Spring 2015

Propensity score matching in higher education assessment

Heather D. Harris
James Madison University

Propensity Score Matching in Higher Education Assessment

Heather Harris

A thesis submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

for the degree of

Master of Arts

Psychological Sciences

May 2015

Dedication

"Here's to the crazy ones. The misfits. The rebels. The troublemakers. The round pegs in the square holes. The ones who see things differently. They're not fond of rules. And they have no respect for the status quo. You can quote them, disagree with them, glorify or vilify them. About the only thing you can't do is ignore them. Because they change things. They push the human race forward. And while some may see them as the crazy ones, we see genius. Because the people who are crazy enough to think they can change the world, are the ones who do."

- Rob Siltanen & Lee Clow

Acknowledgements

First and foremost, I would like to thank Dr. Jeanne Horst. Jeanne, I am incredibly grateful to have worked with you the past two years. Your patient guidance and expertise on this project (among others) was an essential part of my growth as an academic. Because of your dedication and thoughtful direction, I was able to complete a project I'm not only proud of, but one that has spurred countless future research ideas! Your careful feedback has helped me to grow immensely both professionally and personally, and for that I am incredibly grateful. I hope I can one day be half the researcher you are and I am ecstatic I get to work with you over the next three years!

Second, I would like thank my committee members, Dr. Keston Fulcher and Dr. Monica Erbacher. Keston and Monica, thank you for your willingness to help me throughout the entire process! Your careful feedback and insight were an invaluable part of this project. Thank you for supporting me personally and professionally; I'm so appreciative of you both. Working with you has been inspiring. I learned a great deal and my final product was considerably better thanks to your guidance along the way!

I also want to thank my academic cohort. Kate and Liz, thank you both for being such an important part of my graduate experience. It has been truly an honor going through the program with both of you and having opportunities to support each other along the way. I know you'll both do incredible things in the future; I can't wait to see what sort of mountains you move. I hope we continue to collaborate in our future careers and I'm so happy to count you both as friends.

I must also thank the best support system and family a girl could ask for! Mom and Dad, thank you for supporting me over the past 28 years. So much of who I am today

is a result of the opportunities you provided me with growing up. I'm incredibly lucky to have you in my life; I'm especially proud to have you both as parents. Together, you blazed a new trail in life without guarantees as to how things would work out. You're courageous and caring, and quite frankly I hope it's genetic. Jenny and Ben, thank you for supporting me in all of my academic (and sometimes non-academic) pursuits. Jenny, I'm proud of you for all you've accomplished and I look forward to a lifetime full of opportunities to learn from one another. Ben, I cannot thank you enough for supporting me unconditionally. You are a truly wonderful and talented human being and I am incredibly proud to have you as my brother. I must also thank my "father-in-law," Scott Driscoll and my future sister-in-law, Karla Thole. Thank you both for supporting me over the past few years and for being genuinely interested in my research. I am incredibly lucky to have you both in my life. I must also thank my friends who are family, especially Debbie Ernie. Debbie, you are an unfailingly sunny spot in my life and I appreciate you so much. Thank you for being such a great friend. I'm lucky to know you.

Last, but certainly not least, I must thank my best friend, boyfriend, comrade, and biggest supporter. Casey Patrick Driscoll, you are easily the best thing that has ever happened to me. Over the past nine and a half years, you have helped me push myself out of my comfort zone more times than I can count. You support me unconditionally and challenge me to continuously grow as a person. You're also the coolest person I know and you are unfailingly talented at executing a joke (something I'm still working on…). I don't know what sort of righteous karma I accrued in a past life to have you in my corner, but I am certainly not complaining. Thank you.

Table of Contents

List of Tables

List of Figures

Abstract

The applied nature of higher education assessment does not lend itself to rigorous experimental research designs. However, assessment practitioners would like to make claims about the influence of educational programs on student learning outcomes. Propensity score matching (PSM) methods are quasi-experimental techniques that allow researchers to control for known confounding variables. In the context of higher education, PSM techniques allow assessment practitioners to control for confounding variables related to students' self-selected participation in university programs. Research and recommendations on how to apply PSM techniques are scattered throughout several disciplines. However, additional research is needed to evaluate how well PSM techniques control self-selection bias in the context of educational assessment. To couch PSM techniques within the framework of higher education assessment, the current study first summarized common practices and recommendations from literature across several disciplines, then evaluated the application of common PSM techniques via an applied example of honors program assessment. Specifically, the study applied common PSM techniques to compare students in the honors program with students either not invited into the program or students who decided not to participate in the program. Data analyses four research questions: 1) Do honors students differ from students not in the honors program on motivation variables?; 2) How well do different common PSM techniques create quality comparison groups of students?; 3) How well do different common PSM techniques retain honors students in the comparison of program outcomes?; and 4) Do honors students differ from students not in the honors program on outcomes after PSM techniques are applied? Honors students did not differ from students not in the honors program on motivation variables thought to be related to self-selection into the program.

Across matching methods applied in the current study, the quality of propensity score matches was near optimal. Average scores on the outcomes of interest to the honors program did not significantly differ by group. However, decreased sample sizes resulted in a loss of minority student representation in the honors sample and different practically significant results on program outcomes. Recommendations and implications for applied assessment practitioners are offered.

Chapter One

Introduction

Universities implement a wide variety of programs to promote student learning and development. The assessment of such programs is important to determine whether students are indeed learning and developing as a function of program participation. Moreover, sound assessment practices help determine the extent to which students' learning and development can be attributed to university programs rather than simply maturation associated with life experiences (Chickering, 1999).

While assessment of college learning is on the rise, academic settings do not lend themselves to strong experimental designs (Kember, 2003; Kuh, Jankowski, Ikenberry & Kinzie, 2014). Ideally, researchers would *like to* make causal claims about the impact of their programs; however, the characteristics of strong experimental research designs are not easily implemented within applied educational contexts. For example, research designs in assessment are frequently hindered by time and contextual factors within education, such as the length of semester and student self-selection into classes (Kember, 2003).

In addition to logistical constraints, students who choose to participate in university programs may be qualitatively different from students in the general student population. Academic interventions often target student learning outcomes that are important to the university (e.g., science literacy). However, participation in academic interventions may vary systematically as a function of students' interest in the program or student eligibility. Therefore, key to this type of research is differentiating the effect of the program from self-selection bias.

Selection bias is defined as systematic differences in baseline characteristics due to self-selection at the individual level (Winship & Mare, 1992). Such bias limits the inferences one can make about a program's efficacy. For example, there may be individual differences specific to students who decide to participate in a university honors program. If some students who are eligible for the program decide to participate while others who are eligible decide *not* to participate, the underlying reasons for students' decisions may make participants qualitatively unique from non-participants. Thus, there are likely systematic differences between the two groups that make it difficult to accurately draw inferences about group differences in learning outcomes.

Typically, the "gold standard" for making causal inferences is a true experiment via the random assignment of participants to either intervention or control conditions (Ho et al., 2007; Kember, 2003; Luellen, Shadish, & Clark, 2005). In a true experiment, participants are randomly assigned to either intervention or control conditions. In such designs, participants vary only randomly across the two conditions on both observed and unobserved baseline variables. In contrast, when students self-select to participate in a university program, they may qualitatively differ from students who do not participate. Thus, differences in baseline characteristics between participants and nonparticipants may confound the ability to make valid inferences as the two groups do not differ randomly from one another.

Shadish, Cook, and Campbell (2002) defined *confound* as "an extraneous variable that covaries with the variable of interest" (p. 506). For example, students who enroll in certain programs may be more motivated than students who do not. If academic motivation is related to both students' decisions to join an honors program and academic

performance (i.e., program outcomes), it is difficult to parse apart the contribution of academic motivation on academic performance from the actual impact of the program itself. In other words, the self-selection baseline characteristic of academic motivation is a confound.

Random assignment is the primary way that researchers address confounding variables and the threat they pose to internal validity. In a typical "true experiment" (i.e., randomized control design), internal validity is essentially the extent to which changes in the dependent variable can be attributed causally to the manipulated independent variable (Shadish et al., 2002). Because of ethical and logistical concerns related to randomly assigning students to programs, researchers employ quasi-experimental research designs as an alternative (Shadish et al., 2002). However, inherent in quasi-experimental research designs is the validity threat posed by confounding variables.

Fortunately, propensity score matching (PSM) techniques can account for such confounding variables (Austin, 2010a; Ho et al., 2007; Stuart, 2010; Stuart & Rubin, 2008a). However, research on PSM is scattered throughout several disciplines including economics (Czajka, Hirabayashi, Little, & Rubin, 1992), medicine (D'Agostino, 1998; Rubin, 2004), statistics (Rosenbaum, 2002; Rubin, 2006; Stuart, 2010), marketing, (Lu, Zanutto, Hornik, & Rosenbaum, 2001) and applied educational assessment (Agodini & Dynarski, 2004; Dehejia & Wahba, 1999; Frisco, Muller, & Frank, 2007; Hansen, 2004; Heckman, Ichimura, & Todd, 1998). Though the terminology varies, PSM is used in relatively the same manner across disciplines and offers a solute on to situations in which random assignment is not possible. In the context of higher education assessment,

however, more research on the variables related to students' self-selection into university programs is needed.

The current study situates PSM techniques within an applied educational context. Because implementing PSM techniques involves a series of decisions, a review of the literature will first be summarized. Specifically, common practices found in the PSM literature will be outlined and recommendations will be summarized for implementing PSM techniques in higher education assessment.

An applied example of honors program assessment will follow, adhering to recommendations outlined in the literature review. At the author's institution, a select number of incoming first-year students are invited to participate in the honors program based on their academic performance in high school. However, only 20% of invited students join the honors program. Therefore, students who participate in the honors program might be qualitatively different from students who decided not to participate or who were not invited to participate in the program on variables such as academic motivation. To highlight how the series of PSM decisions may influence the inferences drawn from program assessment, several common PSM techniques will be applied and evaluated.

Chapter Two

Review of the Literature

In the context of educational assessment, practitioners frequently attempt to draw causal inferences about the impact of their programs. Specifically, assessment professionals would like to claim that their programs or interventions directly impact student learning. However, given the quasi-experimental nature of the research, the extent to which one can make causal inferences in applied contexts is limited (Holland, 1986; Winship & Morgan, 1999).

Ideally, when attempting to make causal inferences about the impact of some variable, researchers randomly assign participants to conditions. However, the applied context of education often means random assignment to programs or interventions is neither feasible nor ethical. Because it mimics the strengths of true experimental designs, propensity score matching (PSM) provides an appealing alternative (Luellen, Shadish, & Clark, 2005). The current study introduces the concept of PSM, describes best practices for conducting PSM studies, and provides an applied example situated within the educational context.

What are Propensity Scores?

The first step in accounting for variables related to self-selection is to calculate propensity scores. A propensity score is a balancing score that is calculated to create a matched comparison group that is similar to the intervention group on a set of baseline characteristics (Austin, 2011; Stuart & Rubin, 2008a). Propensity scores are simply the "conditional probability of exposure to a treatment given observed covariates" (Joffe & Rosenbaum, 1999, p. 327). The propensity score can be thought of as the combination of

multiple factors that the researcher believes are related to the reason that participants join the intervention (e.g., level of interest, motivation or extraversion).

Mathematically speaking, propensity scores are typically created via logistic regression, and are the predicted probability that a student will participate in the intervention given a set of covariates (Luellen, Shadish, & Clark, 2005). Students' probability of participation is represented by a single calculated score that represents the probability of participating in an intervention, given the covariates. Students with the same propensity score, regardless of whether or not they were in the intervention or comparison group, have identical distributions on the set of covariates (Austin, 2011; Caliendo & Kopeinig, 2005; Ho et al., 2007; Stuart, 2010). Austin (2011) described propensity scores succinctly:

> First, the propensity score is a balancing score: conditional on the propensity score, the distribution of observed baseline covariates is similar between treated and untreated subjects. Thus, just as randomization will, on average, result in both measured and unmeasured covariates being balanced between treatment groups, so conditioning on the propensity score will, on average, result in *measured* [Author emphasis] baseline covariates being balanced between treatment groups. (p. 419)

Austin (2011) also emphasized that *unmeasured* covariates (i.e., variables not used in the propensity matching model) will *not* be balanced between intervention and comparison groups. Thus, only measured confounding variables (also referred to as "observed" covariates in the PSM literature) will be accounted for when PSM models are employed, a difference between PSM and random assignment.

The propensity score allows researchers to control for the systematic bias of measured confounding variables in order to render a more precise estimate of treatment effects (Rosenbaum & Rubin, 1983b; Rosenbaum & Rubin, 1984). If all factors related to participants' self-selection were known, the bias associated with self-selection could be ameliorated (Steyer, Gabler, von Davier, & Nachtigall, 2000). However, because we never know all of the possible factors and underlying motivations for students' participation in university programs, we are unable to confirm that all confounding variables have been accounted for.

The assessment of a freshman seminar program on student attrition serves as an applied higher education example (Clark & Cundiff, 2011). Researchers compared measured outcomes (e.g., attrition) with and without employing PSM techniques. Program effects were only significant once propensity scores were incorporated for covariates associated with both students' participation in and student attrition from the freshmen seminar program. Because the estimates of the program's efficacy were conflated with confounding factors prior to implementing PSM, practitioners may have drawn incorrect inferences from their assessment of this program, that it was ineffective. The inferences drawn from higher education assessment data are strengthened by isolating the impact of the intervention from confounding factors (Clark & Cundiff, 2011). Thus, effective interventions are more likely to be identified as such, despite self-selection confounds.

Several important implementation considerations for PSM have been noted by previous researchers (Austin, 2011; Caliendo & Kopeinig, 2005; Stuart, 2010; Stuart & Rubin, 2008a). Specifically, the process of conducting PSM involves a series of decisions

including the choice of covariates, models for creating propensity scores, matching distances and algorithms, estimation of treatment effects, diagnosing the quality of matches, and common issues and approaches for dealing with missing data (e.g., Caliendo & Kopeinig, 2008; Gu & Rosenbaum, 1993; Ho, King, & Stuart, 2007; Steiner, Shadish, Cook, & Clark, 2010; Stuart, 2010; Stuart & Rubin, 2008a). Because recommendations in the literature are numerous and come from a diverse assembly of disciplines, this paper highlights "best practices" relevant to applied university assessment.

Choice of Covariates for PSM Models

Ideally, propensity scores are created from variables related to self-selection into an intervention (i.e., potential confounds). Moreover, the inclusion or exclusion of key covariates affects the accuracy of inferences made about intervention efficacy (Brookhart et al., 2006; Steiner et al., 2010). Careful consideration should therefore be given to the selection of covariates, as matches will only be made based on the multivariate composite of the specific covariates the researcher decides to include in the model. Covariates not included in the model may systematically vary between groups and therefore lead to biased estimates and a lack of internal validity (Steiner et al., 2010, 2011). In particular, differences in the sets of covariates used to create the propensity score can affect variance and the associated error when estimating treatment effects (Brookhart et al., 2006). Additionally, including covariates that are not related to the outcome may also negatively affect the estimate of treatment effects (Brookhart et al., 2006). For example, if motivation is included as a covariate but is not related to students' self-selection into an honors program, estimates of the intervention's effects on students and student outcomes

may be attenuated. Conversely, if a confounding variable (covariate) is *not* included in the creation of the propensity scores, the bias in estimated treatment effects tends to increase as confounding effects increase (Drake, 1993).

Any covariate that almost completely accounts for the assignment to the intervention or control conditions should be viewed cautiously as it may indicate an underlying issue when attempting to make causal inferences (Stuart, 2010). If one covariate fully accounts for the assignment to either the intervention or non-intervention condition, it may be indicative of a third factor that is driving participation in that program. For example, if participation in a university program targeting underage drinking can be completely predicted by attendance, yet all students attend because of disciplinary requirements, then the occurrence of disciplinary action is what is driving attendance and, plausibly, behavioral change. In this example, a comparison group cannot be formed using this covariate (i.e., attendance) because the entire population of individuals matching the selection criteria (i.e., disciplinary action for underage drinking) is required to participate in the intervention.

Other important considerations regarding the use of covariates in PSM models include whether the variables can be accurately measured and the stability of covariates over time. Moreover, there is a distinction between covariates that are observable traits (e.g., personality traits via a personality inventory) rather than covert, unknown traits (e.g., unreported life events; Dehejia & Wahba, 1999). Pre-intervention variable characteristics are also important to consider including the length of time covariates were present prior to the intervention. For example, there may be notable differences between students who have felt efficacious their entire lives and individuals who have recently

increased to the same level of self-reported academic self-efficacy. Despite the same level of recent self-efficacy, time-related factors may go unmeasured.

Confounding Variables. In PSM modeling, only variables that influence both the decision for individuals to participate in the program and the outcome variable *in tandem* should be included as covariates (Caliendo & Kopeinig, 2008; Steiner et al., 2010; Steiner et al., 2011). For example, one criteria for students to qualify for a university honors program could be whether they scored above a specified benchmark on a standardized test (e.g., SAT). However, it is also plausible that their aptitude as measured by the test would be related to both their participation in the honors program and also their general performance on outcomes targeted by the honors program (e.g., GPA). In this example, the variables related to students' self-selection into a university program might directly contribute to their success on outcomes of interest. Thus, it would be inappropriate to presume that the intervention can be credited for desirable outcomes without accounting for confounding variables. Not only should one consider which variables are potential confounds, but the stability of covariates over time should also be considered.

Stability of Covariates. The stability of covariates over time should also be considered, as unstable covariates may not reliably reduce systematic bias related to self-selection. Findings from simulation studies have suggested that unstable covariates can appreciably reduce the ability of propensity scores to remove bias (Steiner et al., 2011). One example of an unreliable covariate would be a measure of affect (e.g., happiness). If the variable has the potential of changing frequently, it is therefore not exclusively a confounding variable related to self-selection and may introduce error if used to create a

comparison group. The covariates used to create a matched sample should be relatively stable across time in order to provide a valid and reliable measure of baseline characteristics. If covariates lack reliability, the model is unstable and may lead to invalid inferences about the effects of an intervention on participants (Shadish et al., 2002).

Additionally, variables that could be influenced by participation in the treatment should not be included in the creation of propensity scores (Greenland, 2003). This is particularly important if the covariates are used *retrospectively* and measured at the end of the intervention. For example, one covariate could be students' sense of belonging. However, if it is measured after students' participation in the program, their sense of belonging could have changed as a function of their participation. To ensure that the model does not include variables that have changed following participation, covariates should be measured *a priori* for both participants and nonparticipants (Austin, 2011; Stuart, 2010). In order to create the highest quality matches possible, not only is it important to consider the stability of covariates, but researchers also need to consider whether one covariate should be weighted differently than another.

Weighting Covariates. To ensure quality matches, some covariates may be weighted differently than others. For example, a researcher may feel that gender is an important matching variable. In this instance, exact matches can be made prior to creating propensity scores to ensure that individuals are directly matched on important covariates like gender or ethnicity (Caliendo & Kopeinig, 2008; Dehejia, 2013; Lechner, 2002). To use exact matching, subsamples are created prior to creating propensity scores (e.g., all White women are included in one subsample). The only possible matches that can be made will be from the subsample of similar nonparticipants (e.g., other White women).

Propensity scores are then calculated separately by subsample using other covariates associated with self-selection into the intervention. If exact matching is used to match participants and nonparticipants on specific covariates (e.g., gender), the covariates should not also be included in the creation of propensity scores.

