(Pb2, type="response")
#describe(finaldata)
#describeBy(finaldata, finaldata$group)
pairs.panels(finaldata)
#head(finaldata)
write.table(finaldata, file="Cond6SAME.csv", sep=""," row.names=FALSE)

# 2. Propensity Score Matching (NN w/o caliper)
# CONDUCTING PSM WITH TRUE PROPENSIT
m.out1=matchit(finaldata$group~finaldata$L1+finaldata$L2+finaldata$X3+finaldata$X4+finaldata$X5, data=finaldata, method="nearest", ratio=1)
#m.out1
#m.out1$distance
#plot(m.out1, type = "jitter") # propensity score locations
#plot(m.out1, type = "hist")
dataNNT=match.data(m.out1)
#describe(dataNNT)

# CONDUCTING PSM WITH NAIVE PROPENSITY SCORES
m.out2=matchit(finaldata$group~finaldata$X1+finaldata$X2+finaldata$X3+finaldata$X4+finaldata$X5, data=finaldata, method="nearest", ratio=1)
#m.out2
#m.out2$distance
#plot(m.out2, type = "jitter") # propensity score locations
#plot(m.out2, type = "hist")
dataNNN=match.data(m.out2)
#describe(dataNNN)

# 3. Propensity Score Matching (NN with 0.2 caliper)
# CONDUCTING PSM WITH TRUE PROPENSITY SCORES
ps.sd = sd(finaldata$TRUEprop)
#ps.sd
m.out3=matchit(finaldata$group~finaldata$L1+finaldata$L2+finaldata$X3+finaldata$X4+finaldata$X5, data=finaldata, method="nearest", caliper=0.2*ps.sd, ratio=1)
#m.out3$distance
#plot(m.out3, type = "jitter") # propensity score locations
#plot(m.out3, type = "hist")
dataNNCT=match.data(m.out3)
#describe(dataNNCT)

# CONDUCTING PSM WITH NAIVE PROPENSITY SCORES
ps.sd2 = sd(finaldata$NAIVEprop)
#ps.sd
m.out4=matchit(finaldata$group~finaldata$X1+finaldata$X2+finaldata$X3+finaldata$X4+finaldata$X5, data=finaldata, method="nearest", caliper=0.2*ps.sd2, ratio=1)
#m.out4$distance
#plot(m.out4, type = "jitter") # propensity score locations
#plot(m.out4, type = "hist")
dataNNCN=match.data(m.out4)
#describe(dataNNCN)

# 4. Propensity Score Matching (Optimal)
# CONDUCTING PSM WITH TRUE PROPENSITY SCORES
m.out5=matchit(finaldata$group~finaldata$L1+finaldata$L2+finaldata$X3+finaldata$X4+finaldata$X5, data=finaldata, method="optimal", ratio=1)
#m.out5$distance
#plot(m.out5, type = "jitter") # propensity score locations
#plot(m.out5, type = "hist")

dataoptT=match.data(m.out5)
#describe(dataoptT)

# CONDUCTING PSM WITH NAIVE PROPENSITY SCORES
m.out6=matchit(finaldata$group~finaldata$X1+finaldata$X2+finaldata$X3+finaldata$X4+finaldata$X5, data=finaldata, method="optimal", ratio=1)

#m.out6$distance
#plot(m.out6, type = "jitter") # propensity score locations
#plot(m.out6, type = "hist")

dataoptN=match.data(m.out6)
#describe(dataoptN)

# 5. Propensity Score Matching (Mahal Matching)
# CONDUCTING PSM WITH TRUE PROPENSITY SCORES
m.out7=matchit(finaldata$group~finaldata$X1+finaldata$X2+finaldata$X3+finaldata$X4+finaldata$X5+finaldata$TRUEprop, data=finaldata, method="nearest", distance="mahalanobis")
dataMahaltT=match.data(m.out7)
#describe(dataMahaltT)

m.out8=matchit(finaldata$group~finaldata$X1+finaldata$X2+finaldata$X3+finaldata$X4+finaldata$X5+finaldata$NAIVEprop, data=finaldata, method="nearest", distance="mahalanobis")
dataMahaltN=match.data(m.out8)
#describe(dataMahaltN)

# 3. Saving Out Diagnostics
# All means and SDs for variables as simulated.
#AvgX1Treat[i]  <-  mean(data$X1[data$group==1])
#AvgX2Treat[i]  <-  mean(data$X2[data$group==1])
#AvgL1Treat[i]  <-  mean(data$L1[data$group==1])
#AvgL2Treat[i]  <-  mean(data$L2[data$group==1])
#AvgX3Treat[i]  <-  mean(data$X3[data$group==1])
#AvgX4Treat[i]  <-  mean(data$X4[data$group==1])
#AvgX5Treat[i]  <-  mean(data$X5[data$group==1])
#AvgYTreat[i]   <-  mean(data$Y[data$group==1])
#AvgX1Cont[i]   <-  mean(data$X1[data$group==0])
#AvgX2Cont[i]   <-  mean(data$X2[data$group==0])
#AvgL1Cont[i]   <-  mean(data$L1[data$group==0])
#AvgL2Cont[i]   <-  mean(data$L2[data$group==0])
#AvgX3Cont[i]   <-  mean(data$X3[data$group==0])
#AvgX4Cont[i]   <-  mean(data$X4[data$group==0])
#AvgX5Cont[i]   <-  mean(data$X5[data$group==0])
pAvgYCont[i] <- mean(data$Y[data$group==0])
pSDX2Cont[i] <- sd(data$X2[data$group==1])
pSDL2Treat[i] <- sd(data$L2[data$group==1])
pSDX3Treat[i] <- sd(data$X3[data$group==1])
pSDL5Treat[i] <- sd(data$Y[data$group==1])
pSDLX1Cont[i] <- sd(data$X1[data$group==0])
pSDLX2Cont[i] <- sd(data$X2[data$group==0])
pSDL1Cont[i] <- sd(data$L1[data$group==0])
pSDL2Cont[i] <- sd(data$L2[data$group==0])
pSDL3Cont[i] <- sd(data$X3[data$group==0])
pSDL4Cont[i] <- sd(data$X4[data$group==0])
pSDL5Cont[i] <- sd(data$X5[data$group==0])
pSDYCont[i] <- sd(data$Y[data$group==0])
pCorX1.X2[i] <- cor(data$X1, data$X2)
pCorX1.X3[i] <- cor(data$X1, data$X3)
pCorX1.X4[i] <- cor(data$X1, data$X4)
pCorX1.X5[i] <- cor(data$X1, data$X5)
pCorX2.X3[i] <- cor(data$X2, data$X3)
pCorX2.X4[i] <- cor(data$X2, data$X4)
pCorX2.X5[i] <- cor(data$X2, data$X5)
pCorX3.X4[i] <- cor(data$X3, data$X4)
pCorX3.X5[i] <- cor(data$X3, data$X5)
pCorX4.X5[i] <- cor(data$X4, data$X5)
pCorY.X1[i] <- cor(data$Y, data$X1)
pCorY.X2[i] <- cor(data$Y, data$X2)
pCorY.X3[i] <- cor(data$Y, data$X3)
pCorY.X4[i] <- cor(data$Y, data$X4)
pCorY.X5[i] <- cor(data$Y, data$X5)
pPropMeanTT[i] <- mean(data$TRUEprop[data$group==1])
pPropMeanCT[i] <- mean(data$TRUEprop[data$group==0])
pPropsDTT[i] <- sd(data$TRUEprop[data$group==1])
pPropsDTN[i] <- sd(data$TRUEprop[data$group==0])
pPropMeanCN[i] <- mean(data$NAIVEprop[data$group==1])
pPropSDTN[i] <- sd(data$NAIVEprop[data$group==0])
pPropSDCN[i] <- sd(data$NAIVEprop[data$group==0])
pCorL1X1 <- cor(L1, X1)
pCorL2X2 <- cor(L2, X2)
# All variables AFTER matching True Prop NN
NNTAvgX1Treat[i] <- mean(data$X1[data$2group==1])
NNTAvgX2Treat[i] <- mean(data$X2[data$2group==1])
NNTAvgL1Treat[i] <- mean(data$L1[data$2group==1])
NNTAvgL2Treat[i] <- mean(data$L2[data$2group==1])
NNTAvgX3Treat[i] <- mean(data$X3[data$2group==1])
NNTAvgX4Treat[i] <- mean(data$X4[data$2group==1])
NNTAvgX5Treat[i] <- mean(data$X5[data$2group==1])
NNTAvgX1Cont[i] <- mean(data$X1[data$2group==0])
NNTAvgX2Cont[i] <- mean(data$X2[data$2group==0])
NNTAvgL1Cont[i] <- mean(data$L1[data$2group==0])
NNTAvgL2Cont[i] <- mean(data$L2[data$2group==0])
NNTAvgX3Cont[i] <- mean(data$X3[data$2group==0])
NNTAvgX4Cont[i] <- mean(data$X4[data$2group==0])
NNTAvgX5Cont[i] <- mean(data$X5[data$2group==0])
NNTAvgYCont[i] <- mean(data$Y[data$2group==0])
NNTSDX1Treat[i] <- sd(data$X1[data$2group==1])
NNTSDX2Treat[i] <- sd(data$X2[data$2group==1])
NNTSDL1Treat[i] <- sd(data$L1[data$2group==1])
NNTSDL2Treat[i] <- sd(data$L2[data$2group==1])
NNTSDX3Treat[i] <- sd(data$X3[data$2group==1])
NNTSDX4Treat[i] <- sd(data$X4[data$2group==1])
NNTSDX5Treat[i] <- sd(data$X5[data$2group==1])
NNTSDYTreat[i] <- sd(data$Y[data$2group==1])
NNTSDX1Cont[i] <- sd(data$X1[data$2group==0])
NNNSDX2Cont[i] <- sd(data2$X2[data2$group==0])
NNNSDL2Cont[i] <- sd(data2$L2[data2$group==0])
NNNSDX3Cont[i] <- sd(data2$X3[data2$group==0])
NNNSDX4Cont[i] <- sd(data2$X4[data2$group==0])
NNNSDX5Cont[i] <- sd(data2$X5[data2$group==0])
NNNTSDYCont[i] <- sd(data2$Y[data2$group==0])
NNNTCorX1.X2[i] <- cor(data2$X1, data2$X2)
NNNTCorX1.X3[i] <- cor(data2$X1, data2$X3)
NNNTCorX1.X4[i] <- cor(data2$X1, data2$X4)
NNNTCorX1.X5[i] <- cor(data2$X1, data2$X5)
NNNTCorX2.X3[i] <- cor(data2$X2, data2$X3)
NNNTCorX2.X4[i] <- cor(data2$X2, data2$X4)
NNNTCorX2.X5[i] <- cor(data2$X2, data2$X5)
NNNTCorX3.X4[i] <- cor(data2$X3, data2$X4)
NNNTCorX3.X5[i] <- cor(data2$X3, data2$X5)
NNNTCorX4.X5[i] <- cor(data2$X4, data2$X5)
NNNTCorY.X1[i] <- cor(data2$Y, data2$X1)
NNNTCorY.X2[i] <- cor(data2$Y, data2$X2)
NNNTCorY.X3[i] <- cor(data2$Y, data2$X3)
NNNTCorY.X4[i] <- cor(data2$Y, data2$X4)
NNNTCorY.X5[i] <- cor(data2$Y, data2$X5)
NNTMachedN[i] <- length(data2$Y[data2$group==0])
NNTPropMeanT <- mean(data2$TRUEprop[data2$group==1])
NNTPropMeanC <- mean(data2$TRUEprop[data2$group==0])
NNTPropVarT <- (sd(data2$TRUEprop[data2$group==1]))^2
NNTPropVarC <- (sd(data2$TRUEprop[data2$group==0]))^2

