Psychological Methods
2016, Vol. 21, No. 3, 427–445

1082-989X/16/$12.00 http://dx.doi.org/10.1037/met0000076

Propensity Score Analysis With Missing Data

Heining Cham
Fordham University

Stephen G. West
Arizona State University

Propensity score analysis is a method that equates treatment and control groups on a comprehensive set of measured confounders in observational (nonrandomized) studies. A successful propensity score analysis reduces bias in the estimate of the average treatment effect in a nonrandomized study, making the estimate more comparable with that obtained from a randomized experiment. This article reviews and discusses an important practical issue in propensity analysis, in which the baseline covariates (potential confounders) and the outcome have missing values (incompletely observed). We review the statistical theory of propensity score analysis and estimation methods for propensity scores with incompletely observed covariates. Traditional logistic regression and modern machine learning methods (e.g., random forests, generalized boosted modeling) as estimation methods for incompletely observed covariates are reviewed. Balance diagnostics and equating methods for incompletely observed covariates are briefly described. Using an empirical example, the propensity score estimation methods for incompletely observed covariates are illustrated and compared.

Keywords: propensity score, nonrandomization, missing data, machine learning

In the prototypical observational study, participants are measured on a set of covariates at baseline (pretreatment), they then receive a treatment or control intervention, and, finally, one or more outcome variables are measured (Rosenbaum, 2010). Given *nonrandom* assignment to treatment (t) or control (c) groups, there are typically baseline covariates that are common causes of the assignment to t versus c and the outcome variable. These variables are termed *confounders*. Lack of proper statistical adjustment for confounders will produce a biased estimate of the causal effect in an observational study.

To equate the t and c groups at baseline, Rubin (2001) has recommended that a comprehensive set of covariates that are potential confounders be measured because the actual confounders are unknown. Moser, West, and Hughes (2012) measured 72 covariates in a study comparing the subsequent academic achievement of children who were retained in first grade (t group) or promoted to second grade (c group) at the end of their first school year. West et al. (2014) measured 98 covariates in a study investigating the well-being of patients who received versus those who did not receive telephone counseling. Traditional exact matching methods will fail when more than a few covariates are measured because appropriate matches cannot be created (see Chapin, 1947, for an example). Traditional analysis of covariance may produce inappropriate results if there are nonlinear covariate-outcome relationships and the baseline distributions of the covariates of the two groups do not fully overlap (Schafer & Kang, 2008). Propensity score analysis, originally developed by Rosenbaum and Rubin (1983), can be used to equate the distributions of a *large number of measured covariates* in the t and c groups, producing a less biased estimate of the causal effect. Comprehensive reviews of the use of propensity score analysis with complete data are available (e.g., Austin, 2011; Caliendo & Kopeinig, 2008; Dehejia & Wahba, 2002; Schafer & Kang, 2008; Stuart, 2010; West et al., 2014). In practice, the measured covariates (putative confounders) and outcome often have missing values. To the authors' knowledge, there has been no comprehensive review of methods for addressing these problems in propensity score analysis. This is the primary goal of the present article.

The structure of the article is as follows. First, we briefly review of the theory of propensity scores when the covariates are completely observed, followed by a presentation of the theory for incompletely observed covariates. Second, we consider methods of estimating propensity scores, beginning with the traditional logistic regression method followed by newer machine learning methods. Machine learning methods, often used with big data sets, can make correct predictions (here, of the propensity score for treatment assignment) without specifying each of the covariates and their relationship to the outcome. Third, we discuss balance diagnostics, equating methods, and sensitivity analysis for incomplete covariates. Fourth, we present an illustrative example that applies these methods to an empirical data set with incompletely observed covariates. Fifth, we discuss propensity score analysis when the outcome variable has missing values. Finally, we discuss other issues in propensity score analysis with covariates having missing data. Our focus in this article is on the prototypical observational study in which the participants are nonrandomly assigned to t or c and whose goal is to estimate the causal effect of the treatment on the outcome variable Y.

This article was published Online First March 10, 2016.

Heining Cham, Department of Psychology, Fordham University; Stephen G. West, Department of Psychology, Arizona State University.