Evaluating Covariates. When choosing covariates, it is important to include theoretically sound variables (Brookhart et al., 2006; Steiner et al., 2010). For example, standardized test scores are important to include as covariates when assessing an intervention aimed at fostering academic success, such as a university honors program. If standardized scores are a determinant of whether or not someone is admitted into an honors program, then without accounting for student performance on standardized tests, it is difficult to disentangle the impact of the program from students' incoming abilities.

Cross-validation has been championed as a convenient approach for deciding upon covariates (Frolich, 2004). Because researchers often do not know all of the factors associated with students' self-selection into a university intervention, the leave-one-out cross-validation approach affords researchers the flexibility to attempt multiple variations of covariate subsets to create balanced groups. Leave-one-out cross-validation may also be used to create *different* comparison groups using different sets (or "blocks") of variables (Caliendo & Kopeinig, 2008). With cross-validation, matched groups of participants and nonparticipants are created for each block of covariates, and then compared to one another on the overall quality of matches created. However, to avoid allegations of researcher bias in covariate selection, the covariates should be decided upon prior to estimating the effects the intervention on outcomes (Ho et al., 2007; Rubin, 2001). Similar to cross-validation, the current study will use two sets of covariates to

create matches then diagnose the quality of the matches. However, propensity scores must first be created.

Creating Propensity Scores

One uniform requirement for PSM, regardless of the method used, is that every individual must have a nonzero probability of participation in the intervention (Austin, 2011). If an individual has a propensity score of zero, it indicates that they have zero probability of having participated in the intervention conditional upon the covariates. In higher education, probabilities of zero can be problematic because it might mean that a confounding variable is *inhibiting* students from participating in a program. For example, female students may have a calculated probability of zero for participating in a fraternity based off of a set of covariates including gender. In this example, female students did not necessarily decide not to join a fraternity; rather, their gender determined their eligibility for participating.

Propensity scores may be calculated using various techniques (e.g., logistical regression, discriminant analysis, multiple regression, etc.) to create a multivariate composite of the covariates (Rosenbaum & Rubin, 1983; Stuart, 2010; Stuart & Rubin, 2008a). Several methods exist depending on the number or levels of programs offered (e.g., one intervention offered versus two variations of the same intervention). The most frequently used method is logistic regression (Austin, 2011; Stuart, 2010), which is the method that will be applied in the current study. Thus, the primary focus of the current paper will be on PSM using logistic regression.

Logistic Regression. The most commonly employed method for estimating propensity scores in the PSM literature is logistic regression (Austin, 2011; Stuart, 2010).

Logistic regression is a robust statistical technique that allows the use of continuous and categorical predictors (Tabachnick & Fidell, 2013). Because the covariates predict a binary variable (i.e., participation or not participation), the generalized linear model is used and a link function transforms the binary dependent variable into a binomial distribution (i.e., an "s" shaped distribution).

To calculate propensity scores using logistic regression, the set of covariates (i.e., confounding variables) are entered as predictors of the binary outcome. Because the errors are no longer normally distributed, ordinary least squares (OLS) is not appropriate. Logistic regression employs maximum likelihood estimation (MLE) to explain the maximum amount of deviance in the model (Azen & Walker, 2011). The amount of deviance left unexplained by the estimated parameters is the extent to which the data cannot be explained by the model (Azen & Walker, 2011).

Propensity scores are the predicted probability of participating in a program, given the set of variables related to self-selection. Scores are calculated for all students regardless of their participation in the intervention. Predicted probabilities (i.e., propensity scores) are continuous values bounded between 0 and 1. From probabilities, the odds of a student participating can be calculated to remove the upper bound.

Odds are the probability of an event occurring (e.g., a student participating in an honors program) divided by the probability of the event not occurring. However, odds are not an ideal transformation of the variable as negative odds values are not possible. Therefore, the natural logarithm of the odds provides a mathematical solution as it unbounds the lower asymptote and has no restriction of range (Osborne, 2012).

The natural logarithm of the odds (also known as the logit) is a transformation of the probability of participation (Osborne, 2012). Students' predicted participation in a program becomes the logit(*y*), the logit or log-odds of participating in the program, and the regression equation is simply:

$$\text{Logit}(y) = \text{a} + b_1 x_1$$

The logistic regression equation for predicted probability can therefore be obtained by working backwards from the equation for predicted logits.

$$\widehat{p_i} = \frac{1}{1+e^{-(B_1 X_2 + B_0)}} = \frac{e^{(B_1 X_2 + B_0)}}{1+ e^{(B_1 X_2 + B_0)}}$$

In the equation, $\widehat{p_i}$ equals the predicted probability of a student participating in the honors program, $e^{(-B_1 X_2 + B_0)}$ equals the rate of change in log odds, and B_i are the unique contributions of the covariates related to students' self-selection into the honors program (Cohen, Cohen, West, & Aiken, 2003, p. 486).

The creation of the propensity score is based upon the concept of the *counterfactual* (Winship & Morgan, 1999). Scores are calculated for all students predicting the probability of program participation regardless of whether students actually participated. For example, a propensity score of .5 indicates a 50/50 chance of students participating in a program (Cohen, 2003).

It is important to note that propensity scores are created for the sole purpose of producing a *balancing score* rather than for making inferences back to a population. We are simply using logistic regression as a "general method of nonparametric preprocessing, suitable for improving any parametric method" (Ho et al., 2007, p. 202). Because the covariates are related to students' self-selection into the program, it is reasonable to expect participants to have higher propensity scores than nonparticipants. The propensity

scores are then used to create a comparison group of nonparticipants who have a similar propensity for treatment as participants (Stuart, 2010; Stuart & Rubin, 2008a).

Matching Methods

Once propensity scores are computed, there are numerous approaches for creating a comparison group of nonparticipants including exact matching, nearest neighbor (NN) matching, optimal matching, and nearest neighbor with caliper adjustment (Austin, 2011; Caliendo & Kopeinig, 2005; Stuart, 2010; Stuart & Rubin, 2008b). However, the most commonly used approaches, and the methods applied in the current study, are NN and NN with caliper. Additional considerations include the number of nonparticipants to be matched to each participant and also whether replacement (i.e., matching nonparticipants multiple times to participants) is allowed.

Exact Matching. Exact matching is when "perfect" matches are created on specific covariates. It can be used as the only technique, or can be used in tandem with other matching methods. Exact matching on important covariates is ideal whenever possible to ensure a high quality comparison group (Austin, 2011). However, exact matching requires large sample sizes and homogeneous populations of participants and nonparticipants. Therefore, it is not frequently used as unmatched participants are dropped from the sample and subsequently reduce the sample size (Stuart, 2010). Thus, approaches such as NN are frequently employed.

Nearest Neighbor (NN). As it is often not possible to create a matched group that is exactly matched to the intervention group, algorithms can be used to find the closest possible match (Austin, 2011; Stuart, 2010). One approach to creating matches is the use of a "nearest neighbor" (NN) design (Gu & Rosenbaum, 1993). The NN design employs

a greedy algorithm that matches each individual in the intervention condition sequentially with the nearest possible nonparticipant's propensity score (Stuart, 2010; Stuart & Rubin, 2008a). Starting with the first participant, the greedy algorithm picks the best match out of the pool of possible matches, then moves on to the next participant. Because the greedy algorithm moves sequentially through the list of participants, it is possible for later matches to be less than ideal. Specifically, if a match is made late in the sequence, the most similar propensity scores may have already been "assigned." The algorithm does not "backup" and form a match if a nonparticipant was assigned in a previous iteration (Stuart, 2010).

Although, NN is one of the most commonly used matching methods, the use of the NN matching algorithm can result in bias and poor quality matches (Smith, 1997). The NN design does not allow for control of quality over the potential matches as matches will be made regardless of how much the nonparticipant's score differs from the participant's score. Rather, the matches are merely the "best option" out of all possible options within the pool of potential matches. Therefore, it is recommended that additional adjustments be made (Smith, 1997).

Matching with replacement is one option for overcoming the limitation of poor quality matches (Caliendo & Kopeinig, 2005; Stuart, 2010). Matching with replacement allows propensity scores of nonparticipants paired during a previous iteration to remain in the pool of potential matches. Although matching with replacement allows the highest quality matches possible for each iteration, pairing nonparticipants in multiple matches creates a potential issue as it violates the assumption of independence of observations (i.e., that each of the matches are unrelated to one another). This violation of

independence may also be problematic when estimating the effects of the intervention, as nonparticipants that are matched twice should only be counted as one individual assigned to the comparison group when conducting inferential statistics on the outcome variables (Stuart, 2010). To ensure high quality matches are made using a one-to-one matching ratio, additional adjustments may be used.

Nearest Neighbor (NN) with Caliper Adjustment. The use of a caliper adjustment has been frequently implemented to ensure a high quality of matches between the intervention and comparison groups (Austin, 2011; Caliendo & Kopeinig, 2005; Stuart, 2010; Stuart & Rubin, 2008a). A caliper is a specified distance within which the matches on the propensity score are considered acceptable and outside of which the matches are not acceptable and therefore not allowed. Using a caliper adjustment, non-participants are only matched to a participant if their propensity score falls within a designated distance (in propensity score standard deviations) from the participant's propensity score. The appropriate distance at which to set the caliper can be difficult to know *a priori* as researchers often do not usually know the distribution of possible covariates (let alone, the composite used to create the propensity score) prior to conducting analyses (Smith & Todd, 2005).

Although caliper distances have not been studied specifically in the context of educational program assessment, they have been examined within the medical context (Austin, 2009). Monte Carlo simulations indicated matches within a caliper distance of 0.2 and were optimal to estimate treatment effects (Austin, 2011). Although numerous matching algorithms exist, NN matching with a recommended caliper width distance of

0.2 standard deviations is most commonly recommended in the PSM literature (Austin, 2009, 2011; Stuart, 2010).

NN with a caliper adjustment compensates for some of the issues mentioned when only using the NN matching method. However, because NN matching with a caliper restricts the range of potential matches, NN (without a caliper adjustment) may be implemented when a researcher is unsure of the width of caliper that should be used to create matches. The present study will use both NN and NN with a predetermined caliper and will evaluate the quality of matches made using both techniques. However, before doing so, there are several other methods that are worth mentioning.

Optimal Matching. Similar to NN matching, optimal matching uses a greedy algorithm to pick the closest match from the pool of nonparticipants (Rosenbaum, 2002; Stuart, 2010). However, unlike NN matching, optimal matching allows for previously paired matches within the sample to be repaired with a different match based on the global fit and quality of all matches without matching with replacement (Stuart, 2010). Therefore, a match made during an earlier iteration may be broken to reassign a nonparticipant to a better match in an effort to increase the overall quality of propensity score matches.

Other Methods. Genetic matching has become a popular approach for creating quality matches (Diamond & Sekhon, 2013). Genetic matching techniques employ an iterative process to create matches. The weight of each covariate is readjusted within the composite scores to improve the overall balance of matches between the participant and nonparticipant groups. Simulation studies concluded that the use of this “evolutionary search algorithm” is more effective than NN at reducing selection bias in the sample as it

allows for the adjustment of covariate weights according to their contribution to self-selection and the outcome (Diamond & Sekhon, 2013). For the purposes of the current study, genetic matching will not be used because it is more complex and less practical than NN matching techniques. Additionally, limited recommendations in the literature and software options are available for using this technique. Regardless of which matching technique is used, the balance between the participants and nonparticipants should be compared to ensure that the distribution of propensity scores in each group is even. Once the intervention and comparison groups are evenly matched using PSM, the outcomes are compared between groups using typical inferential statistical analyses (Ho et al., 2007; Stuart, 2010). Prior to doing so, however, researchers should evaluate the quality of matches.

Comparing Balance

Once the matches are made, the quality of the matches is diagnosed to ensure the comparison group has a distribution of propensity scores similar to participants. Several approaches exist to diagnose matches including comparing the balance numerically and visually (Caliendo & Kopeinig, 2005; Stuart, 2010). There is a lack of consensus in the literature regarding the use of null hypothesis significance testing (NHST) analyses (e.g., *t*-tests) to diagnose the quality of matches on the covariates and composite propensity scores (Caliendo & Kopeinig, 2005; Ho et al., 2007; Rosenbaum & Rubin, 1985; Stuart, 2010).

Numeric Balance. Despite the common use of NHST analyses to compare the distribution of covariates and propensity scores in the PSM literature (e.g., *t*-tests), use of NHST for this purpose has been criticized in recent work (e.g., Ho et al., 2007; Stuart,

2010). Though the approach of using *t*-tests to compare balance is accessible to many researchers, the use of *p* values to compare balance is not appropriate for two primary reasons: 1) statistically significant differences may be found between groups because of power due to large sample sizes, and 2) there are no inferences being made in relation to a population as the comparison is only comparing the properties within the samples (Ho et al., 2007; Stuart, 2010).

To appropriately compare the balance of participants and nonparticipants, other approaches have been suggested. Stuart (2010) advised comparing the covariate balance (i.e., balance of propensity scores) by comparing the standardized difference of group propensity score means. Additionally, Stuart suggested comparing the ratio of variances between participants and nonparticipants on the propensity score and on each individual covariate. A researcher should also compare the mean of both groups on each covariate to determine whether the groups differ on any of the individual covariates to a degree greater than one-fourth of a standard deviation (Ho et al., 2007). In addition to numeric comparisons of balance, a visual inspection of the data also allows for further balance diagnosis.

Visual Balance. Several visual aids can be used to diagnose propensity score balance between groups (i.e., participants versus nonparticipants). Graphics used for this purpose include quantile-quantile (QQ) plots and jitter graphs (Ho et al, 2007; Stuart, 2010; Stuart & Rubin, 2008a). The visual inspection of these graphs simply involves the researcher "eyeballing" the distribution of propensity scores for each group across different criteria. For example, QQ plots display propensity scores across a probability distribution that is divided into quantiles. When visuals are pivotal in determining

whether the two groups are balanced, they may be included in the results to provide additional evidence of the balance between groups. Once the quality of matches is evaluated, the effects of the intervention can be estimated.

Intervention Treatment Effects

Depending on the research question, estimates of the treatment effects can be made for either 1) the impact of the intervention for only the participants (average treatment effect on the treated), or 2) to make inferences about the potential impact of the program for the overall student population (average treatment effect; Caliendo & Kopeinig, 2005; Ho et al., 2007). If the goal is to estimate treatment effects for only the individuals who participated, then the average treatment effect on the treated (ATT) can be easily estimated. In the context of ATT, the entire population of individuals of interest has data available to analyze on relevant outcomes (Austin, 2011; Imbens, 2004). Alternately, the goal might be to make inferences regarding the effects of an intervention as it would generalize to the overall population of students, regardless of whether they received treatment. In this situation, the average treatment effect (ATE) is estimated as the average effects weighted by the overall population baseline characteristics (see IPTW section; Ho et al., 2007).

To calculate ATT, the impact of a program (or "treatment effect" in the PSM literature) is estimated for participants only (Ho et al., 2007, p. 204). Where the average treatment effect (T) is the expected value of the outcome (Y_i) for the treated [Y_i (1)] minus the observed value of the outcome [Y_i (0)] for the untreated conditional upon the covariate (X_i).

$$\text{ATT} = \frac{1}{\sum_{i=1}^{n} T_i} \sum_{i=1}^{n} T_i E[Y_i\,(1) - Y_i\,(0) |\, X_i]$$

$$= \frac{1}{\sum_{i=1}^{n} T_i} \sum_{i=1}^{n} T_i[\mu_1\,(X_i) - \mu_0\,(X_i)]$$

Whereas to calculate ATE, the mean impact of the program is calculated for all of the individuals in the sample (Ho et al., 2007). Where the average treatment effect (ATE) is the expected value of the outcome (Y_i) for the treated [Y_i (1)] minus the expected value of the outcome [Y_i (0)] for the untreated conditional upon the covariate (X_i).

$$\text{ATE} = \frac{1}{n} \sum_{i=1}^{n} E[Y_i\,(1) - Y_i\,(0) |\, X_i]$$

$$= \frac{1}{n} \sum_{i=1}^{n} \mu(X_i) - \mu_0\,(X_i)$$

For example, program coordinators may want to estimate the impact of an honors program on honors students at the university. In this situation, the entire population of students of interest is served by the program. Thus, only the estimate of the program on the participants is relevant (i.e., ATT). However, if the program coordinators were interested in estimating how the program would generalize to a larger population of students, they would want to estimate the average effects of the honors program (i.e., ATE).

Once researchers have decided on the type of intervention effect that is most relevant to their study (i.e., either ATT or ATE), an estimate of the intervention effects can be calculated. The distinguishing characteristic of ATE (from ATT) is that it again brings us back to the idea of the counterfactual. Specifically, ATE is used to estimate the effects on nonparticipants for whom no intervention effects are measured. Though the distinctions between ATT and ATE are important, it has been suggested in the literature that the estimation of one tends to be a good estimator of the other (Ho et al., 2007). Typically, researchers are interested in the ATT, which will also be the focus of the current study. The impact of the honors program on participants in the current study will be compared using η^2 and Cohen's *d* effect sizes.

Models for Estimating Intervention Effects

Various other non-matching methods for estimating the effects of an intervention have been explored in the PSM literature. Though the present study will focus solely on PSM, other approaches use the propensity score to adjust estimates rather than as a matching parameter. Such models include: covariate adjustment via regression, stratification (also referred to as subclassification in the PSM literature), and inverse probability of treatment weighting (Austin, 2011; D'Agostino, 1998). Inverse probability of treatment weighting (IPTW) affords researchers the ability to estimate both the ATE and the ATT. Thus, the average estimated impact of the intervention can be calculated as it might generalize to nonparticipants (Austin, 2011). However, the most straightforward way of estimating treatment effects is to create the balanced groups and conduct traditional inferential comparison. Because other estimation models may be suitable to answer other research questions, three models will be briefly described.

Covariate Adjustment. In the covariate adjustment approach, intervention outcomes are regressed upon the propensity scores using either a linear model (for continuous outcomes) or logistic model (for dichotomous outcomes). The individual propensity scores are used as predictors in this approach (Austin, 2011; Hade & Lu, 2011). Similar to analysis of covariance (for continuous models), the means of the outcomes (e.g., GPA) can be adjusted using the propensity scores for participants and nonparticipants. For dichotomous outcomes, the outcome is adjusted using an odds ratio in logistic regression (Austin, 2011). The adjusted outcomes can then be compared using an independent t-test (comparing means between groups) or chi-square (comparing ratios between groups). Another approach that does not use propensity scores as a matching parameter is stratification.

Stratification. Also known in the literature as "subclassification," the stratification approach does not include matching. Rather, participants and nonparticipants are grouped into strata based on researcher-defined propensity score cutoff points (Austin, 2011; Stuart, 2010). The strata are created using a predetermined number of subgroups set by the researcher (Stuart, 2010). Individuals who have similar propensity scores are essentially grouped together so participants and nonparticipants within each stratum can be compared. Though stratification allows for comparisons at each strata, the estimates are still dependent upon whether individuals participated in the intervention. Inverse probability of treatment weighting, on the other hand, allows for an estimate of intervention effects that is *independent* of individuals' participation in the intervention.