#All variables AFTER matching Naive Prop NN
NNNavgX1Treat[i] <- mean(data3$X1[data3$group==1])
NNNavgX2Treat[i] <- mean(data3$X2[data3$group==1])
NNNavgX3Treat[i] <- mean(data3$X3[data3$group==1])
NNNavgX4Treat[i] <- mean(data3$X4[data3$group==1])
NNNavgX5Treat[i] <- mean(data3$X5[data3$group==1])
NNNavgYTreat[i] <- mean(data3$Y[data3$group==1])
NNNavgX1Cont[i] <- mean(data3$X1[data3$group==0])
NNNavgX2Cont[i] <- mean(data3$X2[data3$group==0])
NNNavgX3Cont[i] <- mean(data3$X3[data3$group==0])
NNNavgX4Cont[i] <- mean(data3$X4[data3$group==0])
NNNavgX5Cont[i] <- mean(data3$X5[data3$group==0])
NNNavgYCont[i] <- mean(data3$Y[data3$group==0])
NNNSDX1Treat[i] <- sd(data3$X1[data3$group==1])
NNNSDX2Treat[i] <- sd(data3$X2[data3$group==1])
NNNSDX3Treat[i] <- sd(data3$X3[data3$group==1])
NNNSDX4Treat[i] <- sd(data3$X4[data3$group==1])
NNNSDX5Treat[i] <- sd(data3$X5[data3$group==1])
NNNSDYTreat[i] <- sd(data3$Y[data3$group==1])
NNNSDX1Cont[i] <- sd(data3$X1[data3$group==0])
NNNSDX2Cont[i] <- sd(data3$X2[data3$group==0])
NNNSDX3Cont[i] <- sd(data3$X3[data3$group==0])
NNNSDX4Cont[i] <- sd(data3$X4[data3$group==0])
NNNSDX5Cont[i] <- sd(data3$X5[data3$group==0])
NNNSDYCont[i] <- sd(data3$Y[data3$group==0])
NNNTCorX1.X2[i] <- cor(data3$X1, data3$X2)
NNNTCorX1.X3[i] <- cor(data3$X1, data3$X3)
NNNTCorX1.X4[i] <- cor(data3$X1, data3$X4)
NNNTCorX1.X5[i] <- cor(data3$X1, data3$X5)
NNNTCorX2.X3[i] <- cor(data3$X2, data3$X3)
NNNTCorX2.X4[i] <- cor(data3$X2, data3$X4)
NNNTCorX2.X5[i] <- cor(data3$X2, data3$X5)
NNNTCorX3.X4[i] <- cor(data3$X3, data3$X4)
NNNTCorX3.X5[i] <- cor(data3$X3, data3$X5)
NNNTCorX4.X5[i] <- cor(data3$X4, data3$X5)
#All variables AFTER matching Naive Prop NN w/ 0.2 Caliper

```
NNCTcorY.X1[i] <- cor(data3$Y, data3$X1)
NNCTcorY.X2[i] <- cor(data3$Y, data3$X2)
NNCTcorY.X3[i] <- cor(data3$Y, data3$X3)
NNCTcorY.X4[i] <- cor(data3$Y, data3$X4)
NNCTcorY.X5[i] <- cor(data3$Y, data3$X5)
NNNMatchedN[i] <- length(data3$Y[data3$group==0])
NNNPropMeanC <- mean(data3$NAIVEprop[data3$group==1])
NNNPropMeanT <- mean(data3$NAIVEprop[data3$group==0])
NNNPropVarC <- (sd(data3$NAIVEprop[data3$group==1]))^2
NNNPropVarT <- (sd(data3$NAIVEprop[data3$group==0]))^2

# All variables AFTER matching True Prop NN w/ 0.2 Caliper

```

165
#All variables AFTER matching True Prop Optimal

```
#NAIVEprop <- mean(data5$NAIVEprop[data5$group==1])
#NAIVEpropC <- mean(data5$NAIVEprop[data5$group==0])
#NAIVEpropVarT <- (sd(data5$NAIVEprop[data5$group==1]))^2
#NAIVEpropVarC <- (sd(data5$NAIVEprop[data5$group==0]))^2

optTAvgX1Treat[i] <- mean(data6$X1[data6$group==1])
optTAvgX1Treat[i] <- mean(data6$X1[data6$group==0])
optTAvgL1Treat[i] <- mean(data6$L1[data6$group==1])
optTAvgL1Treat[i] <- mean(data6$L1[data6$group==0])
optTAvgL2Treat[i] <- mean(data6$L2[data6$group==1])
optTAvgL2Treat[i] <- mean(data6$L2[data6$group==0])
optTAvgX3Treat[i] <- mean(data6$X3[data6$group==1])
optTAvgX3Treat[i] <- mean(data6$X3[data6$group==0])
optTAvgX4Treat[i] <- mean(data6$X4[data6$group==1])
optTAvgX4Treat[i] <- mean(data6$X4[data6$group==0])
optTAvgX5Treat[i] <- mean(data6$X5[data6$group==1])
optTAvgX5Treat[i] <- mean(data6$X5[data6$group==0])
optTAvgL1Cont[i] <- mean(data6$L1[data6$group==0])
optTAvgL2Cont[i] <- mean(data6$L2[data6$group==0])
optTAvgL3Cont[i] <- mean(data6$L3[data6$group==0])
optTAvgL4Cont[i] <- mean(data6$L4[data6$group==0])
optTAvgL5Cont[i] <- mean(data6$L5[data6$group==0])
optTAvgL6Cont[i] <- mean(data6$L6[data6$group==0])
```

```
# All variables AFTER matching Naive Prop Optimal

```r
defines:
  optTPropMeanC <- mean(data6$TRUEprop[data6$group==1])
  optTPropVarT <- (sd(data6$TRUEprop[data6$group==1]))^2
  optTPropVarC <- (sd(data6$TRUEprop[data6$group==0]))^2
```

```
# optNAvgX1Treat[i] <- mean(data7$X1[data7$group==1])
optNAvgX2Treat[i] <- mean(data7$X2[data7$group==1])
optNAvgX3Treat[i] <- mean(data7$X3[data7$group==1])
optNAvgX4Treat[i] <- mean(data7$X4[data7$group==1])
optNAvgX5Treat[i] <- mean(data7$X5[data7$group==1])
```

```
# optNAvgX1Cont[i] <- mean(data6$X1[data6$group==0])
optNAvgX2Cont[i] <- mean(data6$X2[data6$group==0])
optNAvgX3Cont[i] <- mean(data6$X3[data6$group==0])
optNAvgX4Cont[i] <- mean(data6$X4[data6$group==0])
optNAvgX5Cont[i] <- mean(data6$X5[data6$group==0])
```
optNSDX3Cont[i] <- sd(data7$X3[data7$group==0])
optNSDX4Cont[i] <- sd(data7$X4[data7$group==0])
optNSDX5Cont[i] <- sd(data7$X5[data7$group==0])
optNSDYCont[i] <- sd(data7$Y[data7$group==0])
optNCorX1.X2[i] <- cor(data7$X1, data7$X2)
optNCorX1.X3[i] <- cor(data7$X1, data7$X3)
optNCorX1.X4[i] <- cor(data7$X1, data7$X4)
optNCorX1.X5[i] <- cor(data7$X1, data7$X5)
optNCorX2.X3[i] <- cor(data7$X2, data7$X3)
optNCorX2.X4[i] <- cor(data7$X2, data7$X4)
optNCorX2.X5[i] <- cor(data7$X2, data7$X5)
optNCorX3.X4[i] <- cor(data7$X3, data7$X4)
optNCorX3.X5[i] <- cor(data7$X3, data7$X5)
optNCorX4.X5[i] <- cor(data7$X4, data7$X5)
optNCory.X1[i] <- cor(data7$Y, data7$X1)
optNCory.X2[i] <- cor(data7$Y, data7$X2)
optNCory.X3[i] <- cor(data7$Y, data7$X3)
optNCory.X4[i] <- cor(data7$Y, data7$X4)
optNCory.X5[i] <- cor(data7$Y, data7$X5)
optNMatchedN[i] <- length(data7$Y[data7$group==0])
optNPropMeanT <- mean(data7$NAIVEprop[data7$group==1])
optNPropVarT <- mean(data7$NAIVEprop[data7$group==0])
optNPropVarT <- (sd(data7$NAIVEprop[data7$group==1]))^2
optNPropVarC <- (sd(data7$NAIVEprop[data7$group==0]))^2

# All variables AFTER matching True Prop Mahal

MahalTavgX1Treat[i] <- mean(data8$X1[data8$group==1])
MahalTavgX2Treat[i] <- mean(data8$X2[data8$group==1])
MahalTavgL1Treat[i] <- mean(data8$L1[data8$group==1])
MahalTavgX3Treat[i] <- mean(data8$X3[data8$group==1])
MahalTavgX4Treat[i] <- mean(data8$X4[data8$group==1])
MahalTavgX5Treat[i] <- mean(data8$X5[data8$group==1])
MahalTavgYTTreat[i] <- mean(data8$Y[data8$group==1])
MahalTavgX1Cont[i] <- mean(data8$X1[data8$group==0])
MahalTavgX2Cont[i] <- mean(data8$X2[data8$group==0])
MahalTavgL1Cont[i] <- mean(data8$L1[data8$group==0])
MahalTavgX3Cont[i] <- mean(data8$X3[data8$group==0])
MahalTavgX4Cont[i] <- mean(data8$X4[data8$group==0])
MahalTavgX5Cont[i] <- mean(data8$X5[data8$group==0])
MahalTavgYTC[i] <- mean(data8$Y[data8$group==0])
MahalTSDX1Treat[i] <- sd(data8$X1[data8$group==1])
MahalTSDX2Treat[i] <- sd(data8$X2[data8$group==1])
MahalTSDX3Treat[i] <- sd(data8$X3[data8$group==1])
MahalTSDX4Treat[i] <- sd(data8$X4[data8$group==1])
MahalTSDX5Treat[i] <- sd(data8$X5[data8$group==1])
MahalTSDX1Cont[i] <- sd(data8$X1[data8$group==0])
MahalTSDX2Cont[i] <- sd(data8$X2[data8$group==0])
MahalTSDX3Cont[i] <- sd(data8$X3[data8$group==0])
MahalTSDX4Cont[i] <- sd(data8$X4[data8$group==0])
MahalTSDX5Cont[i] <- sd(data8$X5[data8$group==0])
MahalTSDYCont[i] <- sd(data8$Y[data8$group==0])
MahalTCorX1.X2[i] <- cor(data8$X1, data8$X2)
MahalTCorX1.X3[i] <- cor(data8$X1, data8$X3)
MahalTCorX1.X4[i] <- cor(data8$X1, data8$X4)
MahalTCorX1.X5[i] <- cor(data8$X1, data8$X5)
MahalTCorX2.X3[i] <- cor(data8$X2, data8$X3)
MahalTCorX2.X4[i] <- cor(data8$X2, data8$X4)
MahalTCorX2.X5[i] <- cor(data8$X2, data8$X5)
MahalTCorX3.X4[i] <- cor(data8$X3, data8$X4)
MahalNPropVarC <- cor(data8$X5, data8$X5)
MahalNPropVarT <- cor(data8$Y, data8$X1)
MahalNPropMeanC <- cor(data8$Y, data8$X2)
MahalNMatchedN <- cor(data8$Y, data8$X3)
MahalNCorY.X5 <- cor(data8$Y, data8$X4)
MahalNCorY.X4 <- cor(data8$Y, data8$X2)
MahalNCorY.X3 <- cor(data8$Y, data8$X3)
MahalNCorY.X4 <- cor(data8$Y, data8$X4)
MahalNCorY.X5 <- cor(data8$Y, data8$X5)
MahalNMatchedN <- cor(data8$Y, data8$X5)
MahalNPropVarT <- mean(data8$TRUEprop[data8$group==1])
MahalNPropMeanT <- mean(data8$TRUEprop[data8$group==0])
MahalNPropVarT <- (sd(data8$TRUEprop[data8$group==1]))^2
MahalNPropVarC <- (sd(data8$TRUEprop[data8$group==0]))^2

#All variables AFTER matching Naive Prop Mahal
MahalNAvgX1Treat <- mean(data9$X1[data9$group==1])
MahalNAvgX2Treat <- mean(data9$X2[data9$group==1])
MahalNAvgL1Treat <- mean(data9$L1[data9$group==1])
MahalNAvgX3Treat <- mean(data9$X3[data9$group==1])
MahalNAvgX4Treat <- mean(data9$X4[data9$group==1])
MahalNAvgX5Treat <- mean(data9$X5[data9$group==1])
MahalNAvgX1Cont <- mean(data9$X1[data9$group==0])
MahalNAvgX2Cont <- mean(data9$X2[data9$group==0])
MahalNAvgL1Cont <- mean(data9$L1[data9$group==0])
MahalNAvgX3Cont <- mean(data9$X3[data9$group==0])
MahalNAvgX4Cont <- mean(data9$X4[data9$group==0])
MahalNAvgX5Cont <- mean(data9$X5[data9$group==0])
MahalNAvgYCont <- mean(data9$Y[data9$group==0])
MahalNSDX1Treat <- sd(data9$X1[data9$group==1])
MahalNSDX2Treat <- sd(data9$X2[data9$group==1])
MahalNSDL1Treat <- sd(data9$L1[data9$group==1])
MahalNSDX3Treat <- sd(data9$X3[data9$group==1])
MahalNSDX4Treat <- sd(data9$X4[data9$group==1])
MahalNSDX5Treat <- sd(data9$X5[data9$group==1])
MahalNSDYTreat <- sd(data9$Y[data9$group==1])
MahalNSDX1Cont <- sd(data9$X1[data9$group==0])
MahalNSDX2Cont <- sd(data9$X2[data9$group==0])
MahalNSDL1Cont <- sd(data9$L1[data9$group==0])
MahalNSDX3Cont <- sd(data9$X3[data9$group==0])
MahalNSDX4Cont <- sd(data9$X4[data9$group==0])
MahalNSDX5Cont <- sd(data9$X5[data9$group==0])
MahalNSDYCont <- sd(data9$Y[data9$group==0])
MahalNCorX1.X2 <- cor(data9$X1, data9$X2)
MahalNCorX1.X3 <- cor(data9$X1, data9$X3)
MahalNCorX1.X4 <- cor(data9$X1, data9$X4)
MahalNCorX1.X5 <- cor(data9$X1, data9$X5)
MahalNCorX2.X3 <- cor(data9$X2, data9$X3)
MahalNCorX2.X4 <- cor(data9$X2, data9$X4)
MahalNCorX2.X5 <- cor(data9$X2, data9$X5)
MahalNCorX3.X4 <- cor(data9$X3, data9$X4)
MahalNCorX3.X5 <- cor(data9$X3, data9$X5)
MahalNCorX4.X5 <- cor(data9$X4, data9$X5)
MahalNCorY.X1 <- cor(data9$Y, data9$X1)
MahalNCorY.X2 <- cor(data9$Y, data9$X2)
MahalNCorY.X3 <- cor(data9$Y, data9$X3)
MahalNCorY.X4 <- cor(data9$Y, data9$X4)
MahalNCorY.X5 <- cor(data9$Y, data9$X5)
MahalNMatchedN <- length(data9$Y[data9$group==0])
MahalNPropMeanT <- mean(data9$TRUEprop[data9$group==1])
MahalNPropMeanC <- mean(data9$TRUEprop[data9$group==0])
MahalNPropVarT <- (sd(data9$TRUEprop[data9$group==1]))^2
MahalNPropVarC <- (sd(data9$TRUEprop[data9$group==0]))^2

#Removing the data after each iteration
rm(data, data2, data3, data4, data5, data6, data7, data8, data9,
#In this step, we make a final data set by binding together all of the above variables.