We thank Jan Hughes, Texas A&M University, and a grant from the National Institute of Child Health and Human Development (grant number HD039367), for generously providing the data upon which the illustrative example is based.

Correspondence concerning this article should be addressed to Heining Cham, Department of Psychology, Fordham University, 441 E. Fordham Road, Bronx, NY 10458. E-mail: hcham@fordham.edu

Theory of Propensity Score Analysis

Potential Outcomes Model

Propensity score analysis was originally developed in the framework of the potential outcomes model (also known as Rubin causal model; Holland, 1986; Rubin, 1974; West & Thoemmes, 2010). Each participant i has a potential outcome that would occur if he or she were assigned to the treatment group t, $Y_t(i)$, and a second potential outcome that would occur if he or she were assigned to the control group c, $Y_c(i)$. There are two different causal effects of interest that may typically be of interest in an observational study. The first is the average treatment effect (*ATE*), which is defined as $ATE = E[Y_t(i)] - E[Y_c(i)]$, where $E[\cdot]$ is the expectation function. The second is the average treatment effect on the treated (*ATT*), which is defined as $ATT = E[Y_t(i) \mid Z_i = t] - E[Y_c(i) \mid Z_i = t]$. The binary variable Z_i represents the *actual* treatment group assignment for participant i ($Z_i = t$ or c; Rosenbaum & Rubin, 1985). *ATE* is the average of the hypothetical causal effects across all participants in the population, regardless whether they are assigned to the t or c group. *ATT* is the causal effect on average across all participants in the population who would actually receive the treatment when assigned to the t group. In a randomized experiment with complete data and full treatment compliance, *ATE* will always equal *ATT* (Austin, 2011; Dehejia & Wahba, 2002). In an observational study, *ATE* generally does not equal *ATT*. One example would be a study comparing the efficacy of a medication treatment group and a no-treatment control group in reducing blood pressure. In the study, there is a subgroup of the control group participants with baseline systolic blood pressure less than 140 mmHg. This subgroup would never receive the medication intervention.

Strong Ignorability Assumption

Unbiased *ATE* or *ATT* estimates are possible in observational studies when the strong ignorability assumption is met[1] (Rosenbaum & Rubin, 1983, 1984, 1985). This assumption is closely related to the assumption that data are missing at random (MAR) in missing data theory (discussed later). This assumption consists of two propositions (Austin, 2011; Caliendo & Kopeinig, 2008). The first proposition is expressed as Equation 1,

$$Y_t(i), Y_c(i) \perp Z_i \mid \mathbf{X}_i, \tag{1}$$

where $\perp$ is the independence function and $\mathbf{X}_i$ is a row vector of measured baseline (pretreatment) covariates for participant i. Recall that confounders are common causes of (a) the assignment to t versus c, and (b) the outcome variable. Equation 1 states that the participant i's potential outcomes of t and c groups are conditionally independent of the actual treatment group assignment Z_i, given that all of the confounders are represented in $\mathbf{X}_i$. When all baseline confounders are measured and successfully equated, the observational study resembles the randomized experiment. The use of a comprehensive set of covariates helps minimize the bias of the *ATE* or *ATT* estimates (Rubin, 2001). The comprehensive set of covariates reduces the chance that unmeasured confounders exist that provide unique contributions to the confounding effect, once the effect of the measured covariates has been removed.[2]

The second proposition of the strong ignorability assumption is expressed as Equation 2,

$$0 < Pr(Z_i = t \mid \mathbf{X}_i) < 1, \tag{2}$$

where $Pr(\cdot)$ is a probability function. This proposition states that participant i has a nonzero probability of being in the t group (or in the c group), conditional on the set of confounders in $\mathbf{X}_i$.