Inverse Probability of Treatment Weighting (IPTW). The inverse probability of treatment technique is calculated as the inverse probability of the intervention the individual actually received (Austin, 2011a). The equation for calculating IPTW scores is as follows (Austin, 2011a, p. 408):

$$w_i = \frac{Z_i}{e_i} + \frac{(1 - Z_i)}{1 - e_i}$$

Where w_i refers to the inverse probability of treatment weight; Z_i refers to treatment (Z =1 refers to treated; Z = 0 refers to untreated) and e_i is the propensity score for each individual. If a person is treated, the second part of the equation falls off (1-1 = 0 in the numerator). If untreated, the first part of the equation falls off (0 in the numerator). IPTW models allow researchers to create a sample of both participants and nonparticipants that have weights on baseline characteristics independent of their participation. Intervention effects can then be estimated after adjusting for baseline characteristics in a specific population (Austin, 2011). IPTW is one method of calculating ATE; however, ATT rather than ATE is of interest in the current study. Therefore, the IPTW approach will not be employed. However, prior to comparing outcomes using ATT, participants and non-participants must have overlapping propensity score distributions.

Common Support

The extent to which participants and nonparticipants overlap in their distribution of propensity scores is referred to in the PSM literature as the area of "common support" (Caliendo & Kopeinig, 2005; Stuart, 2010). Differences in the distribution of propensity scores can be problematic and may restrict the number of nonparticipant matches with similar propensity scores (Caliendo & Kopeinig, 2005). Because NN matching with a caliper only creates matches within a predetermined range of scores, a lack of common support can result in fewer matched pairs. A lack of common support across participants and nonparticipants may also lead to a loss of information. Individuals who are qualitatively different across the groups might be excluded from the analyses because of the inability to find appropriate matches (Caliendo & Kopeinig, 2005; Stuart, 2010).

Figure 1 shows an example of the area of common support across propensity score distributions (ranging from 0 to 1). The area where there are propensity scores for both the intervention and comparison groups is indicated in the dashed window. A lack of common support can lead to difficulty matching nonparticipants to participants using NN matching with a caliper. A lack of common support can also lead to issues estimating the effects of an intervention. Specifically, when ATE estimates are of interest, a lack of common support may indicate that ATE cannot be estimated because participants and nonparticipants vary too greatly from one another to allow for a reliable estimate (Stuart, 2010). In situations when ATT is of interest, common support is needed to ensure that the estimation of intervention effects is reliable and representative of the group. Additionally,

there may be qualitative differences in the participants who are excluded from the comparison due to the unavailability of similar propensity scores (Stuart, 2010).

Outcome Variables

Once a quality subsample of nonparticipants is created as a comparison group, the analyses become quite simple. Preprocessing of the data to create a comparison group allows researchers to conduct simple inferential tests on the outcomes (Caliendo & Kopeinig; Gu & Rosenbaum, 1993; Ho et al, 2007; Stuart, 2010; Stuart & Rubin, 2008a). In sum, the tests researchers use following PSM techniques are *exactly the same* as they would be if simply comparing independent groups.

Outcome variables should be compared between groups only after matches are created and the quality of balance between participants and nonparticipants has been evaluated. Once the nonparametric preprocessing steps are finalized (i.e., propensity scores are created and participants and nonparticipants are matched), the threat of researcher bias in creating groups is no longer an issue. One way of ensuring the outcomes did not impact a researcher's decisions is to merge on the outcome variables *only after* all of the PSM preprocessing steps have been completed. Stuart and Rubin (2008a) noted that the inclusion of outcome variables after all matches have been made is critical for following PSM best practices.

Purpose of the Current Study

Because the literature regarding best practices for implementing PSM spans multiple disciplines, more research is needed on the application of PSM in higher education assessment. One recent study used PSM techniques to estimate the effects of an honors program on student GPA and time to degree completion (Keller & Lacy, 2013). The covariates included in the honors program PSM study, however, did not include

motivation measures as covariates. At the institution in which the current study was conducted, each year a select group of entering first year students are invited to join the honors program, based on entering SAT scores and high school grades. The honors program requires students to complete several rigorous honors' courses. In the final year of the program, students complete an honors thesis, requiring rigor above and beyond the typical undergraduate experience. The decision to join the honors program is purely voluntary, so one might expect that students who elect to join the program may differ from those who opt out of the program. It is anticipated that there may be self-selection bias related to academic motivation. Thus, honors program assessment provides a unique opportunity to investigate the use of PSM in higher education for two reasons:

1) Honors programs often require an additional investment of time and energy from the student. Thus, it is feasible that motivation may be related to self-selection into the program and important to include as a PSM covariate.

2) Because motivation may differ between honors students and non-honors students, the area of common support may not be as robust as for other university programs (e.g., general education courses).

Applying PSM best practices to the applied honors program example, the research questions for the current study are four-fold:

1) Do motivation characteristics differ systematically between students in the honors program and students who qualified for the honors program but chose not to participate? Additionally, do honors students and the general student population differ in motivation?

2) How does the quality of propensity score matches differ when created from two different sets of covariates? Specifically, matched groups will be created from two

different covariate sets – one that includes standardized test scores and one that includes a university-created score, based on high school performance.

3) Which matching condition results in high quality matches while preserving information and retaining honors students in the final comparison? Specifically, there were three different matching conditions: NN and NN with caliper distances of 0.1 and 0.2 standard deviations. Because use of a strict caliper can result in loss of information, it was important to evaluate any loss of information resulting from the application of the calipers.

4) Do matched groups of honors students significantly and/or practically differ from non-honors students and students in the general population on outcomes targeted by the honors program?

Chapter Three

Methods

Participants

Participants in the current study were enrolled in a mid-sized public university in the Mid-Atlantic U.S. Participants completed a battery of cognitive and noncognitive assessments at two time points – during orientation to the university, and again when they had completed between 45-70 credits. Data from three subsamples of students were compared: 1) students enrolled in the university honors program, referred to as the "Honors Program Sample" or "the participants"; 2) students who qualified for the honors program but did not accept the offer of admission to the program, referred to as the "Non-Honors sample" or "non-participants'; and 3) students who did not qualify for the honors program, referred to as the "General Student Population." Matched comparison groups were selected from each of these three groups.

Honors Program Sample (i.e., Participants). Honors students were 181 first-semester undergraduate students enrolled in the honors program at James Madison University in the fall semester of 2012. There were a higher number of females (57.9%) than males enrolled in the honors program and a high percentage of students in the honors program identified as White (84.2%). The average age of students in the honors group in fall of 2012 was 18.38 (SD = 0.35).

Non-Honors Sample (i.e., Non-participants). Non-participants were 836 non-honors first-semester undergraduate students also attending James Madison University in the fall semester of 2012. Similar to the honors program sample, there were a higher number of females (60.4%) than males and a high percentage of students in the non-honors program identified as White (86.6%). The average age of students in the non-

honors group in the fall of 2012 was 18.42 (*SD* = 0.35). The importance of identifying the non-honors sample was that they were offered admission into the honors, yet opted not to participate in the program.

General Student Population (GSP). The general student population (GSP) consisted of 2,836 undergraduate students attending James Madison University. Like the honors and non-honors samples, all students in the general university sample were also first-semester students. A similar percentage of students in the GSP were female (61.5%) and identified as White (87.5%) as the honor and non-honors samples. The average age of students in the general university sample in the fall of 2012 was 18.45 (*SD* = 0.48). It is important to note that students in the GSP were not eligible for the honors program and not invited to participate.

Procedure

Data were collected during two university-wide Assessment Days. Assessment Day is a university-wide class-exempt day during which students complete multiple cognitive and attitudinal measures for university assessment purposes. The first Assessment Day was during the fall semester 2012 orientation to the university. The second Assessment Day was in the spring of 2014 when students were midway through their sophomore year and had accrued 45-70 credits. The covariate measures were collected during the first Assessment Day and via archival institutional data. Pre-test scores on three cognitive tests (outcome measures) were collected on the first Assessment Day. Post-test scores for the three cognitive tests were collected on the second Assessment Day.

Covariate Measures

Covariates were selected *prior* to comparing group means on the outcome variables. Two sets of covariates were included in separate analyses, for purposes of creating the propensity scores. The first covariate set included demographic variables, motivation for general education coursework, SAT Math scores, SAT Verbal scores, and the number of transfer credits accepted by the university. The second covariate set included demographic variables, motivation for general education coursework, the number of transfer credits accepted by the university, and a university-computed honors rating score (ARS scores). Because the only difference between the two sets is whether SAT or ARS scores were included, the two covariate sets will heretofore be referred to as 1) the SAT covariate set and 2) the ARS covariate set.

Demographic Variables. Because the non-honors and GSP samples consisted of a higher proportion of females and White students than the honors sample, gender and ethnicity were included in both covariate sets. Gender was dummy-coded as 0 (male) and 1 (female). Ethnicity was dummy-coded (using 0 and 1) for each of the ethnicity classifications as identified at the institution as separate variables including "White," "Hispanic," "Native American," "Pacific Islander," "African American," and "Asian." Also, note that the ethnicity groupings are not mutually-exclusive. For example, someone could self-identify as both "White" and as "Pacific Islander."

Student Motivation. Student motivation was operationally defined via an expectancy-value framework (Eccles, 1983). According to expectancy-value theory, students' motivations are a function of their *expectancies* regarding their ability to complete academic tasks (e.g., "I expect to do well in my classes this semester.") and the

value they place on the tasks (e.g., "I think my classes this semester are worthwhile"; Eccles, 1983). The *cost* associated with completing a task (e.g., "I think my classes require too much time and effort for me to do well.") is negatively related to motivation (Wigfield & Eccles, 2002). Specifically, students' motivation for their undergraduate coursework was measured via three subscales: expectancy (4 items), value (6 items) and cost (6 items). Students responded to each of the items on a 1-8 scale (from "completely disagree" to "completely agree"), and no items on the three subscales were reverse-scored. Items were summed to create total scores on each of the three subscales. Scores on the expectancy subscale ranged from 4-32, and scores on the value and cost subscales ranged from 6-48. The internal consistency reliability estimates of scores for each of the three motivation subscales are indicated in Table 1.

Standardized Test Scores. Standardized test scores were predominantly SAT Math and SAT Verbal scores because of the geographic location of the institution. Possible SAT Math and Verbal scores range from 200-800. If students had data on ACT scores rather than SAT scores, the ACT-SAT Concordance was used to convert ACT scores to the SAT scale (ACT, 2013). Standardized test scores were only included in the SAT covariate set.

Transfer Credits. Students' incoming credits were obtained via institutional research data. Incoming credits are the number of college-level credits that were accepted and transferred to the institution from students' previous coursework or Advanced Placement exams. Students with no incoming credits were assigned a 0.

Honors Program Score (i.e., ARS Scores). ARS scores are computed using high school grades from classes that are deemed relevant to collegiate coursework. Relevant

classes are then assigned points for the letter grade earned in each class (e.g., 12 points for an A, 11 points for an A-, etc.) then averaged to create a final score out of 12 possible points. Scores are rounded to the nearest possible whole number (e.g., a score of 11.4 would round to an ARS score of 11). ARS scores were only included in the ARS covariate set, and were only available for the honors and non-honors samples, and not for the GSP sample.

Outcome Measures

Four measures served as outcome measures: American Experience test (AMEX), Global Experience test (GLEX), the Natural World test (NW; version 9), and sophomore grade point average (GPA). The pre-test scores for the AMEX, GLEX, and the NW were collected during the fall 2012 Assessment Day when students were first entering the university. The post-test scores for the AMEX, GLEX, and NW were collected during spring 2014 assessment day when the students accrued 45-70 credits.

American Experience (i.e., AMEX). The American Experience test is a 40-item selected response test written to assess students' knowledge of American history and politics. The average internal consistency of scores from repeated administrations of the AMEX across over a decade of administration at the current university has been approximately .88 (Cronbach's alpha). Items on the AMEX are scored as either correct or incorrect and range from 0-40. The correlations between the AMEX test and general education American history and politics course grades has been documented at the institution as moderate and positive (DeMars, 2014).

Global Experience (GLEX). The Global Experience test is a 32-item selected response test written to assess students' global and political knowledge. The average

internal consistency of scores from repeated administrations of the GLEX across over a decade of administration at the current university has been approximately .76 (Cronbach's alpha). Items on the GLEX are scored as either correct or incorrect and range from 0-32. The correlations between the GLEX test and general education global studies course grades have been documented at the institution as moderate and positive (DeMars, 2014).

Natural World Test (NW). The Natural World Test (NW) is a 66-item selected response test written to assess scientific and quantitative reasoning (Sundre, 2008). Items are scored as either correct or incorrect and range from 0-66. The correlations between the NW test and general education science course grades has been documented at the institution as moderate and positive (Johnston, Hathcoat, & Sundre, 2014).

Student GPA. The honors program director identified cumulative GPA as an outcome of interest. Student GPAs were provided via institutional research data at the time of post-test and were reported on a 4.0 scale. Therefore, the range of possible values was 0 to 4.0. To remain in good academic standing, students are required to maintain a cumulative GPA of 2.0 or greater during their undergraduate career.

Data Screening

Prior to creating propensity score matches, the data were visually screened for outliers and response set. Individuals who responded with answers outside of the range of the response scales were recoded as missing data. Listwise deletion was conducted for two reasons. First, propensity scores could only be computed for cases with complete data. Second, missing covariate and outcome data were missing at random, given that

students were randomly assigned to assessment testing rooms, and completed the tests assigned to their particular room.

Chapter Four

Results

All data processing and analyses were completed in R Version 3.1.1 (R Core Team, 2013) and in IBM SPSS Statistics (Version 21). Prior to creating matches, all of the covariates were graphed using the ggplot2 package in R (Wickham, 2009). Covariate density distributions were compared across the three groups (see Figures 2-7). Overall, the three student groups had similar distributions on each of the covariates. That is, visually, the groups' distributions appeared similar on each of the individual covariates.

Similar to the density plots, descriptive information for each of the three student groups also indicated that the groups' distributions were similar on each of the covariates (see Table 1). Groups differed most on the number of transfer credits that were accepted by the university. For each of the three groups, the standard deviations for the number of transfer credits accepted at the university were high relative to the other covariates.

Table 2 includes demographic information for each of the three student groups. Students were able to identify as more than one ethnicity; thus, the number of students represented in the columns in Table 2 is higher than the number of students indicated in each student group. Because propensity scores would be computed via logistic regression, the data were screened for sparseness prior to conducting the analyses. Additionally, there was overlap in the distributions for each of the covariates across each of the three groups suggesting that there was adequate common support.

Research Question 1: Do motivation characteristics differ systematically between students in the honors program and students who qualified for the honors program but chose not to participate? Additionally, do honors students and the general student population differ in motivation?

Prior to creating propensity score matches, it was important to determine whether groups differed systematically on motivation. That is, if groups did differ on motivation, it could indicate that students who chose to participate in the honors program are more motivated than the non-honors students who were invited to participate but chose not to. Three analyses of variance were conducted to evaluate whether the three groups of students differed on the motivation subscales: expectancy, value and cost. The three levels of the independent variable included three student groups: honors students, non-honors students, and the GSP. If the motivation of the groups differed, it could indicate that motivation is associated with students' participation in the honors program.

Because the groups were so discrepant in size, careful attention was paid to the assumption of heterogeneity of variances. Variances were nearly identical in each instance. All possible variance ratios (F-max) between groups were near 1.

Overall, there were no significant differences among any of the student groups on the three motivation subscales. The three groups did not differ on average expectancy for general education $F(2, 4042) = .212$, $p = .809$. Additionally, the three groups did not differ on average value for general education coursework $F(2, 4024) = .142$, $p = .868$. Finally, the three groups did not differ on perceived cost of general education coursework $F(2, 3975) = .560$, $p = .571$. The relationship between student groups and each of the three motivation subscales was miniscule ($\eta^2 < .001$), with student groups accounting for

less than 0.1% of the variance in general education course expectancy, value, and cost. Overall, there were no differences among honors students, non-honors students, and the GSP on the three motivation subscales. Thus, the absence of differences among student groups could indicate that motivation was not related to honors students' self-selection into the program. Nonetheless, slight differences in motivation may still contribute to propensity scores (i.e., probability of participation). Thus, despite the lack of group differences on academic motivation, it was included in the creation of propensity scores in Research Question 2.

Research Questions 2: Does using a different set of covariates to created matches (i.e., SAT scores versus ARS scores) provide a higher quality of propensity score matches?

To answer the second research question, a series of nine PSM models were conducted via logistic regression (MatchIt; Ho et al., 2007). Thus, there were nine different conditions in this study resulting in nine different pairs of matched groups. Six conditions involved forming a matched comparison group of Non-Honors students. Of those six conditions, three groups were formed from the SAT covariate set and three matched groups were formed from the ARS covariate set. Within the two covariate set conditions, there were three matching conditions: NN, NN with a caliper of 0.2 standard deviations and NN with a caliper of 0.1 standard deviations. The other three conditions involved forming matched comparison groups of students from the GSP. Each of the three conditions were formed from the SAT covariate set, as ARS scores were not available for the GSP. The GSP comparison groups were also formed under the same

three matching conditions: NN, NN with a caliper of 0.2 standard deviations, and NN with a caliper of 0.1 standard deviations.

Numerical and visual inspections were conducted and compared across the nine conditions. The numeric diagnosing of matches included a comparison of the mean differences between the honors group and the comparison group on each of the individual covariates. Additionally, the variance ratio and standardized mean difference of the propensity scores between groups were examined. Visual diagnosis of matches included an inspection of individual covariates using a Quantile-Quantile (QQ) plot and a visual inspections of the distribution of propensity scores using jitter graphs and histograms.

Though students were matched on the multivariate composite (i.e., propensity score) created from the covariates, it was important to determine how similar the matched groups were on each of the individual covariates. Table 3 displays the means, standard deviations, and standardized mean differences (i.e., effect sizes) on individual covariates across the nine matching conditions. Overall, both covariate sets resulted in groups that were similar on the individual covariate means and standard deviations for each of the three matching conditions.

The standardized mean differences for individual covariates were compared for matches created in each of the six honors/non-honors matching conditions (Table 3). For the SAT covariate set, the standardized mean difference between honors and non-honors students' means on each of the individual covariates was under $d = 0.10$. Matched pairs created from the ARS covariate set also had similar means on each of the covariates ($d < 0.10$), with the exception of the cost variable ($d = -0.10$). In the nearest neighbor with 0.2 caliper matching condition, groups differed more on individual covariates when matches

were created from the SAT covariate set than when matches were created from the ARS covariate set. Specifically, standardized mean differences on SAT Verbal (d = 0.15) and cost (d = 0.12) were higher than that of any of the standardized mean differences for the covariates in the ARS set. In the nearest neighbor with 0.1 caliper matching condition, standardized mean differences were similar across the two covariate sets. For the SAT covariate set, the largest the standardized mean differences were for SAT Math (d = 0.16) and SAT Verbal (d = 0.18). Matched groups created from the ARS covariate set differed on expectancy (d = -0.16) and value (d = -0.14).

The quality of matches created in the three honors/GSP conditions was comparable to the quality of matches created in the six non-honors conditions. The means and standard deviations were also similar across the three GSP matching conditions (see Table 3). In the NN matching condition, the two groups were the most disparate on standardized mean differences for SAT Verbal (d = 0.11) and value (d = -0.11). In the NN with 0.2 caliper matching condition, the honors and GSP groups differed more than the honors and non-honors matched samples. Specifically, the effect size for the mean differences between groups for value (d = -0.13), and the number of transfer credits accepted by the university (d = -0.15) were the most disparate. In the NN with 0.1 caliper matching condition, there were no notable standardized mean differences (i.e., above d = 0.10).

Overall, the standardized mean differences for individual covariates across the nine matching conditions were small. The largest standardized mean differences on individual covariates still fell under the benchmark for small effect sizes (Cohen, 1992; Normand, Landrum, Guadagnoli, Ayanian, Ryan, Cleary, & McNeil, 2001) and were less

than the recommended .25 standard deviations (Ho et al., 2007). Thus, the effect sizes associated with even the most disparate mean differences on individual covariates indicated that quality matches were made.

