Published in final edited form as:

Psychol Methods. 2010 September ; 15(3): 234–249. doi:10.1037/a0019623.

Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research

Valerie S. Harder, M.H.S., Ph.D.,
University of Vermont College of Medicine

Elizabeth A. Stuart, Ph.D., and
Johns Hopkins Bloomberg School of Public Health

James C. Anthony, Ph.D.
Michigan State University College of Human Medicine

Abstract

There is considerable interest in using propensity score (PS) statistical techniques to address questions of causal inference in psychological research. Many PS techniques exist, yet few guidelines are available to aid applied researchers in their understanding, use and evaluation. This study gives an overview of available techniques for PS estimation and PS application. It also provides a way to help compare PS techniques, using the resulting measured covariate balance as the criterion for selecting between techniques. The empirical example for this study involves the potential causal relationship linking early-onset cannabis problems and subsequent negative mental health outcomes, using data from a prospective cohort study. PS techniques are described and evaluated based on their ability to balance the distributions of measured potentially confounding covariates for individuals with and without early-onset cannabis problems. This paper identifies the PS techniques that yield good statistical balance of the chosen measured covariates within the context of this particular research question and cohort.

Keywords

observational studies; causal inference; standardized bias; cannabis

INTRODUCTION

Observational studies, also known as non-experimental or non-randomized studies, often are the only viable options in many psychological research studies that intend to address cause and effect questions. For example, in drug dependence epidemiology, there are open questions about the effects of early drug involvement, but it is unethical and likely impossible to randomly assign individuals to levels or types of drug involvement (Lipsey & Cordray, 2000). However, a complication in these observational studies is that drug involvement is not a random event;

CORRESPONDING AUTHOR: Valerie S. Harder, University of Vermont, 1 South Prospect Street, Burlington, Vermont 05401.

Publisher's Disclaimer: The following manuscript is the final accepted manuscript. It has not been subjected to the final copyediting, fact-checking, and proofreading required for formal publication. It is not the definitive, publisher-authenticated version. The American Psychological Association and its Council of Editors disclaim any responsibility or liabilities for errors or omissions of this manuscript version, any version derived from this manuscript by NIH, or other third parties. The published version is available at www.apa.org/pubs/journals/met

therefore, many differences exist between individuals with and without drug involvement, resulting in the need to account for many potentially influential confounding covariates. Standard epidemiologic and statistical methods, such as multivariable regression models, often are used to control for some of these confounding covariates, but these models have limitations, often including a need for parsimony. Epidemiologists have been searching for more effective means of dealing with large numbers of confounding covariates and statisticians have developed such methods.

For more than two decades, advanced statistical methods known as propensity score (PS) techniques, have been available to aid in the evaluation of cause-effect hypotheses in observational studies (Rosenbaum & Rubin, 1983a). Nonetheless, PS techniques have not yet been used widely in psychological research. This absence may, in part, be due to a lack of understanding of the underlying principles of PS techniques. Perhaps more to the point, there are few clear guidelines to aid applied researchers in their use of the many available PS techniques.

In this manuscript we aim to 1) help psychological researchers become more familiar with PS techniques, and 2) provide details of one way to evaluate the performance of PS techniques. We describe some of the common PS techniques and their underlying principles and considerations. We also present an empirical example to illustrate one strategy for evaluating PS techniques. While it has limitations, psychological researchers in the field of drug use and psychiatric disorders may find our empirical example particularly useful; it explores a suspected causal relationship linking early-onset cannabis problems with negative mental health outcomes. We step through the process of estimating the PS for early-onset cannabis problems, applying the PS to balance the distributions of measured, potentially confounding, covariates (from here on referred to as measured covariates), and estimating the effect of early-onset cannabis problems on negative mental health outcomes in young adulthood.

We separate the description of PS techniques into two main steps: estimation and application. The PS itself is an estimate of the probability of observing a given value of the exposure variable, which might be an intervention or a treatment. In our empirical example, the exposure variable is the existence of early-onset cannabis problems. We choose to sustain the vocabulary used in the causal inference literature for estimation of the effects of a ‘treatment’. As such, we refer to our exposure (early-onset cannabis problems) as a ‘treatment’ throughout. This probability of treatment is predicted for each individual using a model of the treatment received as a function of measured covariates; this is the propensity score model. These measured covariates are chosen based on their empirical and/or theorized potential to confound the relationship between the treatment and the outcome. Causal claims rely on an assumption, discussed below further, that there are no unmeasured confounders once we condition on the measured covariates. The prediction of the treatment received from this model yields the PS for each individual: their propensity to have been ‘treated.’

The second step, referred to as PS application, involves the use of the estimated PS to help make the treated and non-treated groups comparable. Methods for applying the PS presented in this manuscript include matching (Rosenbaum & Rubin, 1985; Dehejia & Wahba, 1998, 1999; Ho et al., 2007a, 2007b; Stuart, 2009), subclassification (Rosenbaum & Rubin, 1984; Lunceford & Davidian, 2004), and weighting (Hirano et al., 2003; Robins et al., 2000). Adding the PS linearly as a covariate to the final regression model is also an option, although there is mixed evidence regarding its performance (Rubin, 2001; Rubin, 2004; Shadish et al., 2008); while simply including the PS itself does not generally work well, very recent work has indicated that including non-linear PS terms to the outcome regression model can work well (Schafer & Kang, 2008). However, since the method of adding the PS as a covariate to the final regression does not allow us a straightforward way of separating the design from the

analysis of our statistical model such that we can use our chosen criterion to evaluate the performance of PS techniques (checking the balance of measured covariates) after the PS is applied, we do not illustrate its use. More complete descriptions of the PS estimation and application steps and our chosen criterion for assessing the performance of the PS techniques illustrated in this manuscript are provided in the Methods section below.

Any PS estimation can be combined with any PS application, but previous research has generally not highlighted this important distinction between the estimation and application steps and the key point that the pairing of different estimation and application methods leads to unique PS techniques, with varying performance. We are interested in exploring these two steps jointly to see how different pairings affect our results. We pair three PS estimation methods with five PS application methods for a total of 15 different PS techniques. There are also other estimation and application methods that are not illustrated in this manuscript, and the number of such methods is likely to increase over time as new methods are developed. The goals of this manuscript are on helping applied researchers understand how to implement any of the PS techniques illustrated in this manuscript and to evaluate any chosen PS technique regardless of whether or not it is discussed here. We do not provide a taxonomy of all possible PS techniques, and we do not provide a list of the only PS techniques that a researcher should consider.

Each PS technique presented in this manuscript is evaluated on its ability to achieve balance on the measured covariates after PS application between the treatment and comparison groups. Of course there are other methods for evaluating PS techniques; we make no claim that this is the best or the only way to evaluate PS techniques. It should be noted that the ultimate goal of any cause-effect analysis is an accurate estimate of the true causal effect, not simply the balance of measured covariates. A key assumption to get us there is that of strongly ignorable treatment assignment (also known as strong ignorability; Rosenbaum & Rubin, 1983a). The main component of the assumption of strong ignorability is that conditional on the measured covariates, there are no unmeasured confounders of the association between the treatment and the outcome. In other words, the measured confounders that have been included in the estimation of the PS account for the overt bias, and there is no hidden bias due to unmeasured confounders.

The empirical example presented in this manuscript comes from a study not designed to examine the causal effects of adolescent drug use using PS techniques, so although we try our best to account for the confounders that are measured, we are not able to control for all possible confounders--unmeasured confounders likely remain, and thus it is likely that this example violates the assumption of strong ignorability. Shadish et al. (2008) show that treatment effect estimates are sometimes biased when the set of measured confounders is not sufficient (e.g., when only demographics or other “variables of convenience” are included) and strong ignorability does not hold. Therefore, we fully admit that even our most balanced PS techniques in our empirical example may still yield biased treatment effects. While outside the scope of this paper, which aims to illustrate propensity score techniques, a full substantive investigation of the effects of adolescent drug use from these data should include a sensitivity analysis to assess sensitivity to unmeasured confounders (e.g., Rosenbaum & Rubin, 1983b). The main point to keep in mind when researching questions of causal inference within a PS framework is that achieving balance on the measured covariates may not be sufficient if the strong ignorability assumption is not satisfied.

This method of evaluating the PS technique based on its ability to achieve balance on measured covariates relies on the assumption that balancing measured covariates reduces overt bias (bias due to the measured covariates; Rosenbaum, 2002) in the treatment effect estimate (Rubin & Thomas, 1996; Ho et al., 2007a; Imai et al., 2008; Imai & van Dyk, 2004; Hansen & Kleyman,

2008; Rosenbaum, 2009). Therefore, we assume that a treatment effect estimate generated from a PS technique that does not at least achieve balance on the measured covariates is more likely to be biased. We do not prove this assumption in this manuscript but utilize it to judge whether our models for the PS are good models based on whether their application results in balancing observed covariates; this is in line with suggestions made by Ho et al. (2007a) and Rosenbaum (2009). Furthermore, we support reporting only those treatment effect estimates that result from PS techniques that achieve balance on measured covariates, rather than all possible effect estimates. We elaborate on these ideas and discuss assumptions more thoroughly throughout the manuscript.