```
Final.Sim.Data<-cbind(
  m.out1, m.out2, m.out3, m.out4, m.out5, m.out6, m.out7, m.out8, ps.sd, ps.sd2, finaldata, dataNNT, dataNNN, dataNCT, dataNNCN, dataoptT, dataoptN, dataMahalT, dataMahalN)
)
```

```
Final.Sim.Data<-cbind(
  m.out1, m.out2, m.out3, m.out4, m.out5, m.out6, m.out7, m.out8, ps.sd, ps.sd2, finaldata, dataNNT, dataNNN, dataNCT, dataNNCN, dataoptT, dataoptN, dataMahalT, dataMahalN)
)
```

```
Final.Sim.Data<-cbind(
  m.out1, m.out2, m.out3, m.out4, m.out5, m.out6, m.out7, m.out8, ps.sd, ps.sd2, finaldata, dataNNT, dataNNN, dataNCT, dataNNCN, dataoptT, dataoptN, dataMahalT, dataMahalN)
)
```
### Final step: Save it out AS A CSV FILE :)

```r
head(Final.Sim.Data)
str(Final.Sim.Data)
describe(Final.Sim.Data)
```

```r
write.table(Final.Sim.Data, file="Cond1SAME.csv", sep=" ", row.names=FALSE)
```
# Appendix B

## # ~~~~~~~~~~~~~~~~~~~~~~~~~~  Dissertation Code - Diff ME ~~~~~~~~~~~~~
## # ~~~~~~~~~~~~~~~~~~~~~~~~~~  Heather Harris ~~~~~~~~~~~~~~~~~~~~~~~~~~~
## # ~~~~~~~~~~~~~~~~~~~~~~~~~~                                  ~~~~~~~~~~~~~

### Quick note: Because I adapted this code from Christine and Jeanne's code
### I left (most) of the previous comments as one "##" so that they can be distinguished
### from
### my comments.

### Follow-up, because the script has been edited numerous times, this may no longer be
### the case.

### Before starting, I set my working directory to my desktop R folder
### because I plan on saving out a final PDF file of my plot(s) there.
setwd("C:/Users/heather.harris/Desktop/Run Two")
### In this step, I double-check that it worked
getwd()

install.packages("mvtnorm")
require(mvtnorm)
library(mvtnorm)
install.packages("optmatch")
require(optmatch)

### In the following steps, I compare the propensity scores created above with those
### calculated by
### the MatchIt package.
### I would first install the package then require it (if I didn't already have it)
install.packages("MatchIt")
require(MatchIt)
require(psych)

### In this step, I create the for loop for simulated data. First, I create places for the
### values I want to save out.

# Numeric diagnostics for PSM WITH COVARIATES
PropMeanTreatW <- rep(NA, 1000)
PropMeanContrW <- rep(NA, 1000)
PropVarTreatW  <- rep(NA, 1000)
PropVarContrW  <- rep(NA, 1000)

# Numeric diagnostics for PSM WITHOUT COVARIATES
PropMeanTreatWO <- rep(NA, 1000)
PropMeanContrWO <- rep(NA, 1000)
PropVarTreatWO  <- rep(NA, 1000)
PropVarContrWO  <- rep(NA, 1000)

# All variables BEFORE matching
pAvgX1Treat  <- rep(NA, 1000)
pAvgX2Treat  <- rep(NA, 1000)
pAvgL1Treat  <- rep(NA, 1000)
pAvgL2Treat  <- rep(NA, 1000)
pAvgX3Treat  <- rep(NA, 1000)
pAvgX4Treat  <- rep(NA, 1000)
pAvgX5Treat  <- rep(NA, 1000)
pAvgY1Treat  <- rep(NA, 1000)
pAvgX1Cont   <- rep(NA, 1000)
pAvgX2Cont   <- rep(NA, 1000)
pAvgL1Cont   <- rep(NA, 1000)
pAvgL2Cont   <- rep(NA, 1000)
<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>pAvgX3Cont</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>pAvgX4Cont</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>pAvgX5Cont</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>pAvgYCont</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>pSDX1Treat</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>pSDX2Treat</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>pSDL1Treat</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>pSDL2Treat</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>pSDX3Treat</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>pSDX4Treat</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>pSDX5Treat</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>pSDYTreat</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>pSDX1Cont</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>pSDX2Cont</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>pSDL1Cont</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>pSDL2Cont</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>pSDX3Cont</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>pSDX4Cont</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>pSDX5Cont</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>pSDYCont</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>pCorX1.X2</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>pCorX1.X3</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>pCorX1.X4</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>pCorX1.X5</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>pCorX2.X3</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>pCorX2.X4</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>pCorX2.X5</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>pCorX3.X4</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>pCorX3.X5</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>pCorX4.X5</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>pCorY.X1</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>pCorY.X2</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>pCorY.X3</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>pCorY.X4</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>pCorY.X5</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>pPropMeanTT</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>pPropMeanCT</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>pPropSDTT</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>pPropSDCT</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>pPropMeanTN</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>pPropMeanCN</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>pPropsDSTN</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>pPropsDCN</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>pCorL1X1</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>pCorL2X2</td>
<td>rep(NA, 1000)</td>
</tr>
</tbody>
</table>

# All variables AFTER matching True Prop NN

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NNTAvgX1Treat</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>NNTAvgX2Treat</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>NNTAvgX3Treat</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>NNTAvgX4Treat</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>NNTAvgX5Treat</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>NNTAvgX1Cont</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>NNTAvgX2Cont</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>NNTAvgX3Cont</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>NNTAvgX4Cont</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>NNTAvgX5Cont</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>NNTAvgYCont</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>NNTSDX1Treat</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>NNTSDX2Treat</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>NNTSDX3Treat</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>NNTSDX4Treat</td>
<td>rep(NA, 1000)</td>
</tr>
<tr>
<td>NNTSDX5Treat</td>
<td>rep(NA, 1000)</td>
</tr>
</tbody>
</table>
#All variables AFTER matching Naive Prop NN