In addition to comparing the covariates individually, the quality of matches was evaluated by examining each groups' propensity scores on the multivariate composite (i.e., the propensity score). Table 4 displays the variance and means of each group's propensity scores by covariate set and matching condition, and includes variance ratios and standardized mean differences between groups in each condition. The variance ratio is calculated by dividing the variance of the honors group (i.e., participants) propensity scores by the non-honors or GSP (i.e., non-participants) propensity scores. Ideally, the variance ratio (VR) should be close to one (Stuart & Rubin, 2008a). And, indeed, the VR was one in each of the nine conditions (see Table 4). Thus, the quality of matches based on the VR at each level was optimal.

The SMD is calculated by subtracting the standardized mean propensity score value of the honors group (i.e., participants) from that of the non-honors and GSP groups (i.e., non-participants). Ideally, the standardized mean difference (SMD) between the propensity scores of two matched groups should be near zero. And indeed, the SMD was near zero for each of the conditions (see Table 4). Thus, based on both VR and SMD, the quality of matches was optimal across the nine conditions again suggesting high quality matches across each of the nine conditions.

A final method for comparing the quality of matches involved visually diagnosing the matches. As an example, Figure 7 is a quantile-quantile QQ plot produced by the MatchIt package in R (Ho et al., 2007). In the QQ plot, the covariate (listed on the left)

for the treatment group is plotted against the same covariate for the untreated group. Note in the left half of Figure 7, the entire sample for the treated group is plotted by quantile against the untreated group. Note in the right half of Figure 7, the two matched groups' scores by quantile are again plotted for each covariate. If the scores fall along the 45 degree line, it indicates that there are no differences in the empirical distribution of scores (Ho et al., 2011).

The QQ plot allows the researcher to visually compare how similar each group is at each quantile in the group's distribution on each of the covariates overall (left column) and after creating matches (right column). Note that the majority of points remain near the center line for the matched QQ plots. This pattern indicates that participants and nonparticipants at each quantile in the distribution had similar scores on the covariates. Visual inspection of the QQ plots for each of nine conditions indicated that the groups were balanced on the individual covariates used to create propensity scores.

A second method of diagnosis involves a visual inspection of jitter graphs. Figure 8 displays the jitter graphs for the nine matching conditions (created via the MatchIt package in R; Ho et al., 2007). Note the propensity score scale along the x axis. Given that propensity scores serve as a method of providing a matched (balanced) comparison group, the jitter graphs provide a visual method for examining the balance of propensity score distributions across matched groups. The jitter graphs can also be used to preliminarily examine loss of information, which will be addressed in the third research question. For example, note in the lower left hand graph in Figure 8 that the honors student with the highest propensity score was not included in the matched sample.

Nonetheless, in each of the nine conditions, a comparison group was created that had a similar distribution of propensity scores.

The distribution of propensity scores for the original full samples and the matched samples can also be visually compared via histograms. Figure 9 displays histograms for diagnosing matches created in the MatchIt package in R (Ho et al., 2007). Once again, note the similarity of the distributions of propensity scores of the matched groups, suggesting that high quality matches were made across the nine conditions.

In sum, based on numeric and visual inspection, each condition resulted in quality matches. However, next it was important to evaluate any loss of information that may have resulted from the various matching conditions, particularly those that employed strict calipers. Thus, the amount of information lost when a stricter caliper was employed (i.e., 0.1 standard deviations) was evaluated in the third research question.

Research Question 3: Can one of the different matching techniques (e.g., NN and NN with caliper distances of 0.1 and 0.2 standard deviations) provide high quality matches while preserving information and retaining honors students in the comparison?

To answer the third research question, the demographic representativeness of students retained in the nine matching conditions was explored. Given the high quality of matches across all nine conditions in the second research question, it paved the way for further evaluation. Specifically, loss of information was evaluated in this third research question.

Loss of information was operationalized in this study as the retention of minority representation in the honors group. Table 5 displays the number of individuals by gender

and ethnicity across each of the nine conditions. Most notably, the number of Black students retained in the honors group was most impacted (in the honors-non-honors SAT covariate set condition). Of the original five Black males and ten Black females, only one male and one female were retained in the NN with 0.1 caliper condition. Conversely, 52 of the original 65 White males and 73 of the original 89 White females were retained. Although students were lost from both ethnicities, proportionately more Black than White students were non-matched; thereby resulting in loss of information from the Black students. Other losses of minority representation in this condition included: two of five Asian males, the only Pacific Islander male, one of two American Indian females, two of three Hispanic males, and three of eight of Hispanic females. Similar trends in loss of information occurred in the other two matched groups as the caliper width decreased to 0.2 and 0.1.

Because several minority groups included only a few honors students, the use of a caliper resulted in the forfeit of minority group representation. Therefore, the final comparison of outcomes was no longer representative of the original honors population. Thus, estimates of the impact of the honors program may only apply to the population of students retained in the comparison group. Because the quality of matches was already high using NN matching, the loss of information that occurred as a result of the strict caliper matching was not compensated for by an increase in the quality of propensity score matches. In fact, the quality of all matches was high, regardless of matching condition. Therefore, given that there was no substantial loss of information for the NN condition, it made sense to champion the NN conditions.

Research Question 4: Are honors students significantly and/or practically different from non-honors students and students in the general population on outcomes targeted by the honors program?

To answer the fourth research question, matched honors students and the comparison groups were compared on outcome measures across each of the nine conditions. Prior to comparing student groups on the outcomes, density distributions by student group were plotted for each of the outcome variables (see Figures 10-16). Overall, the groups' were similar on both the fall 2012 and spring 2014 scores on the AMEX, GLEX, and the NW tests. The three student groups also had similar density distributions for spring 2014 cumulative GPAs.

The outcome variables were *not* included in the data set prior to creating matches. Rather, all outcome variables were merged on to the data set only *after* creating matches. The internal consistency reliability estimates of scores for each of the outcome measures are indicated in Table 6. Also note that each of the matched samples across the nine conditions may include different subsamples of students.

None of the students in the sample were assigned to complete all three of the tests during Assessment Day because of testing time constraints and cognitive fatigue considerations. Thus, by design, no students had complete data for the three tests. However, because students were randomly assigned to the particular tests they completed for Assessment Day, students with scores on the three cognitive tests (i.e., the AMEX, GLEX, and NW) represented a random subsample of students.

Three 3x2 mixed ANOVAs were conducted for the three outcomes of interest to the honors program (AMEX, GLEX, and NW) for each of the nine conditions.

Independent *t*-tests were conducted to compare average sophomore-level GPA's across the matched groups in each of the nine conditions. Because multiple ANOVAs and independent *t*-tests were conducted, the critical value was set at the $\alpha = .01$ level to reduce the risk of making Type I errors.

Table 7 includes the AMEX means, standard deviations, and eta squared values for each of the matched groups. Overall, the honors and the comparison groups' average AMEX scores were similar in each of the nine conditions. The interaction effect of "group" by "time" was also not significant at the $\alpha = .01$ level. The main effect of "group" was not significant at the $\alpha = .01$ level, indicating that students in the honors group did not perform better than the comparison group on the AMEX. Finally, the main effect of "time" was also not significant, indicating that both honors and students in the comparison group did not perform significantly better on the post-test (i.e., spring 2014) than they had on pre-test. Overall, only 3% or less of the variance in AMEX scores was accounted for by the main effects and interaction effects across the nine conditions.

Table 8 includes the GLEX means, standard deviations, and eta squared values for each of the matched groups. Again, the honors and the comparison groups' average GLEX scores were similar in each of the nine conditions. The interaction effect of "group" by "time" was not significant at the $\alpha = .01$ level. The main effect of "group" was again not significant at the $\alpha = .01$ level, indicating that students in the honors group did not perform better than the comparison group on the GLEX. Finally, the main effect of "time" was also not significant, indicating that both honors and students in the comparison group did not perform significantly better on the second test (i.e., spring 2014). Overall, 5% or less of the variance in GLEX scores was accounted for by the main

effects and interaction effects across the nine conditions. For the honors and non-honors ARS covariate set matched groups, there was a small practically significant main effect for time that increased as the caliper distance decreased.

Table 9 includes the NW means, standard deviations, and eta squared values for each of the matched groups. The sample sizes of honors, non-honors, and GSP students for whom there was data was not ideal across the nine conditions. Once again, the honors and the comparison groups' average NW scores were similar in each of the nine conditions. The interaction effect of "group" by "time" was again not significant at the $\alpha = .01$ level. The main effect of "group" was again not significant at the $\alpha = .01$ level, indicating that students in the honors group did not on average perform better than the comparison group on the NW. Finally, the main effect of "time" was not statistically significant, indicating that both honors and students in the comparison group overall did not perform significantly better on the second test (i.e., spring 2014). However, note that although the interaction was not statistically significant, there was a large practical effect (Kirk, 1996). That is, 15% of the variance in NW-9 scores could be attributed to the interaction.

Although the scores for the overall honors sample started near those of the overall GSP sample (see Table 6), the subsamples of students included in this matched condition are not reflective of the original samples (see Table 9 and Figure 17). Specifically, after matching, the subsample of 10 honors students (Honors 0.2) selected in this condition scored similarly well on both the pre-test and post-test. However, given that the overall honors sample scored approximately four points lower than the honors subsample on the NW test and slightly increased on the post-test, the matched honors subsample was no

longer representative of the original honors sample. Thus, the honors subsamples' scores on the NW were no longer reflective of the overall honors groups' scores on the NW. One would expect the GSP subsample to differ from the original GSP sample because we were creating a matched group on their propensity for treatment, the counterfactual. However, given that the honors groups' participation in treatment is *known*, one would hope that their average scores would not be altered through matching.

Note that averages on the NW test across the nine conditions represented outcomes of potentially *different groups of students.* Recall that students from the original honors or non-honors/GSP pool of students may or may not have ended up in the final matched samples. Thus, a practically significant interaction was present for the honors versus GSP comparison in the 0.2 caliper condition *only after* two honors students and two GSP students were cut from the NN comparison groups. Moreover, note that the practical significance of the interaction term was attenuated as additional students were dropped from the sample in the 0.1 caliper condition.

Table 10 includes GPA means, standard deviations, and effect sizes (i.e., Cohen's *d*) for each of the nine conditions. Honors students' GPAs did not significantly differ from the comparison groups at the $\alpha = .01$ level across each of the nine conditions. However, in the honors/non-honors SAT covariate set NN condition, there was a small effect size with the non-honors subsample of students having higher average GPAs than the honors student subsample.

Overall, there were no statistically significant differences between the honors and comparison groups across the nine conditions on the AMEX, GLEX, NW, and GPA. However, in one condition, the interaction effect for the NW was practically significant

and there were small to medium practically significant increases in GLEX scores across time for the ARS covariate set conditions. Additionally, there was a small effect size between the honors and non-honors subsamples on GPA for the SAT covariate set conditions. Thus, the findings across the nine conditions were not the same.

Because practically significant differences were not found systematically across the nine conditions, a researcher might draw different inferences depending on the condition he examines. It is important to keep in mind that each of the outcome comparisons was conducted using different subsamples of students. As the caliper decreased, the sample size decreased as well. Therefore, the representativeness of the subsample to the original full sample of students is an important consideration as inferences are drawn.

Chapter Five

Discussion

The purpose of this study was to evaluate propensity score matching (PSM) techniques in the context of higher education assessment. Specifically, the four research questions addressed implementation considerations and decisions that assessment practitioners would be faced with when using PSM techniques to assess university programs. The first research question compared honors students, non-honors students, and students in the general student population (GSP) on motivation measures prior to creating matches. The second research question compared the quality of matches for different comparison groups created from two different covariate sets. The third research question weighed the improvement in quality of matches against loss of student demographic information as caliper distances became stricter (e.g., 0.2 and 0.1 standard deviations). Finally, the fourth research question compared matched groups across the nine conditions on academic outcomes of interest to the honors program.

Research Question #1

Prior to creating propensity score matches from a covariate set that included three motivation subscales, the three student groups' motivation scores were compared via one-way ANOVA. Honors students, non-honors students, and GSP students did not significantly or practically differ on any of the three motivation subscales. The finding that students did not differ could indicate that motivation was not related to honors students' self-selection into the program. If motivation was a confounding variable, levels of expectancy, value, and cost for the honors students would be expected to differ systematically from the other two groups. Additionally, a lack of differences on average

motivation scores or in the density distribution plots for each of the covariates may have foreshadowed the results of the second and third research questions. Specifically, the fact that students did not differ on the motivation measures suggested that motivation was not related to honors students' self-selection into the program

Research Question #2

The quality of matches was diagnosed numerically and visually across the nine conditions. Overall, the quality of matches was high. The SAT and ARS covariate sets did not result in any noticeable differences in the quality of the matches created. Numerically, the mean values for individual covariates were well-balanced across the nine conditions. The variance ratios and standardized mean differences were also either at or near optimal levels, indicating that groups had similar distributions of propensity score values. Finally, visual inspection of the QQ plots, jitter graphs, and histograms suggested high quality, balanced matches across all of the nine conditions.

Although the purpose of implementing a caliper distance was to increase the quality of matches, no improvements in quality of matches were observed as the caliper distance was reduced to a distance of 0.1 standard deviations. Moreover, because the quality of matches was near optimal using only NN matching, the quality of matches did not improve when a caliper distance was used. The creation of stricter matches within the caliper distance of 0.2 and 0.1 standard deviations merely resulted in the exclusion of students from the honors and comparison matched groups. Thus, it was important to determine *which students* were excluded from analyses when the stricter caliper distance was used to create matches.

Research Question #3

To evaluate the loss of demographic information, the number of minority students excluded from the sample at each caliper distance was inspected. Overall, students from minority groups (e.g., Asian, Pacific Islander, etc.) were excluded at a proportionately higher rate than White students. Because there were few minority students in the original honors program sample, the loss of these students notably altered the composition of the comparison groups. Thus, the comparison groups created using a caliper distance of 0.2 and 0.1 standard deviations were not representative of the original honors student population.

Overall, minority students were at a higher risk than the majority students of being excluded because ethnicity was dummy coded and used to calculate propensity scores. As the caliper distance was reduced, the number of potential matches that had a similar multivariate pattern of scores decreased. Thus, it became more difficult to match minority students and they were consequently dropped from the comparison sample. Additionally, because a higher proportion of honors students were female, males were also more likely to be excluded as the caliper distance was reduced.

The exclusion of minority students from the final outcome comparisons resulted in honors samples that were not wholly representative of the original honors population. Thus, the quality of inferences an assessment practitioner is able to make regarding the impact of the honors program on *all participants* may be depreciated when a strict caliper is applied. For example, an assessment practitioner may be interested in generalizing the estimated impact of the honors program on student GPA to *all honors students*. However, he must first take into account the fact that the demographic composition of the comparison groups is no longer representative of the overall honors student population. If

differences in the outcome (e.g., GPA) are also correlated with students' demographic characteristics, it would be inappropriate to generalize the estimate to the entire honors population.

The loss of minority students in outcome comparisons also resulted in a loss of information regarding how specific minority students performed on outcome measures. For example, an assessment practitioner could be interested in how the program outcome of GPA differs between White and Asian honors students. However, if the composition of the comparison groups again differs from the overall honors population, the estimate may not be representative of the two ethnic groups in the original honors population. One solution for the practitioner interested in preserving the representation of all demographic sub-groups would be to exact match on those sub-groups. For example, if the sample size is large enough, the researcher may want to create exact matches for Black females and Black males, and then create propensity scores for the exact matched sub-groups.

Finally, there is a cost/benefit decision that must be made regarding the use of calipers and the potential loss of information. In the present study, the quality of matches did not increase as the caliper distance was reduced. Hence, the loss of information did not come with any benefit. Moreover, the comparison of individual covariates indicated that the NN with a caliper of 0.1 actually formed the worst quality of matches. Had the quality of matches improved markedly through the use of a caliper, the quality of inferences might have also been improved. Thus, the benefit associated with using a caliper might be outweighed by the cost associated with losing minority representation in the final comparison of outcomes.

Research Question #4

To answer the fourth research question, 2x2 mixed ANOVAs and independent samples *t*-tests were conducted to compare student performance on the outcome measures. Outcome measures were compared for the student groups in each of the nine conditions. The main effects of time (fall 2012 and spring 2014) and group (honors versus comparison groups) were investigated as well as an interaction term (time*group).

Overall, there were no significant differences between groups on the outcomes of the AMEX, GLEX, NW, or on student GPA. However, of practical significance were the small to moderate increases in GLEX scores across the ARS covariate set conditions, the interaction of NW scores for honors and GSP students when a caliper of 0.2 was implemented, and the small effect between honors and GSP students in the NN SAT covariate set condition. Because the honors and GSP averages changed as a function of sampling, it is possible that sampling bias was introduced as the original sample of honors students was reduced. This finding may also shed light on ways in which estimates can be biased based on the matching choices a researcher makes.

The impact of losing information (i.e., students) from the honors versus GSP condition using a caliper of 0.2 introduced sampling bias. Particularly concerning was a change in the pattern of the honors subsamples' means from the original sample's pattern of means. Specifically, the honors subsample average NW pre-test scores were four points higher than the overall honors sample NW pre-test scores (See Tables 6 and 9). One might expect the GSP scores to differ after matching, because the matched pairs were created on their propensity for participation in the honors program, the counterfactual. However, given that honors student participation was known, altering their scores through matching was not ideal. Thus, sampling bias could potentially lead to

different, yet inaccurate, conclusions after matching. Because the actual performance in the overall sample of honors students on the NW test is known, it is clear that the subsample of students did not reflect the performance of the overall honors group.

Figure 17 shows the honors and GSP groups' NW scores overall (solid line and long-dashed line) and for the two subgroups of students in the 0.2 caliper condition (dotted line and small-dashed line). The honors and GSP pre-test means are further apart for the subsamples of students in the caliper of 0.2 condition. The sampling bias associated with the reduced samples resulted in a practically significant interaction. However, it is important to note that only eleven honors students and 21 non-honors students remained in this condition for the NW comparison. In the honors versus GSP caliper of 0.1 condition, the loss of additional students resulted in additional sampling bias and no practically significant findings.

Summary

Overall, the honors, non-honors, and GSP groups had very similar distributions on each of the covariates and outcome measures (see Figures 1-6 and 10-16). Thus, there was a high level of common support to creating quality matches using the covariates selected for the current study. Because the three student groups did not differ in their distributions on the motivation subscales, it's possible that motivation is not a confounding variable related to self-selection into the honors program. Therefore, motivation is likely not appropriate to use as a covariate when creating propensity score matches. The university at which this study was conducted is fairly selective. Thus, it is possible that *all students* at the university are similarly highly motivated resulting in similar group averages on the motivation subscales. Additionally, it is possible that unknown factors are driving students' participation in the honors program. Because only

measured covariates are accounted for in PSM, unmeasured confounding variables result in biased estimates of intervention effects.

Best practices in PSM include the use of theoretically-sound variables as covariates (Brookhart et al., 2006; Steiner et al., 2010). In practice, however, covariates that are theoretically related to self-selection might not actually drive student participation. In the current study, approximately 20% of eligible students decided to participate in the honors program and 80% decided not to participate despite being eligible. Thus, it was anticipated that the motivation subscales of expectancy, value and cost would be related to students' self-selection into the honors program. However, similar score distributions on the motivation measures indicated that the student groups did not differ systematically. Thus, it may be that motivation for general coursework is not different for the three groups, but that other underlying factors related to motivation are different. For example, because there is a monetary award (scholarship) attached to the honors program, a different form of motivation could be related to students' participation. For example, perhaps those who opt in to the honors program are more extrinsically-motivated than those who opt out. Moreover, there is the title attached to being an "honors student;" perhaps there is a form of motivation attached to the status. In sum, it is possible that motivation was operationally defined in the current study in a way that is not pertinent.