Most published studies using PS techniques report a treatment effect estimate calculated using a single PS estimation method paired with a single PS application method. One applied study uses a single estimated PS paired with multiple PS application methods and reports all resulting treatment effect estimates (Kurth et al., 2006), and although they do report a large variability in results, they do not present criteria for comparing or selecting between their results, such as evaluating the balance of measured covariates. We are aware of only a few researchers that have systematically reviewed the effect of varying PS application methods (Schafer & Kang, 2008, Lunceford & Davidian, 2004, Morgan & Winship, 2007, Stuart & Green, 2008). However, to the best of our knowledge, there is no published study that has systematically reviewed several PS estimation methods in combination with several PS application methods and evaluated the ability of the combinations to balance measured covariates. Although we describe 15 different PS techniques and evaluate each technique's ability to balance measured covariates, we are not suggesting that applied researchers must follow all our steps as a template. We aim to convey to the applied researcher a range of PS techniques available, the importance of checking the balance of measured covariates, and illustrate that not all PS techniques balance all measured covariates. The last step of whether balancing the measured covariates adequately reduces the bias in the treatment effect is not addressed in this manuscript as we do not know the true treatment effect in our empirical example (as is the case for nearly any empirical observational study and should be noted as a limitation).

Our study uses the standpoint of clearly separating “design” and “analysis” phases for causal inference, as recommended by Rubin and colleagues (Rubin, 2001; Stuart & Rubin, 2007; Rubin, 2008). In particular, we focus on the design phase by selecting PS techniques that balance measured covariates, before exploring the relationship between treatment and outcome. Although the outcome may be referenced indirectly in the selection of the measured covariates, the outcome itself does not directly inform the PS estimation or application steps. Once the PS technique is selected, the researcher can move on to the analysis phase: building final cause-effect models. For the interest of readers, we briefly discuss some results from the analysis phase. Full details of our research questions and final treatment effect estimates from our analyses have been published elsewhere (Harder et al., 2008).

METHODS

Empirical Example

Study Population—The study population originated from a prospective cohort study that began in 1985-1986 when the Prevention Research Center at Johns Hopkins University enrolled 2,311 first-grade children in a randomized trial of two preventive interventions (Kellam et al., 1991, 1992). For children who remained in the school system until 1989, a series of personal interviews were conducted every year until the spring of 1994 (mean age = 14 years old; Kellam & Anthony, 1998). Between 1999 and 2002, as the participants were entering young adulthood, attempts were made to re-contact all of the original 2,311 enrollees. Participants who could be located were interviewed in-person and/or by telephone; this resulted in survey responses from over 75 percent of the original cohort. After excluding individuals

due to missing data on cannabis use (the treatment of interest) and/or young adulthood mental disorder (the outcome of interest), the resulting sample for this study included 1,494 individuals.

Measures—The treatment is defined as having a cannabis use disorder (abuse or dependence) before age 17 (early-onset). First instance of cannabis abuse or cannabis dependence before the age of 17 is based on the Diagnostic and Statistical Manual-IV criteria, as assessed during the young adult interview (DSM-IV, American Psychiatric Association, 1994). These structured diagnostic interviews, based on modules from the Composite International Diagnostic Interview (CIDI; World Health Organization, 1990), involve retrospective self-reports of degree of cannabis use and age of onset of cannabis problems. Participants are classified as having early-onset cannabis problems from self-report during the in-person or telephone interviews. We consider a dichotomous treatment variable, indicating early-onset cannabis problems versus no early-onset problem use. In our study sample, 217 individuals report early-onset cannabis problems, leaving 1,277 individuals as potential comparison individuals. The comparison group contains individuals who have never used cannabis as well as individuals who have used cannabis but do not report problem use before age 17.

The measured covariates used in this study include demographics (sex and race), socioeconomic status (family income and free lunch status), other drug use (alcohol or cigarette problem use and any other illegal drug use besides cannabis), early socially maladaptive behavior and psychological well-being (aggressive behavior, shyness, concentration, depression and anxiety), prevention trial intervention status, and parental monitoring variables. These covariates were measured at the beginning of the prevention trial (e.g., teacher-rated aggressive behavior and shyness levels soon after entry into primary school) or during questionnaire or interview assessments conducted on multiple occasions between entry into primary school and age 17 (e.g., level of parental monitoring and supervision of child and adolescent behavior). Further details on the broader study data collection activities are in Kellam & Anthony (1998), Reed et al. (2002), Ialongo et al. (2004), and Reed et al. (2007).

These measured covariates are chosen for their theorized association with both the treatment and the outcome and because they are not in the causal pathway between treatment and outcome. It is advisable to use covariates that are measured temporally before the treatment to avoid the possibility of them being affected by the treatment. Fixed covariates, such as race, can be measured at any time in a study. In our study, all covariates, except other drug use, were measured prior to age 12. For treated individuals, other drug use is coded as ‘present’ if it occurs at an age prior to their age of reported first cannabis problems. For comparison individuals, other drug use is coded as ‘present’ if it occurs prior to age 17. As discussed above, as in most observational studies, hidden bias likely remains due to unmeasured confounders, and therefore, our critical assumption of strong ignorability for claims of causal inference may not hold. Sensitivity analyses exist to determine how sensitive the results are to an unmeasured confounder (Rosenbaum, 2002). Since the goals of this manuscript do not include conclusions of whether our treatment and outcome are causally related, sensitivity analyses are beyond the scope of this manuscript.

All of the covariates used in this study were categorical (Table 1). We re-categorized some covariates that had been continuous variables created from averaged ordered categorical items, such as childhood emotional and behavior problems. These were categorized based on the ordered categories of the items that made up the scale. By retaining the categorical structure rather than utilizing these as continuous variables, we were able to allow for non-linear relationships for every unit increase in these measured covariates. Another option might have been to keep them as continuous covariates and to add quadratic or cubic terms to allow for non-linear modeling of the covariates. PS techniques handle all types of covariates, and choices

in modeling covariates should be made by researchers based on their own understanding of their datasets just as is done in any statistical modeling procedure.

There were some missing covariate values in the data. We used a missing indicator approach in the estimation of the PS, as suggested by Rosenbaum and Rubin (appendix, 1984) and Rosenbaum (2009) and applied recently by Haviland et al. (2007). Although this approach is not a generally effective method of dealing with missing data in many statistical models (Allison, 2002; Graham, 2009), it works well in this PS context and has a theoretical justification (Rosenbaum & Rubin, 1984). In this setting, the missing indicator approach essentially tries to balance both the observed values of the covariates and the missing data patterns. A commonly considered and easy to apply alternative approach is to exclude any subjects with missing data on the covariates. However, since missing data on covariates is often very common, especially when a large set of covariates is being used, this might exclude a potentially large number of subjects from analyses and therefore is generally not recommended in PS analyses. By including a missing value indicator, no individuals are removed from analyses due to missing covariate values. Other principled approaches for handling missing covariate values in PS analyses are pattern mixture models (Rosenbaum & Rubin, 1984, D'Agostino et al., 2001), general location models (D'Agostino & Rubin, 2000), and multiple imputation (Song et al., 2001). Recently, a combination of multiple imputation and the missing indicator approach has been proposed (Qu & Lipkovich, 2009). Few studies examine the influence of a chosen missing data method on the bias and variance of treatment effect estimates when using PS techniques. Recent simulation studies shed some light on this issue (D'Agostino et al., 2001; Qu & Lipkovich, 2009), and although bias on the treatment effect is reported for some missing data methods in the simulation results (including using a missing indicator), interestingly, neither study finds significant differences in treatment effect estimates between the various missing data methods applied to real datasets. While the missing data technique used in this study has been advocated by some, it is certainly not necessarily always the best approach, and many still argue against its use. Researchers should give careful consideration to the method chosen for handling missing data in their datasets as new information emerges.

Statistical Methods

Balance of Measured Covariates—An important step in any PS analysis is to assess the balance of the measured covariates between the treatment and comparison groups, where balance refers to the similarity of the covariate distributions. We use a quantity similar to the effect size, known as the standardized bias, to quantify this balance. The standardized bias for continuous covariates is calculated by dividing the difference in means of the covariate between the treated group and the comparison group by the standard deviation. The standard deviation can be calculated using just one of the treatment groups or the groups combined; the key is to be consistent and to use the same method before and after PS-application. Since all measured covariates in this study are categorical (and thus can be expressed as a set of binary indicators), this study uses the differences in proportions of each level of the measured covariate standardized (divided) by the standard deviation in the treatment group. For instance, if a categorical covariate has four levels, then the standardized bias is calculated for each level, rather than calculating a single mean standardized bias treating the categories as ordered and continuous.

Decision Criterion—Using a single decision criterion, we evaluate all 15 PS techniques in their ability to balance the measured covariates between our treatment and comparison groups by reducing the standardized bias. Our decision criterion (Ho et al., 2007a) considers a covariate balanced if the standardized bias is less than 0.25, although this is a rule of thumb rather than a strict cut-off (Rubin, 2001). Some might suggest that when possible, an even stricter cut-off, perhaps less than 0.10, is preferable in PS analyses. Given that this cutoff is not strict, we

recommend that researchers carefully evaluate their measured covariates whose standardized biases are below, yet close to 0.25. For example, a covariate that is very strongly associated with the outcome and that has a standardized bias of 0.24 after PS application may yield substantial bias in the treatment effect estimate, even though it would technically be below the 0.25 cut-off. The 0.25 value provides some guideline, in part based on results in Cochran and Rubin (1973), but there is no widely accepted absolute standardized bias criterion for assessing when balance occurs, and it is the responsibility of the researcher to consider the specifics of their own study.

In some instances, researchers may face a situation where they have tried several PS techniques, and they have several techniques that meet this decision criterion. We suggest that if more than one technique meets this criterion then researchers may consider a stricter criterion such as 0.10 or the researcher may consider a second criterion in which they focus on the more theoretically or empirically critical confounders and choose the technique that produces the lowest standardized biases on those covariates in addition to the overall low bias on the full set of covariates. In this scenario, where more than one PS technique meets the decision criterion, it may also be appropriate to report the effect estimates utilizing these different PS techniques as a sort of sensitivity analysis: it is assumed that methods that yield similar balance should yield similar effect estimates.