```r
NNNAvgX1Treat <- rep(NA, 1000)
NNNAvgX2Treat <- rep(NA, 1000)
NNNAvgX3Treat <- rep(NA, 1000)
NNNAvgX4Treat <- rep(NA, 1000)
NNNAvgX5Treat <- rep(NA, 1000)
NNNAvgX1Cont <- rep(NA, 1000)
NNNAvgX2Cont <- rep(NA, 1000)
NNNAvgX3Cont <- rep(NA, 1000)
NNNAvgX4Cont <- rep(NA, 1000)
NNNAvgX5Cont <- rep(NA, 1000)
NNNAvgYCont <- rep(NA, 1000)
NNNAvgYTreat <- rep(NA, 1000)
NN NASDX1Cont <- rep(NA, 1000)
NN NASDX2Cont <- rep(NA, 1000)
NN NASDX3Cont <- rep(NA, 1000)
NN NASDX4Cont <- rep(NA, 1000)
NN NASDX5Cont <- rep(NA, 1000)
```

NNNCorX2.X4 <- rep(NA, 1000)
NNNCorX2.X5 <- rep(NA, 1000)
NNNCorX3.X4 <- rep(NA, 1000)
NNNCorX3.X5 <- rep(NA, 1000)
NNNCorX4.X5 <- rep(NA, 1000)
NNNCorY.X1 <- rep(NA, 1000)
NNNCorY.X2 <- rep(NA, 1000)
NNNCorY.X3 <- rep(NA, 1000)
NNNCorY.X4 <- rep(NA, 1000)
NNNCorY.X5 <- rep(NA, 1000)
NNNMatchedN <- rep(NA, 1000)
NNNCohensDW <- rep(NA, 1000)
NNNPropMeanT <- rep(NA, 1000)
NNNPropMeanC <- rep(NA, 1000)
NNNPropVarT <- rep(NA, 1000)
NNNPropVarC <- rep(NA, 1000)

#All variables AFTER matching True Prop NN w/ 0.2 Caliper
NNCTAvgX1Treat <- rep(NA, 1000)
NNCTAvgX2Treat <- rep(NA, 1000)
NNCTAvgL1Treat <- rep(NA, 1000)
NNCTAvgL2Treat <- rep(NA, 1000)
NNCTAvgX3Treat <- rep(NA, 1000)
NNCTAvgX4Treat <- rep(NA, 1000)
NNCTAvgX5Treat <- rep(NA, 1000)
NNCTAvgX5treat <- rep(NA, 1000)
NNCTAvgYtreat <- rep(NA, 1000)
NNCTAvgX1Cont <- rep(NA, 1000)
NNCTAvgX2Cont <- rep(NA, 1000)
NNCTAvgL1Cont <- rep(NA, 1000)
NNCTAvgL2Cont <- rep(NA, 1000)
NNCTAvgX3Cont <- rep(NA, 1000)
NNCTAvgX4Cont <- rep(NA, 1000)
NNCTAvgX5Cont <- rep(NA, 1000)
NNCTAvgYCont <- rep(NA, 1000)
NNCTSDX1Treat <- rep(NA, 1000)
NNCTSDX2Treat <- rep(NA, 1000)
NNCTSDL1Treat <- rep(NA, 1000)
NNCTSDL2Treat <- rep(NA, 1000)
NNCTSDX3Treat <- rep(NA, 1000)
NNCTSDX4Treat <- rep(NA, 1000)
NNCTSDX5Treat <- rep(NA, 1000)
NNCTSDYtreat <- rep(NA, 1000)
NNCTSDX1Cont <- rep(NA, 1000)
NNCTSDX2Cont <- rep(NA, 1000)
NNCTSDL1Cont <- rep(NA, 1000)
NNCTSDL2Cont <- rep(NA, 1000)
NNCTSDX3Cont <- rep(NA, 1000)
NNCTSDX4Cont <- rep(NA, 1000)
NNCTSDX5Cont <- rep(NA, 1000)
NNCTSDYCont <- rep(NA, 1000)
NNCTCorX1.X2 <- rep(NA, 1000)
NNCTCorX1.X3 <- rep(NA, 1000)
NNCTCorX1.X4 <- rep(NA, 1000)
NNCTCorX1.X5 <- rep(NA, 1000)
NNCTCorX2.X3 <- rep(NA, 1000)
NNCTCorX2.X4 <- rep(NA, 1000)
NNCTCorX2.X5 <- rep(NA, 1000)
NNCTCorX3.X4 <- rep(NA, 1000)
NNCTCorX3.X5 <- rep(NA, 1000)
NNCTCorX4.X5 <- rep(NA, 1000)
NNCTCorY.X1 <- rep(NA, 1000)
NNCTCorY.X3 <- rep(NA, 1000)
NNCTCorY.X2 <- rep(NA, 1000)
NNCTCorY.X4 <- rep(NA, 1000)
NNCTCorY.X5 <- rep(NA, 1000)
NNCTMatchedN <- rep(NA, 1000)
NNCTCohensDW <- rep(NA, 1000)
NNCTPropMeanT <- rep(NA, 1000)
NNCTPropMeanC <- rep(NA, 1000)
NNCTPropVarT <- rep(NA, 1000)
NNCTPropVarC <- rep(NA, 1000)

#All variables AFTER matching Naive Prop NN w/ 0.2 Caliper
NNCAvgX1Treat <- rep(NA, 1000)
NNCAvgX2Treat <- rep(NA, 1000)
NNCAvgL1Treat <- rep(NA, 1000)
NNCAvgL2Treat <- rep(NA, 1000)
NNCAvgX3Treat <- rep(NA, 1000)
NNCAvgX4Treat <- rep(NA, 1000)
NNCAvgX5Treat <- rep(NA, 1000)
NNCAvgYTreat <- rep(NA, 1000)
NNCAvgX1Cont <- rep(NA, 1000)
NNCAvgX2Cont <- rep(NA, 1000)
NNCAvgL1Cont <- rep(NA, 1000)
NNCAvgL2Cont <- rep(NA, 1000)
NNCAvgX3Cont <- rep(NA, 1000)
NNCAvgX4Cont <- rep(NA, 1000)
NNCAvgX5Cont <- rep(NA, 1000)
NNCAvgYCont <- rep(NA, 1000)
NNCNSDX1Treat <- rep(NA, 1000)
NNCNSDX2Treat <- rep(NA, 1000)
NNCNSDL1Treat <- rep(NA, 1000)
NNCNSDL2Treat <- rep(NA, 1000)
NNCNSDX3Treat <- rep(NA, 1000)
NNCNSDX4Treat <- rep(NA, 1000)
NNCNSDX5Treat <- rep(NA, 1000)
NNCNSDYTreat <- rep(NA, 1000)
NNCNSDX1Cont <- rep(NA, 1000)
NNCNSDX2Cont <- rep(NA, 1000)
NNCNSDL1Cont <- rep(NA, 1000)
NNCNSDL2Cont <- rep(NA, 1000)
NNCNSDX3Cont <- rep(NA, 1000)
NNCNSDX4Cont <- rep(NA, 1000)
NNCNSDX5Cont <- rep(NA, 1000)
NNCNSDYCont <- rep(NA, 1000)
NNCorX1.X2 <- rep(NA, 1000)
NNCorX1.X3 <- rep(NA, 1000)
NNCorX1.X4 <- rep(NA, 1000)
NNCorX1.X5 <- rep(NA, 1000)
NNCorX2.X3 <- rep(NA, 1000)
NNCorX2.X4 <- rep(NA, 1000)
NNCorX2.X5 <- rep(NA, 1000)
NNCorX3.X4 <- rep(NA, 1000)
NNCorX3.X5 <- rep(NA, 1000)
NNCorX4.X5 <- rep(NA, 1000)
NNCorY.X1 <- rep(NA, 1000)
NNCorY.X2 <- rep(NA, 1000)
NNCorY.X3 <- rep(NA, 1000)
NNCorY.X4 <- rep(NA, 1000)
NNCorY.X5 <- rep(NA, 1000)
NMatchedN <- rep(NA, 1000)
NCorCohensD <- rep(NA, 1000)
NPropMeanT <- rep(NA, 1000)
NPropMeanC <- rep(NA, 1000)
NPropVarT <- rep(NA, 1000)
NPropVarC <- rep(NA, 1000)

#All variables AFTER matching True Prop Optimal
optTAvgX1Treat <- rep(NA, 1000)
optTAvgX2Treat <- rep(NA, 1000)
optTAvgL1Treat <- rep(NA, 1000)
optTAvgL2Treat <- rep(NA, 1000)
optTAvgX3Treat <- rep(NA, 1000)
optTAvgX4Treat <- rep(NA, 1000)
optTAvgX5Treat <- rep(NA, 1000)
optTAvgYTreat <- rep(NA, 1000)
optTAvgX1Cont <- rep(NA, 1000)
optTAvgX2Cont <- rep(NA, 1000)
optTAvgL1Cont <- rep(NA, 1000)
optTAvgL2Cont <- rep(NA, 1000)
optTAvgX3Cont <- rep(NA, 1000)
optTAvgX4Cont <- rep(NA, 1000)
optTAvgX5Cont <- rep(NA, 1000)
optTAvgYCont <- rep(NA, 1000)
optTSDX1Treat <- rep(NA, 1000)
optTSDX2Treat <- rep(NA, 1000)
optTSDX3Treat <- rep(NA, 1000)
optTSDX4Treat <- rep(NA, 1000)
optTSDX5Treat <- rep(NA, 1000)
optTSDYTreat <- rep(NA, 1000)
optTSDX1Cont <- rep(NA, 1000)
optTSDX2Cont <- rep(NA, 1000)
optTSDX3Cont <- rep(NA, 1000)
optTSDYCont <- rep(NA, 1000)
optTCorX1.X2 <- rep(NA, 1000)
optTCorX1.X3 <- rep(NA, 1000)
optTCorX1.X4 <- rep(NA, 1000)
optTCorX1.X5 <- rep(NA, 1000)
optTCorX2.X3 <- rep(NA, 1000)
optTCorX2.X4 <- rep(NA, 1000)
optTCorX2.X5 <- rep(NA, 1000)
optTCorX3.X4 <- rep(NA, 1000)
optTCorX3.X5 <- rep(NA, 1000)
optTCorX4.X5 <- rep(NA, 1000)
optTCorY.X1 <- rep(NA, 1000)
optTCorY.X2 <- rep(NA, 1000)
optTCorY.X3 <- rep(NA, 1000)
optTCorY.X4 <- rep(NA, 1000)
optTCorY.X5 <- rep(NA, 1000)
optTMatchedN <- rep(NA, 1000)
optTCohensDW <- rep(NA, 1000)
optTPropMeanT <- rep(NA, 1000)
optTPropMeanC <- rep(NA, 1000)
optTPropVarT <- rep(NA, 1000)
optTPropVarC <- rep(NA, 1000)

#All variables AFTER matching Naive Prop Optimal
optNAvgX1Treat <- rep(NA, 1000)
optNAvgX2Treat <- rep(NA, 1000)
optNAvgL1Treat <- rep(NA, 1000)
optNAvgL2Treat <- rep(NA, 1000)
optNAvgX3Treat <- rep(NA, 1000)
optNAvgX4Treat <- rep(NA, 1000)
optNAvgX5Treat <- rep(NA, 1000)
optNAvgYTreat <- rep(NA, 1000)
optNAvgX1Cont <- rep(NA, 1000)
optNAvgL1Cont <- rep(NA, 1000)
optNAvgL2Cont <- rep(NA, 1000)
optNAvgX3Cont <- rep(NA, 1000)
optNAvgX4Cont <- rep(NA, 1000)
optNAvgX5Cont <- rep(NA, 1000)
optNAvgYCont <- rep(NA, 1000)
optNSDX1Treat <- rep(NA, 1000)
optNSDX2Treat <- rep(NA, 1000)
optNSDYTreat <- rep(NA, 1000)
optNSDX1Treat <- rep(NA, 1000)
optNSDYTreat <- rep(NA, 1000)
optNSDX1Cont <- rep(NA, 1000)
optNSDX2Cont <- rep(NA, 1000)
optNSDL1Cont <- rep(NA, 1000)
optNSDX2Cont <- rep(NA, 1000)
optNSDX3Cont <- rep(NA, 1000)
optNSDX4Cont <- rep(NA, 1000)
optNSDX5Cont <- rep(NA, 1000)
optNSDYCont <- rep(NA, 1000)
optNCorX1.X2 <- rep(NA, 1000)
optNCorX1.X3 <- rep(NA, 1000)
optNCorX1.X4 <- rep(NA, 1000)
optNCorX1.X5 <- rep(NA, 1000)
optNCorX2.X3 <- rep(NA, 1000)
optNCorX2.X4 <- rep(NA, 1000)
optNCorX2.X5 <- rep(NA, 1000)
optNCorX3.X4 <- rep(NA, 1000)
optNCorX3.X5 <- rep(NA, 1000)
optNCorX4.X5 <- rep(NA, 1000)
optNCorY.X1 <- rep(NA, 1000)
optNCorY.X2 <- rep(NA, 1000)
optNCorY.X3 <- rep(NA, 1000)
optNCorY.X4 <- rep(NA, 1000)
optNCorY.X5 <- rep(NA, 1000)
optNMatchedN <- rep(NA, 1000)
optNCohensDW <- rep(NA, 1000)
optNPropMeanT <- rep(NA, 1000)
optNPropMeanC <- rep(NA, 1000)
optNPropVarT <- rep(NA, 1000)
optNPropVarC <- rep(NA, 1000)

#All variables AFTER matching True Prop Mahal
MahaltAvx1Treat <- rep(NA, 1000)
MahaltAvx2Treat <- rep(NA, 1000)
MahaltAvx1Treat <- rep(NA, 1000)
MahaltAvx2Treat <- rep(NA, 1000)
MahaltAvx3Treat <- rep(NA, 1000)
MahaltAvx4Treat <- rep(NA, 1000)
MahaltAvx5Treat <- rep(NA, 1000)
MahaltAvyTreat <- rep(NA, 1000)
MahaltAvx1Cont <- rep(NA, 1000)
MahaltAvx2Cont <- rep(NA, 1000)
MahaltAvx3Cont <- rep(NA, 1000)
MahaltAvx4Cont <- rep(NA, 1000)
MahaltAvx5Cont <- rep(NA, 1000)
MahaltAvyCont <- rep(NA, 1000)
MahaltSDX1Treat <- rep(NA, 1000)
MahaltSDX2Treat <- rep(NA, 1000)
MahaltSDL1Treat <- rep(NA, 1000)
MahaltSDL2Treat <- rep(NA, 1000)
MahaltSDX3Treat <- rep(NA, 1000)
MahaltSDX4Treat <- rep(NA, 1000)
MahaltSDX5Treat <- rep(NA, 1000)
MahaltSDYTreat <- rep(NA, 1000)
MahaltSDX1Cont <- rep(NA, 1000)
MahaltSDX2Cont <- rep(NA, 1000)
MahaltSDL1Cont <- rep(NA, 1000)
MahaltSDL2Cont <- rep(NA, 1000)
MahaltSDX3Cont <- rep(NA, 1000)
MahaltSDX4Cont <- rep(NA, 1000)
MahaltSDX5Cont <- rep(NA, 1000)
MahaltSDYCont <- rep(NA, 1000)
MahaltCorX1.X2 <- rep(NA, 1000)
MahaltCorX1.X3 <- rep(NA, 1000)
MahaltCorX1.X4 <- rep(NA, 1000)
MahaltCorX1.X5 <- rep(NA, 1000)
MahaltCorX2.X3 <- rep(NA, 1000)
MahaltCorX2.X4 <- rep(NA, 1000)
MahaltCorX2.X5 <- rep(NA, 1000)
MahaltCorX3.X4 <- rep(NA, 1000)
MahalTCorX3.X5 <- rep(NA, 1000)
MahalTCorX4.X5 <- rep(NA, 1000)
MahalTCorY.X1 <- rep(NA, 1000)
MahalTCorY.X2 <- rep(NA, 1000)
MahalTCorY.X3 <- rep(NA, 1000)
MahalTCorY.X4 <- rep(NA, 1000)
MahalTCorY.X5 <- rep(NA, 1000)
MahalTMatchedN <- rep(NA, 1000)
MahalTCohenSDW <- rep(NA, 1000)
MahalTPropMeanT <- rep(NA, 1000)
MahalTPropMeanC <- rep(NA, 1000)
MahalTPropVarT <- rep(NA, 1000)
MahalTPropVarC <- rep(NA, 1000)

# All variables AFTER matching Naive Prop Mahal
MahalNAvgX1Treat <- rep(NA, 1000)
MahalNAvgX2Treat <- rep(NA, 1000)
MahalNAvgX1L1Treat <- rep(NA, 1000)
MahalNAvgX2L1Treat <- rep(NA, 1000)
MahalNAvgX3Treat <- rep(NA, 1000)
MahalNAvgX4Treat <- rep(NA, 1000)
MahalNAvgX5Treat <- rep(NA, 1000)
MahalNAvgY.Treat <- rep(NA, 1000)
MahalNAvgX1Cont <- rep(NA, 1000)
MahalNAvgX2Cont <- rep(NA, 1000)
MahalNAvgX1L1Cont <- rep(NA, 1000)
MahalNAvgX2L1Cont <- rep(NA, 1000)
MahalNAvgX3Cont <- rep(NA, 1000)
MahalNAvgX4Cont <- rep(NA, 1000)
MahalNAvgX5Cont <- rep(NA, 1000)
MahalNAvgX1T <- rep(NA, 1000)
MahalNAvgX2T <- rep(NA, 1000)
MahalNAvgX1L1T <- rep(NA, 1000)
MahalNAvgX2L1T <- rep(NA, 1000)
MahalNAvgX3T <- rep(NA, 1000)
MahalNAvgX4T <- rep(NA, 1000)
MahalNAvgX5T <- rep(NA, 1000)
MahalNSDX1Treat <- rep(NA, 1000)
MahalNSDX2Treat <- rep(NA, 1000)
MahalNSDX1L1Treat <- rep(NA, 1000)
MahalNSDX2L1Treat <- rep(NA, 1000)
MahalNSDX3Treat <- rep(NA, 1000)
MahalNSDX4Treat <- rep(NA, 1000)
MahalNSDX5Treat <- rep(NA, 1000)
MahalNSDY.Treat <- rep(NA, 1000)
MahalNSDX1Cont <- rep(NA, 1000)
MahalNSDX2Cont <- rep(NA, 1000)
MahalNSDX3Cont <- rep(NA, 1000)
MahalNSDX4Cont <- rep(NA, 1000)
MahalNSDX5Cont <- rep(NA, 1000)
MahalNSDYCont <- rep(NA, 1000)
MahalNCorX1.X2 <- rep(NA, 1000)
MahalNCorX1.X3 <- rep(NA, 1000)
MahalNCorX1.X4 <- rep(NA, 1000)
MahalNCorX1.X5 <- rep(NA, 1000)
MahalNCorX2.X3 <- rep(NA, 1000)
MahalNCorX2.X4 <- rep(NA, 1000)
MahalNCorX2.X5 <- rep(NA, 1000)
MahalNCorX3.X4 <- rep(NA, 1000)
MahalNCorX3.X5 <- rep(NA, 1000)
MahalNCorX4.X5 <- rep(NA, 1000)
MahalNCorY.X1 <- rep(NA, 1000)
MahalNCorY.X2 <- rep(NA, 1000)
MahalNCorY.X3 <- rep(NA, 1000)
MahalNCorY.X4 <- rep(NA, 1000)
MahalNCorY.X5 <- rep(NA, 1000)
MahalNMatchedN <- rep(NA, 1000)
MahalNCohensD.W <- rep(NA, 1000)
MahalNPropMeanT <- rep(NA, 1000)
MahalNPropMeanC <- rep(NA, 1000)
MahalNPropVarT <- rep(NA, 1000)
MahalNPropVarC <- rep(NA, 1000)
# In case I want to figure out how long it will take. :)