Given that honors students are required to complete rigorous coursework and complete an honors thesis as a part of the program, perhaps a motivation-related variable, such as work-avoidance, would be more strongly related to self-selection than the variables selected for the current study. In the future, it might be worthwhile for

assessment practitioners to consider alternate covariates when creating matched groups from similarly highly motivated students. Additionally, practitioners may want to examine the distribution of potential covariates between participants and nonparticipants prior to creating matches even when the covariates are theoretically sound.

Although not currently suggested in the propensity score literature, assessment practitioners may wish to evaluate the relative contribution of each covariate in the creation of propensity scores. For propensity scores created via logistic regression, practitioners can conduct a logistic regression model including the covariates as independent variables to predict participation (dependent variable). Within the model, the relative contribution of each covariate can be evaluated to determine how much each covariate contributes to predicting student participation. Additionally, conducting the logistic regression model also allows assessment practitioners to evaluate how well the *set of covariates* predicts participation via examination of the null deviance explained.

Limitations

Several limitations to the current study exist. Of particular interest is that the AMEX, GLEX and NW measures were collected under low-stakes testing conditions at both time points. Although small increases are consistently measured pre-to-post at the institution at which this study was conducted, there are known motivation issues associated with low-stakes tests (e.g., Wise & DeMars, 2005). Specifically, sophomore students tend to report lower test-taking motivation during the second Assessment Day than on the first Assessment Day. Thus, scores on the three tests administered during the second Assessment Day may not be reflective of students' actual knowledge but rather their willingness to put forth effort on the tests. Future studies may want to consider

incorporating outcome measures that are course/program-embedded or conducted under different stakes.

Another limitation to the current study is that not all honors students may have completed general education courses that aligned with the outcome measures (i.e., AMEX, GLEX, and NW tests). Thus, honors students might not have received the "treatment" of taking courses with smaller class sizes before completing the post-tests as a sophomore. Because course completion was not taken into consideration for the purposes of the current study, it is unknown whether honors students' scores on the outcome measures is representative of how students would perform upon completing the general education courses.

The reduction in sample size, and consequently the number of individuals included in the final comparisons, is another limitation to this study. Particularly on the NW test, all of the subsamples created in the caliper conditions had fewer than 20 students. Thus, as students were dropped from the comparisons, omitted scores were more likely to have a greater impact on the group means than if the sample were large. For example, in the SAT covariate set condition of honors versus GSP with a caliper of 0.2, the loss of two students from both groups resulted in a practically significant interaction on the NW test. Particularly concerning was the distortion of the honors' means on the NW test, which after matching were no longer representative of the original means. However, the results were not practically significant as additional students were lost in the 0.1 caliper condition. Therefore, the 0.2 caliper condition may illustrate issues with sampling error that can occur with the reduction of sample size. However, additional research is needed regarding the use of PSM techniques in higher education assessment.

Future Research

More research is needed on PSM techniques in the context of higher education assessment. Although applied PSM studies within the educational context have become more frequent in recent years, further research is needed regarding how well these techniques operate within the social sciences. Specifically, it is unknown which covariates are important to include across different educational contexts. For example, it may be important to include extraversion as a covariate when evaluating educational programs that require high levels of social interaction. On the other hand, openness might be an important covariate to include for study abroad assessment. Because the potential reasons underlying self-selection may be as varied as the university programs students select into, it is important to emphasize that covariates are best chosen on a program-by-program basis.

For the future assessment of honors programs, researchers may wish to investigate other forms of motivation than expectancy, value, and cost for students' coursework. Because honors students were not different from other students on the form of academic motivation employed in the current study, it is likely that it did not contribute to students' self-selection into the honors program. Student motivation related to status (i.e., being an "honors student") or financial benefits (scholarships), however, might differentially motivate student participation. Additionally, there may be differences in work avoidance among honors and non-honors students, particularly if the honors program requirement of completing a thesis deterred students from participating.

Finally, simulation studies could shed light on how well PSM techniques perform in the social sciences. Though numerous simulation studies have been conducted in the

context of economics and medicine (e.g., Austin, 2009b; 2011), studies have yet to be replicated using covariates related to self-selection in educational research. For example, a researcher could simulate data with covariates adjusted at varying degrees of relatedness to self-selection and the program outcomes. Until such studies are conducted, it is unknown whether PSM techniques operate differently in the context of social science research.

Implications

In the current study, ARS scores did not prove useful above and beyond the standardized test scores when creating matches using the two sets of covariates. Across the six non-honors matching conditions, similarly high-quality matches were made. Additionally, there were no discernable differences in student performance on the honors program outcomes using either the ARS or the SAT covariate sets. Thus, it might be worthwhile to explore using SAT scores in addition to other variables as an honors program eligibility criterion.

Additionally, the underlying reasons that approximately 80% of eligible student opt not to participate in the honors program remain unknown. Because student motivation was similar among honors and non-honors students, there are likely other (unobserved) variables contributing to student participation. If the mechanisms related to self-selection into a program are unmeasured or not known, PSM techniques cannot effectively be used. However, if variables related to self-selection are known, then it may also be possible to intervene in order to promote student participation. If interested in exploring reasons that students opt into the honors program, it may behoove honors program administration to

conduct focus groups in order to shed light on some of the factors associated with students' decisions to participate.

Conclusions

Although more research is needed on PSM in the context of higher education assessment, it is a promising method that offers a way of accounting for confounding variables in applied contexts. Because the use of PSM techniques has become more frequent in recent years, it is important to investigate how to best use these techniques within the realm of assessment. Specifically, the underlying motivations for students' participation in university programs may vary widely and uniquely with each program. Thus, additional research is needed to understand how to best use PSM techniques in educational assessment in order to better estimate the impact of university programs on students.

Table 1.

Means, Standard Deviations, and Reliability for Covariates Used to Make Propensity Scores by Student Group

Means (SD)				Internal Consistency Reliability (α)
	Honors (n = 181)	Non-Honors (n = 836)	GSP (n = 2,836)	
Motivation Subscales				
Expectancy	25.45 (3.43)	25.27 (3.49)	25.32 (3.43)	0.84
Value	36.73 (5.32)	36.70 (5.66)	36.81 (5.45)	0.88
Cost	20.36 (5.35)	20.67 (5.49)	20.47 (5.31)	0.70
SAT Math	578.47 (63.41)	578.82 (68.11)	580.84 (66.36)	
SAT Verbal	568.10 (67.61)	571.10 (67.88)	571.05 (67.93)	
Transfer Credits	1.85 (5.36)	2.95 (8.68)	2.44 (7.74)	
ARS Scores	9.96 (1.97)	9.67 (2.05)		

Note: GSP is the General Student Population.

Table 2.

Demographic Information for Honors, Non-Honors, and General Student Population

Student Group	White		Asian		Pacific Islander		American Indian		Black		Hispanic		Not Specified	
	Males	Females	Males	Females	Males	Females	Males	Females	Males	Females	Males	Females	Males	Females
Honors (n = 181)	70	100	5	5	2	1	0	2	5	10	3	8	3	4
Non-Honors (n = 836)	339	512	29	29	3	3	6	9	16	29	17	29	8	23
GSP (n = 2,836)	982	1616	91	101	9	14	6	36	50	84	70	88	36	48
Overall (n = 3,853)	1391	2228	125	135	14	18	12	47	71	123	90	125	47	75

Table 3

Numeric Diagnosing of Matched Samples Including Means (SD's) and Cohen's d Effect Sizes for Each Set of Covariates at Different Matching Distances

	Honors vs Non-Honors (NH) SAT Covariate Set			Honors vs Non-Honors (NH) ARS Covariate Set			Honors vs General Student Population (GSP) SAT Covariate Set		
Nearest Neighbor	Honors (n = 181)	NH (n = 181)	d	Honors (n = 171)	NH (n = 171)	d	Honors (n = 181)	GSP (n = 181)	d
SAT Math	579.28 *(63.64)*	578.12 *(68.59)*	0.02				579.28 *(63.64)*	575.52 *(66.90)*	0.06
SAT Verbal	567.24 *(67.11)*	561.44 *(69.72)*	0.08				567.24 *(67.11)*	560.11 *(67.70)*	0.11
Expectancy	25.60 *(3.50)*	25.64 *(3.36)*	-0.01	25.61 *(3.35)*	25.50 *(3.36)*	0.03	25.60 *(3.50)*	25.80 *(3.51)*	-0.03
Value	36.69 *(5.37)*	36.44 *(5.61)*	0.05	36.68 *(5.39)*	36.66 *(5.80)*	0.00	36.69 *(5.37)*	37.28 *(5.35)*	-0.11
Cost	20.29 *(5.38)*	20.14 *(5.16)*	0.03	20.50 *(5.29)*	21.05 *(5.20)*	-0.10	20.29 *(5.38)*	20.59 *(5.60)*	-0.05
ARS Scores				9.96 *(1.97)*	9.80 *(2.11)*	0.08			
Transfer Credits	2.01 *(5.62)*	2.56 *(7.16)*	-0.09	2.04 *(5.70)*	2.08 *(6.62)*	-0.01	2.01 *(5.62)*	2.33 *(9.50)*	-0.04
Caliper 0.2	Honors (n = 154)	NH (n = 154)	d	Honors (n = 148)	NH (n = 148)	d	Honors (n = 154)	GSP (n = 154)	d
SAT Math	582.79 *(57.67)*	583.96 *(64.15)*	-0.02				581.56 *(60.72)*	587.47 *(66.33)*	-0.09
SAT Verbal	571.88 *(65.97)*	562.60 *(61.15)*	0.15				570.52 *(67.68)*	570.58 *(69.75)*	0.00
Expectancy	25.76 *(3.46)*	25.49 *(3.40)*	0.08	25.50 *(3.20)*	25.66 *(3.15)*	-0.05	25.48 *(3.31)*	25.76 *(3.43)*	-0.08
Value	36.75 *(5.52)*	36.77 *(5.59)*	0.00	36.78 *(5.30)*	36.57 *(5.92)*	0.04	36.96 *(4.83)*	37.62 *(5.53)*	-0.13
Cost	20.45 *(4.97)*	19.82 *(5.68)*	0.12	20.69 *(5.01)*	20.64 *(5.60)*	0.01	20.17 *(4.97)*	20.36 *(5.53)*	-0.04
ARS Scores				9.75 *(1.53)*	9.82 *(1.82)*	-0.04			
Transfer Credits	1.81 *(5.23)*	2.29 *(6.79)*	-0.08	1.74 *(4.46)*	1.98 *(6.65)*	-0.04	1.49 *(4.23)*	2.51 *(8.82)*	-0.15
Caliper 0.1	Honors (n = 137)	NH (n = 137)	d	Honors (n = 135)	NH (n = 135)	d	Honors (n = 145)	GSP (n = 145)	d
SAT Math	583.07 *(57.17)*	573.50 *(64.78)*	0.16				582.48 *(60.47)*	579.66 *(60.74)*	0.05
SAT Verbal	574.38 *(67.43)*	562.85 *(61.44)*	0.18				571.45 *(65.06)*	576.34 *(63.89)*	-0.08
Expectancy	25.77 *(3.42)*	25.69 *(3.20)*	0.02	25.49 *(3.17)*	25.99 *(3.12)*	-0.16	25.59 *(3.42)*	25.62 *(3.62)*	-0.01
Value	36.80 *(5.52)*	36.88 *(5.32)*	-0.01	36.78 *(5.34)*	37.55 *(5.91)*	-0.14	37.05 *(4.96)*	37.19 *(5.31)*	-0.03
Cost	20.52 *(5.02)*	20.26 *(4.68)*	0.05	20.70 *(5.10)*	20.30 *(5.58)*	0.07	20.26 *(5.10)*	20.70 *(5.91)*	-0.08
ARS Scores				9.69 *(1.23)*	9.67 *(1.60)*	0.01			
Transfer Credits	1.77 *(5.24)*	2.09 *(6.64)*	-0.05	1.55 *(4.06)*	1.42 *(4.26)*	0.03	1.69 *(5.14)*	1.72 *(7.10)*	0.00

Note. To calculate Cohen's *d* values, the non-honors and GSP values were subtracted from the honors values.

Table 4

Numeric Diagnosing of Matched Samples on the Multivariate Composite (i.e., Propensity Score) for Different Covariates and Matching Distances

	Honors vs Non-Honors SAT Covariate Set			Honors vs Non-Honors (NH) ARS Covariate Set			Honors vs General Student Population (GSP) SAT Covariate Set		
	Honors	NH		Honors	NH		Honors	GSP	
Nearest Neighbor									
Variance	0.001	0.001		0.002	0.002		<0.001	<0.001	
Mean	0.182	0.181		0.179	0.179		0.069	0.069	
VR			1.000			1.000			1.000
SMD			0.032			0.000			0.000
Caliper 0.2									
Variance	0.001	0.001		0.001	0.001		<0.001	<0.001	
Mean	0.175	0.175		0.170	0.170		0.065	0.065	
VR			1.000			1.000			1.000
SMD			0.000			0.000			0.000
Caliper 0.1									
Variance	0.001	0.001		0.001	0.001		<0.001	<0.001	
Mean	0.174	0.174		0.170	0.170		0.064	0.064	
VR			1.000			1.000			1.000
SMD			0.000			0.000			0.000

Note. VR indicates the variance ratio between the treatment and honors groups, which should be as close to 1 as possible. SMD indicates the standardized mean difference between groups on the propensity scores, which should be as close to 0 as possible.

Table 5

Representation of Ethnic Groups by Gender for Each Set of Covariates and Matching Distance

Honors vs. Non-Honors	White		Asian		Pacific Islander		American Indian		Black		Hispanic		Not Specified	
SAT Covariate Set	M	F	M	F	M	F	M	F	M	F	M	F	M	F
Nearest Neighbor														
Honors (n = 181)	65	89	5	4	1	1	0	2	5	10	3	8	1	2
Non-Honors (n = 181)	73	83	4	3	2	0	0	1	4	10	4	6	0	2
Caliper 0.2														
Honors (n = 154)	60	79	5	4	1	1	0	1	1	2	1	5	1	2
Non-Honors (n = 154)	64	74	4	4	1	0	0	0	0	3	1	5	0	4
Caliper 0.1														
Honors (n = 137)	52	73	3	4	0	1	0	1	1	1	1	5	1	2
Non-Honors (n = 137)	59	64	4	4	0	0	0	0	0	3	0	5	1	2
Honors vs. Non-Honors	White		Asian		Pacific Islander		American Indian		Black		Hispanic		Not Specified	
ARS Covariate Set	M	F	M	F	M	F	M	F	M	F	M	F	M	F
Nearest Neighbor														
Honors (n = 171)	59	85	4	5	1	1	0	2	3	10	3	7	3	2
Non-Honors (n = 171)	67	74	7	5	1	0	2	0	4	12	5	5	0	4
Caliper 0.2														
Honors (n = 148)	55	75	3	5	1	0	0	2	1	4	1	5	3	2
Non-Honors (n = 148)	52	80	4	4	1	1	1	0	1	2	4	2	1	4
Caliper 0.1														
Honors (n = 135)	51	66	3	5	1	0	0	2	1	4	1	4	3	2
Non-Honors (n = 135)	55	69	3	2	0	2	1	0	2	1	3	2	1	3
Honors vs General Student	White		Asian		Pacific Islander		American Indian		Black		Hispanic		Not Specified	
Population (GSP)	M	F	M	F	M	F	M	F	M	F	M	F	M	F
Nearest Neighbor														
Honors (n = 181)	65	89	5	4	1	1	0	2	5	10	3	8	1	2
GSP (n = 181)	58	93	3	5	1	0	1	2	6	14	5	9	2	4
Caliper 0.2														
Honors (n = 154)	59	84	5	3	1	0	0	2	1	2	1	8	0	1
GSP (n = 154)	59	78	4	5	2	1	0	0	1	3	2	5	1	2
Caliper 0.1														
Honors (n = 145)	55	80	5	3	1	1	0	2	1	1	1	8	0	1
GSP (n = 145)	53	78	4	4	1	0	0	1	2	2	1	4	0	4

Note. Males are in the columns indicated with an "M" and females are in the columns marked with an "F.

Table 6

Means, Standard Deviations, and Reliability for Outcome Measures by Student Group

	Honors			Non-Honors			General Student Population			Internal Consistency Reliability (α)	
Measure	Fall '12	Spring '14	N	Fall '12	Spring '14	N	Fall '12	Spring '14	N	Fall '12	Spring '14
AMEX	22.40 *(6.53)*	23.89 *(4.94)*	53	23.22 *(6.62)*	23.98 *(6.14)*	223	23.69 *(5.96)*	24.08 *(5.86)*	618	0.80	0.79
GLEX	21.72 *(4.43)*	22.61 *(4.89)*	90	21.27 *(4.74)*	22.98 *(4.86)*	321	20.95 *(4.75)*	22.52 *(5.14)*	926	0.74	0.79
NW9	45.56 *(7.71)*	48.69 *(8.39)*	60	46.71 *(7.17)*	49.98 *(7.32)*	231	46.05 *(7.18)*	49.21 *(7.97)*	679	0.73	0.79
GPA		3.01 *(0.48)*	187		3.02 *(0.48)*	842		2.99 *(0.52)*	2,539		

Table 7

Means (SD's) and Eta Squared Effect Sizes for Outcomes on the American Experience (AMEX) Test for Each Set of Covariates at Different Matching Distances

	Honors vs Non-Honors SAT Covariate Set				Honors vs Non-Honors ARS Covariate Set				Honors vs General Student Population (GSP) SAT Covariate Set			
	Means *(SD)*			η^2	Means *(SD)*			η^2	Means *(SD)*			η^2
Nearest Neighbor	Honors (n = 87)	NH (n = 70)	Group	0.00	Honors (n = 80)	NH (n = 72)	Group	0.00	Honors (n = 89)	GSP (n = 80)	Group	0.00
AMEX Fall 2012	23.68 *(6.59)*	22.27 *(5.90)*	Time	0.02	24.00 *(6.75)*	23.81 *(7.05)*	Time	0.00	23.45 *(6.58)*	23.34 *(6.44)*	Time	0.00
AMEX Spring 2014	23.55 *(5.71)*	25.00 *(6.26)*	Int.	0.03	23.95 *(5.37)*	23.71 *(6.08)*	Int.	0.00	23.40 *(5.72)*	23.16 *(6.29)*	Int.	0.00
Caliper 0.2	Honors (n = 76)	NH (n = 62)	Group	0.00	Honors (n = 71)	NH (n = 54)	Group	0.00	Honors (n = 78)	GSP (n = 70)	Group	0.00
AMEX Fall 2012	23.89 *(6.45)*	22.95 *(6.24)*	Time	0.00	24.15 *(6.62)*	23.67 *(6.67)*	Time	0.00	23.96 *(6.35)*	24.30 *(5.88)*	Time	0.01
AMEX Spring 2014	23.53 *(5.48)*	24.13 *(6.67)*	Int.	0.01	24.04 *(5.35)*	23.81 *(6.02)*	Int.	0.00	23.41 *(5.85)*	22.99 *(5.96)*	Int.	0.00
Caliper 0.1	Honors (n = 72)	NH (n = 61)	Group	0.00	Honors (n = 65)	NH (n = 47)	Group	0.02	Honors (n = 72)	GSP (n = 61)	Group	0.00
AMEX Fall 2012	23.85 *(6.38)*	23.87 *(6.11)*	Time	0.00	23.95 *(6.77)*	22.98 *(7.00)*	Time	0.00	23.85 *(6.38)*	23.87 *(6.11)*	Time	0.00
AMEX Spring 2014	23.22 *(5.85)*	23.44 *(6.29)*	Int.	0.00	24.06 *(5.43)*	22.85 *(5.99)*	Int.	0.00	23.22 *(5.85)*	23.44 *(6.29)*	Int.	0.00

Note. None of the main effect between groups (noted at "Groups" in the table), between the fall 2012 and spring 2014 tests (noted as "Time" in the table), and the interaction effect (noted as "Int." in the table) were significant at the $\alpha = .01$ level.