Another scenario may occur in which none of the PS techniques used meet this decision criterion. An option in this setting includes trying different PS estimation or application methods (as described in this manuscript or in other literature) to see if this balance criterion is attainable. As referenced later in the Outcome Analyses section, an additional strategy for minimizing residual bias, which might be due to moderate levels of imbalance, is that of incorporating covariance adjustment when estimating the final treatment effect. However, if no PS technique yields adequate balance, this may imply that the research question of interest is not answerable from the available data: that the treatment and comparison groups are just too dissimilar (as found, for example, in Haviland et al., 2007).

Although we present these general guidelines, there are also other ways of assessing balance, and in fact it may be beneficial to use a variety of balance measures. Other statistical tests such as t-tests, Mantel-Haensel tests, and Kolmogorov-Smirnov tests can be used to assess the statistical significance of the imbalance of measured covariates and to help make decisions. However, care needs to be taken when comparing test statistics and p-values before and after the PS application, as changes in balance may be conflated with changes in statistical power (Imai et al., 2008).

The evaluation of balance of measured covariates is illustrated in two places in this manuscript. The first is during the iterative process of building the more complex model for parametric PS estimation (Model 2), as discussed further in the next section. The second is immediately following the PS application step; evaluating the balance on the measured covariates between the treatment and comparison groups prior to estimating the treatment effect.

Propensity Score Estimation—We consider two parametric methods and one non-parametric method for estimating the PS. The first parametric method (Model 1) is a multivariable logistic regression predicting treatment status with all of the measured covariates included as predictors. The second parametric approach (Model 2), described by Dehejia and Wahba (1998,1999), after initial conceptualization by Rosenbaum and Rubin (1983a), starts with Model 1, but then adds selected measured covariate product-terms through an iterative process. The third estimation method (Model 3), by McCaffrey, Ridgeway and Morral (2004), is a non-parametric generalized boosted modeling (GBM) technique that uses regression trees.

The iterative process that distinguishes Model 1 from Model 2 requires the evaluation of balance of the measured covariates and their product-terms within subclasses of the PS. The algorithm used for building Model 2 in this empirical example begins by ordering the PS obtained from Model 1 from low to high and then dividing the PS into strata such that each stratum holds an approximately equal number of treated individuals. In this study, five strata are defined because six strata result in very few comparison individuals in the top stratum. The balance of the measured covariates and all of their two-way product-terms is then evaluated within each stratum. Criteria for including a product-term must be established. In this study, if the standardized bias for a product-term term is above 0.25 across two or more strata, then that product-term is added to the PS model. If continuous covariates are included in a PS estimation model, then researchers may consider checking the balance of higher order polynomial terms of those covariates (e.g., X^2 , X^3) and potentially including those terms in the model if they are unbalanced (Dehejia & Wahba, 1999). After the incorporation of product-terms, we re-estimate the PS, and the strata-wise standardized biases are re-calculated to check the balance of each covariate and the product-terms. If the overall balance of the measured covariates and their product-terms has improved, the new PS model is used. Otherwise, different product-terms are incorporated into the model and the process is repeated. This process is iterated until the balance no longer improves when adding terms. To be consistent with the recommendations of Dehejia & Wahba (1999) we use subclasses to check the balance regardless of the PS application method that will be used; an interesting refinement of the method could potentially tailor the balance checks at this stage to the application method that will be used.

The estimation of the PS using Model 3 is similar to Model 2 in that it can include product-terms, but it does so in a non-parametric manner. The method used here, GBM (Ridgeway, 1999), is a variant of boosting. Boosting is an algorithmic approach to fitting a non-linear surface predicting treatment assignment and can incorporate a large number of measured covariates. Friedman (2001) and Madigan and Ridgeway (2004) have shown that boosting can outperform alternative methods in terms of prediction error. GBM uses regression trees (Breiman et al., 1984), a form of recursive partitioning which starts by splitting the data across values of the one covariate that best predicts treatment as compared to all other measured covariates. Within each new subset formed by the initial split, additional splits are made sequentially, with the same statistical criterion used to select the best predictor out of the remaining measured covariates. Via this process, GBM forms numerous simple regression trees, which then are combined to yield one smoothed function for the overall estimate of the PS. GBM has flexibility in that it allows multiple splits of each simple regression tree and allows for multi-way product-terms between all measured covariates in an attempt to optimize the likelihood function. This method involves tuning parameters that can be adjusted by the user until measured covariate balance is achieved (McCaffrey et al. 2004). McCaffrey et al. (2004) provide statistical details of the non-parametric estimation of the PS through GBM, a simple example of how many and where splits are made in the regression trees, a discussion of the advantage of GBM over alternative parametric approaches, and sensitivity analyses to illuminate the limitations of GBM. The GBM-estimated PS can be used with any PS application method, such as subclassification or matching (described in more detail below), but at least in the GBM-based software provided by McCaffrey et al. (2004), the default PS application method used when tuning parameters are adjusted for balance of measured covariates is weighting by the odds (described in more detail below). As discussed above in relation to using subclassification for Model 2, an interesting future refinement of this method would be to tailor the tuning parameters to the application method that will be used.

After estimating the PS, the overlap (region of common support) of the distributions of the estimated PS for the treatment and comparison groups should be examined graphically before moving on to the PS application step. There are three possible scenarios: first, the PS

distributions might overlap completely in terms of both range and density, meaning the two groups (treatment and comparison) are almost identical with respect to the chosen covariate distributions – i.e., the treatment and the comparison groups have essentially the same propensity to have been treated based on the observed covariate values. Second, there could be absolutely no overlap between the PS distributions. Third, the PS distributions might have some degree of overlap.

The first scenario, which is expected in the context of a randomized controlled trial, is generally unlikely in the context of most observational studies. The second scenario—that of no overlap of the distributions—may indicate that there are too many pre-existing differences between the treated and comparison individuals for reliable causal inference, and thus might lead the research team to stop the analysis. For example, it may be that all individuals in the treatment group are also problem drinkers, leading to difficulties in separating out the effects of the treatment from problem alcohol use. A more complex PS estimation model, such as one with many higher-order covariate terms, is also more likely to yield two PS distributions with less overlap than is a simpler estimation model. Although having fewer covariates in the PS model may appear to improve the overlap of the PS distributions, excluding measured covariates that are confounders violates the strong ignorability assumption and should not be done. The third scenario, some overlap, is the most commonly encountered. The next section describes the PS application methods that aim to help Scenario 3 approach Scenario 1 through matching, weighting, or subclassification.

Propensity Score Application—After PS estimation, the next step is PS application. The goal of the application step is to use the estimated PS to match, weight, or group (“subclassify”) the treated and comparison individuals in preparation for the final outcome model. Here we discuss and apply five PS application methods: 1:1 matching, full matching, weighting by the odds, inverse probability of treatment weights (IPTW), and subclassification.

Before we provide details on each of these five PS application methods, we draw a distinction between two estimands of interest: 1) the average treatment effect on the treated (ATT, Hirano et al., 2003), and 2) the average treatment effect (ATE) in the population (Imbens, 2004; Ho et al., 2007a). The ATT is the average effect that would be seen if everyone in the treated group received the treatment, compared to if no one in the treated group received the treatment. In contrast, the ATE is the average effect that would be seen if all individuals (treated and comparison) received the treatment compared to if none of these individuals (treated and comparison) received the treatment. If the treatment effect is constant or if the treated and comparison groups are only randomly different from one another, as in a randomized controlled trial, the ATT and ATE will be equal. However, if the treatment effect varies across individuals, and in particular varies across individuals in the treated and comparison groups, the ATT and ATE will not necessarily be equal. Different methods may yield different estimates of the treatment effect if one is estimating the ATE while the other estimates the ATT (e.g., as in Kurth et al., 2006). Therefore, each researcher must decide and explain which causal effect estimand is most appropriate to report for their particular research question and use a PS application method that estimates that quantity.

Three PS application methods illustrated in this manuscript estimate the ATT (1:1 matching, full matching, and weighting by the odds) and two estimate the ATE (IPTW and subclassification). The estimand of each method depends on the (implicit or explicit) weighting that is done to the treated and comparison groups, as described in further detail below. Weighting by the odds and IPTW are an example of how PS weighting applications can estimate either ATT or ATE, and although not illustrated in this manuscript, with slight modifications, the PS matching and subclassification application methods can also be used to

estimate either the ATT or the ATE. See Morgan and Winship (2007) and Stuart (2009) for more details on these various options.

The following is an overview of each of the five PS application methods we present here.

1:1 Matching: One-to-one matching selects for each treated individual the comparison individual with the most similar PS (Rosenbaum & Rubin, 1985). Comparison individuals not selected as a match are dropped from subsequent analyses. Conceptually this is very simple, although complications arise when two treated individuals both have the same “closest” comparison individual. In those cases, when a simple “greedy” match is done, which comparison individual is paired to each treated individual will depend on the order in which the treated individuals are matched. More advanced methods, such as optimal matching, where the matches are chosen to minimize the average distance between all the matched pairs (Ho et al., 2007a), or matching with replacement, can also be used. Another variation is 1:k matching, where k comparison individuals are selected for each treated individual; 1:k matching will tend to work particularly well when the pool of comparison individuals is much larger than the set of treated individuals. Because a match (or matches) is obtained for each treated individual, 1:1 or 1:k matching estimates the ATT. Once the matches are obtained, the standardized biases of the measured covariates within these matched treated and comparison groups are calculated to examine the resulting balance.