```r
str(mSDYCont)
proc.time()
```

```r
now<-proc.time()
set.seed(21)
for(i in 1:1000){

# 1. Simulating Examinees

# correlations among covariates

```r
corrX=matrix(c(1, .58, .42, .49, .60,
                .58, 1, .56, .44, .41,
                .42, .56, 1, .59, .44,
                .49, .44, .59, 1, .51,
                .60, .41, .44, .51, 1),5,5)
```

#correlation between each covariate and continuous (latent) propensity;

```r
corrXlp=c(.4, .4, .4, .4, .4)
```

#calculate regression coefficients

```r
Pcoef=solve(corrX) %*% corrXlp #solve means inverse
```

#variance in latent propensity explained by covariates

```r
temp= Pcoef %*% t(Pcoef) * corrX
varExpP=sum(temp) #varExpP
```

```r
library(mvtnorm)
library(psych)
# simulate covariates
Nexaminee=1000
Nrep=1000
```

```r
X=rmvnorm(Nexaminee, rep(0,5), corrX, method="chol")
```

#the portion of latent propensity accounted for by the covariates

```r
noErr=as.vector(X %*% Pcoef)
```

#should be close to varExpP

```r
Rsq=varExpP/(1+varExpP) #because the error variance of a probit is 1
```

#probability that examinee is above cutoff

```r
dist=noErr-mycut/sqrt(1-Rsq)
```

```r
trueprop=pnorm(dist)
```

#randomly assign each person Nrep times

```r
randraw=matrix(runif(Nexaminee*Nrep),nrow=Nexaminee,ncol=Nrep)
```

```r
trueprop=ifelse(trueprop>randraw,1,0)
```

```r
meangrp=rowMeans(group)
```

```r
```
# using first replication, check if Pcoef is estimated well
dataZ <- data.frame(X, group[,1])
#str(dataZ)
#describe(temp1)
#head(temp1)
#describeBy(temp1, temp1$group)
#describe(dataZ)

# Saving out the error-free variables
dataZ$L1 <- dataZ$X1
dataZ$L2 <- dataZ$X2

# Now breaking into to two groups to add differential measurement error.
dataT <- dataZ[which(dataZ$group == 1),]
describe(dataT)
dataC <- dataZ[which(dataZ$group == 0),]
describe(dataC)

# Creating the SD for the error term to create measurement error FOR THE TREATMENT GROUP.
ErrorSDt <- sqrt((0.2)*(1/0.8)) #Reliability is kept at 0.8 across all conditions for the treatment group.

# ERROR FOR X1 TREATMENT GROUP
nT <- nrow(dataT)
v1 <- rnorm(nT, mean = 0, sd = ErrorSDt)
dataT$X1 <- dataT$L1 + v1
cor(dataT$L1, dataT$X1)

# ERROR FOR X2 TREATMENT GROUP
v2 <- rnorm(nT, mean = 0, sd = ErrorSDt)
dataT$X2 <- dataT$L2 + v2
cor(dataT$L2, dataT$X2)

# Creating the SD for the error term to create measurement error FOR THE CONTROL GROUP.
#ErrorSDc <- sqrt((0.1)*(1/0.9)) #Condition 1
#ErrorSDc <- sqrt((0.2)*(1/0.8)) #Condition 2
#ErrorSDc <- sqrt((0.3)*(1/0.7)) #Condition 3
#ErrorSDc <- sqrt((0.4)*(1/0.6)) #Condition 4
#ErrorSDc <- sqrt((0.5)*(1/0.5)) #Condition 5
ErrorSDc <- sqrt((0.6)*(1/0.4)) #Condition 6

# ERROR FOR X1 CONTROL GROUP
nC <- nrow(dataC)
v3 <- rnorm(nC, mean = 0, sd = ErrorSDc)
dataC$X1 <- dataC$L1 + v3
cor(dataC$L1, dataC$X1)

# ERROR FOR X2 CONTROL GROUP
v4 <- rnorm(nC, mean = 0, sd = ErrorSDc)
dataC$X2 <- dataC$L2 + v4
cor(dataC$L2, dataC$X2)

# I'm now combining the two groups back together.
CombData <- rbind(dataT, dataC)
describe(CombData)
head(CombData)

## Now transforming the simulated variables from a mean of 0 and SD of 1 to a mean of 25 and an SD of 5.
L1 <- CombData$L1 * 5 + 25
L2 <- CombData$L2 + 25
X1 <- CombData$X1 + 25
X2 <- CombData$X2 + 25
X3 <- CombData$X3 + 25
X4 <- CombData$X4 + 25
X5 <- CombData$X5 + 25

group <- CombData$group
finaldata <- cbind(L1, L2, X1, X2, X3, X4, X5, group)
finaldata <- as.data.frame(finaldata)
Pb <- glm(formula = group ~ L1 + L2 + X3 + X4 + X5, data = finaldata, family = binomial)

finaldata$TRUEprop <- predict(Pb, type = "response")
finaldata$Y <- Y

# Now adding naive propensity scores using the error-prone covariates
Pb2 <- glm(formula = group ~ X1 + X2 + X3 + X4 + X5, data = finaldata, family = binomial)
finaldata$NAIVEprop <- predict(Pb2, type = "response")

describe(finaldata)
describeBy(finaldata, finaldata$group)
pairs.panels(finaldata[which(finaldata$group == 1),])
pairs.panels(finaldata[which(finaldata$group == 0),])
pairs.panels(finaldata)

#head(finaldata)
write.table(finaldata, file = "Cond6DIFF.csv", sep = "", row.names = FALSE)

# 2. Propensity Score Matching (NN w/o caliper)
# CONDUCTING PSM WITH TRUE PROPENSITY SCORES
m.out1 <- matchit(finaldata$group ~ finaldata$L1 + finaldata$L2 + finaldata$X3 + finaldata$X4 + finaldata$X5, data = finaldata, method = "nearest", ratio = 1)
m.out1

#plot(m.out1, type = "jitter") # propensity score locations
#plot(m.out1, type = "hist")
dataNNT <- match.data(m.out1)
describe(dataNNT)

# CONDUCTING PSM WITH NAIVE PROPENSITY SCORES
m.out2 <- matchit(finaldata$group ~ finaldata$X1 + finaldata$X2 + finaldata$X3 + finaldata$X4 + finaldata$X5, data = finaldata, method = "nearest", ratio = 1)
m.out2

#plot(m.out2, type = "jitter") # propensity score locations
#plot(m.out2, type = "hist")
dataNNN <- match.data(m.out2)
describe(dataNNN)
# 3. Propensity Score Matching (NN with 0.2 caliper)
# CONDUCTING PSM WITH TRUE PROPENSITY SCORES
ps.sd = sd(finaldata$TRUEprop)
#ps.sd
m.out3 = matchit(finaldata$group ~ finaldata$L1 + finaldata$L2 + finaldata$X3 + finaldata$X4 + finaldata$X5, data = finaldata, method = "nearest", caliper = 0.2 * ps.sd, ratio = 1)
#m.out3$distance
#plot(m.out3, type = "jitter") # propensity score locations
#plot(m.out3, type = "hist")
dataNNCT = match.data(m.out3)
#describe(dataNNCT)

# CONDUCTING PSM WITH NAIVE PROPENSITY SCORES
ps.sd2 = sd(finaldata$NAIVEprop)
#ps.sd
m.out4 = matchit(finaldata$group ~ finaldata$X1 + finaldata$X2 + finaldata$X3 + finaldata$X4 + finaldata$X5, data = finaldata, method = "nearest", caliper = 0.2 * ps.sd2, ratio = 1)
#m.out4$distance
#plot(m.out4, type = "jitter") # propensity score locations
#plot(m.out4, type = "hist")
dataNNCN = match.data(m.out4)
#describe(dataNNCN)

# 4. Propensity Score Matching (Optimal)
# CONDUCTING PSM WITH TRUE PROPENSITY SCORES
m.out5 = matchit(finaldata$group ~ finaldata$L1 + finaldata$L2 + finaldata$X3 + finaldata$X4 + finaldata$X5, data = finaldata, method = "optimal", ratio = 1)
#m.out5$distance
#plot(m.out5, type = "jitter") # propensity score locations
#plot(m.out5, type = "hist")
dataoptT = match.data(m.out5)
#describe(dataoptT)

# CONDUCTING PSM WITH NAIVE PROPENSITY SCORES
m.out6 = matchit(finaldata$group ~ finaldata$X1 + finaldata$X2 + finaldata$X3 + finaldata$X4 + finaldata$X5, data = finaldata, method = "optimal", ratio = 1)
#m.out6$distance
#plot(m.out6, type = "jitter") # propensity score locations
#plot(m.out6, type = "hist")
dataoptN = match.data(m.out6)
#describe(dataoptN)

# 5. Propensity Score Matching (Mahal Matching)
# CONDUCTING PSM WITH TRUE PROPENSITY SCORES
m.out7 = matchit(finaldata$group ~ finaldata$L1 + finaldata$L2 + finaldata$X3 + finaldata$X4 + finaldata$X5, finaldata$TRUEprop, data = finaldata, method = "nearest", distance = "mahalanobis")
dataMahalT = match.data(m.out7)
#describe(dataMahalT)