Table 8

Means (SD's) and Eta Squared Effect Sizes for Outcomes on the Global Experience (GLEX) Test for Each Set of Covariates at Different Matching Distances

	Honors vs Non-Honors SAT Covariate Set				Honors vs Non-Honors ARS Covariate Set				Honors vs General Student Population (GSP) SAT Covariate Set			
	Means *(SD)*			η^2	Means *(SD)*			η^2	Means *(SD)*			η^2
Nearest Neighbor	Honors (n = 104)	NH (n = 78)	Group	0.00	Honors (n = 93)	NH (n = 79)	Group	0.00	Honors (n = 104)	GSP (n = 99)	Group	0.00
GLEX Fall 2012	21.20 *(4.64)*	21.54 *(4.59)*	Time	0.03	21.46 *(4.52)*	21.58 *(4.85)*	Time	0.03	21.23 *(4.58)*	21.49 *(4.96)*	Time	0.02
GLEX Spring 2014	22.44 *(4.72)*	22.77 *(4.92)*	Int.	0.00	22.55 *(4.68)*	23.09 *(5.16)*	Int.	0.00	22.46 *(4.60)*	22.27 *(4.83)*	Int.	0.00
Caliper 0.2	Honors (n = 91)	NH (n = 73)	Group	0.01	Honors (n = 84)	NH (n = 61)	Group	0.00	Honors (n = 92)	GSP (n = 82)	Group	0.00
GLEX Fall 2012	21.21 *(4.50)*	22.23 *(4.58)*	Time	0.02	21.48 *(4.51)*	21.25 *(4.63)*	Time	0.04	21.01 *(4.47)*	21.37 *(4.78)*	Time	0.04
GLEX Spring 2014	22.55 *(4.47)*	22.47 *(5.24)*	Int.	0.01	22.46 *(4.40)*	22.85 *(5.21)*	Int.	0.00	22.65 *(4.61)*	22.32 *(4.28)*	Int.	0.00
Caliper 0.1	Honors (n = 86)	NH (n = 75)	Group	0.00	Honors (n = 77)	NH (n = 50)	Group	0.00	Honors (n = 86)	GSP (n = 75)	Group	0.00
GLEX Fall 2012	21.07 *(4.30)*	21.63 *(4.92)*	Time	0.03	21.36 *(4.60)*	21.02 *(5.22)*	Time	0.05	21.07 *(4.30)*	21.63 *(4.92)*	Time	0.03
GLEX Spring 2014	22.56 *(4.53)*	22.28 *(4.65)*	Int.	0.00	22.49 *(4.46)*	22.92 *(4.44)*	Int.	0.00	22.56 *(4.53)*	22.28 *(4.65)*	Int.	0.00

Note. None of the main effect between groups (noted at "Groups" in the table), between the fall 2012 and spring 2014 tests (noted as "Time" in the table), and the interaction effect (noted as "Int." in the table) were significant at the $\alpha = .01$ level.

Table 9

Means (SD's) and Eta Squared Effect Sizes for Outcomes on the Natural World (NW) Test for Each Set of Covariates at Different Matching Distances

	Honors vs Non-Honors SAT Covariate Set				Honors vs Non-Honors ARS Covariate Set				Honors vs General Student Population (GSP) SAT Covariate Set			
	Means *(SD)*			η^2	Means *(SD)*			η^2	Means *(SD)*			η^2
Nearest Neighbor	Honors (n = 13)	NH (n = 24)	Group	0.00	Honors (n = 15)	NH (n = 15)	Group	0.00	Honors (n = 13)	GSP (n = 23)	Group	0.03
NW Fall 2012	48.77 *(7.27)*	47.54 *(6.69)*	Time	0.11	48.00 *(7.28)*	48.87 *(6.99)*	Time	0.01	48.77 *(7.27)*	45.57 *(6.39)*	Time	0.06
NW Spring 2014	49.54 *(8.96)*	50.79 *(4.61)*	Int.	0.04	49.00 *(8.59)*	49.60 *(7.60)*	Int.	0.00	49.53 *(9.00)*	48.34 *(7.99)*	Int.	0.02
Caliper 0.2	Honors (n = 11)	NH (n = 14)	Group	0.00	Honors (n = 13)	NH (n = 15)	Group	0.00	Honors (n = 11)	GSP (n = 21)	Group	0.03
NW Fall 2012	49.36 *(7.72)*	48.00 *(6.84)*	Time	0.07	47.62 *(7.72)*	47.60 *(6.80)*	Time	0.03	49.36 *(7.72)*	44.00 *(6.76)*	Time	0.16
NW Spring 2014	49.45 *(9.77)*	50.86 *(5.32)*	Int.	0.06	48.69 *(9.22)*	49.20 *(7.11)*	Int.	0.00	49.45 *(9.77)*	50.48 *(6.06)*	Int.	0.15
Caliper 0.1	Honors (n = 10)	NH (n = 16)	Group	0.02	Honors (n = 11)	NH (n = 20)	Group	0.00	Honors (n = 10)	GSP (n = 16)	Group	0.00
NW Fall 2012	48.70 *(7.80)*	47.06 *(6.23)*	Time	0.05	47.55 *(7.66)*	47.20 *(7.24)*	Time	0.01	48.70 *(7.80)*	47.06 *(6.23)*	Time	0.05
NW Spring 2014	48.20 *(9.32)*	51.00 *(9.26)*	Int.	0.09	47.27 *(9.03)*	48.95 *(7.65)*	Int.	0.02	48.20 *(9.32)*	51.00 *(9.26)*	Int.	0.09

Note. None of the main effect between groups (noted at "Groups" in the table), between the fall 2012 and spring 2014 tests (noted as "Time" in the table), and the interaction effect (noted as "Int." in the table) were significant at the α = .01 level.

Table 10

Means (SD's) and Cohen's d Effect Sizes for Students' Spring 2014 GPAs for Each Set of Covariates at Different Matching Distances

	Honors vs Non-Honors SAT Covariate Set			Honors vs Non-Honors ARS Covariate Set			Honors vs General Student Population (GSP) SAT Covariate Set		
	Means *(SD)*			Means *(SD)*			Means *(SD)*		
Nearest Neighbor	Honors (n = 166)	NH (n = 151)	*d*	Honors (n = 162)	NH (n = 156)	*d*	Honors (n = 166)	GSP (n = 160)	*d*
GPA Spring 2014	2.92 *(0.48)*	3.02 *(0.47)*	-0.21	2.96 *(0.48)*	2.93 *(0.47)*	0.06	2.92 *(0.48)*	2.92 *(0.56)*	0.00
Caliper 0.2	Honors (n = 142)	NH (n = 120)	*d*	Honors (n = 140)	NH (n = 138)	*d*	Honors (n = 144)	GSP (n = 139)	*d*
GPA Spring 2014	2.92 *(0.48)*	2.97 *(0.50)*	-0.10	2.94 *(0.49)*	2.96 *(0.49)*	-0.04	2.90 *(0.48)*	2.94 *(0.50)*	-0.08
Caliper 0.1	Honors (n = 127)	NH (n = 115)	*d*	Honors (n = 127)	NH (n = 121)	*d*	Honors (n = 136)	GSP (n = 126)	*d*
GPA Spring 2014	2.93 *(0.48)*	3.00 *(0.49)*	-0.14	2.92 *(0.49)*	2.93 *(0.46)*	-0.02	2.90 *(0.47)*	2.93 *(0.54)*	-0.05

Note. To calculate Cohen's *d* values, the non-honors and GSP values were subtracted from the honors values. None of between group comparisons were significant at the $\alpha = .01$ level.

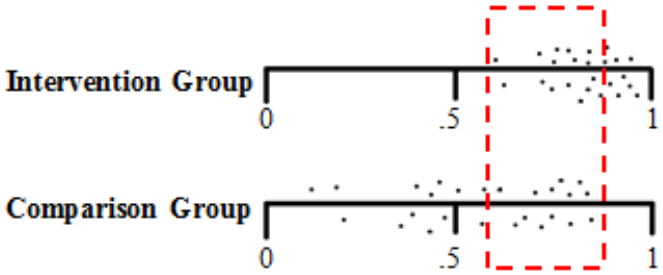

Figure 1. Area of common support across propensity score distributions (ranging from 0 to 1). The area of common support is indicated in the red dashed window.

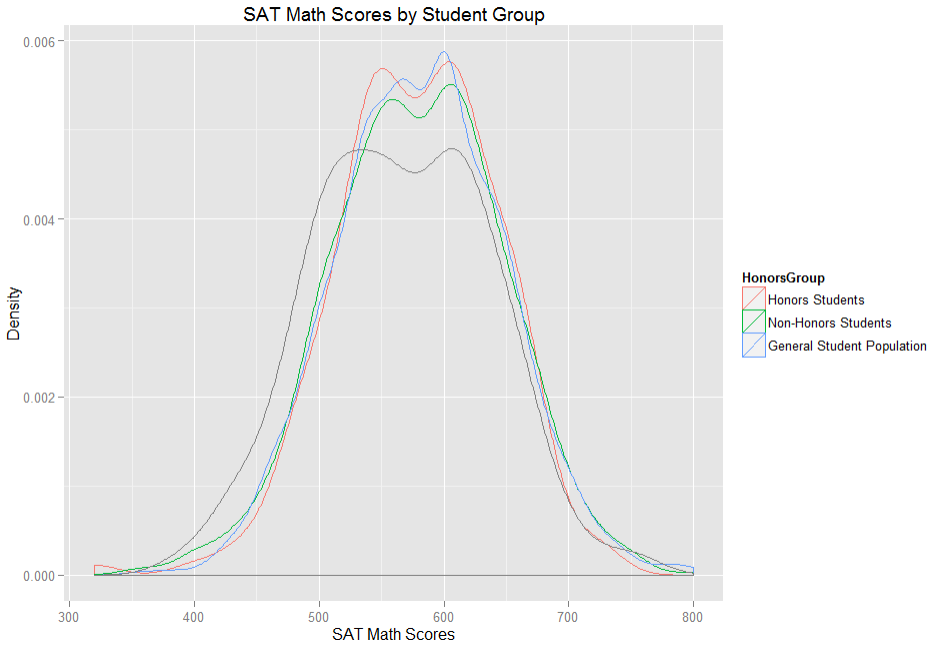

Figure 2. Density plot of SAT Math scores plotted for Honors, Non-Honors, and the General Student Population student groups.

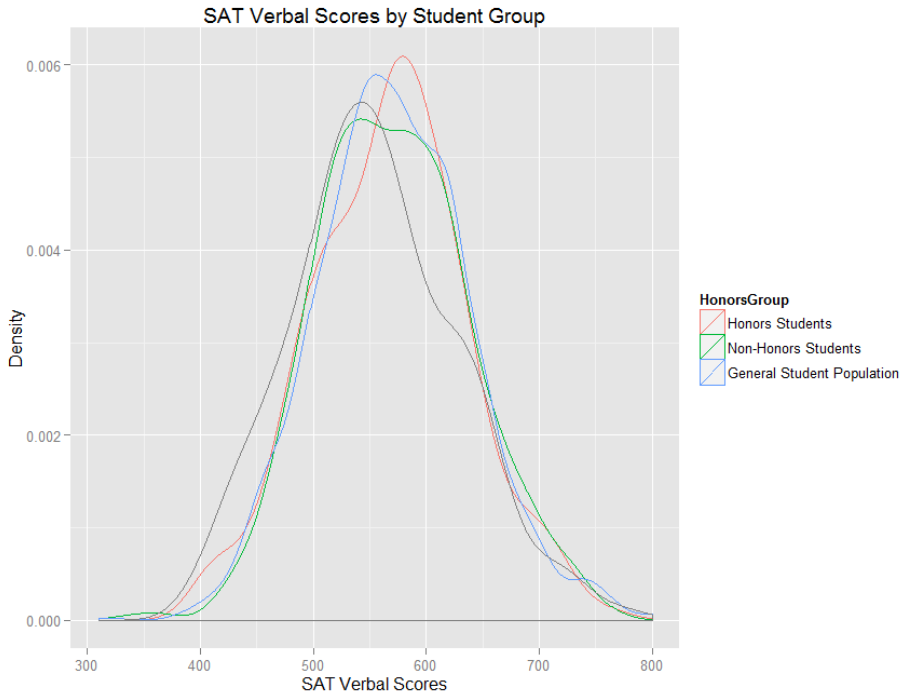

Figure 3. Density plot of SAT Verbal scores plotted for Honors, Non-Honors, and the General Student Population student groups.

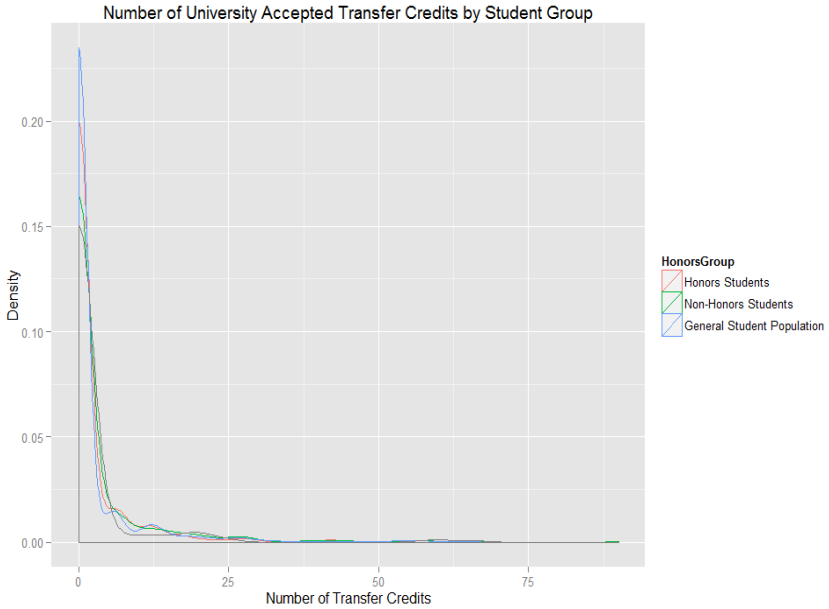

Figure 4. Density plot of transfer credits accepted at the university plotted for Honors, Non-Honors, and the General Student Population student groups.

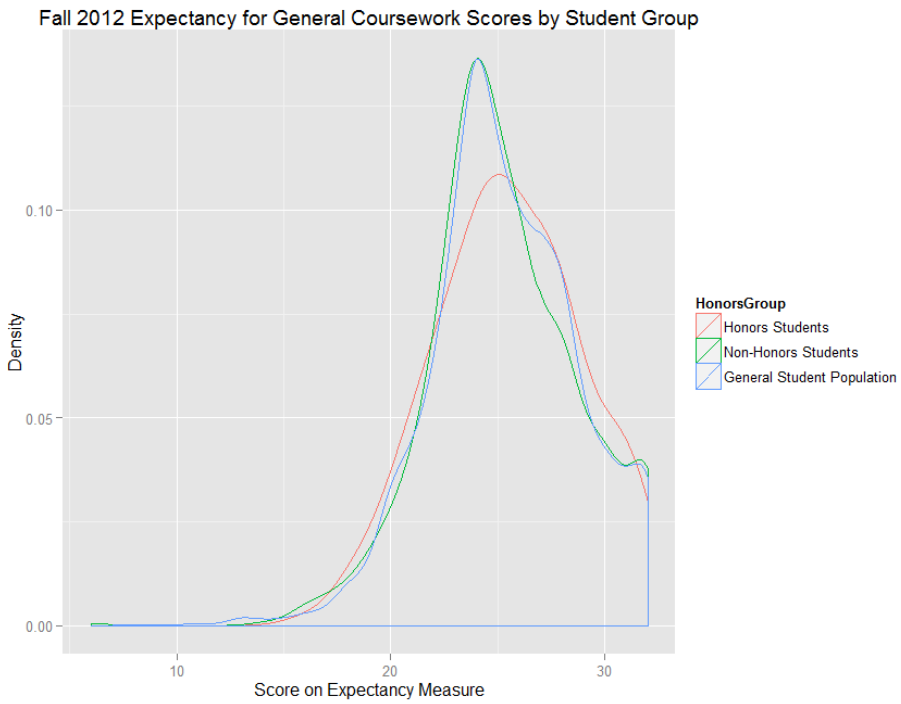

Figure 5. Density plot of General Education expectancy scores plotted for Honors, Non-Honors, and the General Student Population student groups.

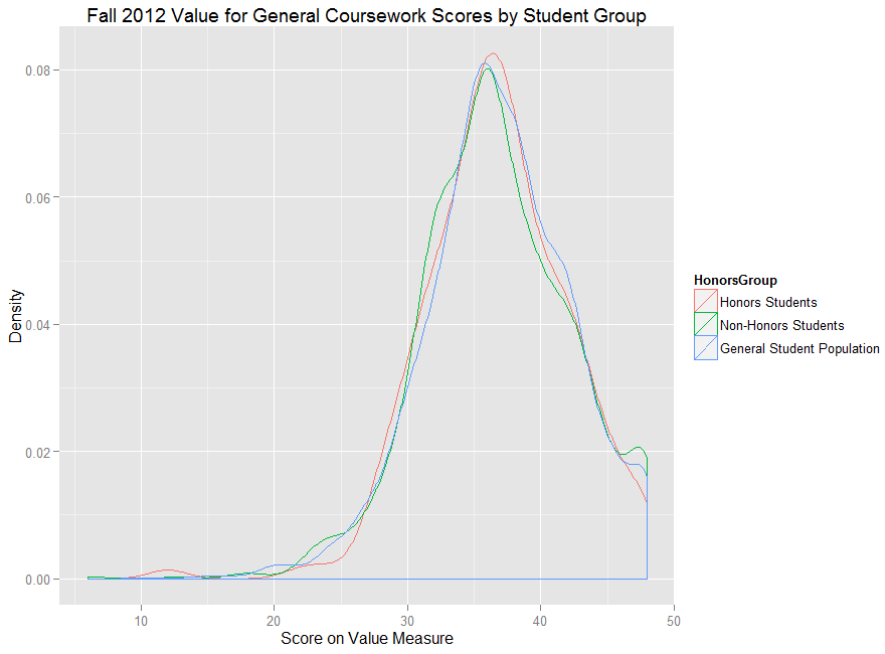

Figure 6. Density plot of General Education value scores plotted for Honors, Non-Honors, and the General Student Population student groups.

