Challenges With Propensity Score Strategies in a High-Dimensional Setting and a Potential Alternative

Article *in* Multivariate Behavioral Research · May 2011

DOI: 10.1080/00273171.2011.570161

CITATIONS
17

READS
306

3 authors, including:

Jennifer Lynn Hill
New York University
69 PUBLICATIONS **10,036** CITATIONS

SEE PROFILE

Fuhua Zhai
Fordham University
40 PUBLICATIONS **1,146** CITATIONS

SEE PROFILE

Multivariate Behavioral Research

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hmbr20

Challenges With Propensity Score Strategies in a High-Dimensional Setting and a Potential Alternative

Jennifer Hill [a], Christopher Weiss [b] & Fuhua Zhai [c]

[a] New York University

[b] Columbia University

[c] State University of New York, Stony Brook

Available online: 09 Jun 2011

To cite this article: Jennifer Hill, Christopher Weiss & Fuhua Zhai (2011): Challenges With Propensity Score Strategies in a High-Dimensional Setting and a Potential Alternative, Multivariate Behavioral Research, 46:3, 477-513

To link to this article: http://dx.doi.org/10.1080/00273171.2011.570161

Multivariate Behavioral Research, 46:477–513, 2011

ISSN: 0027-3171 print/1532-7906 online
DOI: 10.1080/00273171.2011.570161

Psychology Press
Taylor & Francis Group

Challenges With Propensity Score Strategies in a High-Dimensional Setting and a Potential Alternative

Jennifer Hill
New York University

Christopher Weiss
Columbia University

Fuhua Zhai
State University of New York
Stony Brook

This article explores some of the challenges that arise when trying to implement propensity score strategies to answer a causal question using data with a large number of covariates. We discuss choices in propensity score estimation strategies, matching and weighting implementation strategies, balance diagnostics, and final analysis models. We demonstrate the wide range of estimates that can result from different combinations of these choices. Finally, an alternative estimation strategy is presented that may have benefits in terms of simplicity and reliability. These issues are explored in the context of an empirical example that uses data from the Early Childhood Longitudinal Study, Kindergarten Cohort to investigate the potential effect of grade retention after the 1st-grade year on subsequent cognitive outcomes.

The benefits of propensity score strategies for causal inference have been explored in many articles (for more basic introductions and reviews see, e.g., D'Agostino, 1998; Gelman & Hill, 2007; Rosenbaum, 2009; Stuart, 2010) including many of the other articles in this special issue. Less frequently discussed are some of the challenges involved in a thoughtful implementation

Correspondence concerning this article should be addressed to Jennifer Hill, New York University, The Steinhardt School of Culture, Education, and Human Development, 246 Greene Street, New York, NY 10003. E-mail: jennifer.hill@nyu.edu

of a propensity score approach. This article grew out of an applied research experience encountered by us as we attempted to use propensity score strategies to answer a causal question of interest in education policy. This example is used throughout the article to illustrate some of the issues that may be encountered when using propensity scores for causal inference, particularly in settings with a large number of potential confounding covariates.

The motivating research question was the effect on future test scores of holding children back in school (i.e., retaining children in grade) immediately after their first-grade year. Specifically, we were interested in estimating this effect for those children who were in fact retained in grade at this time (the so-called effect of the treatment on the treated). Children who are retained tend to be quite different on average from children who are promoted to the next grade, not only in terms of the cognitive skills they display on test scores but additionally with regard to a number of other characteristics. Therefore a necessity arises to condition on quite a large number of covariates in order to justify the ignorability (selection on observables, all confounders measured) assumption often invoked when making causal inferences in the context of observational studies.[1,2] We describe several complications and roadblocks encountered in attempting to apply standard propensity score matching approaches in this setting: standard models for the propensity score appear to be inadequate; the best choice of how to use the estimated propensity scores (e.g., matching or weighting methods) is unclear; and balance checks with many covariates are complicated, making it difficult to discriminate between strategies and thus treatment effect estimates. An alternative estimation strategy is introduced that has far fewer complications yet may still be reliable.

STATISTICAL NOTATION AND ASSUMPTIONS

To formalize our inferential goal we define notation and assumptions. Let Z indicate retention in grade ($Z = 1$) versus promotion ($Z = 0$). Sometimes we refer to this variable as the "treatment" variable. X denotes our vector of available pretreatment covariates. Y denotes our observed outcomes of interest, which will be student test scores (discussed further in the section on

[1]For another study that looks at the effect of first-grade retention on cognitive outcomes using different data but also conditioning on a relatively large number of covariates (72) see Wu, West, and Hughes (2008a, 2008b).

[2]In theory some sort of natural experiment might be used to answer this question. This has been done by a few other authors (notably Jacob & Lefgren, 2004) with access to different data. Such analyses, however, require use of data to which it is often quite difficult to get access. Moreover they tend to answer much more limited kinds of questions. For instance, the Jacob and Lefgren paper can only reliably make inferences about children close to the threshold for promotion.

data). Crucial for discussing causal inference we need to define corresponding "potential outcomes," $Y(0)$ and $Y(1)$. Potential outcomes represent the test score outcomes that would occur under promotion and retention conditions, respectively; in other words these represent what would have happened in both the "factual" (outcome under observed treatment) and "counterfactual" (outcome under treatment not received) states. Therefore for the ith person we can conceive of an individual-level causal effect defined as the difference between these potential outcomes, $Y_i(0) - Y_i(1)$. However, we can only ever observe one of these outcomes, $Y_i(Z_i = z)$; we do not directly observe the counterfactual outcome, $Y_i(Z_i = (1 - z))$.

A completely randomized experiment has the property $Y(0), Y(1) \perp Z$. In the absence of such an experiment, researchers typically attempt to adjust for all confounding covariates (i.e., variables that predict both treatment assignment and the outcome). The assumption that such adjustments are sufficient for unbiased estimates of the treatment effect can be expressed as "ignorability of the treatment assignment" (also known as selection on observables or all confounders measured), formally $Y(0), Y(1) \perp Z|X$ (Rubin, 1978). A stronger version of this assumption additionally requires the existence of "common support," $0 < \Pr(Z = 1|X) < 1)$, also known as sufficient overlap. Conceptually the idea is that for every region of the covariate space defined by the vector X there needs to be some positive probability that both treated and control units might reside there. In practice researchers might take a stronger stance and require that for every "neighborhood" (variously defined) of covariate space where we observe any units, we must observe both treated and control units if we are to perform inference for any units in that neighborhood. We examine the empirical evidence in support of this condition for these data.

Our goal is to estimate the effect of the treatment on the treated, formally, $E[Y(1) - Y(0)|Z = 1]$ (where $E[\cdot]$ denotes the average over some population or distribution of interest). In social and education policy analysis we are often more interested in this estimand than in the more traditional average treatment effect, $E[Y(1) - Y(0)]$ because the former focuses on effects for the type of person who might actually self-select into the program or policy of interest (or be selected into such a program by a teacher, administrator, or social worker) whereas the latter may average the effect over people who might have little to no possibility of ever being exposed to this program or policy.

THE EMPIRICAL EXAMPLE

Current longitudinal data with detailed information regarding grade retention (particularly the specific grades in which the child was retained), rich information on covariates and other compensatory educational policies related to retention,

and information on later life educational and behavioral outcomes is vital for retention analyses but is not necessarily easy to come by. The Early Childhood Longitudinal Study, Kindergarten cohort (ECLS-K; http://nces.ed.gov/ecls/kindergarten.asp) has these features, with additional information on children's teachers and schools, and it contains data collected for a national population. Moreover, the ECLS-K uses cognitive assessments that, for the most part, are not administered by the school district or as part of a school-based assessment. Therefore, we do not have to worry about effects of teachers who "teach to the test" or seek to influence student performance to enhance their performance review (Jacob & Levitt, 2003). In addition, these performance data do not allow for reclassification of students into special education or other similar categories of students often exempted from testing, as happens increasingly in high-stakes testing environments (Booher-Jennings, 2005).

The ECLS-K is a nationally representative longitudinal panel study of children that started with a group of about 22,000 kindergarten children (from both public and private schools) in 1998–1999. Children are nested within schools that are themselves nested within primary sampling units (counties or groups of counties). This study collects information on an incredibly rich array of individual-, household-, teacher-, and school-level measures. The individual-level measures include children's standardized test scores (the same test for all students regardless of grade); behavioral measures; sociodemographic characteristics; childcare history; attendance and transcript information; teacher assessments of children's cognitive abilities, social skills, and behavioral issues as well as whether the child was retained; whether the child has an Individualized Education Program—indicates special education status—and the disability classification; and the other compensatory programs in which the child has participated (tutoring, small group, English as a Second Language [ESL], special education, counseling) and whether that program took place during or outside of the regular school day. Parent characteristics include income, employment history, race/ethnicity, educational attainment, expectations for their child, and involvement in the school.

An important feature of the ECLS-K is that it also collects information about each student's teachers and school. Teacher-level measures include age, experience, educational attainment, salary, and attitudes toward certain teaching philosophies and practices. School-level measures include public/private designation, some average sociodemographic information regarding the school community, and programs available at the school. One of the strongest criticisms of past research is the failure to adequately control for teacher and school measures; the ECLS-K allows us to address this concern.

For these analyses, we excluded those children who were classified as special education in their kindergarten, first-grade, or third-grade years. We also excluded Native American children. Schools that serve Native Americans primarily or exclusively may be quite different from the typical school in our sample, and

likewise the relationship between kindergarten and first-grade characteristics and subsequent retention may be quite different from the rest of our sample. Small sample sizes of Native American children preclude us from estimating these differences adequately so we chose instead not to make inferences for this population. We excluded children who had missing information of grade or grade retention in first or third grade due to attribution or unresponsiveness in data collection (4,407 children). We perform listwise deletion with respect to our outcome variables (this excludes 5,469 children).[3] After these exclusions our remaining sample has 9,226 students, 233 of whom were retained immediately after their first-grade year.

Further exclusions were made to ameliorate problems with computational stability, as discussed in more detail later. This brought the total sample size to 6,900 students of whom, again, 233 had been retained in grade immediately after their first-grade year.

Variables

Outcomes. We used several outcome variables in our analyses based on students' scores on the ECLS-K standard assessment of math and reading measured in spring 2002 (at the end of their expected third-grade year) and spring 2004 (at the end of their expected fifth-grade year). These were standardized to have a mean value of 100 and standard deviation of 15 across our original sample (therefore the mean and standard deviation may be slightly different in our restricted analysis sample). Because the spring 2002 tests were intended to be taken by third graders (in the sense that children progressing typically through school would have been in third grade in that spring) and the spring 2004 tests were similarly intended to be taken by fifth graders, we refer to them as such (even though the retained students taking the exam will not be in third or fifth grade, respectively). Using these outcome variables leads us to compare students of the same age, rather than students of the same grade, using the same test—a practice that many researchers consider an unfair comparison (see, e.g., Greene & Winters, 2004). Therefore we also tip the scales in the opposite direction by constructing a variable that uses 2002 test scores from children never retained and 2004 test scores (note, on the same test) for children who were retained in first grade; we refer to this as the 3rd/5th grade comparison. This gives the retained children an extra year's advantage over the nonretained children.