# CONDUCTING PSM WITH NAIVE PROPENSITY SCORES
```r
m.out8 <- matchit(finaldata$group ~ finaldata$X1 + finaldata$X2 + finaldata$X3 + finaldata$X4 + finaldata$X5 + finaldata$SNIVFprop, data = finaldata, method = "nearest", distance = "mahalanobis")

dataMahalN <- match.data(m.out8)
# describe(dataMahalN)

# 3. Saving Out Diagnostics
# 4. ---------------------------------------------

data1 <- finaldata
data2 <- dataNNN
data3 <- dataNNCT
data4 <- dataNCN
data5 <- dataNCN
data6 <- dataoptT
data7 <- dataoptN
data8 <- dataMahalT
data9 <- dataMahalN

# All means and SDs for variables as simulated.
pAvgX1Treat[i] <- mean(finaldata$X1[finaldata$group==1])
pAvgX2Treat[i] <- mean(finaldata$X2[finaldata$group==1])
pAvgX1Cont[i] <- mean(finaldata$X1[finaldata$group==0])
pAvgX2Cont[i] <- mean(finaldata$X2[finaldata$group==0])
pAvgX1Cont[i] <- mean(finaldata$X1[finaldata$group==1])
pAvgX2Cont[i] <- mean(finaldata$X2[finaldata$group==0])
pAvgX1Treat[i] <- mean(finaldata$X1[finaldata$group==1])
pAvgX2Treat[i] <- mean(finaldata$X2[finaldata$group==1])
pAvgX1Cont[i] <- mean(finaldata$X1[finaldata$group==1])
pAvgX2Cont[i] <- mean(finaldata$X2[finaldata$group==0])
pAvgX1Treat[i] <- mean(finaldata$X1[finaldata$group==1])
pAvgX2Treat[i] <- mean(finaldata$X2[finaldata$group==1])
pAvgX1Cont[i] <- mean(finaldata$X1[finaldata$group==1])
pAvgX2Cont[i] <- mean(finaldata$X2[finaldata$group==0])
pAvgX1Treat[i] <- mean(finaldata$X1[finaldata$group==1])
pAvgX2Treat[i] <- mean(finaldata$X2[finaldata$group==1])
pAvgX1Cont[i] <- mean(finaldata$X1[finaldata$group==1])
pAvgX2Cont[i] <- mean(finaldata$X2[finaldata$group==0])
pAvgX1Treat[i] <- mean(finaldata$X1[finaldata$group==1])
pAvgX2Treat[i] <- mean(finaldata$X2[finaldata$group==1])
pAvgX1Cont[i] <- mean(finaldata$X1[finaldata$group==1])
pAvgX2Cont[i] <- mean(finaldata$X2[finaldata$group==0])
pAvgX1Treat[i] <- mean(finaldata$X1[finaldata$group==1])
pAvgX2Treat[i] <- mean(finaldata$X2[finaldata$group==1])
pAvgX1Cont[i] <- mean(finaldata$X1[finaldata$group==1])
pAvgX2Cont[i] <- mean(finaldata$X2[finaldata$group==0])
pAvgX1Treat[i] <- mean(finaldata$X1[finaldata$group==1])
pAvgX2Treat[i] <- mean(finaldata$X2[finaldata$group==1])
pAvgX1Cont[i] <- mean(finaldata$X1[finaldata$group==1])
pAvgX2Cont[i] <- mean(finaldata$X2[finaldata$group==0])
pAvgX1Treat[i] <- mean(finaldata$X1[finaldata$group==1])
pAvgX2Treat[i] <- mean(finaldata$X2[finaldata$group==1])
pAvgX1Cont[i] <- mean(finaldata$X1[finaldata$group==1])
pAvgX2Cont[i] <- mean(finaldata$X2[finaldata$group==0])
pAvgX1Treat[i] <- mean(finaldata$X1[finaldata$group==1])
pAvgX2Treat[i] <- mean(finaldata$X2[finaldata$group==1])
pAvgX1Cont[i] <- mean(finaldata$X1[finaldata$group==1])
pAvgX2Cont[i] <- mean(finaldata$X2[finaldata$group==0])
pAvgX1Treat[i] <- mean(finaldata$X1[finaldata$group==1])
pAvgX2Treat[i] <- mean(finaldata$X2[finaldata$group==1])
pAvgX1Cont[i] <- mean(finaldata$X1[finaldata$group==1])
pAvgX2Cont[i] <- mean(finaldata$X2[finaldata$group==0])
pAvgX1Treat[i] <- mean(finaldata$X1[finaldata$group==1])
pAvgX2Treat[i] <- mean(finaldata$X2[finaldata$group==1])
pAvgX1Cont[i] <- mean(finaldata$X1[finaldata$group==1])
pAvgX2Cont[i] <- mean(finaldata$X2[finaldata$group==0])
pAvgX1Treat[i] <- mean(finaldata$X1[finaldata$group==1])
pAvgX2Treat[i] <- mean(finaldata$X2[finaldata$group==1])
pAvgX1Cont[i] <- mean(finaldata$X1[finaldata$group==1])
pAvgX2Cont[i] <- mean(finaldata$X2[finaldata$group==0])
pAvgX1Treat[i] <- mean(finaldata$X1[finaldata$group==1])
pAvgX2Treat[i] <- mean(finaldata$X2[finaldata$group==1])
pAvgX1Cont[i] <- mean(finaldata$X1[finaldata$group==1])
pAvgX2Cont[i] <- mean(finaldata$X2[finaldata$group==0])
```

#All variables AFTER matching True Prop NN

#All variables AFTER matching Naive Prop NN

```
pPropMeanTT[i] <- mean(data$TRUEprop[data$group==1])
pPropMeanTN[i] <- mean(data$TRUEprop[data$group==0])
pPropSDTT[i] <- sd(data$TRUEprop[data$group==1])
pPropSDTN[i] <- sd(data$TRUEprop[data$group==0])
pPropMeanNN[i] <- mean(data$NAIVEprop[data$group==1])
pPropMeanCN[i] <- mean(data$NAIVEprop[data$group==0])
pPropsDTN[i] <- sd(data$NAIVEprop[data$group==1])
pPropsDCN[i] <- sd(data$NAIVEprop[data$group==0])
pCorX1X1[i] <- cor(L1, X1)
pCorL2X2[i] <- cor(L2, X2)
```

#All variables AFTER matching Naive Prop NN

```
NNTAvgX5Treat[i] <- mean(data2$X5[data2$group==1])
NNTAvgL4Treat[i] <- mean(data2$L4[data2$group==1])
NNTAvgL3Treat[i] <- mean(data2$L3[data2$group==1])
NNTAvgL2Treat[i] <- mean(data2$L2[data2$group==1])
NNTAvgL1Treat[i] <- mean(data2$L1[data2$group==1])
NNTAvgX5Cont[i] <- mean(data2$X5[data2$group==0])
NNTAvgL4Cont[i] <- mean(data2$L4[data2$group==0])
NNTAvgL3Cont[i] <- mean(data2$L3[data2$group==0])
NNTAvgL2Cont[i] <- mean(data2$L2[data2$group==0])
NNTAvgL1Cont[i] <- mean(data2$L1[data2$group==0])
NNTAvgX5[i] <- mean(data2$X5[data2$group==0])
NNTAvgL4[i] <- mean(data2$L4[data2$group==0])
NNTAvgL3[i] <- mean(data2$L3[data2$group==0])
NNTAvgL2[i] <- mean(data2$L2[data2$group==0])
NNTAvgL1[i] <- mean(data2$L1[data2$group==0])
NNTAvgX5Treat[,1] <- mean(data2$X5[which(data2$group==1)])
NNTAvgL4Treat[,1] <- mean(data2$L4[which(data2$group==1)])
NNTAvgL3Treat[,1] <- mean(data2$L3[which(data2$group==1)])
NNTAvgL2Treat[,1] <- mean(data2$L2[which(data2$group==1)])
NNTAvgL1Treat[,1] <- mean(data2$L1[which(data2$group==1)])
NNTAvgX5Cont[,1] <- mean(data2$X5[which(data2$group==0)])
NNTAvgL4Cont[,1] <- mean(data2$L4[which(data2$group==0)])
NNTAvgL3Cont[,1] <- mean(data2$L3[which(data2$group==0)])
NNTAvgL2Cont[,1] <- mean(data2$L2[which(data2$group==0)])
NNTAvgL1Cont[,1] <- mean(data2$L1[which(data2$group==0)])
NNTAvgX5[,1] <- mean(data2$X5[which(data2$group==0)])
NNTAvgL4[,1] <- mean(data2$L4[which(data2$group==0)])
NNTAvgL3[,1] <- mean(data2$L3[which(data2$group==0)])
NNTAvgL2[,1] <- mean(data2$L2[which(data2$group==0)])
NNTAvgL1[,1] <- mean(data2$L1[which(data2$group==0)])
NNTAvgX5Treat[,2] <- mean(data2$X5[which(data2$group==0)])
NNTAvgL4Treat[,2] <- mean(data2$L4[which(data2$group==0)])
NNTAvgL3Treat[,2] <- mean(data2$L3[which(data2$group==0)])
NNTAvgL2Treat[,2] <- mean(data2$L2[which(data2$group==0)])
NNTAvgL1Treat[,2] <- mean(data2$L1[which(data2$group==0)])
NNTAvgX5Cont[,2] <- mean(data2$X5[which(data2$group==1)])
NNTAvgL4Cont[,2] <- mean(data2$L4[which(data2$group==1)])
NNTAvgL3Cont[,2] <- mean(data2$L3[which(data2$group==1)])
NNTAvgL2Cont[,2] <- mean(data2$L2[which(data2$group==1)])
NNTAvgL1Cont[,2] <- mean(data2$L1[which(data2$group==1)])
NNTAvgX5[,2] <- mean(data2$X5[which(data2$group==1)])
NNTAvgL4[,2] <- mean(data2$L4[which(data2$group==1)])
NNTAvgL3[,2] <- mean(data2$L3[which(data2$group==1)])
NNTAvgL2[,2] <- mean(data2$L2[which(data2$group==1)])
NNTAvgL1[,2] <- mean(data2$L1[which(data2$group==1)])
NNTPropMeanT <- mean(data2$TRUEprop[data2$group==1])
NNTPropMeanC <- mean(data2$TRUEprop[data2$group==0])
NNTPropVarT <- (sd(data2$TRUEprop[data2$group==1]))^2
NNTPropVarC <- (sd(data2$TRUEprop[data2$group==0]))^2
```
### R Code

```r
# Load necessary packages
library(dplyr)
library(ggplot2)

# Define variables
NAIVEprop <- data3$NAIVEprop
data3$group <- ifelse(data3$group == 0, 0, 1)

# Calculate means and SDs
mean_y <- mean(data3$Y)
mean_x <- mean(data3$X)

# Matched sample
matched <- length(data3$Y) == length(data3$X)

# Propensity scores
prop_mean <- mean(NAIVEprop)
prop_var <- var(NAIVEprop)

# Summary statistics
summary(data3)
```

### Notes
- The code above assumes that `data3` is a data frame containing variables `Y`, `X`, and `group`.
- `NAIVEprop` is a variable indicating the propensity score.
- The code calculates the mean and SD of `Y` and `X`, and checks if the sample is matched.
- Propensity scores are calculated and used in various statistical tests.
- Summary statistics are provided for the matched sample.

#All variables AFTER matching True Prop NN w/ 0.2 Caliper

```r
# Additional code

# Further analysis

### Additional Notes
- Further analysis could include visualization of propensity scores or other statistical tests.
- The code snippet is part of a larger script for propensity score matching.
```
### Variables after matching Naive Prop NN w/ 0.2 Caliper

- `NCNCTSDX1Treat[i]` <- `sd(data4$X1[data4$group==1])`
- `NCNCTSDX2Treat[i]` <- `sd(data4$X2[data4$group==1])`
- `NCNCTSDX3Treat[i]` <- `sd(data4$X3[data4$group==1])`
- `NCNCTSDX4Treat[i]` <- `sd(data4$X4[data4$group==1])`
- `NCNCTSDX5Treat[i]` <- `sd(data4$X5[data4$group==1])`
- `NCNCTSDX3Cont[i]` <- `sd(data4$X3[data4$group==0])`
- `NCNCTSDX2Cont[i]` <- `sd(data4$X2[data4$group==0])`
- `NCNCTSDX1Cont[i]` <- `sd(data4$X1[data4$group==0])`

### Correlation Coefficients

- `NCNCTCorX1.X2` <- `cor(data4$X1, data4$X2)`
- `NCNCTCorX1.X3` <- `cor(data4$X1, data4$X3)`
- `NCNCTCorX1.X4` <- `cor(data4$X1, data4$X4)`
- `NCNCTCorX1.X5` <- `cor(data4$X1, data4$X5)`
- `NCNCTCorX2.X3` <- `cor(data4$X2, data4$X3)`
- `NCNCTCorX2.X4` <- `cor(data4$X2, data4$X4)`
- `NCNCTCorX2.X5` <- `cor(data4$X2, data4$X5)`
- `NCNCTCorX3.X4` <- `cor(data4$X3, data4$X4)`
- `NCNCTCorX3.X5` <- `cor(data4$X3, data4$X5)`
- `NCNCTCorX4.X5` <- `cor(data4$X4, data4$X5)`

### Mean and Variance

- `NCNCTMeanX1[i]` <- `mean(data4$X1)`
- `NCNCTMeanX2[i]` <- `mean(data4$X2)`
- `NCNCTMeanX3[i]` <- `mean(data4$X3)`
- `NCNCTMeanX4[i]` <- `mean(data4$X4)`
- `NCNCTMeanX5[i]` <- `mean(data4$X5)`

### SD

- `NCNCTSDX1` <- `sd(data4$X1)`
- `NCNCTSDX2` <- `sd(data4$X2)`
- `NCNCTSDX3` <- `sd(data4$X3)`
- `NCNCTSDX4` <- `sd(data4$X4)`
- `NCNCTSDX5` <- `sd(data4$X5)`