Full Matching: Unlike 1:1 matching, which discards a number of comparison individuals, full matching uses all individuals by forming matched sets that have at least one treated and at least one comparison individual in each matched set (Rosenbaum, 1991; Hansen, 2004; Stuart & Green, 2008). The matched sets are created in a way that minimizes the global PS difference, defined as the sum of the distances between the PS of all pairs of treated and comparison individuals within each matched set, across all matched sets (details in Stuart & Green, 2008). After the matched sets are formed, a weighting approach is applied where each treated individual receives a weight of 1. The comparison individuals within each matched set receive a weight proportional to the number of treated individuals in the matched set divided by the number of comparison individuals in that set. For example, in a matched set with five treated individuals and two controls, the two controls each get a weight proportional to 5/2 (see the documentation of the MatchIt software package (Ho et al., 2007b) for more details). Because of this weighting where the comparison individuals are weighted to the treated individuals, the full matching method that we implement estimates the ATT. The standardized bias of each measured covariate is then calculated using weighted proportions and weighted standard deviations.

Weighting by the odds: With weighting by the odds, treated individuals receive a weight equal to 1 and comparison individuals receive a weight equal to their PS (p_i), converted to the odds scale ($= p_i / (1 - p_i)$; Hirano et al., 2003). This weighting effectively up-weights comparison individuals whose measured covariate values (propensity scores) best match those of the treated individuals and down-weights comparison individuals whose measured covariate values are dissimilar from treated individuals. Both the treatment and comparison groups are weighted to the treatment group, thus estimating the ATT. One way of thinking about weighting by the odds is that the comparison individuals are first weighted to the entire population, using $1 / (1 - p_i)$, and are then weighted to the treatment group, using p_i . Similar to full matching, the standardized biases of the measured covariates are calculated using these weighted treatment and comparison groups.

Inverse Probability of Treatment Weights (IPTW): An alternative weighting technique, IPTW, may be used when the desired estimand is the ATE. With IPTW, each individual is weighted by the inverse probability of receiving the treatment that they actually received. In

particular, treated individuals are given an IPTW = $1 / p_i$ and the comparison individuals are given an IPTW = $1 / (1-p_i)$ (Robins et al., 2000). In this way, each group (treated and comparison) is weighted up to represent the full sample population (treated and comparison), thus estimating the ATE. Once again, after the IPTW weights are calculated, the standardized bias for each measured covariate is calculated using the weighted samples.

A concern with IPTW may arise when some weights are very large and thus influential, possibly resulting in biased (or at least very imprecise) estimates of the treatment effect (Schafer & Kang, 2008, Robins et al., 2000). It is not recommended to simply remove treated individuals with large weights because those individuals are generally the best predictors of the outcome under comparison due to the fact that a large IPTW weight results from a small PS. To reduce the variability of the IPTW weights and give individuals with extreme weights less influence, Robins and colleagues (2000) discuss a technique they refer to as stabilization. Stabilization is accomplished by multiplying the treatment and comparison weights (separately) by a constant, equal to the expected value of being in the treatment or comparison groups, respectively (Equations 1 & 2). Because the IPTW weights in each group are multiplied by a constant, stabilization does not affect the point estimate of the treatment effect but it does decrease the variance (Robins, 1999; Robins, 1998). When the PS model is saturated (i.e., it includes all covariates and possible product-terms) the variance of the final outcome estimate will be the same with and without stabilization (Robins et al., 2000; Robins, 1999; Robins, 1998), but in our setting with an unsaturated PS model, stabilization reduces the variability of the IPTW weights and reduces the variance of our treatment effect estimates. Equation 1 and Equation 2 for stabilizing the treatment and comparison group weights, respectively, are:

$$\frac{\sum_{i=1}^{N_T} PS_i}{N_T} * \frac{1}{PS_i} \quad (1)$$

$$\frac{\sum_{j=1}^{N_C} (1 - PS_j)}{N_C} * \frac{1}{(1 - PS_j)} \quad (2)$$

where i and j denote treated person i and comparison person j , N_T and N_C are the total number of treated and comparison individuals, respectively, and PS is the propensity score.

In this empirical example, we first use stabilization, and then we also use a second technique known as trimming to minimize the influence of a few remaining outlier weights. Trimming is often done in survey sampling where the sampling weights are truncated (not removing individuals but limiting the size of their weights) to be within a pre-specified range, selected to minimize the mean square error (Potter, 1990, 1993). The procedure we use trims the stabilized weights; weights are set equal to 0.10 if the stabilized weight is less than 0.10 and set equal to 10 if the stabilized weight is greater than 10. Trimming weights in this manner will affect the treatment effect estimate itself. We note here that while these adjustments can also be applied to weights generated from the weighting by the odds method, in our experience, the weights generated by weighting by the odds are generally less extreme and it is less likely that they would require stabilization or trimming. This stabilization and trimming method of adjusting IPTW weights is often ad hoc due to the lack of formalized guidelines and is another area where further research is warranted.

Subclassification: As implemented here, the final PS application method discussed also estimates the ATE. Subclassification (Rosenbaum & Rubin, 1984) divides individuals into many groups (or “subclasses”) based on their PS values; in this way it is similar to full matching but creates many fewer groupings. In studies utilizing subclassification, various numbers of subclasses have been chosen. Early work in subclassification (Cochran, 1968) is often cited as providing justification for choosing five subclasses. However, the optimal number of subclasses will depend on sample size and the amount of overlap between the treatment and comparison group PS. Unfortunately there is no one method for deciding how to delineate the subclasses, and ad hoc methods are often used by data analysts. Therefore, each dataset and its PS distributions require critical thought and decision making by the researcher. In this study, 10 subclasses were initially defined, with an equal number of treated individuals in each subclass in an attempt to ensure adequate numbers of treated individuals within each subclass. However, this initial approach yielded very small numbers of comparison individuals in the top four subclasses. This would have resulted in those few comparison individuals being given large influence in the final causal model estimates. Therefore, we decided to collapse the top four subclasses into one, yielding seven subclasses. Once the subclasses were determined, the standardized biases of the measured covariates were calculated within each subclass and then averaged over all the subclasses, weighted by the proportion of the full sample population within each subclass out of the total number in the sample (this weighting is what leads our version of subclassification to estimate the ATE; Rosenbaum & Rubin, 1984). Some researchers may prefer to adjust the delineation or number of subclasses in an attempt to minimize the standardized biases. Again, there is no one standard methodology.

Statistical Software—Many readily accessible statistical software packages can be used to conduct PS analyses. Several software packages, including Stata (StataCorp, 2005), SAS/STAT software (SAS, 2004), and R (R Development Core Team, 2007), have programs and commands written specifically to conduct PS analyses. For this study, we use two software programs written for R: ‘MatchIt’ (Ho et al., 2007b) and ‘twang’ (Ridgeway et al., 2006). MatchIt is used to estimate PS Models 1 and 2 (parametric). Twang, which calls the ‘GBM’ package (Ridgeway, 2007) is used to estimate PS Model 3 (nonparametric). MatchIt is also used for the PS application methods involving matching (1:1 and full) and subclassification, and twang is used for the two weighting application methods (odds and IPTW). For full matching, MatchIt calls the R ‘optmatch’ package (Hansen, 2009). The calculation of standardized bias is built into both MatchIt and twang, but this calculation is also easily done outside these programs. We examine the standardized biases through both graphical and numerical methods built into MatchIt and twang, as well as through R code written exclusively for this study. The R statistical software is particularly useful because of automated features of MatchIt and twang which facilitate checking balance of measured covariates and because of the ease and transparency of having open source code.

Outcome Analyses

After PS estimation and application, the third and final step in using PS techniques is to obtain and interpret effect estimates. For each of the 15 PS techniques, a generalized linear regression model can be built applying the PS technique. In our example, we use a logistic regression model with the binary outcome (a 0/1 indicator of reporting a specific mental disorder over the age of 17) and a treatment term (a 0/1 indicator for early-onset cannabis problems) and covariates as predictors. Many researchers have shown the benefits of combining PS methods and outcome regression models (Rubin, 2001; Ho et al., 2007a, Robins & Rotnitzky, 2001), and so our outcome analyses include the covariates from the PS estimation models (Harder et al., 2008).

With 1:1 matching, the final outcome regression is estimated using only the 1:1 matched sample. With full matching, weighting by the odds, and IPTW, the regression is estimated using the weights described above for each method. Lastly, with subclassification, an effect is first estimated within each subclass and then the subclass-specific effects are averaged, with the subclass-specific estimates weighted by the proportion of individuals in that subclass out of the total number in the sample. When applying the subclassification method, because we have a dichotomous outcome, the effect estimate is averaged across subclasses before it is transformed into an odds ratio, and the overall variance of the averaged effect estimate is calculated using a multivariate delta method (Casella & Berger, 2002). Continuous outcomes do not require a multivariate delta method since the variance of a linear combination is a linear combination of the variances. Although delta method calculations are common for statisticians, statistical code may not be readily available to calculate the variance of the treatment effect generated from a subclassification analysis with a binary outcome. Further improvements to existing software are needed to make this method more readily accessible for applied researchers.