[3]If the focus of this article was not methodological we would likely perform multiple imputation rather than listwise deletion to address these substantial missing data issues. We were loathe to muddy the water by introducing further methods that themselves could inspire controversy, however, so we decided instead to keep things simple.

Treatment. The treatment or causing variable of interest in this study is retention in grade immediately after the first-grade year. We use teacher's report of grade retention as well as student's reported grade for each analysis year to construct the variable.

Covariates. The assumption of ignorability is always difficult to justify in the absence of a randomized experiment. However, in the social and behavioral sciences experiments simply are not possible for many questions of interest. Nevertheless, many scholars are justifiably wary of this assumption. Therefore the onus is on the researcher to include as many potentially relevant pretreatment variables as possible to increase the plausibility of this assumption. Fortunately, in this example we had access to an extremely rich set of pretreatment covariates. In these analyses we use 236 pretreatment predictors (where each multicategory variable has been transformed into corresponding binary variables for all but one of the categories) measured in kindergarten and first grade. These variables fall into several broad categories:

- Characteristics of the child, including demographics and behavioral measures;
- Child scores in tests of math, reading, and general knowledge in kindergarten and first grade (though kindergarten scores are used only indirectly as discussed later);
- Characteristics of parents and the family, including race/ethnicity, income and poverty measures, and parental education level;
- Parents' expectations and participation in child's education (at both home and school);
- Teacher and parent assessments of the child's effort and ability levels;
- Teachers' cognitive and behavioral ratings of the child;
- Programs in which the child participates;
- Teacher characteristics; and
- School characteristics.

These variables are described in greater detail in Appendix A.

Complications With the Data

These data are subject to two additional complications that we want to address before pursuing a causal analysis. We describe our strategies for handling these complications for completeness and transparency and to aid anyone who might want to replicate our analyses. These choices may not be optimal (particularly for the missing data); however, they also were not the focus of this methodological inquiry. Given that they are not integral to our discussion regarding the impli-

cations of choice of methodological approach for causal analysis in this setting, we decided to use relatively simple solutions to avoid further complication and controversy.

Missing data. Missing data are endemic to virtually all studies with human participants. For each categorical predictor with missing data we simply added an additional category representing "value missing" (for a helpful discussion, see Jones, 1996). With the exception of kindergarten and first-grade test scores, missing values for the relatively small percentage of continuous variables were addressed by imputing the mean and adding an indicator variable reflecting missingness for each of these variables. Although this is far from an ideal solution, these missing values represent a very small proportion of the data set. Pre-"treatment" test scores were handled in a more principled manner due to their importance in the subsequent analyses. These test scores were (singly) imputed by predicting scores using a regression model (with all other variables as predictors) and adding "noise" in the form of a draw from a normal random variable with mean zero and variance equal to the maximum of the individual-level forecast errors from that regression. Observations missing grade retention data or outcome data were deleted.

Measurement error. It is well known that if a regression model includes a predictor that has been measured with error the coefficient on this variable will be biased toward zero (Klepper & Leamer, 1984). More important for our analysis, however, this measurement error can also induce bias in the coefficients of any variables correlated with this predictor. What are the implications of this for propensity score strategies? It is unclear that measurement error is a problem for propensity score estimation per se. The more obvious problem occurs if we try to perform postmatching (or weighting) adjustments for covariates, through, for example, a linear regression. This is fairly common practice and is widely advocated in the propensity score literature (Rubin, 1973, 1979; Rubin & Thomas, 2000) and can be thought of as an approximation to more formal "double-robust" estimators (Robins & Ritov, 1997).[4] Preretention reading, math, and general knowledge test scores, for example, are critical predictors in our analyses that we would be likely to include in such an analysis given their importance as a predictor of the outcomes. However, these variables are almost surely measured with error and are highly correlated with retention; therefore their inclusion would

[4]Doubly robust estimators gain strength by modeling both the treatment assignment mechanism ($E[Z|X]$) and the response surface ($E[Y|Z, X]$). This class of estimators has the property that if either of these mechanisms is estimated without bias then the overall estimator will have no bias. Propensity score matching followed by covariance adjustment (typically in the form of a linear model) is an informal way of accomplishing the same goal.

bias our treatment effect estimate and the direction of the bias is not known (due to the complicated correlation structures among all the variables in our model). Therefore we attempt to correct for this error. To do so we take advantage of the fact that we have multiple pretreatment test scores for each of reading, math, and general knowledge (from kindergarten and spring of first grade). We use the mean of the kindergarten scores as an instrument for the first-grade score (similar strategies have been used and discussed by Schwartz, Stiefel, & Zabel, in press) in an attempt to purge the first-grade test scores of measurement error.

ANALYSIS STRATEGY AND PROBLEMS ENCOUNTERED

In observational study settings where causal estimates are nonetheless desired, researchers are increasingly making use of propensity scores. The propensity score is defined as the probability of being treated ($Z = 1$), conditional on observed pretreatment covariates: $e(X) = Pr(Z = 1|X_1, \ldots, X_k)$. When ignorability holds, Rosenbaum and Rubin (1983, 1984) have shown that the propensity score can act as a one-number summary of the covariates, such that conditioning on the propensity score is sufficient for conditioning on the full set of covariates. This makes possible certain important simplifications. For instance, if the goal is to estimate the effect of the treatment on the treated, $E[Y(1) - Y(0)|Z = 1]$, the tricky part is capturing $E[Y(0)|Z = 1]$. If ignorability holds we can attack the problem by realizing that $E[Y(0)|Z = 1, X] = E[Y(0)|Z = 0, X] = E[Y|Z = 0, X]$. Intuitively, we can conceptualize this equivalence as saying that for observations with identical X values the observed outcome for the control observation represents what we *would have seen* for a given treated observation had it instead been assigned to the control group. Conditioning on a potentially high-dimensional vector of covariates, X (in order to evaluate $E[Y|Z = 0, X]$) can be nontrivial, however. Therefore the simplification allowed by the propensity score, in this case that $E[Y(0)|Z = 0, X] = E[Y(0)|Z = 0, e(x)] = E[Y|Z = 0, e(x)]$ is, in theory, a crucially important advantage; it has the potential to allow the researcher to control for many covariates without having to specify a parametric model for the outcome conditional on the treatment indicator and all the covariates. Rather we condition on the propensity score by matching on it or by using it to reweight one group to look like another before estimating treatment effects.

How to Estimate the Propensity Score

The propensity score typically is estimated using logistic or probit regression. However, these models are not ideal across all settings. Our setting is com-

plicated by the large number of covariates. In theory, access to substantial information on potential confounders is advantageous for identifying a causal effect; the richer the pretreatment information we can condition on the more we may be willing to believe the ignorability assumption.[5] In practice, however, we can run into real computational issues in this setting. However, standard logistic or probit regression models can fall apart or simply perform poorly due to problems of perfect separation. Even if the model runs (and provides nondegenerate estimates), if it is overfit to the data this can make it difficult to identify comparables. As an illustration, in our first attempt at answering this research question, we attempted to adjust for over 500 covariates. When using standard logistic regression software all propensity scores were estimated at either 1 or 0 (when rounding at the 15th decimal place). Even more sophisticated models such as Bayesian versions of logistic regression (as described later) had the same problems. Therefore we pared down our set of controls to the current set of 236.

Our situation is also somewhat complicated by the low percentage of treated observations relative to controls. Although this is typically preferable to the reverse situation (when estimating the effect of the treatment on the treated) in terms of the availability of potential controls, it can pose its own computational challenges such as increasing the likelihood of problems with separation. With the original data set of 9,226 observations there were still some problems with computational stability. Moreover the huge number of controls forces the majority of the distribution of propensity scores to be in a very small region making it difficult to discriminate between observations. Given that a nonnegligible subset of the control group appears to have substantially more resources and higher test scores than the retained students, it makes sense to discard such students first. We did this by fitting a very limited propensity score model that included only (measurement-error-corrected) math, reading, and general knowledge test scores from first grade plus child age at time of test, sex, language spoken at home, an indicator for health insurance coverage,

[5] Pearl (2010) provides an argument against the common advice to simply control for as many pretreatment covariates as possible. He demonstrates that if one of those covariates is in fact a true instrument, such conditioning can lead to more rather than less bias. Part of the early work on the article in fact was an attempt to find an instrument lurking among our rich set of variables; we are reasonably certain that no such instrument exists. However, even if one does exist (in the sense of satisfying the untestable assumptions of being randomized and satisfying exclusion), then we have at least determined that no *strong* instrument exists. If a researcher does have access to such an instrument then by all means it should be used in an instrumental variables analysis (in addition perhaps to a standard analysis with the instrument excluded from the set of confounding covariates). Finally, we also argue more generally that when controlling for such a huge number of covariates it is likely that whatever conditional independence relationship might have existed, it would more than likely be destroyed by the conditioning that occurs on some subset of the rest of the pretreatment variables.

and region of the country. The logit of the propensity scores were then used to determine a cutoff (less than −7.3) for discarding plausibly irrelevant controls (2,326 were discarded); this is equivalent to discarding comparison observations with estimated propensity scores less than .000675.

Diagnostics for Model Misspecification

How can we tell if the model is reasonable? With so many covariates it is difficult to determine if we have misspecified the functional form of some of the variables. Binned residuals plots for all of the models display lack of fit, but the logistic and probit regression models (both the standard and Bayesian versions) are by far the worst offenders. How should we go about fixing these models? Should we be logging or otherwise transforming some of our continuous variables? Should we be including interactions? With so many variables it is difficult to determine how best to correct each model.

Diagnostics for Overfit

Given the number of confounders in the propensity score model we are still likely to overfit. If our model overfits it will be less useful in finding matches that will yield good balance, even if they exist. Stepwise regression used to be suggested as a strategy to deal with large numbers of covariates in this setting; this has been shown to be problematic as well (see, e.g., Austin, Grootendorst, & Anderson, 2007) because this strategy differentially favors covariates that strongly predict the treatment whereas the researcher should be most concerned with variables that strongly predict the outcome (Pearl, 2010). How can we diagnose a degree of overfit that could be problematic for this setting?