### PropVar

- `NCNCTPropVarC` <- `sd(data4$Y)`
- `NCNCTPropMeanT` <- `mean(data4$TRUEprop)`
- `NCNCTPropMeanC` <- `mean(data4$TRUEprop == 0)`
- `NCNCTPropVarC` <- `sd(data4$TRUEprop == 0)`

---

All variables AFTER matching Naive Prop NN w/ 0.2 Caliper.
# All variables AFTER matching True Prop Optimal

```r
NNCCorX1.X2[i] <- cor(data5$X1, data5$X2)
NNCCorX1.X3[i] <- cor(data5$X1, data5$X3)
NNCCorX1.X4[i] <- cor(data5$X1, data5$X4)
NNCCorX1.X5[i] <- cor(data5$X1, data5$X5)
NNCCorX2.X3[i] <- cor(data5$X2, data5$X3)
NNCCorX2.X4[i] <- cor(data5$X2, data5$X4)
NNCCorX2.X5[i] <- cor(data5$X2, data5$X5)
NNCCorX3.X4[i] <- cor(data5$X3, data5$X4)
NNCCorX3.X5[i] <- cor(data5$X3, data5$X5)
NNCCorX4.X5[i] <- cor(data5$X4, data5$X5)
NNCCorY.X1[i] <- cor(data5$Y, data5$X1)
NNCCorY.X2[i] <- cor(data5$Y, data5$X2)
NNCCorY.X3[i] <- cor(data5$Y, data5$X3)
NNCCorY.X4[i] <- cor(data5$Y, data5$X4)
NNCCorY.X5[i] <- cor(data5$Y, data5$X5)
NNCMatched[i] <- length(data5$Y[data5$group==0])
NNNPropMeanT <- mean(data5$NAIVEprop[data5$group==1])
NNNPropMeanC <- mean(data5$NAIVEprop[data5$group==0])
NNNPropVarT <- (sd(data5$NAIVEprop[data5$group==1]))^2
NNNPropVarC <- (sd(data5$NAIVEprop[data5$group==0]))^2
```

# All variables AFTER matching True Prop Optimal
```r
#All variables AFTER matching Naive Prop Optimal

optNAvgX1Treat[1] <- mean(data7$X1[data7$group==1])
optNAvgX2Treat[1] <- mean(data7$X2[data7$group==1])
optNAvgL1Treat[1] <- mean(data7$L1[data7$group==1])
optNAvgL2Treat[1] <- mean(data7$L2[data7$group==1])
optNAvgX3Treat[1] <- mean(data7$X3[data7$group==1])
optNAvgX4Treat[1] <- mean(data7$X4[data7$group==1])
optNAvgX5Treat[1] <- mean(data7$X5[data7$group==1])
optNAvgYTreat[1] <- mean(data7$Y[data7$group==1])
optNAvgX1Cont[1] <- mean(data7$X1[data7$group==0])
optNAvgX2Cont[1] <- mean(data7$X2[data7$group==0])
optNAvgL1Cont[1] <- mean(data7$L1[data7$group==0])
optNAvgL2Cont[1] <- mean(data7$L2[data7$group==0])
optNAvgX3Cont[1] <- mean(data7$X3[data7$group==0])
optNAvgX4Cont[1] <- mean(data7$X4[data7$group==0])
optNAvgX5Cont[1] <- mean(data7$X5[data7$group==0])
optNAvgYCont[1] <- mean(data7$Y[data7$group==0])
optNSDX1Treat[1] <- sd(data7$X1[data7$group==1])
optNSDX2Treat[1] <- sd(data7$X2[data7$group==1])
optNSDX3Treat[1] <- sd(data7$X3[data7$group==1])
optNSDX4Treat[1] <- sd(data7$X4[data7$group==1])
optNSDX5Treat[1] <- sd(data7$X5[data7$group==1])
optNSDX1Cont[1] <- sd(data7$X1[data7$group==0])
optNSDX2Cont[1] <- sd(data7$X2[data7$group==0])
optNSDX3Cont[1] <- sd(data7$X3[data7$group==0])
optNSDX4Cont[1] <- sd(data7$X4[data7$group==0])
optNSDX5Cont[1] <- sd(data7$X5[data7$group==0])
optNSDYCont[1] <- sd(data7$Y[data7$group==0])
optNCorX1.X2[1] <- cor(data7$X1, data7$X2)
optNCorX1.X3[1] <- cor(data7$X1, data7$X3)
optNCorX1.X4[1] <- cor(data7$X1, data7$X4)
optNCorX1.X5[1] <- cor(data7$X1, data7$X5)
optNCorX2.X3[1] <- cor(data7$X2, data7$X3)
optNCorX2.X4[1] <- cor(data7$X2, data7$X4)
optNCorX2.X5[1] <- cor(data7$X2, data7$X5)
optNCorX3.X4[1] <- cor(data7$X3, data7$X4)
optNCorX3.X5[1] <- cor(data7$X3, data7$X5)
optNCorX4.X5[1] <- cor(data7$X4, data7$X5)
optNCorY.X1[1] <- cor(data7$Y, data7$X1)
optNCorY.X2[1] <- cor(data7$Y, data7$X2)
optNCorY.X3[1] <- cor(data7$Y, data7$X3)
optNCorY.X4[1] <- cor(data7$Y, data7$X4)
optNCorY.X5[1] <- cor(data7$Y, data7$X5)
optNMatchedN[1] <- length(data7$Y[data7$group==0])
optNPPropMeanT <- mean(data7$NAIVEprop[data7$group==1])
optNPPropMeanC <- mean(data7$NAIVEprop[data7$group==0])
optNPPropVarT <- (sd(data7$NAIVEprop[data7$group==1]))^2
optNPPropVarC <- (sd(data7$NAIVEprop[data7$group==0]))^2

# All variables AFTER matching True Prop Mahal

MahalTAvx1Treat[1] <- mean(data8$X1[data8$group==1])
```
MahalTAvgX2Treat[i] <- mean(data8$X2[data8$group==1])
MahalTAvgL2Treat[i] <- mean(data8$L2[data8$group==1])
MahalTAvgX3Treat[i] <- mean(data8$X3[data8$group==1])
MahalTAvgX4Treat[i] <- mean(data8$X4[data8$group==1])
MahalTAvgX5Treat[i] <- mean(data8$X5[data8$group==1])
MahalTAvgX1Cont[i] <- mean(data8$X1[data8$group==0])
MahalTAvgL2Cont[i] <- mean(data8$L2[data8$group==0])
MahalTAvgL2Cont[i] <- mean(data8$L2[data8$group==0])
MahalTAvgX3Cont[i] <- mean(data8$X3[data8$group==0])
MahalTAvgX4Cont[i] <- mean(data8$X4[data8$group==0])
MahalTAvgX5Cont[i] <- mean(data8$X5[data8$group==0])
MahalTAvgyCont[i] <- mean(data8$Y[data8$group==0])
MahalTSDX1Treat[i] <- sd(data8$X1[data8$group==1])
MahalTSDX2Treat[i] <- sd(data8$X2[data8$group==1])
MahalTSDL1Treat[i] <- sd(data8$L1[data8$group==1])
MahalTSDL2Treat[i] <- sd(data8$L2[data8$group==1])
MahalTSDX3Treat[i] <- sd(data8$X3[data8$group==1])
MahalTSDX4Treat[i] <- sd(data8$X4[data8$group==1])
MahalTSDX5Treat[i] <- sd(data8$X5[data8$group==1])
MahalTSDYCont[i] <- sd(data8$Y[data8$group==0])
MahalTCorX1.X2[i] <- cor(data8$X1, data8$X2)
MahalTCorX1.X3[i] <- cor(data8$X1, data8$X3)
MahalTCorX1.X4[i] <- cor(data8$X1, data8$X4)
MahalTCorX1.X5[i] <- cor(data8$X1, data8$X5)
MahalTCorX2.X3[i] <- cor(data8$X2, data8$X3)
MahalTCorX2.X4[i] <- cor(data8$X2, data8$X4)
MahalTCorX2.X5[i] <- cor(data8$X2, data8$X5)
MahalTCorX3.X4[i] <- cor(data8$X3, data8$X4)
MahalTCorX3.X5[i] <- cor(data8$X3, data8$X5)
MahalTCorX4.X5[i] <- cor(data8$X4, data8$X5)
MahalTCorY.X1[i] <- cor(data8$Y, data8$X1)
MahalTCorY.X2[i] <- cor(data8$Y, data8$X2)
MahalTCorY.X3[i] <- cor(data8$Y, data8$X3)
MahalTCorY.X4[i] <- cor(data8$Y, data8$X4)
MahalTCorY.X5[i] <- cor(data8$Y, data8$X5)
MahalTMatchedN[i] <- length(data8$Y[data8$group==0])
MahalTPropMeanC <- mean(data8$TRUEprop[data8$group==1])
MahalTPropMeanC <- mean(data8$TRUEprop[data8$group==0])
MahalTPropVarT <- (sd(data8$TRUEprop[data8$group==1]))^2
MahalTPropVarC <- (sd(data8$TRUEprop[data8$group==0]))^2

#All variables AFTER matching Naive Prop Mahal

MahalNAvgX1Treat[i] <- mean(data9$X1[data9$group==1])
MahalNAvgX2Treat[i] <- mean(data9$X2[data9$group==1])
MahalNAvgL1Treat[i] <- mean(data9$L1[data9$group==1])
MahalNAvgL2Treat[i] <- mean(data9$L2[data9$group==1])
MahalNAvgX3Treat[i] <- mean(data9$X3[data9$group==1])
MahalNAvgX4Treat[i] <- mean(data9$X4[data9$group==1])
MahalNAvgX5Treat[i] <- mean(data9$X5[data9$group==1])
MahalNAvgYTreat[i] <- mean(data9$Y[data9$group==1])
MahalNAvgX1Cont[i] <- mean(data9$X1[data9$group==0])
MahalNAvgX2Cont[i] <- mean(data9$X2[data9$group==0])
MahalNAvgL1Cont[i] <- mean(data9$L1[data9$group==0])
MahalNAvgL2Cont[i] <- mean(data9$L2[data9$group==0])
MahalNAvgX3Cont[i] <- mean(data9$X3[data9$group==0])
MahalNAvgX4Cont[i] <- mean(data9$X4[data9$group==0])
MahalNAvgX5Cont[i] <- mean(data9$X5[data9$group==0])
MahalNAVgYCont[i] <- mean(data9$Y[data9$group==0])
MahalNSDX1Treat[i] <- sd(data9$X1[data9$group==1])
MahaNSDX2Treat[i] <- sd(data9$X2[data9$group==1])
MahaNSDX1Treat[i] <- sd(data9$L1[data9$group==1])
MahaNSDL2Treat[i] <- sd(data9$L2[data9$group==1])
MahaNSDX4Treat[i] <- sd(data9$X4[data9$group==1])
MahaNSDX5Treat[i] <- sd(data9$X5[data9$group==1])
MahaNSDYtreat[i] <- sd(data9$Y[data9$group==1])
MahaNSDL1Cont[i] <- sd(data9$L1[data9$group==0])
MahaNSDL2Cont[i] <- sd(data9$L2[data9$group==0])
MahaNSDX3Cont[i] <- sd(data9$X3[data9$group==0])
MahaNSDX4Cont[i] <- sd(data9$X4[data9$group==0])
MahaNSDX5Cont[i] <- sd(data9$X5[data9$group==0])
MahaNSDYCont[i] <- sd(data9$Y[data9$group==0])
MahaLXcorX1.X2[i] <- cor(data9$X1, data9$X2)
MahaLXcorX1.X3[i] <- cor(data9$X1, data9$X3)
MahaLXcorX1.X4[i] <- cor(data9$X1, data9$X4)
MahaLXcorX1.X5[i] <- cor(data9$X1, data9$X5)
MahaLXcorX2.X3[i] <- cor(data9$X2, data9$X3)
MahaLXcorX2.X4[i] <- cor(data9$X2, data9$X4)
MahaLXcorX3.X4[i] <- cor(data9$X3, data9$X4)
MahaLXcorX3.X5[i] <- cor(data9$X3, data9$X5)
MahaLXcorX4.X5[i] <- cor(data9$X4, data9$X5)
MahaLXcorY.X1[i] <- cor(data9$Y, data9$X1)
MahaLXcorY.X2[i] <- cor(data9$Y, data9$X2)
MahaLXcorY.X3[i] <- cor(data9$Y, data9$X3)
MahaLXcorY.X4[i] <- cor(data9$Y, data9$X4)
MahaLXcorY.X5[i] <- cor(data9$Y, data9$X5)
MahaLXmatchednD[i] <- length(data9$Y[data9$group==0])
MahaLXpropMeanT <- mean(data9$NAIVEprop[data9$group==1])
MahaLXpropMeanC <- mean(data9$NAIVEprop[data9$group==0])
MahaLXpropVarT <- (sd(data9$NAIVEprop[data9$group==1]))^2
MahaLXpropVarC <- (sd(data9$NAIVEprop[data9$group==0]))^2