RESULTS

Balance of measured covariates

Before PS application, several measured covariates are not well balanced between the treatment and comparison groups (Table 1). The measured covariates with the largest imbalances are daily tobacco use (67% of treated vs. 13% of comparison), alcohol problem use (43% of treated vs. 3% of comparison), concentration problems (39% of treated vs. 29% of comparison), and behavior problems (16% of treated vs. 6% of comparison). These differences mean that, before PS application, many of the measured covariate standardized biases are greater than 0.25 (Table 1). This imbalance of the measured covariates motivates the use of PS techniques to help bring into balance the distributions of the measured covariates.

Propensity Score Estimation

Parametric—The simple parametric PS estimation (Model 1) results in overlapping PS distributions between the treated and comparison groups. A majority of individuals in both groups have low PS, and the treated individuals have higher PS on average, as expected (Figure 1). This degree of overlap and the spread of the distribution of the comparison individuals across the full range of potential PS (0 – 1) enhance the possibility of finding appropriate comparison individuals. If there had been no comparison individuals with a high PS (or no treated with a low PS), then the utility of several application methods might have been limited. Or, some treated and/or comparison individuals might have had to be discarded due to failure to match (e.g., as in Dehejia & Wahba, 1999).

As described earlier, additional product-terms are selected to include in the estimation of Model 2 by first dividing the PS generated by Model 1 into five strata. Nine of the measured covariates have standardized biases greater than 0.25 in at least one of the five strata. Two measured covariates have standardized biases greater than 0.25 in more than two of the five strata. However, many of the two-way product-terms are imbalanced in a number of strata. In an attempt to reduce this imbalance, five product-terms having standardized biases greater than 0.25 in more than two strata are included in the revised PS model. The final Model 2 reduces the highest measured covariate standardized bias from 1.16 in one of the PS stratum to 0.45, without increasing the number of measured covariates with standardized biases greater than 0.25.

Non-Parametric—The non-parametric GBM estimation of the PS (Model 3) was implemented using the twang package, as described above. One of the available outputs is a

listing of the measured covariates that have the largest relative influence (by %) on the PS estimation. In this example, these influential covariates are alcohol problem use (46%), daily tobacco use (34%), concentration problems (3%), behavior problems (3%), sex (3%), and shyness (2%). Not surprisingly, these are many of the same covariates that showed large differences between the treated and comparison groups (Table 1) in that they are the strongest predictors of treatment. Each of the remaining measured covariates contributes less than 2% of the variance.

Comparing Estimation Results—As would be expected, the PS distributions from Models 1, 2 and 3 are highly correlated. Models 1 and 3 appear to exhibit the highest correlation (correlation coefficient=0.99), while the correlation between Model 2 with either Model 1 or Model 3 is slightly less (0.91 and 0.92, respectively). Box plots of PS distributions stratified by treatment status (Figure 1) are useful for visualizing the overlap between treated and comparison groups. Model 1 and Model 3 appear to yield more similar estimates of the PS in terms of the median and spread, while Model 2 is slightly different.

Propensity Score Estimation and Application Combined

The distribution of standardized biases resulting from each of the 15 PS techniques is shown in Figure 2. We identify only two PS techniques that meet our decision criterion and balance all of the measured covariates; one estimating the ATT and one the ATE. The PS estimate from Model 3 (GBM) paired with weighting by the odds is the only one of nine procedures estimating the ATT that has all measured covariate standardized biases below the 0.25 cut off (max=0.21, Table 1). Although most of the standardized biases decrease after PS application, the variable with the largest standardized bias after PS application has a very small standardized bias before PS application (Table 1). For the ATE, the PS estimate from Model 2 (MLR with product-terms) paired with subclassification is the only one of six that yields all measured covariate standardized biases below the 0.25 cut off (max=0.22). Median, spread, and number of outlying standardized biases vary across all 15 techniques (Figure 2). The 13 remaining PS techniques all have at least one standardized bias above the 0.25 cutoff, some as extreme as 0.5 (Figure 2).

Treatment Effect Estimates

Detailed exploration of the estimated effect of early-onset cannabis problems on negative mental health outcomes, including subgroup analyses by sex, is reported in Harder et al. (2008). The effect estimates from our chosen ATT and ATE methods lead to qualitatively similar, not statistically significant, conclusions regarding the association of interest. Because of the nature of the research question and the particular dataset, detailed further in the Discussion, we focus on presenting only the ATT estimates. Our balance criterion for our measured covariates suggests that we should feel confident in reporting the effect estimate from the GBM estimation (our Model 3) in combination with weighting by the odds (last row in Table 2). For illustrative purposes, Table 2 presents the effect estimates from all of the techniques estimating the ATT; the values for effect estimates in that table should not be used alone to reach conclusions about the relationship between early-onset cannabis problems and negative mental health outcomes. Our companion manuscript clearly discusses conclusions regarding the causal association between early-onset cannabis problems and later negative mental health outcomes (Harder et al., 2008).

DISCUSSION

To date, there are many PS estimation and application methods to choose from but very little discussion regarding evaluation of these techniques. Many textbooks have been published with regard to model building and model checking for traditional regression techniques (e.g.,

Harrell, 2001), but at present, applied researchers will find very few references for evaluating various PS techniques. This is not the first manuscript to discuss the specifics of PS techniques nor is it the first to discuss the benefits of assessing the balance between treatment and comparison groups and selecting a PS method before the treatment effect is estimated (Rubin, 2001; Schafer & Kang, 2008; Rubin & Thomas, 1996; Ho et al., 2007a; Imai et al., 2008; Imai & van Dyk, 2004; Hansen & Kleyman, 2008). What is unique about this manuscript is that it describes three different estimation methods and five different application methods (clearly separating these two steps), illustrates their 15 combinations in one empirical example, and presents specifics regarding one method for evaluating the ability of each combination to achieve balance on measured covariates. This contribution may help applied researchers who need to or want to explore different PS techniques to evaluate the ability of their chosen technique(s) to balance measured covariates, or to explore the sensitivity of their treatment effect estimate to varying PS techniques.

While in this empirical example, we find only one ATT technique and one ATE technique that meet our decision criterion of having all standardized biases below 0.25, there may be value in presenting effect estimates from several different PS techniques within those that estimate the ATT or the ATE. However, it is important that these estimates be presented only from PS techniques that balance the measured covariates between the treatment and comparison groups. Presenting treatment effect estimates from PS techniques that do not balance the measured covariates can be misleading because of the assumption that the treatment effect estimates from those analyses may still be subject to overt bias.

It is important to review the assumptions and limitations of PS techniques. The assumption of strong ignorability (Rosenbaum & Rubin, 1983a) has two components: 1) conditional on the measured covariates, there are no unmeasured confounders of the association between the treatment and the outcome (i.e., no hidden bias, as discussed earlier), and 2) every individual has a positive probability of receiving the treatment and of not receiving the treatment (i.e., each person could have been a problem cannabis user or not), and thus the effect can be defined for the full sample. The first component of this assumption is usually more problematic. In a randomized controlled trial (assuming an adequately large sample size), measured and unmeasured covariates are balanced between the treated and comparison groups. In our study, as in nearly any observational study, we can only try to directly balance the measured covariates and assume that balance reduces the overt bias. The inability to balance unmeasured confounders is a major limitation of observational studies in general, and neither PS techniques nor traditional multivariable regression methods are able to directly correct for this potential hidden bias. In practice, it is important to assess whether the assumption of strong ignorability is likely to be met before the application of PS techniques: i.e., are there important confounders that are not observed in the study? In our empirical example, we may have been very concerned about this assumption had we not been able to observe childhood measures of behavior and psychopathology and early alcohol problem use or other drug use prior to the early-onset cannabis problem use. As mentioned earlier, another strategy that can be used to assess the assumption of strong ignorability is to test the sensitivity of the results to a potential unmeasured confounder, as recommended by Rosenbaum & Rubin (1983b) and Rosenbaum (2002). This method has been applied relatively rarely, but examples can be found in McCaffrey et al. (2004), Schafer and Kang (2008), and Haviland et al. (2007).

One issue that relates to the second component of strong ignorability is that of the overlap of the PS between the two groups. Checking the overlap of the PS distributions between the treatment and comparison groups may illuminate situations where 1:1 matching will not be feasible because there are too few comparison individuals with PS similar to those of the treated group. To deal with settings where there is not complete overlap, researchers sometimes choose to modify the original dataset by excluding some individuals: those with PS outside the range

of the PS for the other group (outside the region of “common support”; Heckman et al., 1997, 1998). For example, some analyses appear to benefit from removal of the comparison individuals with PS that are outside the range of the treated groups’ PS (Dehejia & Wahba, 1999). One caution in doing this sort of data reduction is that it may affect the generalizability of the treatment effect estimates -- i.e., those to whom the effects being estimated apply. Removing comparison individuals (but not treated individuals) will still enable estimation of the ATT, but it will no longer be possible to estimate the ATE. In our empirical example, the region of common support from all three estimation methods covers almost the entire available range of PS (Figure 1), yet our comparison individuals with high PS are outliers (outside the upper whisker of the box plot). Using comparison individuals that are outliers is fine for the estimation of the ATT, but it may present a problem for the estimation of the ATE, and therefore, this visualization of our PS overlap influenced our decision to present only ATT outcome estimates for these data. Another thing to consider with overlap is whether or not the number of comparison individuals at the higher PS range is adequate to match to the treated individuals. If a PS for a treated individual is not close to any comparison individual’s PS, then “... inferences for the causal effect of treatment on such [an individual] cannot be drawn without making relatively heroic modeling assumptions involving extrapolations” (Rubin, 2001). Since we did have some comparison individuals in the higher PS range, rather than exclude treated individuals to make the matching better, we decided to use the full sample and to allow the balance check to be our decision criterion. Although our decision not to remove any treated individuals from the analyses may be questioned by some, we feel that restricting the analysis to a subset of treatment group members would limit the conclusions of the study. Regardless, we caution against drawing causal inference for the few treated units that may not have had a suitable comparison. If the decision to discard data is made, the researcher must carefully consider the potential implications of such a decision and be clear to whom the estimated effects apply.