One problem with a model that overfits is that it has the potential to make units from the treatment and control groups appear to be quite different from each other even if they are actually quite comparable with respect to the features that are truly important (i.e., with respect to predicting the outcome). In fact, it is possible that even if the treatment variable were completely unassociated with the covariates at the population level, the empirical distribution of estimated propensity scores for the observed treatment group for a given sample might look different from the empirical distribution of estimated propensity scores for the control group for that sample. This suggests a way to get a handle on this aspect of the model. We can fit the proposed model to a constructed data set with all the real covariate information and with a fake treatment variable simulated with the same marginal rate of success for the true treatment variable but generated completely independently of the covariates. Specifically, the user can create a treatment variable by simulating data from the binomial distribution

with probability of treatment for each observation equal to the same marginal probability in the data (in this example that rate was about .033). If the model fit to these data displays sufficient lack of overlap this could indicate that it is not the best model to use for the propensity score estimation.

We used this diagnostic with our data to compare the behavior of six different models for the propensity score. All computations were performed in the R software package. The first two models were just standard implementations of logistic regression and probit regression.[6] The next two propensity score estimation strategies rely on Bayesian versions of the first two models implemented using the function bayesglm() (Gelman, Jakulin, Pittau, & Su, 2008) using the arm package in R (Gelman et al., 2009). The default versions of these models place Student-t prior distributions on the coefficients in the model that helps constrain them to lie in a reasonable range thus stabilizing the resulting predictions.

The fifth strategy makes use of a proposal by McCaffrey, Ridgeway, and Morral (2004) to use generalized boosted models (GBM) to estimate propensity scores. We used the defaults suggested by the authors for the shrinkage parameter (.0005), interaction depth (4), and bag.fraction (.5) and implemented their functions for optimizing the number of trees with respect to an objective function based on covariate balance. The final propensity score estimation strategy uses an algorithm called Bayesian Additive Regression Trees (BART; Chipman, George, & McCulloch, 2007, 2010) that can be loosely conceived of as a Bayesian version of boosted regression trees where the tuning parameters are replaced by clever choices of prior distributions.[7] We used the probit link version of the model and left the settings for the hyperparameters of the prior distributions at their default settings as recommended by Chipman et al. (2007). These last choices represent just two of a number different choices for flexible estimation that originate in the data mining literature (for some other choices see Setoguchi, Schneeweiss, Brookhart, Glynn, & Cook, 2008).

Figure 1 plots the results of this diagnostic. Each plot in the figure displays overlaid histograms of the propensity scores for the control group (shaded in gray) and the propensity scores for the treatment group (unfilled, outlined in black). We also report for each method the percentage of treated observations that are "off support"—in this case that means the percentage of treated observations

[6]This statement glosses over the fact that computationally the algorithms to fit these models can vary a bit between software packages. For instance, the logistic regression function in Stata (as well as the glm function using a logit link) was not able to fit the specified model, which R fit without incident.

[7]BART has also been proposed for causal inference as a strategy for directly estimating the response surface (Hill, 2011) and indeed it is used that way later in this article. In this section, however, BART is merely being used to estimate the propensity score as part of a broader propensity score strategy.

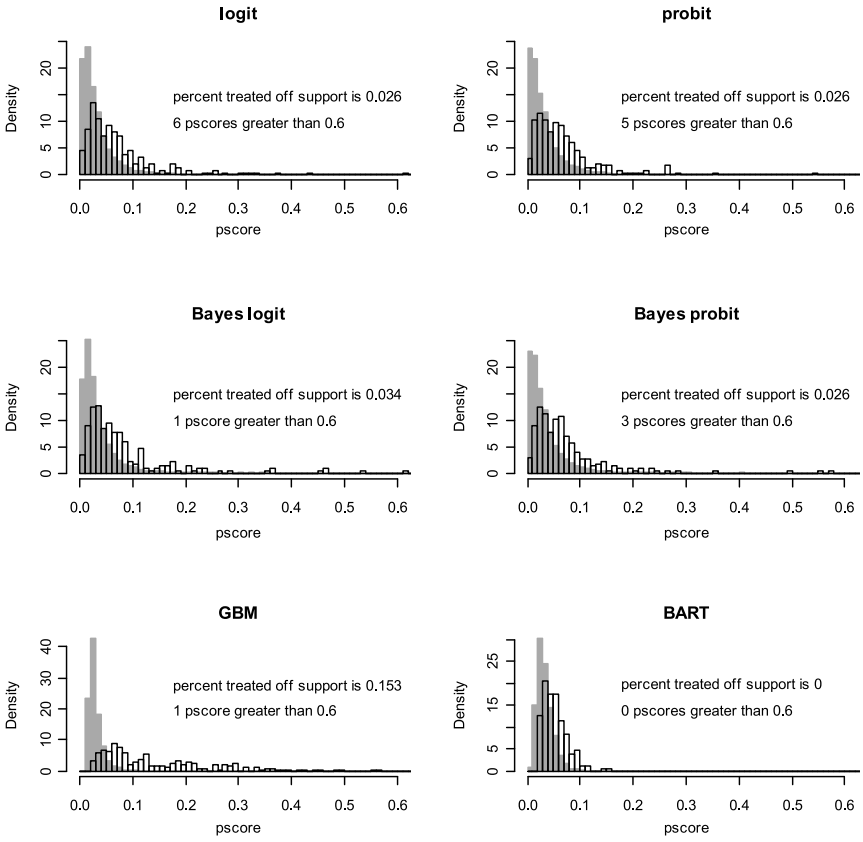

FIGURE 1 Diagnostic histograms to assess the extent of overfit in the propensity score estimation models. Each model was fit to real covariate data and a randomly generated treatment (grade retention) variable with probability of being treated equal to the percentage retained in the real data. The shaded gray histogram displays the empirical distribution of propensity scores for the controls and the hollow histogram with the dark border the same distribution for the treated. Overfit is manifested by wide ranges of propensity scores (because the true probability is the same for all observations) and the lack of overlap between treated and control distributions.

with propensity scores that exceed the maximum propensity score for the control observations. Histograms were plotted on the density scale so that the substantial differential in the number of observations in the promoted (6,667) versus control (233) groups would not distort the plot.

This figure suggests that all of the propensity score estimation models may be overfitting to some degree in the sense that in all plots the treatment and

control propensity score distributions appear to exhibit lack of overlap. There is not much difference between the logit and probit models. Moreover, although the estimates of coefficients and their standard errors are quite different between the Bayesian and standard forms of these models (with standard errors often implausibly large—in the thousands—for the standard models), the differences in their predictions don't appear to be substantial. GBM appears to be the worst offender in the sense that it mistakenly reflects a lack of distributional overlap that we know does not exist in our simulated data. The BART estimates for both treatment groups are also more tightly constrained in general, reflecting less of an overall tendency to overfit. Relatedly, BART appears to reflect the most overlap although the plots may be slightly misleading with respect to this feature given that the plots are all graphed on the same scale for the x-axis (to facilitate comparison); therefore there is less space to distinguish between the BART propensity score distributions. Nevertheless, this figure seems to suggest that BART might be the best choice among these six models, at least with regard to overfit.

Of course there are many modeling choices not even attempted here and such plots might be useful for choosing among an even wider range of models. They also might be used to fine-tune any given model. For instance, the researcher could rerun the GBM procedure with different choices for the tuning parameters to see if this yields more promising results. Further research is necessary to ascertain the more global usefulness of such a diagnostic in this setting. Moreover these results should not be interpreted as a referendum of the usefulness of each of these models beyond the given specification or in data analytic settings outside of this one; for instance, standard logit and probit models are likely to work well in a wide variety of other settings.

We turn now to assessing overlap in our observed data using the diagnostic plots as a guide to allow us to understand the level of overfit present. The plots in Figure 2 have been graphed on the frequency rather than the density scale to help ascertain whether the number of control units appears to be sufficient to act as empirical counterfactuals for the retained units. To avoid the distortions that might be caused by the extreme ratio of promoted to retained, the plot is restricted to observations with propensity scores greater than .2 because there is no problem with finding sufficient numbers of control (promoted) units with similar propensity scores at the lower end of the distribution. Taken on its own, each of these plots would seem to indicate that there is an issue with lack of overlap in that there appear to be treated units at the high end of the propensity score distribution without controls in a close neighborhood. However, given the lack of overlap displayed in the diagnostic plots when we knew such overlap did not in fact exist in reality, it is unclear that there is necessarily a need to restrict our analyses to areas of common support.

Two additional points are worth noting. First, it is troublesome that the assessment of which retained units lack sufficient overlap is so dependent on

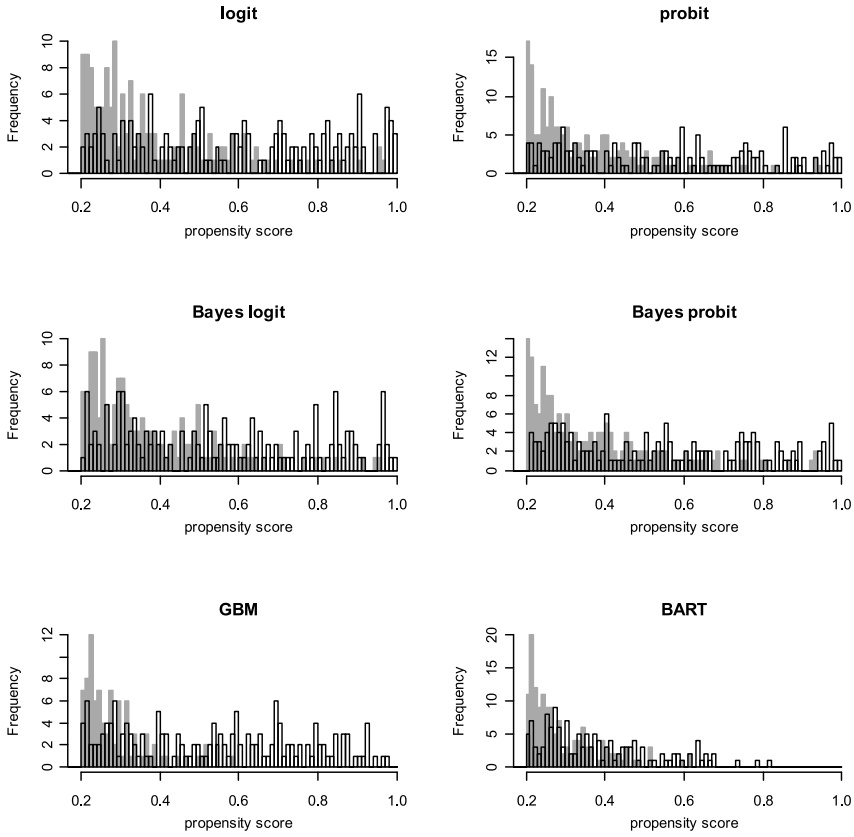

FIGURE 2 Histograms to assess the lack of overlap (common support) across treatment groups. Each plot displays overlaid frequency histograms of propensity scores estimated from a different model. The shaded gray histogram displays the empirical distribution of propensity scores for the controls and the hollow histogram with the dark border the same distribution for the treated. The x-axis is on the logit scale and starts at .2 to facilitate comparisons (we are not concerned with lack of overlap at the lower end of the distribution).

the model used to estimate the propensity score. With a strict definition of how to discard units that ignores our diagnostic plots regarding overfit, GBM would suggest that 71 treated units be discarded whereas the BART propensity score would indicate that 13 treated units should be discarded.