#Removing the data after each iteration
rm(data, data2, data3, data4, data5, data6, data7, data8, data9, m.out1, m.out2, m.out3, m.out4, m.out5, m.out6, m.out7, m.out8, ps.sd, ps.sd2, finaldata, dataNNT, dataNNN, dataNNCT, dataNNCN, dataoptT, dataoptN, dataMahalT, dataMahalN)

proc.time() -> now

####### END OF LOOP ########
warnings()

#In this step, we make a final data set by binding together all of the above variables.
Final.Sim.Data <- cbind(

pAvgX1Treat, pAvgX2Treat, pAvgL1Treat, pAvgL2Treat, pAvgX3Treat, pAvgX4Treat, pAvgX5Treat, pAvgY1Cont, pAvgX2Cont, pAvgL1Cont, pAvgL2Cont, pAvgX3Cont, pAvgX4Cont, pAvgX5Cont, pAvgYCont, pSDX1Treat, pSDX2Treat, pSDL1Treat, pSDL2Treat, pSDX3Treat, pSDX4Treat, pSDX5Treat, pSDX1Cont, pSDX2Cont, pSDX3Cont, pSDX4Cont, pSDX5Cont, pSDYCont, pCorX1.X2, pCorX1.X3, pCorX1.X4, pCorX1.X5, pCorX2.X3, pCorX2.X4, pCorX2.X5, pCorX3.X4, pCorX3.X5, pCorY.X1, pCorY.X2, pCorY.X4, pCorY.X5, pPropMeanTT, pPropMeanCN, pPropSDTN, pPropSDDCN, pCorLX1, pCorLX2,

NNTAvgX1Treat, NNTAvgX2Treat, NNTAvgL1Treat, NNTAvgL2Treat, NNTAvgX3Treat, NNTAvgX4Treat, NNTAvgX5Treat, NNTAvgY1Treat, NNTAvgX2Cont, NNTAvgL1Cont, NNTAvgL2Cont, NNTAvgX3Cont, NNTAvgX4Cont, NNTAvgX5Cont, NNTAvgYCont, NNTSDX1Treat, NNTSDX2Treat, NNTSDX3Treat, NNTSDX4Treat, NNTSDX5Treat, NNTSDX1Cont, NNTSDX2Cont, NNTSDX3Cont, NNTSDX4Cont, NNTSDX5Cont, NNTSDYCont, NNTCorX1.X2, NNTCorX1.X3, NNTCorX1.X4, NNTCorX1.X5, NNTCorX2.X3, NNTCorX2.X4,
head(Final.Sim.Data)
tail(Final.Sim.Data)
str(Final.Sim.Data)
describe(Final.Sim.Data)

### Final step: Save it out AS A CSV FILE :)
Appendix C

The simulated correlation matrix (see Table 1) and vector of correlations between simulated error-free covariates and latent propensity (as a continuous variable, not a probability) of selecting into treatment (all set to 0.4) were used to derive the coefficients via the following formula:

\[ B = (X'X)^{-1}X'Y \]

where \( B \) represents the coefficient weights, \( X'X \) represents the simulated covariate correlation matrix, and \( X'Y \) represents the correlations between the simulated covariates and the continuous latent probability of treatment assignment. The proportion of variance in latent propensity explained by the covariates was thus

\[ R^2 = \frac{BB'(X'X)}{(BB'(X'X) + 1)} \]

because the error variance for probit regression equals one. The intercept was calculated as the Z-value of a standard normal distribution corresponding to .20, divided by \( \sqrt{1 - R^2} \). The resulting values of the intercept and \( B \) are provided in Equation 16. Using the calculated coefficients, each simulee was then assigned a predicted probability of participation based on the simulated covariates.
Appendix D

Figure D1a. Correlations among simulated variables for the same measurement error condition with reliability=0.9.

Figure D1b. Jitter graphs of propensity score matches using true propensity scores (left) and naïve propensity scores (right) via nearest neighbor matching (same measurement error condition with reliability=0.9).
Figure D2a. Correlations among simulated variables for the same measurement error condition with reliability=0.8.

Figure D2b. Jitter graphs of propensity score matches using true propensity scores (left) and naïve propensity scores (right) via nearest neighbor matching (same measurement error condition with reliability=0.8).
Figure D3a. Correlations among simulated variables for the same measurement error condition with reliability=0.7.

Figure D3b. Jitter graphs of propensity score matches using true propensity scores (left) and naïve propensity scores (right) via nearest neighbor matching (same measurement error condition with reliability=0.7).
Figure D4a. Correlations among simulated variables for the same measurement error condition with reliability=0.6.

Figure D4b. Jitter graphs of propensity score matches using true propensity scores (left) and naïve propensity scores (right) via nearest neighbor matching (same measurement error condition with reliability=0.6).
**Figure D5a.** Correlations among simulated variables for the same measurement error condition with reliability=0.5.

**Figure D5b.** Jitter graphs of propensity score matches using true propensity scores (left) and naïve propensity scores (right) via nearest neighbor matching (same measurement error condition with reliability=0.5).
**Figure D6a.** Correlations among simulated variables for the same measurement error condition with reliability=0.4.

**Figure D6b.** Jitter graphs of propensity score matches using true propensity scores (left) and naïve propensity scores (right) via nearest neighbor matching (same measurement error condition with reliability=0.4).
Figure D7a. Correlations among simulated variables for the differential measurement error condition with comparison group reliability=0.9 for the treatment group (top panel), control group (middle panel), and both groups combined (bottom panel).
**Figure D7b.** Jitter graphs of propensity score matches using true propensity scores (left) and naïve propensity scores (right) via nearest neighbor matching (different measurement error condition with comparison group reliability=0.9).
Figure D8a. Correlations among simulated variables for the differential measurement error condition with comparison group reliability=0.8 for the treatment group (top panel), control group (middle panel), and both groups combined (bottom panel).
Figure D8b. Jitter graphs of propensity score matches using true propensity scores (left) and naïve propensity scores (right) via nearest neighbor matching (different measurement error condition with comparison group reliability=0.8).
Figure D9a. Correlations among simulated variables for the differential measurement error condition with comparison group reliability=0.7 for the treatment group (top panel), control group (middle panel), and both groups combined (bottom panel).
Figure D9b. Jitter graphs of propensity score matches using true propensity scores (left) and naïve propensity scores (right) via nearest neighbor matching (different measurement error condition with comparison group reliability=0.7).
Figure D10a. Correlations among simulated variables for the differential measurement error condition with comparison group reliability=0.6 for the treatment group (top panel), control group (middle panel), and both groups combined (bottom panel).
Figure D10b. Jitter graphs of propensity score matches using true propensity scores (left) and naïve propensity scores (right) via nearest neighbor matching (different measurement error condition with comparison group reliability=0.6).
Figure D11a. Correlations among simulated variables for the differential measurement error condition with comparison group reliability=0.5 for the treatment group (top panel), control group (middle panel), and both groups combined (bottom panel).
Figure D11b. Jitter graphs of propensity score matches using true propensity scores (left) and naïve propensity scores (right) via nearest neighbor matching (different measurement error condition with comparison group reliability=0.5).
Figure D12a. Correlations among simulated variables for the differential measurement error condition with comparison group reliability=0.4 for the treatment group (top panel), control group (middle panel), and both groups combined (bottom panel).
Figure D12b. Jitter graphs of propensity score matches using true propensity scores (left) and naïve propensity scores (right) via neighbor matching (differential measurement error condition with comparison group reliability=0.4).
Appendix E

*Comparing measurement error conditions.
GLM Cond1 Cond2 Cond3 Cond4 Cond5 Cond6 BY PScore ErrorType
/WSFACTOR=Condition 6 Polynomial
/METHOD=SSTYPE(3)
/EMMEANS=TABLES(OVERALL)
/EMMEANS=TABLES(PScore) COMPARE ADJ(BONFERRONI)
/EMMEANS=TABLES(Condition) COMPARE ADJ(BONFERRONI)
/EMMEANS=TABLES(PScore*Condition)
/EMMEANS=TABLES(ErrorType*Condition)
/EMMEANS=TABLES(PScore*ErrorType*Condition)
/PRINT=DESCRIPTIVE ETASQ HOMOGENEITY
/CRITERIA=ALPHA(.05)
/WSDESIGN=Condition
/DESIGN=PScore ErrorType PScore*ErrorType.

*Comparing matching methods.
GLM NN NNC OPT MAH BY PScore ErrorType
/WSFACTOR=Method 4 Polynomial
/METHOD=SSTYPE(3)
/EMMEANS=TABLES(OVERALL)
/EMMEANS=TABLES(PScore) COMPARE ADJ(BONFERRONI)
/EMMEANS=TABLES(Method) COMPARE ADJ(BONFERRONI)
/EMMEANS=TABLES(PScore*Method)
/EMMEANS=TABLES(ErrorType*Method)
/PRINT=DESCRIPTIVE ETASQ HOMOGENEITY
/CRITERIA=ALPHA(.05)
/WSDESIGN=Method
/DESIGN=PScore ErrorType PScore*ErrorType.

*Test of the three-way interactions below, followed by the test of the four-way interaction.
GLM NaivePS TruePS BY MEtype Condition
/WSFACTOR=Prop 2 Polynomial
/METHOD=SSTYPE(3)
/POSTHOC=MEtype Condition(TUKEY)
/EMMEANS=TABLES(OVERALL)
/EMMEANS=TABLES(MEtype) COMPARE ADJ(BONFERRONI)
/EMMEANS=TABLES(Condition) COMPARE ADJ(BONFERRONI)
/EMMEANS=TABLES(Prop) COMPARE ADJ(BONFERRONI)
/EMMEANS=TABLES(MEtype*Condition)
/EMMEANS=TABLES(MEtype*Prop)
/EMMEANS=TABLES(Condition*Prop)
/EMMEANS=TABLES(MEtype*Condition*Prop)
/PRINT=DESCRIPTIVE ETASQ OPOWER HOMOGENEITY
/CRITERIA=ALPHA(.05)
/WSDESIGN=Prop
/DESIGN=MEtype Condition MEtype*Condition.

GLM NaivePS TruePS BY MEtype Condition Method
/WSFACTOR=Prop 2 Polynomial
/METHOD=SSTYPE(3)
/POSTHOC=MEtype Condition(TUKEY)
/EMMEANS=TABLES(OVERALL)
/EMMEANS=TABLES(MEtype) COMPARE ADJ(BONFERRONI)
/EMMEANS=TABLES(Condition) COMPARE ADJ(BONFERRONI)
/EMMEANS=TABLES(Method) COMPARE ADJ(BONFERRONI)
/EMMEANS=TABLES(Prop) COMPARE ADJ(BONFERRONI)
/EMMEANS=TABLES(MEtype*Condition)
/EMMEANS=TABLES(MEtype*Prop)
/EMMEANS=TABLES(Condition*Prop)
/EMMEANS=TABLES(MEtype*Method)
/EMMEANS=TABLES(Method*Prop)
/EMMEANS=TABLES(Condition*Method)
/EMMEANS=TABLES(MEtype*Condition*Prop)
/EMMEANS=TABLES(Method*Condition*Prop)
/EMMEANS=TABLES(Method*MEtype*Condition)
/EMMEANS=TABLES(Method*MEtype*Prop)
/EMMEANS=TABLES(MEtype*Condition*Prop*Method)
/PRINT=DESCRIPTIVE ETASQ OPOWER HOMOGENEITY
/CRITERIA=ALPHA(.05)
/WSDESIGN=Prop
/DESIGN=MEtype Condition Method MEtype*Condition Method*Condition Method*MEtype*Condition.
References


Austin, P. C. (2009c). Type 1 error rates, coverage of confidence intervals, and variance estimation in propensity-score matched analyses. *International Journal of Biostatistics, 5*(1), 1-12.


Hong, H., Rudolph, K. E., & Stuart, E. A. (2016). Bayesian approach for addressing
differential covariate measurement error in propensity score methods.
*Psychometrika, 33*, 1-19

Publishing.


matching method on the quantity and quality of matches. *The Journal of
Experimental Education, DOI: 10.1080/00220973.2016.1250209*

Kainz, K., & Pan, Y. (2014). Segregated school effects on first grade reading gains:
Using propensity score matching to disentangle effects for African-American,
Latino, and European-American students. *Early Childhood Research Quarterly,
29*, 531-537.

Bayesian perspective: A review and synthesis. *Large-scale Assessments in


Kim, J., & Steiner, P. M. (2015). Multilevel propensity score methods for estimating
causal effects: A latent class modeling strategy. *Quantitative Psychology
Research, 140*, 293-306.