Although this empirical example does not involve an a priori exclusion of treated individuals from analyses, one PS application method (1:1 matching) automatically excludes comparison individuals who are not selected as a match. In this study, over 1,000 comparison individuals are not matched in the 1:1 matching and therefore are left out of the final outcome analysis. This exclusion of individuals may seem as if it would substantially reduce the statistical power of the study to detect effects. However, this is not always the case because of the improved covariate balance (Snedecor & Cochran, 1980; Smith, 1997). Nevertheless, it is still the case that 1:1 matching may not be a preferred method due to the sometimes drastic reduction in sample size. When researchers are reluctant to discard data in this way, other approaches, such as 1:k or full matching, may be more appropriate, assuming good balance can still be attained with those other approaches.

We present basic versions of each of the PS techniques described. Many variations are possible. For example, some matching methods utilize the Mahalanobis distance, a multivariate distance measure. Mahalanobis matching within propensity score calipers has been found to be quite effective (Gu & Rosenbaum, 1993). In some cases all of the measured covariates are included in the Mahalanobis distance, as recommended by Gu and Rosenbaum (1993). When there are a large number of measured covariates, it is sometimes beneficial to instead include in the Mahalanobis distance a smaller set of especially important covariates on which particularly close balance is desired (Rubin & Thomas, 2000). A recent application of this approach in the psychological literature is Haviland et al. (2007).

Our choice of using a single decision criterion based on the balance of the marginal distribution of measured covariates may also be considered a limitation of this study. For example, in order to facilitate comparison of all 15 PS techniques, we examine balance only on the measured covariates themselves, which are the same for all PS estimation models, not on product-terms

between these covariates, since PS estimation Model 2 and Model 3 incorporate different product-terms. Although not tested in this empirical example, it might be possible that the PS techniques using PS estimation Model 2 and Model 3 yield better balance on the product-terms that they each incorporate than the PS techniques that use Model 1. If multiple PS techniques are found to meet our chosen single balance criterion, then one might consider presenting additional balance checks on theoretically important product-terms to help facilitate choosing between multiple PS techniques. This additional decision criterion could incorporate balance checks on the product-terms included in one of the estimation models, or, to be even more comprehensive, on all possible product-terms. Although this will help researchers examine balance on the joint distribution of the covariates, rather than simply their marginal distributions, looking at the balance of all possible product-terms may raise a new question for researchers regarding which product-terms are more important confounders in their given dataset and research question. We encourage researchers to consider looking at the balance of product-terms, particularly those that are believed to be important in predicting treatment assignment or the outcome and especially if the researcher has identified product-terms that are important confounders. This issue of how to examine balance of the joint distribution of covariates is a topic of ongoing research in the methodological literature (Imai et al., 2008; Hansen & Bowers, 2008). Despite all the aforementioned limitations, one of the benefits of applying this decision criterion for selecting between PS techniques is that variations of existing PS techniques as well as new PS techniques developed in the future may be checked with respect to the resulting balance of measured covariates.

Examining our results in Table 1 more critically, we find that the standardized bias for the “Intervention in school” covariate increases considerably after PS application and approached our chosen cut off. This phenomenon of increasing standardized bias is something researchers should pay special attention to and evaluate based on the relative influence of that measured covariate as a confounder. We are not as concerned about the fact that this particular covariate approaches the cut off because it merely indicates whether or not the child’s school in 1st grade was a control or experimental school; even in the experimental schools only some classrooms received the intervention. In essence, we do not believe this covariate is a strong confounder in that it is likely not strongly related to the outcome. The “Intervention in classroom” covariate is arguably more important as it specifies whether the child’s 1st grade classroom was a control, received the good behavior game intervention, or received the mastery learning intervention. The standardized bias for the “Intervention in classroom” covariate after the PS application did also increase slightly but did not indicate imbalance based on our cut off (Table 1). We encourage researchers to examine their standardized biases of measured covariates before and after the PS application step to check whether they have increased, especially for those covariates that are considered more significant confounders. Again, it is assumed that increasing standardized bias has the potential to add bias to the estimated treatment effect. Therefore, reducing the standardized bias helps to minimize the overt bias due to the measured covariates in the estimated treatment effect (Imai et al., 2008; Rubin & Thomas, 1996).

While other recent studies support the use of balance checks (Haviland et al., 2007; McCaffrey et al., 2004), balance is only a potential indicator of the level of overt bias in the treatment effect estimate, and an ideal evaluation would consider the actual bias of the estimates. Just as the true treatment effect is not known in any real (i.e., non-simulated) study, which PS technique minimizes the bias is also not known. Simulation studies can be used to compare the bias of effect estimates, and are often very helpful (e.g., as in Schafer & Kang, 2008), but simulations require an assumed model of the outcome of interest and the treatment effect itself, and it is never known how well those models and simulation settings approximate reality. Although not illustrated in this study, applied researchers should be aware that simulation studies are becoming more widely used in psychological research and may become a more standard component to support the use of PS techniques.

Choosing the PS technique prior to generating effect estimates also does not allow for the evaluation of the resulting variance of the estimated treatment effect. It would be ideal to minimize both bias and variance; while bias is the primary concern in observational studies, low variance is also an appealing feature of any estimate. In particular, if two PS techniques yield similar point estimates of the treatment effect, the one that yields a smaller variance of that estimated effect may be preferred (Hansen, 2004; McCaffrey et al., 2004; Diamond & Sekhon, 2006; Stuart & Green, 2008).

A more practical limitation of integrating PS techniques into psychological research methods may be the difficulty of implementing these techniques. There are a few statistical software packages that implement PS analyses, as mentioned earlier, but not all are widely used. Only R provides a program for non-parametric estimation of the PS (GBM), and the R packages also provide the largest set of built-in PS diagnostics. A lack of familiarity with certain statistical software packages may thus limit a researcher's ability to implement a variety of PS techniques, which may account, at least in part, for the lack of published research that explores multiple PS techniques. In an attempt to aid applied researchers in implementing the techniques presented in this manuscript, we provide annotated code that can be modified to implement PS analyses in R.

(<http://www.biostat.jhsph.edu/~estuart/HarderStuartAnthony09SampleRCode.R>).

Although it is understandable that applied researchers may choose to use the PS estimation and application methods that are most familiar and easy to implement, one of the key contributions of this manuscript is to stress the importance of considering the interaction of the estimation and application methods. For example, we see that Model 3 (GBM) estimation combined with weighting by the odds yields good balance, but Model 3 performs less well when paired with the other two ATT application methods; 1:1 matching and full matching (Table 2 & Figure 2). Within the weighting by the odds application method, the other two estimation techniques (Model 1: multivariable logistic regression and Model 2: multivariable logistic regression) also do not perform as well based solely on our one decision criterion (Table 1 & Figure 2). Thus, it is important to consider the estimation and application together, and not just assume that an estimation method that works well with one application method (in one particular dataset) will work just as well with another application method, or vice versa.

Our results of effect estimates, as well as results in Ho et al. (2007), show that different conclusions can be obtained depending on the PS technique used. The ATT effect estimate from the one PS technique that meets our balance criterion is not statistically significant. Six of the remaining eight ATT PS techniques yield similar results, but the remaining two yield marginally statistically significant results leading to a qualitatively different conclusion (Table 2; confidence intervals that do not contain the null). In this empirical example, there is no clear relationship between the number of covariates with standardized bias > 0.25 or maximum standardized bias and the magnitude of the final treatment effect estimate (Table 2). One thing to keep in mind is that unbalanced covariates can induce varying amounts of bias that we are not able to measure, and each PS technique has different covariates that it leaves unbalanced to varying degrees. Unfortunately, in this empirical example, we can not directly measure the influence of these combinations of unbalanced covariates or test our assumption that imbalance is associated with bias of the treatment effect estimate. However, we caution researchers from drawing conclusions from the results of PS techniques when measured covariates remain unbalanced.

By presenting results from this one empirical example, we are not suggesting that a single PS technique should be used over all other techniques all of the time, or that these are the 15 PS methods researchers should choose from, regardless of the dataset and estimand of interest. We present a decision criterion to be used during the "design" stage to help applied researchers

explore their own data when applying whichever PS techniques they choose, relying on the assumption that improving balance of measured covariates reduces overt bias in our treatment effect estimate. We choose the PS technique that balances the measured covariates in our empirical example, and we report the resulting treatment effect estimate from our “analysis” stage in this manuscript and elsewhere (Harder et al., 2008).

In conclusion, propensity scores have become an important statistical tool, making possible a more thorough evaluation of and correction for measured covariate imbalances between treated and comparison groups in observational studies. Many PS techniques exist, and more variations are expected, making descriptive manuscripts such as this necessary for applied researchers. Neither the outcome values nor the size or statistical significance of the treatment effect estimate from the final regression models should be brought into play when choosing among PS techniques; the method we outline here is one that gives researchers a way to select a PS technique for their particular dataset and research question. The choice of an optimal PS technique can be expected to vary depending on the dataset, the research question, and the desired generalizability of results. In this context, researchers are well advised to compare and contrast the measured covariate balance achieved by several combinations of PS estimation and PS application methods. We expect decision criteria such as that described here to have continued utility as more and more PS estimation and application methods are developed.