Second, not much guidance exists in the literature about how to identify lack of overlap. Simply excluding treatment units with propensity scores outside the range of propensity scores estimated for the controls, as is commonly implemented in propensity score software packages as the "common support" option,

seems oversimplistic, both because it may be too severe a cutoff and because it ignores the treated units in the middle of the distribution that may also not have a close neighbor. However, it is unclear how far a treated unit's propensity score need be from the closest control before causing concern. Moreover, discarding treatment units is a much more serious choice than discarding control units when the estimand is the effect of the treatment on the treated because this changes the estimand. Yet researchers rarely make the effort to profile either the units dropped or to describe the new inferential population of interest (represented by the remaining treated; for an exception see Mincy, Hill, & Sinkewicz, 2009).

Which Propensity Score Strategy Should I Use?

Suppose that estimating appropriate propensity scores was not an issue. What is the best way to make use of my estimated propensity scores? Propensity scores can be used in many ways including matching, subclassification, and inverse probability of treatment weighting. Each strategy can be adapted to target different estimands (e.g., the effect of the treatment on the treated, the effect of the treatment on the controls, or the average treatment effect). Additional choices remain for each method, however.

For matching there are choices regarding the matching algorithm (for a thorough review see Stuart, 2010) including (but not limited to) whether matching is performed with or without replacement, the number of controls matched to each treated, whether matching is performed within calipers and if so the caliper width and whether kernel methods are used to differentially weight the controls in the caliper or if Mahalanobis matching is performed to choose units within calipers, whether additional exact matching is performed on one or more categorical covariates, and whether the matching algorithm is greedy or optimizes some criterion.

Weighting that makes use of propensity scores (Imbens, 2004; Rosenbaum, 1987) can be performed with stabilized or unstabilized weights and many choices may exist regarding how to stabilize the weights. When using subclassification, decisions need to be made about the number of subclasses used and how the cutoffs between subclasses are made (typically based on quantiles either of the overall propensity score or the propensity score for just the treated or just the controls).

In this article we focus on matching and weighting. We ignore subclassification in this comparative analysis because it creates additional complications in terms of how to combine treatment effects estimates and balance statistics across subclasses.

Matching. We use three different kinds of matching algorithms: a greedy nearest neighbor (N), optimal pair matching (O), and full matching (F; Hansen

and Klopfer, 2006; Rosenbaum, 1991). Greedy nearest neighbor matching orders the treated units (e.g., by the magnitude of the corresponding propensity scores or randomly) and then for each treated unit in turn chooses the comparison group member with the closest propensity score to be the match. Optimal pair matching in contrast chooses the set of matched control units for the treated units that minimizes the total distance across matched pairs. For nearest neighbor and optimal matching we examine matching to one control versus two controls for each treated participant (this is not an option for full matching). For nearest neighbor matching we additionally consider matching with replacement (this is not an option for the other methods). Full matching attempts to use all units and thus partitions the full sample into nonoverlapping stratum such that each contains either one treated and one control unit, one treated unit and multiple controls, or one control and multiple treated units. The goal of the full matching algorithm is to create this partition in such a way that it minimizes the total distance between the units. Unlike the pair matching methods, which are designed to estimate the effect of the treatment on the treated, the full matching algorithm is designed to estimate the average treatment effect across the entire sample.[8] All matching methods and balance checks are implemented using the package MatchIt (Ho, Imai, King, & Stuart, 2010) in R, although MatchIt calls the optmatch program for optimal and full matching (Hansen, 2004).

Weighting. We implement an extension of standard inverse probability of treatment weighting that has been discussed in several other articles (see, e.g., Imbens, 2004; Kurth et al., 2006) that is appropriate for the treatment on the treated estimand. In particular, treated observations receive a weight of 1 whereas control observations receive a weight of $\frac{\hat{e}(X)}{1-\hat{e}(X)}$ (weights are then normalized so that the sum of the weights for the control group equals the observed number of controls); this particular weighting implementation has also been referred to as "weighting by the odds" (Stuart, 2010). A potential problem with weighting occurs when the denominator is very small (in this case when the estimated propensity score is close to 1); then the weights can grow quite large and estimation can become instable. We addressed this potential concern with a common ad hoc solution, which is to truncate the weights. In this case the truncation was set at 30 though the weights were renormalized afterward, which sometimes changed the ceiling a bit (the truncation ceiling is itself a somewhat arbitrary choice that might also affect results). We refer to our weighted estimator

[8]Because our goal in this analysis is to estimate the effect of the treatment on the treated, this distinction might argue for excluding full matching from the strategies tested. On the other hand, the documentation for the popular propensity score matching package used (MatchIt) that implements this algorithm is sufficiently vague on this point that it would be easy for the typical user to misuse it in exactly this way. Also, if treatment effects are additive these two estimands will be equal.

as an Inverse Probability of Treatment Weighting (IPTW) estimator because we think of IPTW as a class of estimators that make use of the propensity score.

Checking Balance

The balance diagnostic is often touted as one of the greatest strengths of the propensity score strategies that incorporate it. This diagnostic provides feedback regarding whether we are successfully creating fair comparisons through our matching, weighting, or subclassification. A particular strength of this diagnostic is that it can be used to choose a propensity score model and matching/weighting strategy before ever calculating a treatment effect estimate. This allows the researcher to assess the adequacy of his propensity score procedure without contaminating his judgment by simultaneously revealing the associated treatment effect estimate (although in practice of course there is nothing to stop the researcher from checking this estimate after each match). Unfortunately, many researchers who have published articles using propensity scores have been lax in performing or at least reporting the results of such checks. Austin (2008) describes such a trend in propensity score matching applications in medical research. Balance diagnostics have not typically been used at all in conjunction with inverse probability of treatment weighing (IPTW), although they can and should be (for an example of where they are used in this context, see Mincy et al., 2009).

A more subtle problem is with the *type* of diagnostics used. Most articles (including a few by the authors of this article) have relied on only very simple balance diagnostics such as comparisons of means rather than comparisons of higher order moments of the full covariate distributions (Austin, 2008). Balancing means of covariate distributions across treatment and control groups would help to reduce bias most in a scenario when the response surface is linear in the covariates. Of course in that case we could get an unbiased estimate of the treatment effect with a simple linear regression. We are using propensity scores presumably because we do not trust that a simple model such as linear regression will accurately capture the response surface and thus we want to rely on a semiparametric approach. To be most effective, then, propensity score adjustments should balance the *functions* of the confounding covariates that are most predictive of the outcome. Thus in order to optimize our balance check to highlight the precise moments of the joint covariate distribution that are most important for predicting the outcome, we would need to know the exact form of the response surface ($E[Y(0), Y(1)|Z, X]$). However, if we knew this we wouldn't have to preprocess (match, weight, subclassify) with propensity scores to begin with; we could just fit the right model!

This reasoning doesn't bring us much closer to best practice in balance checking. It does suggest, however, a broader strategy than simply balancing

means, which is balancing multiple moments of the joint covariate distributions. For instance, a more comprehensive check might include comparisons of the entire distribution of each ordinal/continuous covariate. However, if we have many such covariates it may be difficult to process all this information. Therefore empirical summaries of such comparisons may be useful such as statistics that capture key features of the differences between empirical QQ plots of the covariate distributions across treatment and control groups (see Ho et al., in press; Sekhon, 2010a).[9] Moreover, this broader strategy also implies that we should be checking whether associations between variables are the same across treatment groups (important if interactions between these variables play an important role in the response surface). A few current software packages (Ho et al., in press; Sekhon, 2010b) offer more comprehensive options: graphical comparisons of distributions, comparisons of the variance or standard deviations of distributions, reports of differences in empirical quantiles of the marginal distributions, and comparisons of correlations among confounders. Nevertheless, these diagnostics have yet to become the norm.

The deeper issue, however, is that even if all researchers were to avail themselves of these options, the guidelines regarding what constitutes adequate balance are by no means clear. Short of being able to exactly balance the joint covariates distributions of all variables across treatment groups, how close must any of these statistics be across groups for the researcher to feel reassured? Most scholars reject the idea that statistical tests are useful in this setting (for a discussion, see Austin, 2008; for dissenting ideas, see Hansen, 2008). Sekhon (2010a) offers perhaps the most useful advice, which is, conditional on a given balance metric, to simply choose the matches that get you as close as possible.[10] However, without having a clearly defined balance optimization rule (and an algorithm to implement it) there are almost always still trade-offs to be made across variables in terms of achieving such "close as possible" balance; improving balance in one variable will come at the cost of decreased balance for another variable. How do we make decisions regarding which (functions of) variables are most important to balance? Several authors have noted that it is

[9]Martens (2007) proposes alternative summaries of this information, which could potentially be superior. We focus on the empirical QQ metrics in this article because they are available in existing software and thus are more likely to be used in current practice.

[10]Diamond and Sekhon (2008) have created a "genetic matching" algorithm that can be implemented in the Matching package in R (Sekhon, 2011) that performs this optimization. Given that the algorithm is not technically a propensity score matching approach (it matches using all the covariate data though typically performance is improved by also including the propensity score) and the fact that it can be quite computationally intensive (e.g., it was too big to run on a standard PC with these data), we did not include it among our set of typical propensity score methods attempted. It is a competitor to these methods that should be considered in smaller scale problems which has many of same advantages as the BART algorithm espoused later in the article in terms of simplicity.

most important to balance the variables that are most predictive of our outcome variable (Pearl, 2010), but this in itself may not be an obvious ranking.

Balance metrics used. Balance across treatment groups was calculated using the summary command in MatchIt. We used this function to calculate standardized difference in means for all covariates, the median and maximum difference between empirical QQ plots for all continuous covariates, and the standardized difference in means for the set of all covariates as well as their interactions and squared terms (there were a total of 28,440 of these terms[11]). With 236 covariates it is overwhelming to consider separately evaluating the each of these balance measures for each covariate (even more so when we consider the additional interaction terms). Therefore we created summary measures for each statistic across variables as follows: the mean and maximum of the absolute value of standardized difference in means (std.mn and std.max), the number of absolute value of standardized difference in means that exceeded .1 (std.over.1), the maximum of the median difference in empirical QQ plots (medQQ.max), the mean of the maximum difference in empirical QQ plots (maxQQ.mean), and the number of maximum differences in QQ plots that exceeds .1 (maxQQ.over.1).