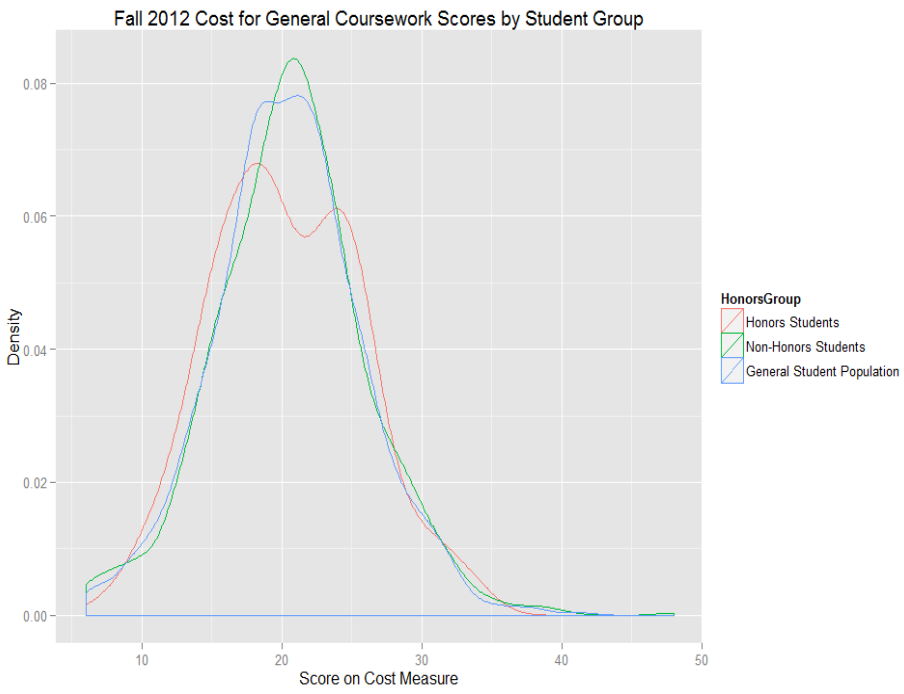

Figure 7. Density plot of General Education cost scores plotted for Honors, Non-Honors, and the General Student Population student groups.

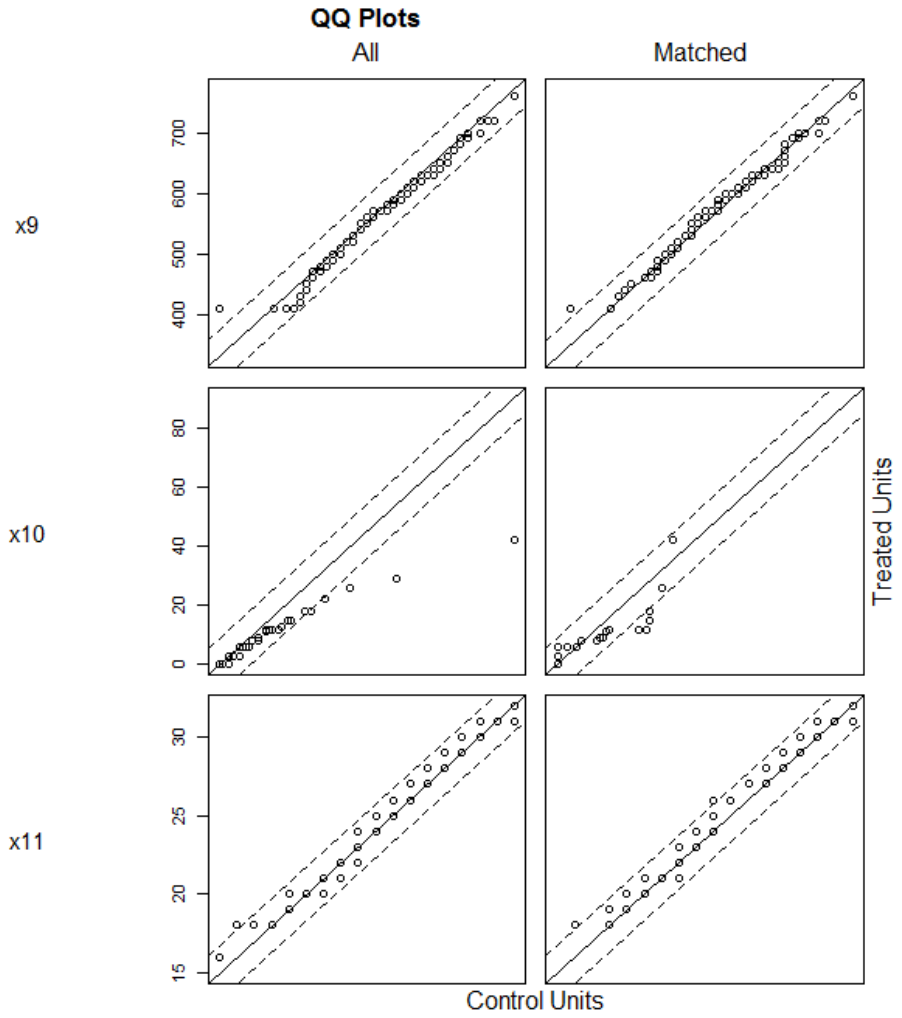

Figure 8. Example of QQ Plots produced by the MatchIt Package in R for visual diagnosing of matches (Ho et al., 2007).

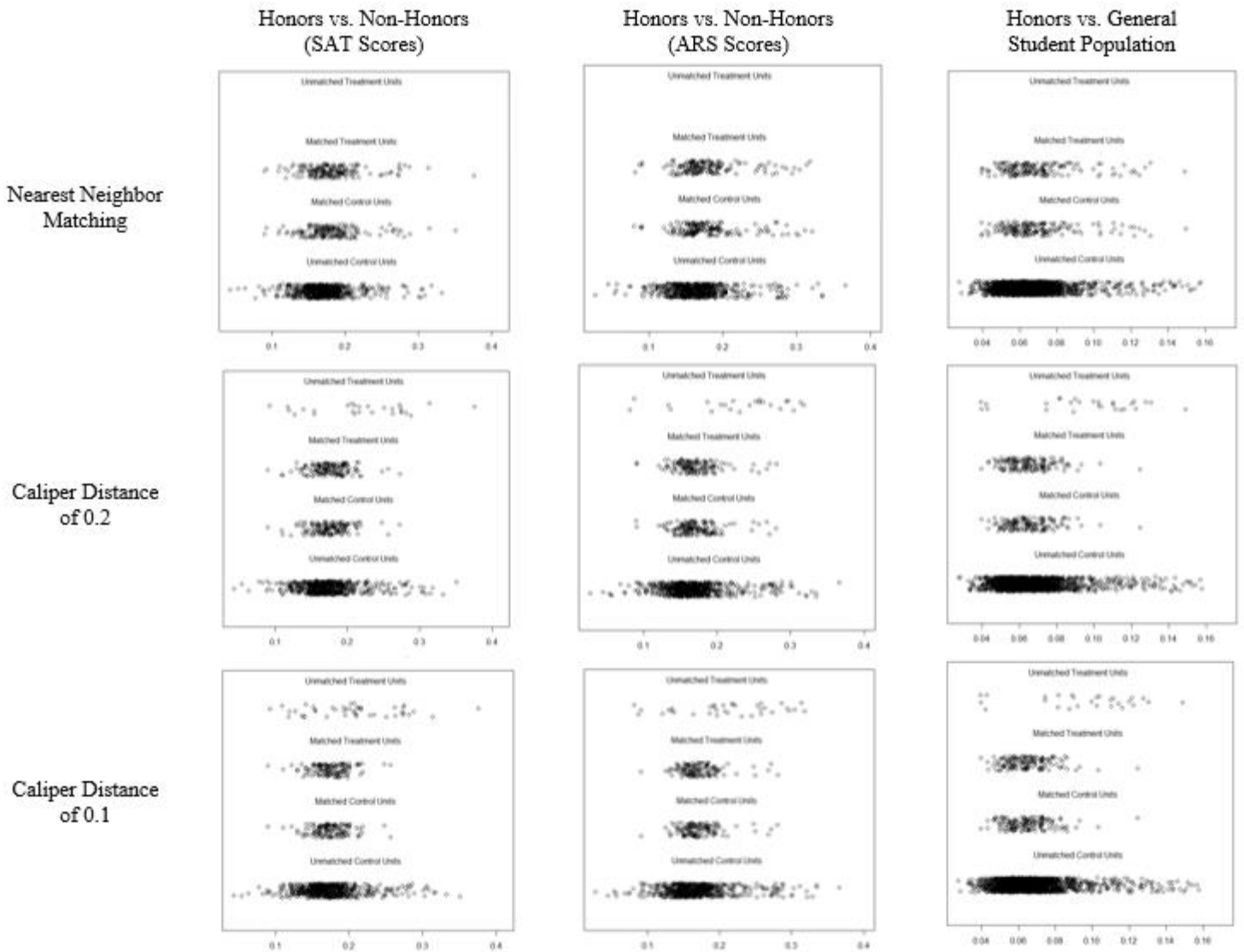

Figure 9. Jitter graphs of propensity scores using each set of covariates and at each matching distance.

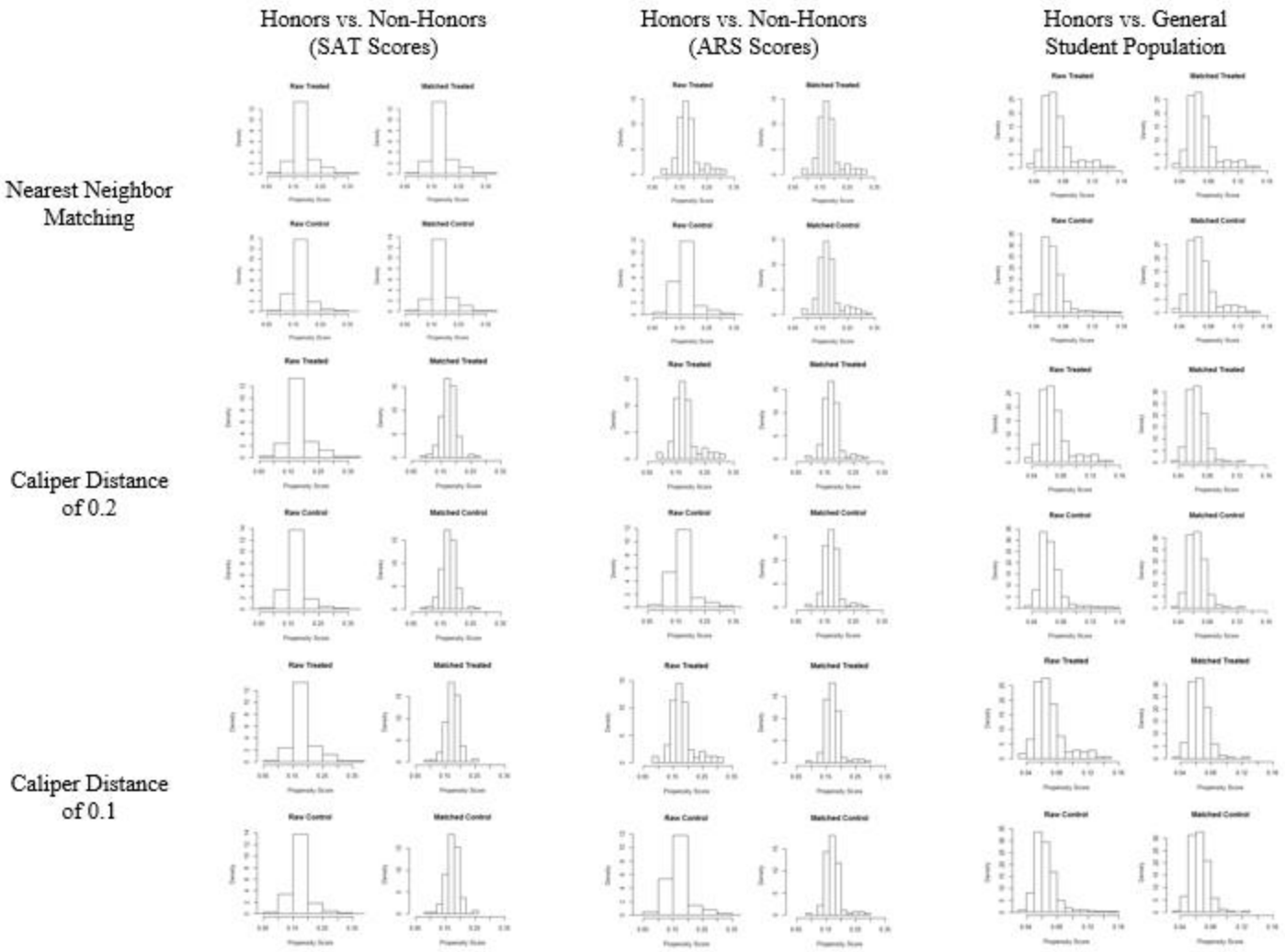

Figure 10. Distribution of propensity scores for raw and matched samples using each set of covariates and at each matching distance.

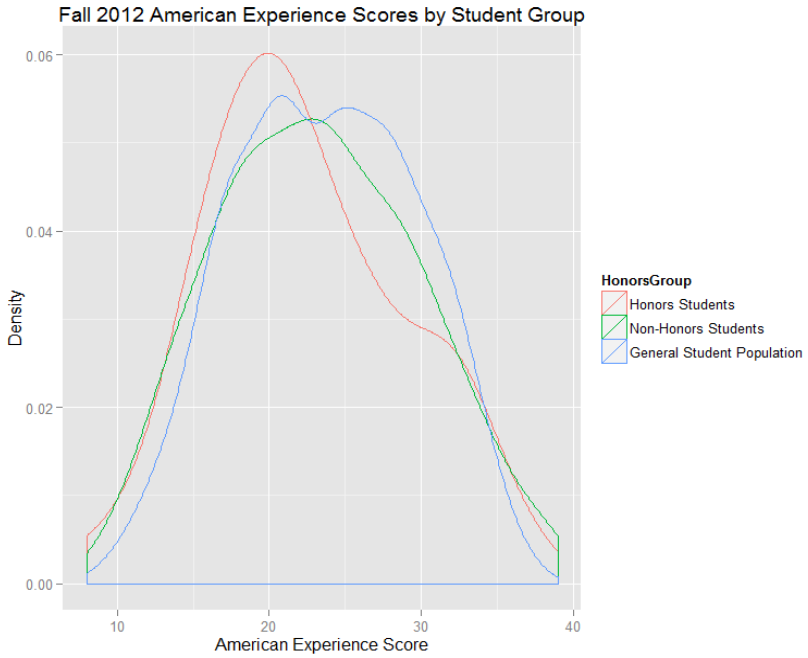

Figure 11. Density plot of Fall 2012 American Experience (AMEX) scores plotted for Honors, Non-Honors, and the General Student Population student groups.

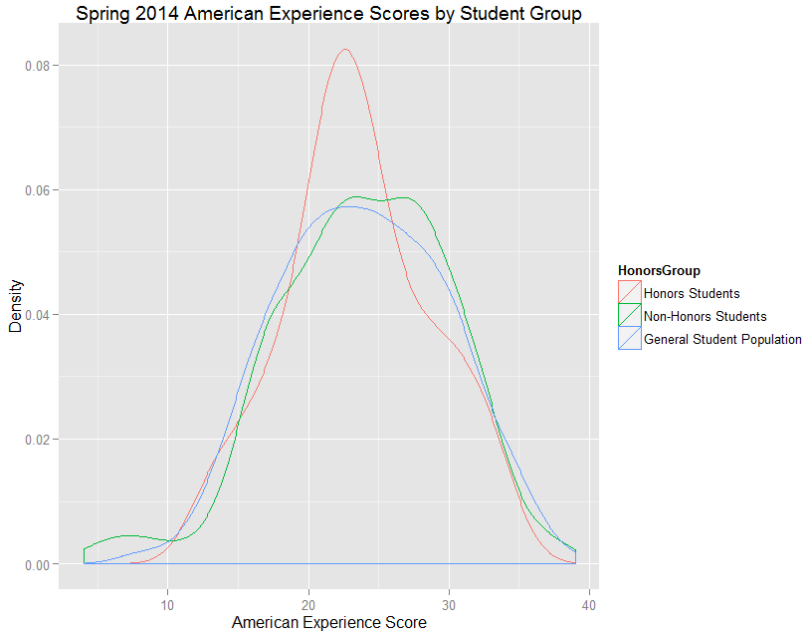

Figure 12. Density plot of Spring 2014 American Experience (AMEX) scores plotted for Honors, Non-Honors, and the General Student Population student groups.

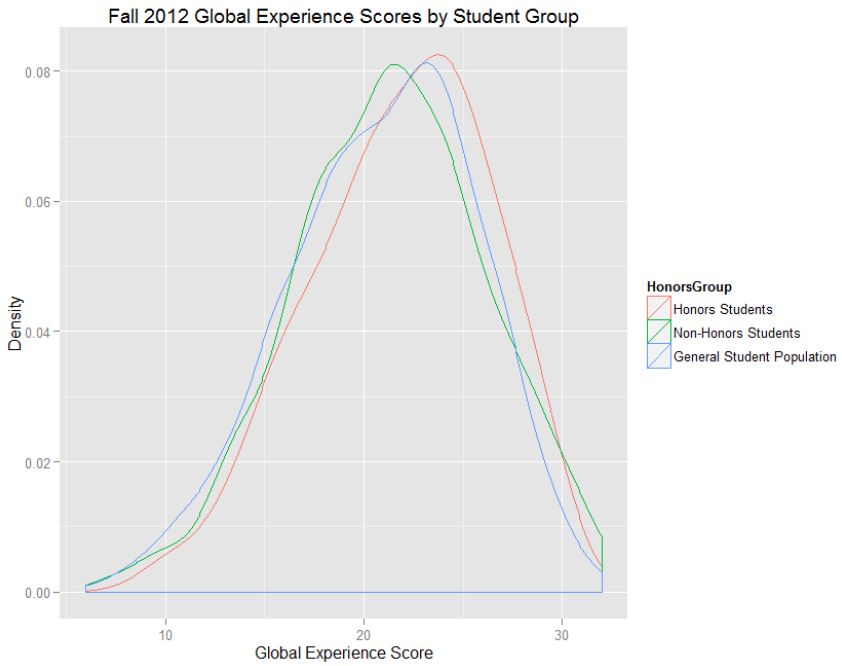

Figure 13. Density plot of Fall 2012 Global Experience (GLEX) scores plotted for Honors, Non-Honors, and the General Student Population student groups.

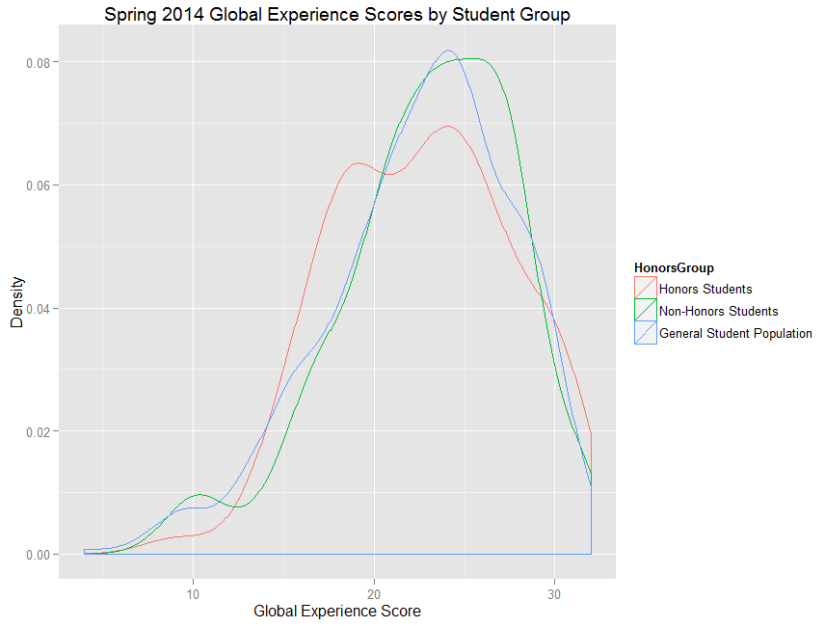

Figure 14. Density plot of Spring 2014 Global Experience (GLEX) scores plotted for Honors, Non-Honors, and the General Student Population student groups.