Acknowledgments

We would like to thank Drs. Andrew Morral, Dan McCaffrey, and Greg Ridgeway from The RAND Corporation for their valuable guidance while working with their generalized boosted modeling estimation of the propensity score and for their support while using their ‘twang’ package in R. In addition, we thank Dr. Constantine Frangakis for his expertise and support in applying subclassification methods. We thank Dr. Nicholas Ialongo for the use of data from the NIMH-sponsored project titled “Development and Malleability from Childhood to Adulthood” (Grant R01MH42968-11A1), and we also must thank the many members of the Johns Hopkins University research teams, the school partners, and the community participants whose efforts produced the data analyzed for the empirical example presented in this manuscript.

SUPPORT

This research and resulting manuscript were supported by an Individual Ruth L. Kirschstein (F31) National Research Service Award, DA021956 (PI: Harder), National Institute of Drug Abuse (NIDA). Dr. Stuart’s time was supported by the Center for Prevention and Early Intervention, jointly funded by the National Institute of Mental Health (NIMH) and NIDA (Grant MH066247, PI: Ialongo), by the Centers for Disease Control and Prevention-funded Center for the Prevention of Youth Violence (Grant 5U49CE000728, PI: Leaf), and by an NIMH K-award (1K25MH083846). Dr. Anthony’s time was supported by his K05 Senior Scientist Award from NIDA (DA015799); he also was a Co-Director of the original Prevention Research Center and served as Principal Investigator for the NIDA research grant awards that supported assessment of cannabis involvement, mental disorders, and other constructs between 1989 and 2002 (R01DA004392 & R01DA009897).

REFERENCES

- Allison, PD. Missing Data. Sage; Thousand Oaks, CA: 2002. Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-136
- American Psychiatric Association. Diagnostic and statistical manual of mental disorders: DSM-IV. 4th ed.. Washington, DC.: 1994.
- Breiman, L.; Friedman, JH.; Olshen, RA.; Stone, CJ. Classification and regression trees. Wadsworth and Brooks/Cole; Pacific Grove, CA: 1984.
- Casella, G.; Berger, RL. Statistical Inference. 2nd ed.. Duxbury, Thomas Learning, Inc.; Pacific Grove, CA: 2002.
- Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* 1968;24:295–313. doi:10.1037/0012-1649.44.2.395. [PubMed: 5683871]
- Cochran WG, Rubin DB. Controlling bias in observational studies: A review. *Sankhya, Ser. A* 1973;35:417–446.

- D'Agostino RB Jr, Lang W, Walkup M, Morgan T. Examining the impact of missing data on propensity score estimation in determining the effectiveness of self-monitoring of blood glucose (SMBG). *Health Services & Outcomes Research Methodology* 2001;2:291–315.
- D'Agostino RB Jr, Rubin DB. Estimating and using propensity scores with partially missing data. *Journal of the American Statistical Association* 2000;95:749–759. doi:10.2307/2669455.
- Dehejia, RH.; Wahba, S. Propensity score matching methods for non-experimental causal studies. NBER working paper No. 6829. 1998. Retrieved from <http://www.nber.org/papers/w6829>
- Dehejia RH, Wahba S. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association* 1999;94:1053–1062. doi: 10.2307/2669919.
- Diamond, A.; Sekhon, JS. Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. Institute of Governmental Studies. 2006. Paper WP2006-35. Retrieved from <http://repositories.cdlib.org/igs/WP2006-35>
- Friedman, JH. Greedy function approximation: A gradient boosting machine; *Annals of Statistics*. 2001. p. 1189–1232. Retrieved from <http://www.jstor.org/stable/2699986>
- Graham JW. Missing data analysis: making it work in the real world. *Annual Review of Psychology* 2009;60:549–576. doi:10.1146/annurev.psych.58.110405.085530.
- Gu X, Rosenbaum PR. Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics* 1993;2:405–420. doi: 10.2307/1390693.
- Hansen BB. Full Matching in an Observational Study of Coaching for the SAT. *Journal of the American Statistical Association* 2004;99:609–618. doi:10.1198/016214504000000647.
- Hansen, BB. R package version 0.3-4. R Foundation for Statistical Computing; Vienna, Austria: 2009. Optmatch: Functions for optimal matching. Retrieved from <http://cran.r-project.org/web/packages/optmatch/index.html>
- Hansen BB, Bowers J. Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science* 2008;23:219–235.
- Hansen, BB.; Kleyman, Y. Deconfounding small quasi-experiments using propensity scores and other dimension reduction techniques; Abstract and presentation at the International Conference on Health Policy Statistics; Philadelphia, PA. January, 2008; 2008. doi:10.1214/08-STS254
- Harder VS, Stuart EA, Anthony JC. Adolescent cannabis problem use and young adult depression: A gender stratified propensity score analysis. *American Journal of Epidemiology* 2008;168:592–601. doi:10.1093/aje/kwn184. [PubMed: 18687663]
- Harrell, FE. Regression modeling strategies. Springer-Verlag New York, Inc.; New York, NY: 2001.
- Haviland A, Nagin DS, Rosenbaum PR. Combining propensity score matching and group-based trajectory analysis in an observational study. *Psychological Methods* 2007;12:247–267. doi: 10.1037/1082-989X.12.3.247. [PubMed: 17784793]
- Heckman JJ, Ichimura H, Todd PE. Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme. *The Review of Economic Studies* 1997;64:605–654. doi: 10.2307/2971733.
- Heckman JJ, Ichimura H, Todd PE. Matching as an Econometric Evaluation Estimator. *The Review of Economic Studies* 1998;65:261–294. doi:10.1111/1467-937X.00044.
- Hirano K, Imbens G, Ridder G. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 2003;71:1161–1189. doi:10.1111/1468-0262.00442.
- Ho D, Imai K, King G, Stuart EA. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* 2007a;15:199–236. doi:10.1093/pan/ mpl013.
- Ho, D.; Imai, K.; King, G.; Stuart, EA. MatchIt: Nonparametric Preprocessing for Parametric Casual Inference. R package version 2.2-13. 2007b. Retrieved from <http://gking.harvard.edu/matchit>
- Ialongo NS, Koenig-McNaught AL, Wagner BM, Pearson JL, McCreary BK, Poduska J, Kellam S. African American children's reports of depressed mood, hopelessness, and suicidal ideation and later suicide attempts. *Suicide and Life-Threatening Behaviors* 2004;34:395–407. doi:10.1521/suli.34.4.395.53743.

- Imai K, van Dyk DA. Causal inference with general treatment regimes: generalizing the propensity score. *Journal of the American Statistical Association* 2004;99:854–866. doi: 10.1198/016214504000001187.
- Imai K, King G, Stuart EA. Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society* 2008;171:481–502. doi:10.1111/j.1467-985X.2007.00527.x. Series A
- Imbens G. Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics* 2004;86:4–29.
- Kellam SG, Werthamer-Larsson L, Dolan LJ, Brown CH, Mayer LS, Rebok GW, Edelsohn G. Developmental epidemiologically based preventive trials: baseline modeling of early target behaviors and depressive symptoms. *American Journal of Community Psychology* 1991;19:563–584. doi: 10.1007/BF00937992. [PubMed: 1755436]
- Kellam, SG.; Rebok, GW. Building developmental and etiological theory through epidemiologically based preventive intervention trials. In: McCord, J.; Tremblay, RE., editors. *Preventing Antisocial Behavior: Interventions from Birth through Adolescence*. Guilford Press; New York, NY: 1992. p. 162-195.
- Kellam SG, Anthony JC. Targeting early antecedents to prevent tobacco smoking: findings from an epidemiologically based randomized field trial. *American Journal of Public Health* 1998;88:1490–1495. doi:10.2105/AJPH.88.10.1490. [PubMed: 9772850]
- Kurth T, Walker AM, Glynn RJ, Chan KA, Gaziano JM, Berger K, Robins JM. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *American Journal of Epidemiology* 2006;163:262–270. doi: 10.1093/aje/kwj047. [PubMed: 16371515]
- Lipsey MW, Cordray DS. Evaluation methods for social intervention. *Annual Review of Psychology* 2000;51:345–375. doi:10.1146/annurev.psych.51.1.345.
- Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine* 2004;23:2937–2960. doi:10.1002/sim.1903. [PubMed: 15351954]
- Madigan D, Ridgeway G. Discussion of ‘least angle regression’ by Efron et al. *Annals of Statistics* 2004;32:465–469.
- McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods* 2004;9:403–425. doi: 10.1037/1082-989X.9.4.403. [PubMed: 15598095]
- Morgan, SL.; Winship, C. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press; New York, NY: 2007.
- Potter FJ. A Study of Procedures to Identify and Trim Extreme Sampling Weights. *Proceedings of the Section on Survey Research Methods of the American Statistical Association* 1990:225–230.
- Potter FJ. The effect of weight trimming on nonlinear survey estimates. *Proceedings of the Section on Survey Research Methods of the American Statistical Association* 1993:758–763.
- Qu Y, Lipkovich H. Propensity score estimation with missing values using a multiple imputation missingness pattern (MIMP) approach. *Statistics in Medicine* 2009;28:1402–1414. doi:10.1002/sim.3549. [PubMed: 19222021]
- R Development Core Team. *R Foundation for Statistical Computing*. Vienna, Austria: 2007. R: A language and environment for statistical computing. Retrieved from <http://www.cran.r-project.org>
- Reed PL, Storr CL, Anthony JC. Drug dependence environments: job strain in the work environment and risk of becoming drug-dependent. *American Journal of Epidemiology* 2002;163:404–411. doi: 10.1093/aje/kwj064. [PubMed: 16421241]
- Reed PL, Anthony JC, Breslau N. Incidence of drug problems in young adults exposed to trauma and posttraumatic stress disorder: do early life experiences and predispositions matter? *Archives of General Psychiatry* 2007;64:1435–1442. doi:10.1001/archpsyc.64.12.1435. [PubMed: 18056552]
- Ridgeway, G.; McCaffrey, D.; Morral, A. R package version 1.0–1. R Foundation for Statistical Computing; Vienna, Austria: 2006. *Twang: Toolkit for Weighting and Analysis of Nonequivalent Groups*. Retrieved from <http://cran.r-project.org/web/packages/twang/index.html>
- Ridgeway G. The state of boosting. *Computing Science and Statistics* 1999;31:172–181.