Using balance metrics to discriminate between approaches. How do we decide upon the right combination of propensity score estimation model and matching/weighting method? One possibility is to try a range of combinations and then to assess superiority using multiple balance checks. In this example we assess all 54 combinations of our six propensity score estimation methods and nine propensity score strategies for each of nine balance metrics.

Figure 3 plots each of these balance summaries for all 54 propensity score strategies (combinations of propensity score model and matching/weighting method). The x-axis organizes the strategies by matching/weighting method and the balance statistic corresponding to each propensity score estimation strategy for that method is printed using a different symbol as described in the caption.

There is a substantial amount of heterogeneity in the performance of these methods with respect to these balance metrics. A few broad trends can be ascertained. First the inverse probability of treatment weighting methods appears to fall apart, no matter the propensity score estimation strategy used, when judged with respect to the empirical QQ plot balance statistics. This suggests that in this setting these methods were not able to balance well aspects of the distributions for the continuous variables beyond simply the mean. Another fairly

[11] Unfortunately this huge number of terms reflects a substantial number of redundancies because the way that this option is implemented does not discriminate between categorical and continuous variables. So, for instance, the squared term for each variable is included even though for binary variables this is equivalent to the original term.

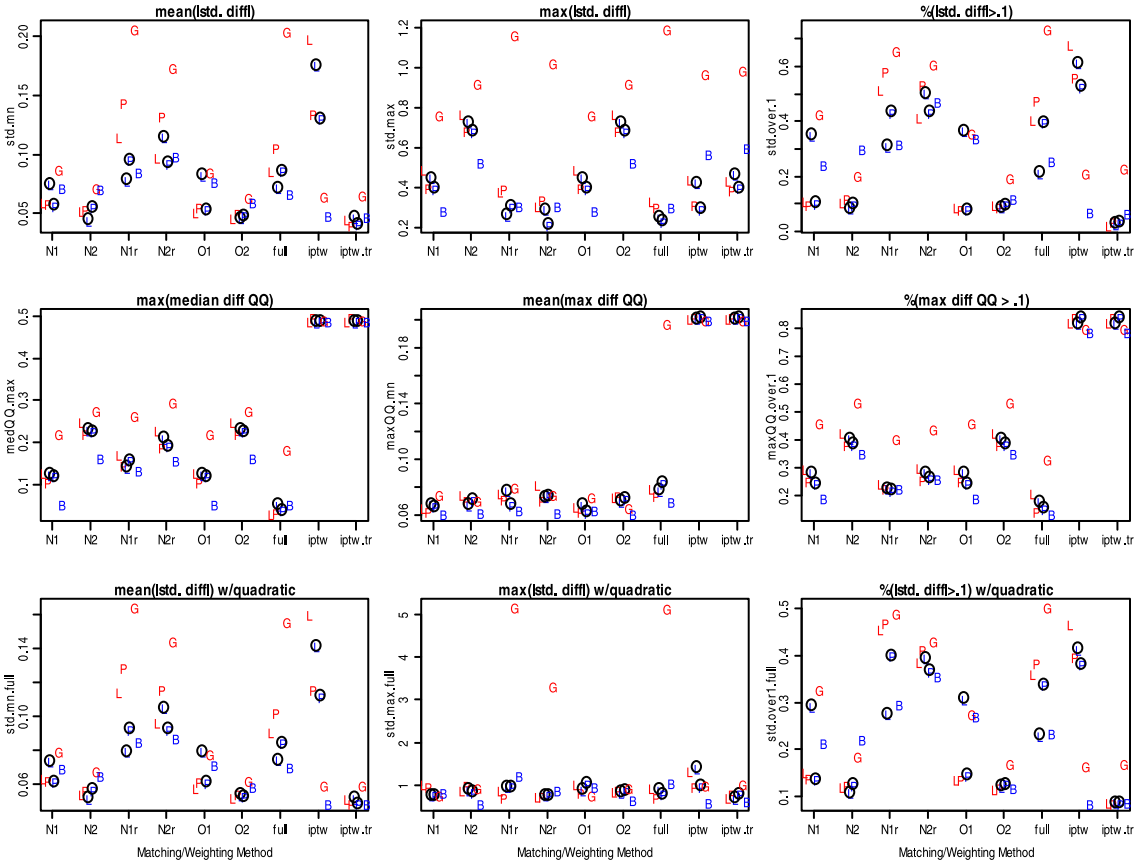

FIGURE 3 (See caption on page 497.)

consistent feature of these plots is the poor performance demonstrated, across methods, when using GBM to estimate propensity scores. The overfit signaled in our diagnostic plots seems to be translating into worse balance for approaches relying on the GBM propensity scores (at least with regard to these metrics) regardless of matching or weighting method.

The lack of fit displayed by these methods and estimation strategies makes it more difficult to differentiate between the other methods so we re-created the balance plot excluding all strategies that use the IPTW methods or the GBM estimation strategy. In Figure 4 it is easier to discriminate among the remaining methods; however, it is now more difficult to identify clear trends. A few patterns exist, however. For instance, there is no clear dominance between the standard and Bayesian versions of the logistic and probit regression models. Moreover, propensity scores estimated using BART tend to outperform other estimates for any given matching method with regard to the QQ balance statistics. However, they are more in the mix for the average (absolute) standardized mean and percent standardized means exceeding .1 statistics.

Treatment Effect Estimates

After creating matches or estimating weights, the researcher uses these to actually estimate the treatment effect. Our goal in this setting is to estimate the effect of the treatment on the treated.[12] At this stage of the analysis still more choices need to be made. In this section we demonstrate the impact of these choices on treatment effect estimate for the third-grade reading score outcome. If the matching or weighting has sufficiently balanced the covariate distributions between the treatment and control groups (and ignorability holds), a simple difference in means estimate is sufficient for unbiased estimate of the

[12]Full matching is actually geared toward estimating the average treatment effect across the entire sample, not just the average effect for the treated. Therefore differences between estimates from this method and the others are a bit more complicated to interpret.

FIGURE 3 (See artwork on page 496.) Balance plots. Each of the nine plots displays one balance summary calculated for the each 54 propensity score strategies. Each tick mark on the x-axis designates a matching or weighting method (N1, N2, N1r, N2r for nearest neighbor matching with one or two controls for each treated and without replacement or with (r); O1 and O2 for optimal matching with one matched control or two; full for full matching; iptw and iptw.tr for inverse probability of treatment weighting without or with truncation). The points above the tick marks display the (slightly jittered) balance results for the six different propensity score estimation models used for each method. L and P denote logistic and probit regression models, respectively; circled L and P denote the Bayesian versions of these models; G denotes GBM; and B denotes BART. The bottom three plots are summaries over all 236 covariates as well as the 28,440 corresponding quadratic terms. (color figure available online)

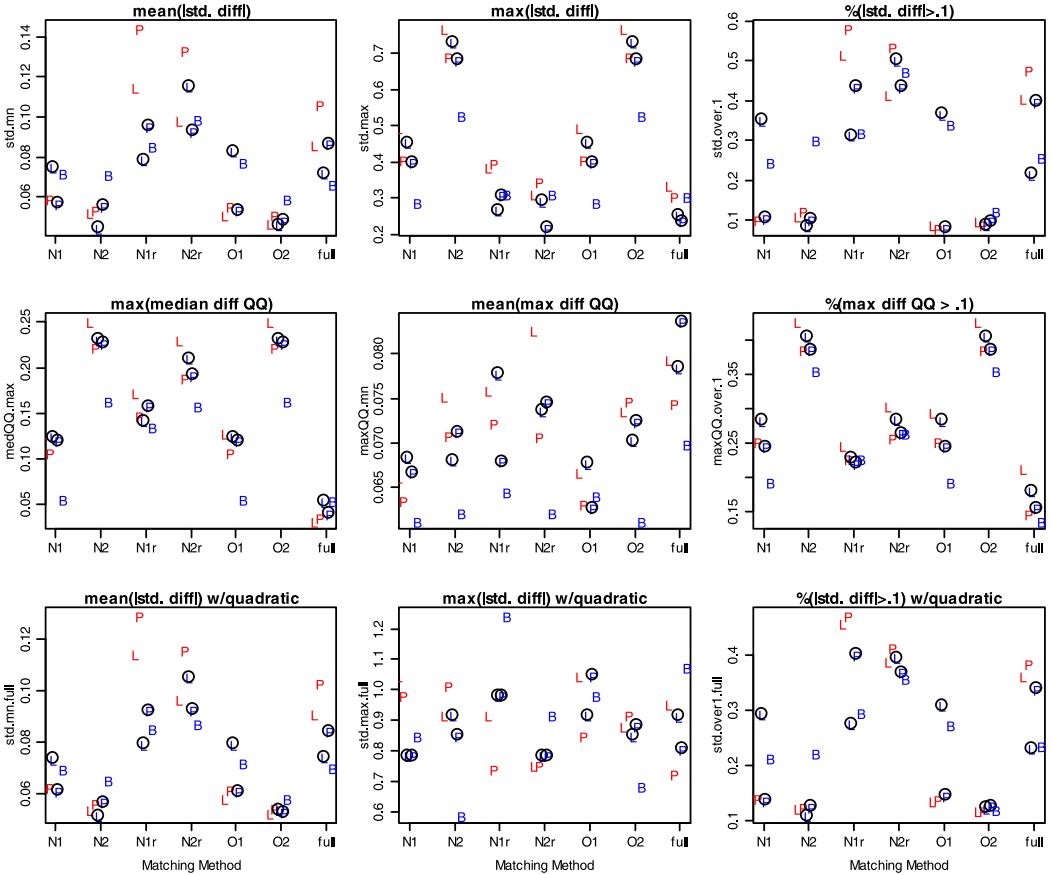

FIGURE 4 (See caption on page 499.)

treatment effect. In this example difference in means estimates ranges from −6.3 to 2.6 across the 54 strategies attempted. Additional covariance adjustment is often preferred, however, in case balance across groups is insufficient or simply to achieve efficiency gains (Rubin, 1979; Rubin & Thomas, 2000). This choice requires specifying a parametric model again[13] (though the increased balance should make these models far more robust to violations of the parametric assumptions). Additionally the decision to make additional adjustments requires the researcher to choose whether to perform this additional adjustment on all of the covariates or just a subset. For this example we tried standard linear regression (on the matched or weighted data) both on just the subset of test scores and also on the full set of covariates; the former choice led to a range of estimates from −6.5 to .8 and the latter to a range of estimates from −3.7 to .7 (standard errors range from .53 to 2.6). These ranges do not change if we exclude the full matching estimates (because they were designed for a different estimand). This array of estimates represents a huge range of uncertainty for treatment effect estimates in terms of both practical and statistical significance.