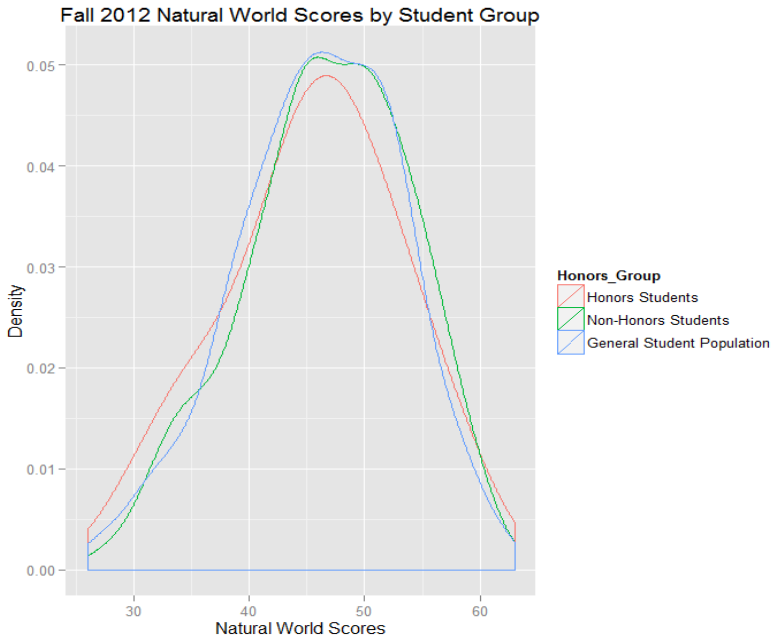

Figure 15. Density plot of Fall 2012 Natural World (NW9) scores plotted for Honors, Non-Honors, and the General Student Population student groups.

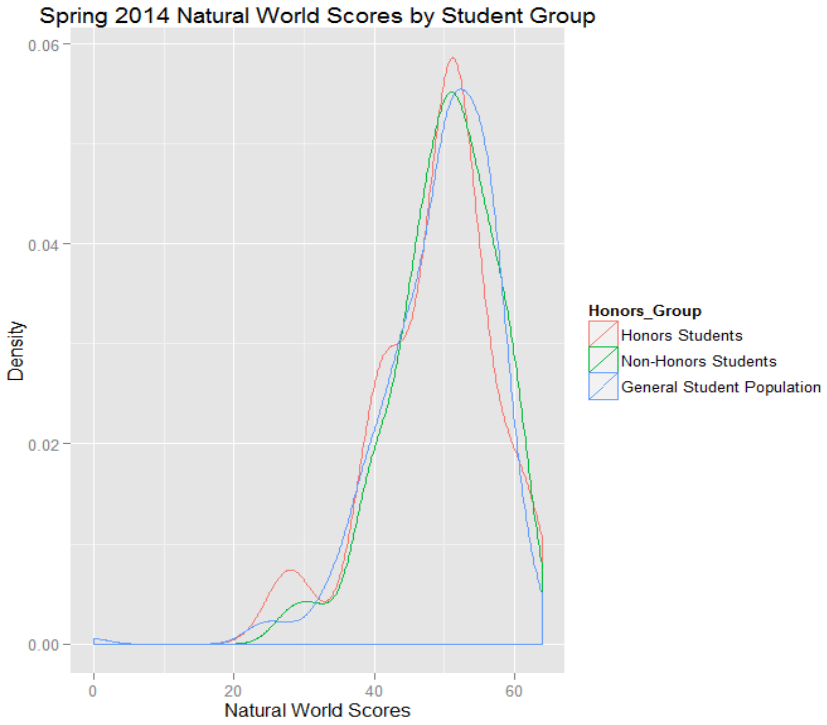

Figure 16. Density plot of Spring 2014 Natural World (NW9) scores plotted for Honors, Non-Honors, and the General Student Population student groups.

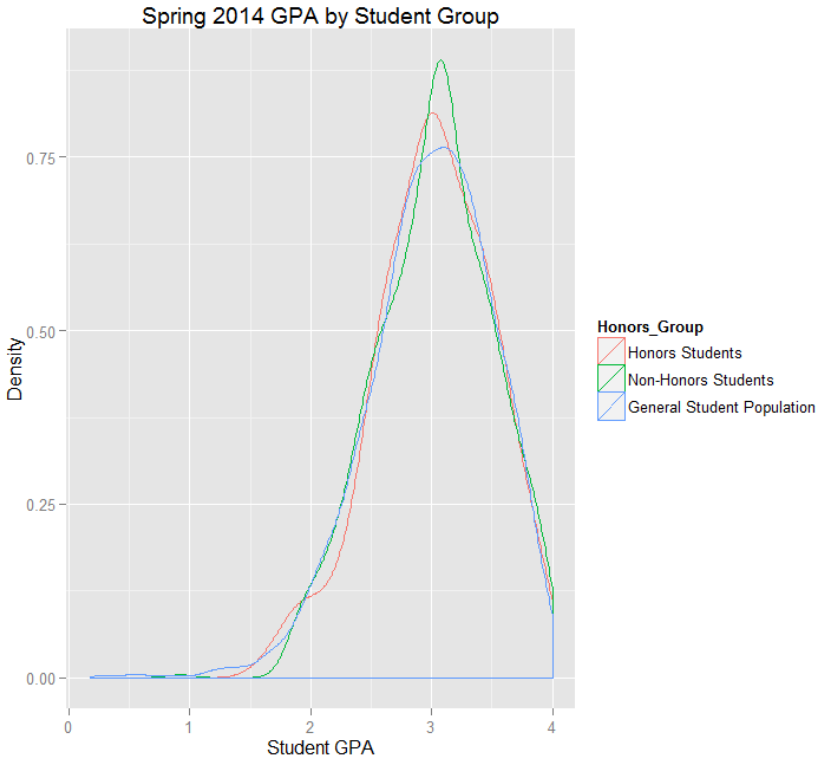

Figure 17. Density plot of Spring 2014 student GPA plotted for Honors, Non-Honors, and the General Student Population student groups.

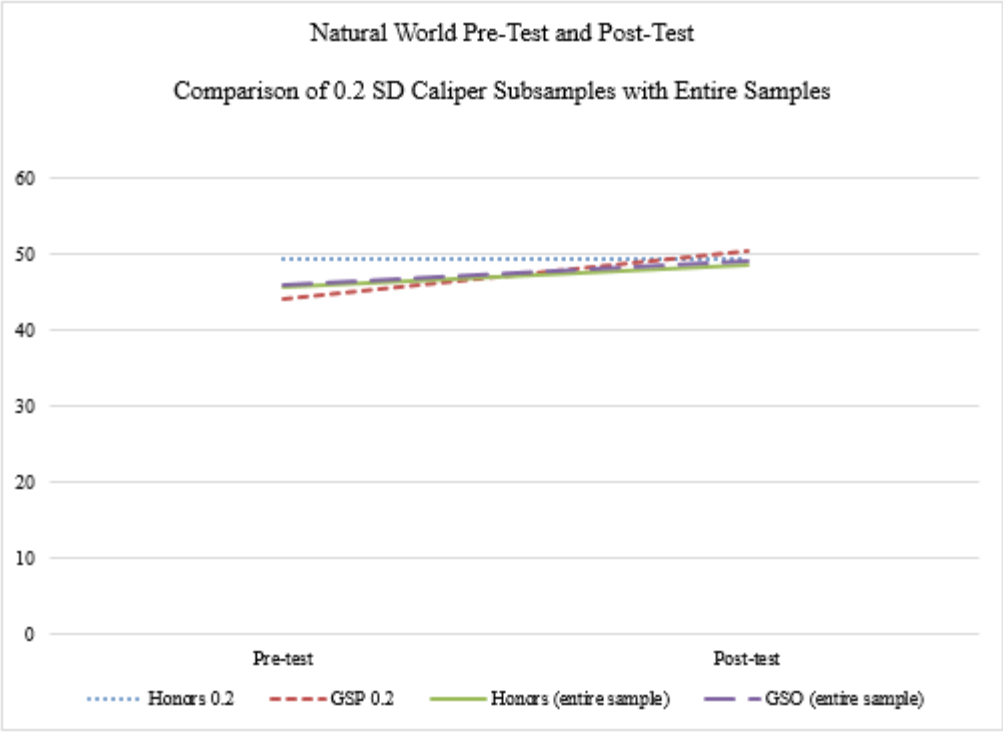

Figure 18. Graph of pre-test and post-test scores on the NW test for Honors and the General Student Population student groups in the NN with a caliper distance of 0.2 standard deviations and the original sample.

References

Agodini, R., & Dynarski, M. (2004). Are experiments the only option? A look at dropout prevention programs. *Review of Economics and Statistics, 86(1),* 180-194.

Allen, J., Robbins, S., Casillas, A., & Oh, I. (2007). Third-year college retention and transfer: Effects of academic performance, motivation, and social connectedness. Research in Higher Education, *49(7),* 647-664.

Althauser, R., & Rubin, D. (1970). The computerized construction of a matched sample. *American Journal of Sociology, 76,* 325-346.

Austin, P. C. (2007). A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine, 27(12),* 2037-2049.

Austin, P. C. (2009). Type I error rates, coverage of confidence intervals, and variance estimation in propensity-score matched analyses. *The International Journal of Biostatistics, 5(1),* 1-21.

Austin, P. C. (2011a). An introduction to propensity score methods for reducing the effects of confounding in observational studies, *Multivariate Behavioral Research, 46(3),* 399-424.

Austin, P. C. (2011b). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics, 10,* 150-161.

Azen, R., & Walker, C.M, (2011). *Categorical data analysis for the behavioral and social sciences.* New York, NY: Routledge.

Boruch, R., & Rui, N. (2008). From randomized controlled trials to evidence grading schemes: Current state of evidence-based practice in social sciences, *Journal of Evidence-Based Medicine, 1(1),* 41-49.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Sturmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology, 163(12),* 1149-1156.

Caliendo, M., & Kopeinig, S. (2005). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys, 22(1),* 31-72.

Chickering, A. W. (1999). Personal qualities and human development in higher education: Assessment in the service of educational goals. In S. J. Messick (Ed.), Assessment in higher education. (pp. 13-33). Mahwah, NJ: Lawrence Erlbaum.

Chowdhury, M. S., & Amin, M. N. (2006). Personality and students' academic achievement: Interactive effects of conscientiousness and agreeableness on students' performance in principles of economics. *Social Behavior and Personality, 34(4),* 381-388.

Clark, M. H., & Cundiff, N. L. (2011). Assessing the effectiveness of a college freshman seminar using propensity score adjustments. *Research in Higher Education, 52,* 616-639.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112(1),* 155-159.

Cohen, J. P., West, S. & Aiken, L. (2003). *Applied Multiple Regression / Correlation Analysis for the Behavioral Sciences (*3rd ed.) Hillsdale: Erlbaum.

Czajka, J. L., Hirabayashi, S. M., Little, R. J. A., & Rubin, D. B. (1992). Projecting from advance data using propensity modeling: An application to income and tax statistics. *Journal of Business and Economic Statistics, 10,* 117-131.

D'Agostino, R. B. Jr. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine, 17,* 2265-2281.

Dehejia, R. (2013). Does matching overcome Lalonde's critique of non-experimental estimators? A postscript.

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluation the evaluation of training programs. *Journal of the American Statistical Association, 94,* 1053-1062.

Dehue, T. (2000). From deception trials to control reagents: The introduction to the control group about a century ago. *American Psychologist, 55(2),* 264-268.

DeMars, C. E. (2014). Cluster four assessment report: Spring 2014. James Madison University, Harrisonburg, VA.

Diamond, A. & Sekhon, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *The Review of Economics and Statistics, 95(3),* 932-945.

Drake, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effects. *Biometrics, 49,* 1231-1236.

Eccles (Parsons), J.S. (1983). Expectancies, values, and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motives* (pp. 75-146). San Francisco: W. H. Freeman and Company.

Ewell, P. T. (2009). Assessment, Accountability, and Improvement: Revisiting the Tension. Occasional Paper #1. National Institute for Learning Outcomes Assessment.

Frisco, M. L., Muller, C., & Frank, K. (2007). Family structure change and adolescents' school performance: A propensity score approach. *Journal of Marriage and Family, 69,* 721-741.

Greenland, S. (2003). Quantifying biases in causal models: Classical confounding vs. collider-stratification bias. *Epidemiology*, *14(3),* 300-306.

Gu, X., & Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics, 2*, 405-420.

Hade, E. M. & Lu, B. (2014). Bias associated with using the estimated propensity score as a regression covariate. *Statistics in Medicine, 33,* 74-87.

Heckman, J. J., Ichimura, H., & Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies, 65,* 261-294.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2006). MatchIt: Nonparametric preprocessing for parametric causal inference. Software for using matching methods in R. Available at http://gking.harvard.edu/matchit/.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis, 15(3),* 199-236.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software, 42(8),* 1-28.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81,* 945-960.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika, 87(3),* 706-710.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity. *The Review of Economics and Statistics, 86(1),* 4-29.

Joffe, M. M., & Rosenbaum, P. R. (1999). Propensity scores. *American Journal of Epidemiology, 150,* 327-333.

Johnston, M., Hathcoat, J., & Sundre, D. (2014). Natural World Cluster 3 Assessment Report: Spring 2014. James Madison University, Harrisonburg, VA.

Kember, D. (2003). To control or not to control: The question of whether experimental designs are appropriate for evaluating teaching innovations in higher education. *Assessment & Evaluation in Higher Education, 28(1),* 89-101.

Kirk. R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56(5),* 746-759.

Komarraju, M. & Karau, S. J. (2005). The relationship between the big five personality traits and academic motivation. *Personality and Individual Differences, 39,* 557-567.

Kuh, G. D. (2009). What student affairs professionals need to know about student engagement. *Journal of College Student Development, 50(6),* 683-706.

Kuh, G. D., Jankowski, N., Ikenberry, S. O., & Kinzie, J. (2014). Knowing what students know and can do: The current state of student learning outcomes assessment in

U.S. colleges and universities, National Institute for Learning Outcomes Assessment.

Lechner, M. (2002). Program heterogeneity and propensity score matching: An application to the evaluation of active labor market policies. *Review of Economics and Statistics, 84,* 205-220.

Lemons, C. J., Fuchs, D., Gilbert, J. K., & Fuchs, L. S. (2014). Evidence-based practices in a changing world: Reconsidering the counterfactual in education research. *Educational Researcher, 43(5),* 242-252.

Leuven, E., & Sianesi, B. (2003). Psmatch2. Stata module to perform full Mahalanobis and propensity score matching.

Lu, B., Zanutto, E., Hornik, R., & Rosenbaum, P. R. (2001). Matching with doses in an observational study of a media campaign against drug abuse. *Journal of the American Statistical Association, 96,* 1245-1253.

Luellen, J. K., Shadish, W. R., & Clark, M. H. (2005). Propensity scores: An introduction and experimental test. *Evaluation Review, 29,* 530-558.

Martin, A. J. (2009). Motivation and engagement across the academic life span: A developmental construct validity study of elementary school, high school, and university/college students. *Educational and Psychological Measurement, 69,* 794-824.

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods, 9(4),* 403-425.

Normand, S. T., Landrum, M. B., Guadagnoli, E., Ayanian, J. Z., Ryan, T. J., Cleary, P. D., & McNeil, B. J. (2001). Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: A matched analysis using propensity scores. *Journal of Clinical Epidemiology, 54,* 387-398.

Osborne, J. (2012). Logits and tigers and bears, oh my! A brief look at the simple math of logistic regression and how it can improve dissemination of results. *Practical Assessment, Research & Evaluation, 17(11),* 1-10.

Pike, G. (1999). The effects of residential learning communities and traditional residential living arrangements on educational gains during the first year of college. *Journal of College Student Development, 38,* 609-621.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Reed, M., & Cochrane, D. (2012). Student Debt and the Class of 2011. *Project on Student Debt*.

Ridgeway, G., McCaffrey, D., & Morral, A. (2006). Twang: Toolkit for weighting and analysis of nonequivalent groups. Software for using matching methods in R. Available at http://cran.r-project.org/src/contrib/Descriptions/twang.html.

Rocconi, L. M. (2011). The impact of learning communities on first year students' growth and development in college. *Research in Higher Education, 52,* 178-193.

Rosenbaum, P. R. (2002). Observational studies (2nd ed.). New York: Springer-Verlag.

Rosenbaum, P. R., & Rubin, D. B. (1983a). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society, 45(2),* 212-218.

Rosenbaum, P. R., & Rubin, D. B. (1983b). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70(1),* 41-55.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79,* 516-524.

Rubin, D. B. (2004). On principles for modeling propensity scores in medical research. *Pharacoepidemiology and Drug Safety, 13,* 855-857.

Rubin, D. B. (2006). *Matched sampling for causal inference*. Cambridge, UK: Cambridge University Press.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental design for generalized causal inference*. Boston: Houghton-Mifflin.

Smith, H. L. (1997). Matching with multiple controls to estimate treatment effects in observational studies. *Sociological Methodology, 27,* 325-353.

Smith, J., & Todd, P. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics, 125,* 305-353.

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

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

Steyer, R., Gabler, S., von Davier, A. A., & Nochtigall, C. (2000). Causal regression models: II. Unconfoundedness and causal unbiasedness. *Methods of Psychological Research Online, 5,* 55-87.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science, 25(1),* 1-21.

Stuart, E. A., & Rubin, D. B. (2008a). Best practices in quantitative methods: 11 Best Practices in Quasi-Experimental Designs: Matching Methods for Causal Inferences.

Stuart, E. A., & Rubin, D. B. (2008b). Matching with multiple control groups with adjustment for group differences. *Journal of Educational and Behavioral Statistics, 33(3),* 279-306.

Stukel, T. A., Fisher, E. S., Wennberg, D. E., Alter, D. A., Gottlieb, D. J., & Vermeulen, M. J. (2007). Analysis of observational studies in the presence of treatment selection bias. *JAMA, 297(3),* 278-285.

Vaughan, A. L., Lalonde, T. L., & Jenkins-Guarnieri, M. A. (2014). Assessing student achievement in large-scale educational programs using hierarchical propensity scores. *Research in Higher Education, 55,* 564-580.

Wickham, H. (2009). ggplot2: elegant graphics for data analysis. Springer New York.

Wigfield, A., & Eccles, J.S. (2002). The development of competence, beliefs, expectancies for success, and achievement values from childhood through

adolescence. In A. Wigfield and J.S. Eccles (Eds.), *Development of Achievement Motivation* (pp. 91-120). New York: Academic Press.

Winship, C., & Mare, R. D. (1992). Models for sample selection bias. *Annual Review of Sociology, 18*, 327-350.

Winship, C., & Morgan, S. L. (1999). The estimations of causal effects from observational data. *Annual Review of Sociology, 25,* 659-706.

Zhao, C., & Kuh, G. D. (2004). Adding value: Learning communities and student engagement. Research in Higher Education, *45(2),* 115-138.