- Ridgeway, G. GBM 1.6-3 package manual. R Foundation for Statistical Computing; Vienna, Austria: 2007. Retrieved from <http://cran.r-project.org>
- Robins, JM. Marginal structural models; 1997 Proceedings of the Section on Bayesian Statistical Science; Alexandria, VA: American Statistical Association. 1998; p. 1-10.
- Robins, JM. Marginal structural models versus structural nested models as tools for causal inference. In: Halloran, E.; Berry, D., editors. *Statistical Models in Epidemiology: The Environment and Clinical Trials*. Springer-Verlag; New York: 1999. p. 95-134.
- Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000;11:550–560. doi:10.1097/00001648-200009000-00011. [PubMed: 10955408]
- Robins JM, Rotnitzky A, Bickel PJ, Kwon J. Comment on “Inference for semiparametric models: some questions and an answer,”. *Statistica Sinica* 2001;11:920–936.
- Rosenbaum PR. Sensitivity Analysis for Matched Case-Control Studies. *Biometrics* 1991;47:87–100. doi:10.2307/2532498. [PubMed: 2049516]
- Rosenbaum, PR. *Observational Studies*. 2nd edition. Springer-Verlag; New York, NY: 2002.
- Rosenbaum, PR. *Design of Observational Studies*. Springer; New York, NY: 2009.
- Rosenbaum PR, Rubin DB. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* 1983a;70:41–55. doi:10.1093/biomet/70.1.41.
- Rosenbaum PR, Rubin DB. Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome. *Journal of the Royal Statistical Society. Series B (Methodological)* 1983b;45:212–218.
- Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 1984;79:516–524. doi: 10.2307/2288398.
- Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* 1985;39:33–38. doi: 10.2307/2683903.
- Rubin DB. Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services & Outcomes Research Methodology* 2001;2:169–188.
- Rubin DB. On principles for modeling propensity scores in medical research. *Pharmacoepidemiology and Drug Safety* 2004;13:855–857. doi:10.1002/pds.968. [PubMed: 15386710]
- Rubin DB. For objective causal inference, design trumps analysis. *The Annals of Applied Statistics* 2008;2:808–840. doi:10.1214/08-AOAS187.
- Rubin DB, Thomas N. Matching using estimated propensity scores: relating theory to practice. *Biometrics* 1996;52:249–264. doi:10.2307/2533160. [PubMed: 8934595]
- Rubin DB, Thomas N. Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association* 2000;95:573–585. doi:10.2307/2669400.
- SAS. Release 9. SAS Institute Inc.; Cary, NC.: 2004.
- Schafer JL, Kang J. Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychological Methods* 2008;13:279–313. doi:10.1037/a0014268. [PubMed: 19071996]
- Shadish WR, Clark MH, Steiner PM. Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random and Nonrandom Assignments. *Journal of the American Statistical Association* 2008;103:1334–1343. doi:10.1198/016214508000000733.
- Smith H. Matching with multiple controls to estimate treatment effects in observational studies. *Sociological Methodology* 1997;27:325–353. doi:10.1111/1467-9531.271030.
- Snedecor, GW.; Cochran, WG. *Statistical methods*. 7th ed.. Iowa State University Press; Ames: 1980.
- Song J, Belin TR, Lee MB, Gao X, Rotheram-Borus MJ. Handling baseline differences and missing items in a longitudinal study of HIV risk among runaway youths. *Health Services & Outcomes Research Methodology* 2001;2:317–329.
- StataCorp. *Stata Statistical Software: Release 9*. StataCorp LP; College Station, TX: 2005.
- Stuart, EA. *Statistical Science. Matching methods for causal inference: A review and a look forward*. in press

- Stuart, EA.; Rubin, DB. Best practices in quasi-experimental designs: Matching methods for causal inference. In: Osborne, J., editor. Best Practices in Quantitative Social Science. Sage Publications; Thousand Oaks, CA: 2007. p. 155-176. Chapter 11
- Stuart EA, Green KM. Using Full Matching to Estimate Causal Effects in Non-Experimental Studies: Examining the Relationship between Adolescent Marijuana Use and Adult Outcomes. *Developmental Psychology* 2008;44:395–406. doi:10.1037/0012-1649.44.2.395. [PubMed: 18331131]
- World Health Organization. Composite International Diagnostic Interview. World Health Organization; Geneva, Switzerland: 1990.

Figure 1.

Propensity score distribution and overlap produced by the three estimation models, stratified by treatment status. Model 1 is the parametric multivariable logistic regression (MLR) estimation of the propensity score, Model 2 is MLR with product-terms, and Model 3 is non-parametric Generalized Boosted Modeling (GBM). (Treatment Status: 0=comparison group, 1=treated group)

Figure 2.

Comparing the distributions of the standardized biases from all 15 models combining the three estimation and the five application methods. (M1=multivariable logistic regression, M2=multivariable logistic regression with product-terms, M3=non-parametric generalized boosted modeling, 1:1=nearest neighbor one to one matching, F=full matching, O=weighting by the odds, S=subclassification, W=inverse probability of treatment weighting).

Table 1
Selected measured covariate standardized biases between treatment and comparison groups
pre- and post- propensity score (PS) adjustment

	Standardized Bias [*] Pre-PS adjustment	Standardized Bias Post-PS adjustment ^{**}
Race		
white	0.27	0.05
black	0.29	0.05
other	0.09	0.03
Gender		
male	0.63	0.004
female	0.63	0.004
Family income		
low	0.01	0.09
middle	0.10	0.09
high	0.06	0.03
missing	0.04	0.06
Alcohol problem use		
no	0.85	0.11
yes	0.80	0.14
missing	0.13	0.06
Other illegal drug use		
no	0.18	0.03
yes	0.16	0.07
missing	0.08	0.04
Daily cigarette user		
no	1.15	0.08
yes	1.16	0.08
missing	0.11	0.01
Free lunch		
no	0.09	0.15
yes	0.09	0.13
missing	0.02	0.05
Parent monitoring		
none	0.02	0.08
low	0.02	0.09
moderate	0.02	0.08
high	0.06	0.14
missing	0.04	0.10
Anxiety symptoms		
low	0.06	0.11
moderate	0.16	0.01
high	0.01	0.08

	Standardized Bias Pre-PS adjustment *	Standardized Bias Post-PS adjustment **
missing	0.12	0.06
Depression symptoms		
low	0.03	0.00
moderate	0.12	0.02
high	0.06	0.07
missing	0.12	0.06
Behavior problems		
none	0.46	0.02
low	0.05	0.02
moderate	0.01	0.10
high	0.21	0.06
higher	0.14	0.08
missing	0.09	0.08
Concentration problems		
none	0.21	0.08
lower	0.16	0.12
low	0.07	0.05
moderate	0.04	0.05
high	0.18	0.10
higher	0.11	0.10
missing	0.09	0.08
Shyness		
none	0.11	0.01
low	0.04	0.11
moderate	0.02	0.06
high	0.03	0.05
higher	0.04	0.002
missing	0.09	0.08
Intervention in class		
none	0.04	0.08
good behavior game	0.04	0.11
mastery learning	0.01	0.01
Intervention in school		
none	0.01	0.10
good behavior game	0.00	0.21
mastery learning	0.01	0.11

* Standardized Bias = Covariate Effect Size = Difference in mean covariate value divided by the standard deviation of treated group

** Resulting standardized bias after the chosen average treatment on the treated PS technique (generalized boosted modeling estimation paired with weighing by the odds application)

Table 2
Information on measured covariate balance and treatment effect estimates for the Average Treatment on the Treated Propensity Score Techniques

Propensity Score Technique (estimation , application)	Number of covariates with standardized bias < 0.25	Maximum standardized bias	Name of covariate with maximum standardized bias	Treatment Effect Estimate	95% Confidence Interval	
					Low	High
Model 1 1 , 1:1	2	0.50	alcohol problems	1.17	0.69	1.99
Model 2 , 1:1	2	0.55	alcohol problems	1.54	0.87	2.74
Model 3 , 1:1	2	0.53	alcohol problems	1.34	0.78	2.31
Model 1 , Full	3	0.32	free lunch	1.43	0.90	2.26
Model 2 , Full	10	0.50	parental monitoring	1.70	1.07	2.70
Model 3 , Full	2	0.39	free lunch	1.79	1.14	2.82
Model 1 , Odds	3	0.30	free lunch	1.14	0.66	1.95
Model 2 , Odds	7	0.32	free lunch	1.12	0.63	1.99
Model 3 , Odds	0	0.21	intervention in school	1.33	0.76	2.33

Estimation Methods = Model 1: Multivariable Logistic Regression, Model 2: Multivariable Logistic Regression with product-terms, Model 3: Generalized Boosted Modeling Average Treatment on the Treated (ATT) application methods: 1:1 = nearest neighbor one to one matching, Full = full matching, Odds = weighting by the odds