We can use the balance summaries, however, to discriminate between methods. Figure 5 plots treatment effect estimates for third-grade reading test scores and 95% confidence intervals for each of the three analysis choices (difference in means, regression on test scores, regression on all covariates) for each of six propensity score strategies that met a set of balance criterion for the full set of covariates (std.mn < .08, std.max < .5, std.over.1 < .4), a set of balance criterion applying to all of the continuous covariates (medQQ.max < .2, maxQQ.mean < .08, maxQQ.over.1 < .3), and a set of balance criterion for the full set of covariates plus quadratic terms (std.mn < .1, std.max < 1, std.over.1 < .3). Two of the "winning" strategies (Probit.N1 and Probit.O1) used propensity scores estimated using standard probit models and relied on nearest neighbor (without replacement) and optimal pair matches, respectively. Two of these strategies (Bprobit.N1 and Blogit.N1) relied on nearest neighbor pair matches (without replacement)—one using the Bayesian logit and the other using the Bayesian

[13]The researcher may be able to specify a semiparametric or nonparametric model at this stage; however, even this would require decisions regarding tuning parameters. The nonparametric choice presented in the next section requires a minimum of such choices (or rather they are prespecified for the user). If the researcher is willing to invest in such a model at this stage, however, why not simply estimate the response surface directly rather than matching/weighting first?

FIGURE 4 (See artwork on page 498.) Balance plots for restricted set of methods. These plots take the same form as in Figure 3; however, they exclude any approach that relies on either the IPTW strategies or the GBM propensity score estimation method. (color figure available online)

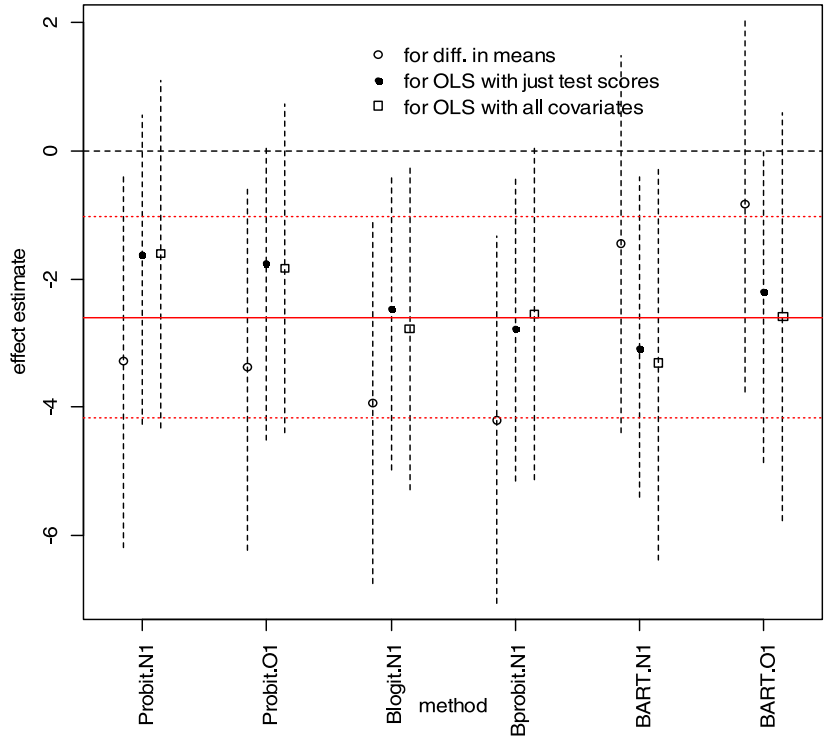

FIGURE 5 Treatment effects and corresponding 95% confidence intervals on third-grade reading test scores from the six propensity score strategies with balance satisfying given criteria (see text). For each strategy difference in means estimates are plotted alongside estimated coefficients on the treatment variable from linear regression models that additionally control for (a) the subset of (measurement-error-corrected) first-grade test scores and (b) the full set of pretreatment covariates. The solid horizontal line displays the estimate of the treatment effect obtained using BART to fit the response surface. The dotted lines above and below it display the corresponding 95% uncertainty interval. (color figure available online)

probit to estimate the propensity score. The other two strategies (BART.N1 and BART.O1) both used BART to estimate the propensity score but, as with the first two described, one used nearest neighbor (without replacement) and the other optimal pair matching.

Limiting to these six strategies still results in a total of 18 separate treatment effect estimates. It is somewhat difficult to further distinguish between these methods because they represent trade-offs in some criteria over others (and some variables over others). However, the range of estimates that they yield, although narrower than the range for the full set of strategies, is still nontrivial.

How should the researcher decide which estimates or set of estimates to report? Or should all the estimates be reported to more accurately reflect our uncertainty about the method?

Finally, a note about standard error calculations for these estimates. The treatment effect estimates presented are from approaches that used matching without replacement thus avoiding some of the complications that can arise in variance estimation when control units are used as matches multiple times (for discussion see, e.g., Hill & Reiter, 2006). However, there remains the complication for matched estimates that the matching process likely induced dependencies across the treatment and control groups. The literature is divided on the best approach to this issue and varies from the extreme of Ho, Imai, King, and Stuart (2007), who suggest ignoring the issue, to those who advise use of pair-matched analyses (Austin, 2008), to more model-based solutions somewhere in the middle (Hill, 2008; Hill & Reiter, 2006). Abadie and Imbens (2006) provide a more elegant solution for variance estimation in the case difference in means estimates from nearest neighbor matching. We avoid these debates in this work by using classical estimates that should closely approximate the truth for the model-based estimates but may be a bit conservative for the difference in means estimates.

ALTERNATIVE STRATEGY AND RESULTS

What alternatives exist for researchers who are wary of using standard models such as linear regression for conditioning on confounding covariates but are understandably confused about "best practice" using propensity scores? We describe here an alternative causal inference strategy proposed by Hill (2011) that has some evidence of superior performance (relative to linear regression, propensity score matching, and inverse probability of treatment weighting) and yet is relatively simple to implement and interpret.

All causal inference methods can be conceived of as attempts to predict counterfactual states. For instance, in this study, we can observe the outcome for a child who participated in the study and was retained after the first grade. However, we cannot directly observe the expected outcome for that child had he or she been promoted to the second grade instead (the counterfactual state).

Matching recovers a form of predicted counterfactual outcome for each treated unit. Weighting creates a pseudopopulation of controls that can be used to create a type of average counterfactual. Another approach, however, is to build a model predicting outcomes based on the treatment status and covariate information and then use that model directly to predict counterfactual outcome values for each treated unit (or the full sample if, e.g., the Average Treatment Effect is desired). This strategy has been criticized in the past because traditionally researchers

were concerned about lack of robustness of the estimates to deviations from the strict parametric assumptions of the popular models at the time (e.g., linear regression). Recent advances in high-dimensional nonparametric modeling, however, address these concerns.

Bayesian Additive Regression Trees (BART)

We take advantage of these advances and fit a very flexible nonparametric model to our data using an algorithm called Bayesian Additive Regression Trees (BART), which has been shown in previous work to have superlative in-sample and out-of-sample prediction properties (Chipman et al., 2010). In fact, BART performed at the top of the heap of competitors in the data mining literature (with regard to both in- and out-of-sample prediction) even when tested across 42 different data sets (Chipman et al., 2007).[14] The key for our purposes is that nonparametric modeling using BART allows for nonlinear relationships between the outcomes and the confounding covariates for which we need to adjust. Crucially, however, to fit the model the researcher need only provide the algorithm with the outcome, retention status (treatment), and the baseline covariates (all hyperparameter settings can be kept at their default values as described in Appendix B). BART is a learning algorithm that will, in essence, *find* the nonlinearities (including interactions). Thus the researcher can avoid the complicated process of diagnosing model fit for linear models in high-dimensional space and trying to find solutions such as performing transformations to appropriate variables or altering the general parametric form of the model.[15] These decisions, in addition to being challenging and tedious (especially with so many covariates), have a tendency to be ad hoc, leading to too many opportunities to "tweak" the model (inadvertently or not) based on the treatment effect estimates they yield.

After fitting the model to the data for the children in our analysis sample, we were able to make counterfactual comparisons by using the model to make predictions for each student who had been retained in grade. That is, we used it to make a prediction about what the test score would have been for that child had he or she been promoted. The algorithm also produces estimates of our uncertainty about that prediction. Then for each child we can get an estimate of

[14]When BART's hyperparameters were chosen via cross validation it performed the best on average. Using the default BART settings for these hyperparameters (the practice we espouse here; also espoused by Hill, 2011), BART performed at least as well with regard to in- and out-of-sample prediction as the strongest current competitors in the data mining literature (neural nets, gradient boosting, random forests) and noticeably better than lasso even though the other methods get to choose their free parameters using cross validation.

[15]Probably no statistical procedure should be used without any sort of diagnostic. Appendix B describes some simple checks that can be used to help ensure that the BART fit is appropriate.

his or her individual level causal effect (the difference between predicted values for observed and counterfactual states)[16] and our uncertainty about that estimate (technically we get a Monte Carlo estimate of the full posterior distribution for this unknown parameter). More usefully, we can average these estimates to create an estimate of the average causal effect for those retained.[17]

The algorithm was implemented in R (in package BayesTree), a freely-available, open-source software package. The model is described in greater detail in Appendix B. Given the size of the data set (both in terms of sample size and number of variables), the algorithm was slightly more challenging to implement than is usual because it took longer to converge than it typically would (as described in Appendix B). But the convergence criterion is clearly defined and again all of the "tuning parameters" in the algorithm were kept at their default settings so we still retained the desired simplicity of avoiding multiple choices regarding the statistical approach. This contrasts sharply to propensity score approaches that require several sets of choices regarding estimation and matching/weighting strategies not to mention ad hoc choices with respect to balance diagnostics.

Results Using BART

The results across the six outcome variables using this strategy are presented in Figure 6. These estimates suggest that 2 years postretention those retained are scoring significantly lower on both reading and math assessments (by about 2.5 points on a scale with standard deviation of 15) than they would have had they been promoted. This difference remains for reading at 4 years after the retention decision but the posterior interval for the math assessment is centered closer to zero and the corresponding interval includes zero. As noted previously, however, these comparisons aren't quite fair because in each case the promoted students are (at least) one grade ahead of the retained students. The 3rd/5th outcome gives the retained students a 1-year advantage; they are being tested 4 years after the retention decision while in the fourth grade whereas the counterfactual condition is what their test scores would have been in third grade (2 years after the retention decision) had they been promoted. In this comparison we cannot detect any noticeable effect of retention.

[16] Although we would caution against interpreting any individual level causal effects as that is probably asking too much from the model.

[17] As described in Hill (2011), this estimate is more directly applicable to the conditional average causal effect for the treated (CATT) as defined by Abadie and Imbens (2002). However, if our sample is representative of the population it should also be unbiased for the population average treatment effect on the treated (PATT) and uncertainty estimates can be augmented to reflect our additional uncertainty in this setting.

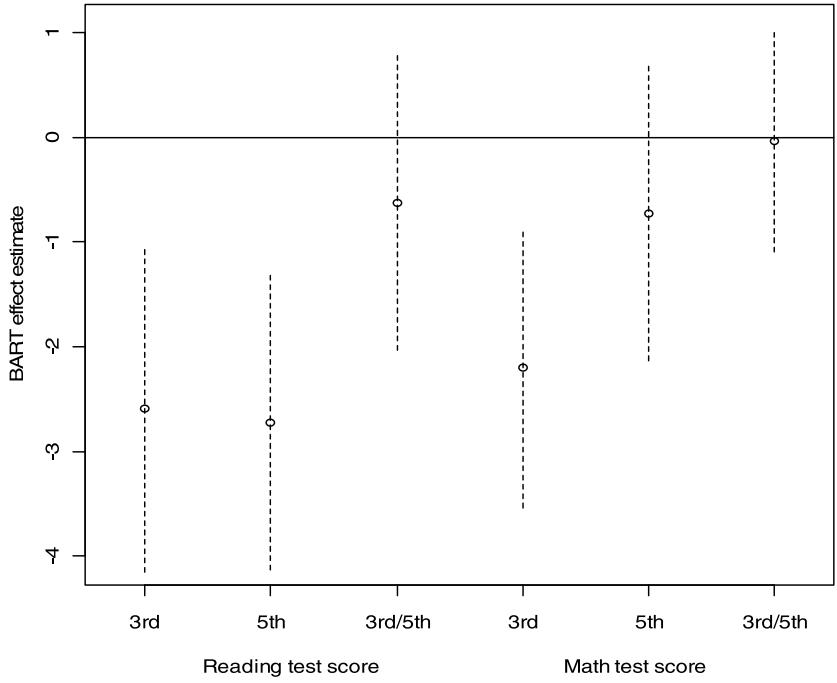

FIGURE 6 Treatment effect estimates and corresponding 95% uncertainty intervals for all six outcome variables calculated using the BART estimation strategy.

DISCUSSION

Propensity score strategies were developed to address a real problem in observational study settings: how best to condition on a (potentially substantial) number of covariates in order to satisfy a given ignorability condition without making undue parametric assumptions. These days, however, there are more obvious trade-offs to be made in terms of complexity of implementation. Effective propensity score analysis requires making choices about (a) how to fit the propensity score model, (b) what type of matching/weighting algorithm will be used, (c) which balance diagnostics to use and how to determine when balance is sufficient, and (d) choice of analysis model. This article has demonstrated in a real, albeit somewhat challenging, example that these choices can have nonnegligible impacts on the resulting estimates.[18]

[18]Moreover, although we have focused on some of the most important choices the researcher faces when implementing propensity score strategies, we have largely skirted the issue of common support and have completely ignored the somewhat contentious issue of variance estimation.

We have presented an alternative estimation approach for this setting that relies on the BART algorithm that eliminates this complexity. This strategy has been demonstrated in previous work (Hill, 2011) to have equal or superior performance compared with some common propensity score strategies in a variety of settings. In this example the point estimate of the effect of the treatment on the treated produced by BART lies near the center of the estimate corresponding to the subset of propensity score approaches that achieve the best balance with these data. More research needs to be done to determine if there are scenarios in which BART may not perform as well. However, it appears to be a potentially promising alternative to propensity score matching, at least in situations with a large number of covariates, and at a minimum is worthy of further investigation and comparison.

ACKNOWLEDGMENTS

We acknowledge support from National Science Foundation (NSF) Grant 0532400.

REFERENCES

Abadie, A., & Imbens, G. W. (2002). Simple and bias-corrected matching estimators for average treatment effects (Working Paper No. 0283). National Bureau of Economic Research, Inc.

Abadie, A., & Imbens, G. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica, 74,* 235–267.

Austin, P. (2008). A critical appraisal of propensity score matching in the medical literature between 1996 and 2003. *Statistics in Medicine, 27,* 2037–2049.

Austin, P. C., Grootendorst, P., & Anderson, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: A Monte Carlo study. *Statistics in Medicine, 26,* 734–753.

Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas accountability system. *American Educational Research Journal, 42,* 231–268.

Chipman, H., George, E., & McCulloch, R. (2007). Bayesian ensemble learning. In B. Scholkopf, J. Platt, & T. Hoffman (Eds.), *Advances in neural information processing systems, 19,* 265–272. Cambridge, MA: MIT Press.

Chipman, H., George, E., & McCulloch, R. (2010). BART: Bayesian Additive Regression Trees. *Annals of Applied Statistics, 4,* 266–298.

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

Diamond, A., & Sekhon, J. (2008). *Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies* (Tech. Rep., Working Paper). Berkeley, CA: U.C. Berkeley.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* New York, NY: Cambridge University Press.

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y. S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics, 2,* 1360–1383.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7,* 457–472.

Gelman, A., Su, Y. S., Yajima, M., Hill, J., Pittau, M. G., Kerman, J., & Zheng, T. (2009). *arm: Data analysis using regression and multilevel/hierarchical models.* R package (Version 1.2–9). Retrieved from http://www.stat.columbia.edu/∼gelman/software/arm

Greene, J. P., & Winters, M. A. (2004). *An evaluation of Florida's program to end social promotion.* New York, NY: Center for Civic Innovation, Manhattan Institute for Policy Research. Education Working Paper No. 7.

Hansen, B. (2004). OPTMATCH: An add-on package for R (Tech. Rep.). Ann Arbor, MI: University of Michigan.

Hansen, B. (2008). The essential role of balance tests in propensity-matched observational studies: Comments on "A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003" by Peter Austin. *Statistics in Medicine, 27,* 2050–2054.

Hansen, B., & Klopfer, S. O. (2006). Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics, 15,* 609–627.

Hill, J. (2008). Discussion of Peter Austin's "A critical appraisal of propensity score matching in the medical literature between 1996 and 2003." *Statistics in Medicine, 27,* 2055–2061.

Hill, J. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics, 20,* 217–240.

Hill, J., & Reiter, J. (2006). Interval estimation for treatment effects using propensity score matching. *Statistics in Medicine, 25,* 2230–2256.

Hill, J., & Su, Y. (2010). Addressing lack of common support in causal inference using Bayesian non-parametrics (Working Paper). New York, NY: New York University.

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

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (in press). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software.*

Imbens, G. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics, 86,* 4–29.

Jacob, B. A., & Lefgren, L. (2004). Remedial education and student achievement: A regression-discontinuity analysis. *Review of Economics and Statistics, 86,* 226–244.

Jacob, B. A., & Levitt, S. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics, 118,* 843–877.

Jones, M. P. (1996). Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American Statistical Association, 91,* 222–230.

Klepper, S., & Leamer, E. (1984). Consistent sets of estimates for regressions with errors in all variables. *Econometrica, 52,* 163–183.

Kurth, T., Walker, A. M., Glynn, R. J., Chan, K. A., Gaziano, J. M., Berger, K., & Robins, J. M. (2006). Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of non-uniform effect. *American Journal of Epidemiology, 163,* 262–270.

Martens, E. P. (2007). *Methods to adjust for confounding: Propensity scores and instrumental variables* (Doctoral thesis, University of Utrecht, Utrecht, The Netherlands).

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

Mincy, R., Hill, J., & Sinkewicz, M. (2009). Marriage: Cause or mere indicator of future earnings growth. *Journal of Policy Analysis and Management, 28,* 417–439.

Pearl, J. (2010). On a class of bias-amplifying covariates that endanger effect estimates. UCLA Cognitive Systems Laboratory. In P. Grunwald & P. Spirtes (Eds.), *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence* (pp. 417–424). Corvallis, OR: Association for Uncertainty in Artificial Intelligence (AUAI).

Robins, J. M., & Ritov, Y. (1997). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in Medicine, 16,* 285–319.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association, 82,* 387–394.

Rosenbaum, P. R. (1991). A characterization of optimal designs for observational studies. *Journal of the Royal Statistical Society, Series B, Methodological, 53,* 597–610.

Rosenbaum, P. R. (2009). *Design of observational studies.* New York, NY: Springer-Verlag.

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

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

Rubin, D. B. (1973). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics, 29,* 185–203.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics, 6,* 34–58.

Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Society, 74,* 318–328.

Rubin, D. B., & Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association, 95,* 573–585.

Schwartz, A. E., Stiefel, L., & Zabel, J. (in press). The appropriate uses of student-level and school-level data for measuring school performance. In J. Hannaway (Ed.), *Learning from longitudinal data in education.* Washington, DC: Urban Institute Press.

Sekhon, J. S. (2010). *Alternative balance metrics for bias reduction in matching methods for causal inference* (Tech. Rep.). Berkeley, CA: University of California.

Sekhon, J. S. (in press). Multivariate and propensity score matching software with automated balance optimization: The Matching package for R. *Journal of Statistical Software*.

Setoguchi, S., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., & Cook, E. F. (2008). Evaluating uses of data mining techniques in propensity score estimation: A simulation study. *Pharmacoepidemiology and Drug Safety, 17,* 546–555.

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

Wu, W., West, S. G., & Hughes, J. N. (2008a). Effect of retention in first grade on children's achievement trajectories over 4 years: A piecewise growth analysis using propensity score matching. *Journal of Educational Psychology, 100,* 727–740.

Wu, W., West, S. G., & Hughes, J. N. (2008b). Short-term effects of grade retention on the growth rate of Woodcock-Johnson III broad math and reading scores. *Journal of School Psychology, 46,* 85–105.

APPENDIX A
List and Definition of Variables Used

Constructs and Covariates	*Definition and Notes*
Child characteristics	
Gender	Dummy variable: boy or not
Age	Continuous variable
Race/Ethnicity	Categorical variable: non-Hispanic white, non-Hispanic black, Hispanic, other

(*continued*)

APPENDIX A
(*Continued*)

Constructs and Covariates	*Definition and Notes*
Disability diagnosed	6 dummy variables: physical disability, learning problem, attention problem, speech problem, hearing problem, vision problem
Health insurance	5 dummy variables: no insurance, private, Medicaid, CHIP, other
Nonparental care prior to kindergarten	Dummy variable
Primary child care arrangement in first grade	Categorical variable: parental care, relative care, nonrelative care, center-based care, other care arrangement
Tutoring programs participated in first grade	14 dummy variables: individual reading/math, pull-out small group reading/math, pull-out/in-class ESL, gifted program in reading/math, Title I program in reading/math/English/ESL/combined subjects, program for problem behaviors
Tutoring programs participated in fall kindergarten	14 dummy variables: individual reading/math, pull-out small group reading/math, pull-out/in-class ESL, gifted program, Title I program in reading/math/English/ESL/combined subjects/other Title I program, program for problem behaviors
Parents' characteristics	
Marital status	Categorical variable: married, separated, divorced, widowed, never married, other
Age	2 continuous variables: mother's age, father's age
Education	2 categorical variables (mother and father's education): below high school, high school or equivalent, vocational school or some college, bachelor, graduate or higher
Employment	2 categorical variables (mother and father's education): 35 hr or more per week, less than 35 hr, unemployed, not in labor force
Home and neighborhood environment	
English spoken at home	Dummy variable
Family type	Categorical variable: two parents and siblings, two parents and no sibling, one parent and siblings, one parent and no sibling, other
Household size	Continuous variable
Number of siblings	Continuous variable
Number of places child lived	Continuous variable
Number of schools child changed	Continuous variable
Times child was late for school last month	Continuous variable
Family annual income	Categorical variable: $5,000 or less, $5,001–$10,000, $10,001–$15,000, $15,001–$20,000, $20,001–$25,000, $25,001–$30,000, $30,001–$35,000, $35,001–$40,000, $40,001–$50,000, $50,001–$75,000, $75,001–$100,000, $100,001–$200,000, $200,001 and higher

(*continued*)

APPENDIX A
(*Continued*)

Constructs and Covariates	*Definition and Notes*
Housing situation	Categorical variable: own/condominium, rent, other arrangement
Receipt of AFDC/TANF last 12 months	Dummy variable
Receipt of food stamps last 12 months	Dummy variable
Rural/urban residency	Categorical variable: large/middle city, large/midsuburb/town, small town & rural
Region of the country	Categorical variable: Northeast, Midwest, South, West
Teacher characteristics (teachers in spring kindergarten and first grade, respectively)	
Gender	Dummy variable: female or not
Age	Continuous variable
Race/Ethnicity	Categorical variable: non-Hispanic white, non-Hispanic black, Hispanic, other
Education	Categorical variable: high school, GED or associate, bachelor, graduate or higher, professional diploma
Bachelor degree major	5 dummy variables: early education, elementary education, special education, other educational major, noneducational major
Graduate degree major	5 dummy variables: early education, elementary education, special education, other educational major, noneducational major
Number of college courses taken	5 continuous variables: early education, elementary education, special education, ESL, child development; methods of teaching reading, math, science
Teaching experience	10 continuous variables: years of teaching in preschool, kindergarten, second to fifth grade, sixth grade and up; years of teaching ESL, bilingual program, special education, physical education, art; years of teaching in current school
Types of teaching certification	Categorical variable: none, temporary/probational, alternative program, regular, highest
Certified areas	3 dummy variables: early education, elementary education, other areas
Teaching preparation hours per week	2 categorical variables (2 or less, 2–5, 5–9, 9–14, 15 hr or more): paid, unpaid
School characteristics (first grade)	
Public/Private school	Dummy variable
Racial components of students	4 categorical variables (none, less than 5%, 5%–15%, 15% or higher): Asian, non-Hispanic black, Hispanic, Native American
Percentage of minority students	Categorical variable: 0%–10%, 10%–25%, 25%–50%, 50%–85%, 85% and higher
School received Title I funds	Categorical variable: yes, no, not offered

(*continued*)

APPENDIX A
(*Continued*)

Constructs and Covariates	*Definition and Notes*
Number of students served by Title I funds	Categorical variable: none, 1–100, more than 100
Percentage of students from neighborhood	Categorical variable: less than 50%, 50%–80%, more than 80%
Any children with limited English proficiency	Dummy variable
Services provided for families of children with limited English proficiency	6 dummy variables: translators, written translation, home visit, outreach workers to help children's first enrollment, conducting special meetings for non-English speaking families, other services
Parents' educational expectation and involvement	
Expected degree of child	Categorical variable: high school or lower, some college, bachelor, graduate or higher
School involvement	6 dummy variables: attending open house, PTA meeting, parent-teacher conference, or school events, acting as volunteers, participating in fund-raising
Assessment of how well school has done	4 categorical variables (very well, just okay, or do not do at all): reporting how child is doing, helping understand children, informing chances to volunteer at school, providing information of helping child
Number of parents talked to regularly	Continuous variable
Frequency of helping child with homework	Categorical variable: never, less than once per week, 1–2 times per week, 3–4 times per week, 5 or more times per week
Activities with child at home	10 categorical variables (never, 1–2 times per week, 3–6 times per week, or every day): telling child stories, singing together, helping child with art, child doing chores, playing games together, teaching child nature, building with child, doing sports together, practicing numbers with child, reading to child
Parents' assessment of child	
Skills compared with others in first grade	2 categorical variables (far below average, below average, average, above average, far above average): reading, math
Rating of child's social skills in fall kindergarten	5 continuous variables: approaches to learning, self-control, social interaction, sadness/loneliness, impulsiveness/ overactivity
Rating of child's social skills in first grade	5 continuous variables: approaches to learning, self-control, social interaction, sadness/loneliness, impulsiveness/ overactivity
Teacher's assessment of child	
Problem behaviors in fall kindergarten	2 continuous variables: internalizing problem behaviors, externalizing problem behaviors
Problem behaviors in first grade	2 continuous variables: internalizing problem behaviors, externalizing problem behaviors

(*continued*)

APPENDIX A
(*Continued*)

Constructs and Covariates	*Definition and Notes*
Skills compared with others in first grade	5 categorical variables (far below average, below average, average, above average, far above average): language skills, science/social study, math, activities in structured play/unstructured play
Rating of child's social skills in first grade	3 continuous variables: approaches to learning, self-control, interpersonal skills
Rating of child's social skills in fall kindergarten	3 continuous variables: approaches to learning, self-control, interpersonal skills
Test scores	
Imputed test scores in kindergarten and first grade	3 continuous variables: math, reading, general knowledge

Note. AFDC = Aid to Families with Dependent Children; TANF = Temporary Assistance to Needy Families; CHIP = Children's Health Insurance Program; ESL = English as a Second Language; GED = General Educational Development; PTA = Parent Teacher Association.

APPENDIX B

Bayesian Additive Regression Trees (BART)

To describe how BART works, we start with a description of the regression tree, a simpler technique that acts as a building block for BART. A regression tree is a simple form of a nonparametric fit. Regression trees partition the data into (often many) subgroups based on their covariate values. This is done through an iterative algorithm that progressively splits the data, based on observed values for the predictors, into more and more homogenous subsets with respect to the outcome. At each step, the algorithm searches through all possible splits of the subgroups determined in the previous step (each based on one predictor), with the goal of minimizing deviance in the outcome within the resulting subgroups. Stopping rules determine when the subgroups are split no more; the final set of subgroups are called the "terminal nodes."

Although regression trees allow a nonparametric fit to data, they also can be inefficient, too easily overfit the data, and are not particularly good at finding additive structures (i.e., these models may "find" interactions or seeming nonlinearities that reflect sample anomalies rather than the true data generating process).

An improvement on simple regression trees is available in the form of boosting algorithms. The basic idea is to iteratively fit many (often 200) small trees and use the sum of the fit from each as the final fit. The tree-fitting can be accomplished by first fitting a small tree (perhaps four subsets), calculating

residuals from the fit of this tree, then fitting another small tree to these residuals, and so on. To avoid overfitting, the fit from each tree is typically shrunk, using a constant (typically chosen using cross validation). Boosting of regression trees has been shown to be almost uniformly superior to simple regression trees.

The drawback of boosting algorithms is that they require many ad hoc choices for the "tuning" parameters, such as the total number of trees fit, the number of terminal nodes allowed in each tree, and the shrinkage parameter. Additionally, because this was developed as a prediction algorithm, the calculation of standard errors is not straightforward and typically requires other algorithms (such as bootstrapping), that can be quite computationally burdensome (nesting an already iterative algorithm within another iterative algorithm).

BART can be thought of (loosely) as a Bayesian form of boosting, in which the model is fit using Gibbs sampling (a simulation-based estimation procedure used commonly with fitting Bayesian models). There are several advantages of BART over boosting. First, rather than ad hoc tuning parameters, choices regarding the number of trees, the number of subsets, and the shrinkage of the individual fit from each tree are incorporated into the prior distribution. The model has been parameterized so that these "hyperparameters" are constrained by the nature of the data at hand. BART's default prior settings have been shown to work well in a wide variety of situations (Chipman et al., 2007).

BART is fit using a Markov Chain Monte Carlo (MCMC) approach (as is typical when fitting Bayesian models; for additional details see Chipman et al., 2007, 2010). Typically this algorithm converges quite quickly and requires running only one chain. Convergence can be checked by monitoring the only identified parameter, the residual standard deviation. In this setting we also monitored convergence using the estimate of the effect of the treatment on the treated. Therefore we used a standard procedure for MCMC convergence assessment, which was to run three chains and use the $\hat{R}$ statistic to determine convergence (based on a threshold of 1.1; Gelman & Rubin, 1992). The code was still relatively simple and is available from the authors upon request. Although this implementation of BART was more difficult than usual because of the large number of covariates and the large sample size (in typical applications we can achieve convergence with only one chain), the key point is that the criterion for convergence was clear and made use of a well-accepted and easy-to-use statistic.

Besides checking convergence, it is generally advisable to perform some simple checks to assess the appropriateness of the predictions from the algorithm. Simple comparisons of predicted versus observed values can be helpful. BART could also be prone to inappropriate predictions in situations where it is being forced to extrapolate far beyond the range of the data; for instance, this could happen if a researcher was making counterfactual predictions for a treatment group based on data from a control group whose covariate distribution didn't sufficiently overlap with that of the treatment group. Work in progress (Hill

& Su, 2010) explores strategies for diagnosing this in settings where lack of common support might be a problem.

Finally, although we advise use of the default settings of BART we did explore sensitivity of the BART results to changes in the settings of the three key hyperparameters. Varying one at a time we explored variation in the "k" parameter (see BayesTree documentation) from 1 to 3, variation in the "power" parameter from 1 to 3, and variation in the "base" parameter from .9 to .99 (these are fairly large changes in the parameter values; it is unlikely one would set a parameter outside of these ranges). None of these changes resulted in a difference in treatment effect estimates of more than .2. We also explored the impact of changing the random seeds used for each of the three chains run. This yielded a difference of about the same magnitude. These differences are small compared with the differences yielded by even the preferred subset of propensity score strategies that yielded the best balance.