Propensity Scores

The vector $\mathbf{X}_i$ can consist of many covariates. The propensity score approach is a data reduction technique for $\mathbf{X}_i$ that bypasses the problems of exact matching with many covariates by allowing researchers to equate the t and c groups on a single variable, the propensity score. The propensity score $e(\mathbf{X}_i)$ is defined as the conditional probability that participant i is assigned to the t group given $\mathbf{X}_i$ (Equation 3; Rosenbaum & Rubin, 1983, 1984):

$$e(\mathbf{X}_i) = Pr(Z_i = t \mid \mathbf{X}_i) \tag{3}$$

Important theorems for the propensity score $e(\mathbf{X}_i)$ are proved in Rosenbaum and Rubin (1983). Two results of these proofs are of special importance. First, $e(\mathbf{X}_i)$ is the function of the confounders $\mathbf{X}_i$, which can be used to equate the distributions of $\mathbf{X}_i$ between t and c groups (Rosenbaum & Rubin, 1983, Theorems 1 and 2). Second, if the strong ignorability assumption in Equations 1 and 2 holds for $\mathbf{X}_i$, the strong ignorability assumption also holds for $e(\mathbf{X}_i)$ (Theorem 3). Rosenbaum and Rubin (Theorem 5) prove that these asymptotic theorems can be carried over to the estimated propensity score. This result implies that propensity scores estimated by researchers can be used to equate the distributions of $\mathbf{X}_i$ between the t and c groups in the sample.

Propensity Score Theory for Incomplete Confounders

The basic strong ignorability assumption and propensity score theory outlined in Equations 1 to 3 assume that the covariates in $\mathbf{X}_i$ are completely observed (no missing values) for all participants. Rosenbaum and Rubin (1984, Appendix B) extended the basic theory to the situation where $\mathbf{X}_i$ have missing data. The vector $\mathbf{X}_i$ may be partitioned into two components: $\mathbf{X}_i = (\mathbf{X}_i^{\mathbf{obs}}, \mathbf{X}_i^{\mathbf{mis}})$, where $\mathbf{X}_i^{\mathbf{obs}}$ refers to participant i's observed values for $\mathbf{X}_i$, and $\mathbf{X}_i^{\mathbf{mis}}$ refers to the participant i's missing values for $\mathbf{X}_i$. Participant i's missingness pattern of $\mathbf{X}_i$ can be described by creating a vector $\mathbf{R}_i$, which has the same dimensions as $\mathbf{X}_i$. The elements of $\mathbf{R}_i$ equal 1 when the corresponding elements in $\mathbf{X}_i$ are observed, and equal 0 when

[1] In order to produce unbiased sample estimates of *ATE* or *ATT* in an observational study, the study needs to fulfill four assumptions in addition to the strong ignorability assumption: (a) all participants adhere completely to the actual treatment group assignment (Sagarin et al., 2014), (b) all participants are measured on the outcome variable after the treatment, (c) the participant's outcome is not affected by the potential outcomes of other participants in the study, and (d) there are no hidden variations of the treatment (West et al., 2014). Except in the final section which considers missing data on the outcome, it is assumed that these four assumptions are fulfilled.

[2] Although we recommend that a comprehensive set of covariates be selected, we do not recommend selecting variables on the basis of empirical associations between treatment assignment and outcome. This distorts the use of propensity score analysis as a design-based tool to mimic a randomized experiment (Austin, 2011; Kelcey, 2011).

Table 1
Illustration of Arbitrary Missingness Patterns

	Measured covariates $\mathbf{X}_i$				Missingness patterns $\mathbf{R}_i$			
Participant	X_{1i}	X_{2i}	X_{3i}	X_{4i}	R_{1i}	R_{2i}	R_{3i}	R_{4i}
1	1.2	4.3	2.6	?	1	1	1	0
2	.9	?	4.9	?	1	0	1	0
3	?	2.8	3.2	5.1	0	1	1	1
4	?	2.6	?	4.7	0	1	0	1
5	?	3.5	3.8	6.5	0	1	1	1

Note. A question mark (?) indicates that the value is not observed.

missing. Table 1 illustrates some arbitrary missingness patterns of four covariates (X_{1i} to X_{4i}) for five participants. In this article, we consider the missingness pattern of $\mathbf{X}_i$ to be arbitrary—any subset of the covariates in $\mathbf{X}_i$ may be missing for any participant i (Schafer & Graham, 2002).

When there are incomplete observations in $\mathbf{X}_i$, the strong ignorability assumption and the definition of propensity scores must be modified (Rosenbaum & Rubin, 1984, Appendix B). The goal of this modification is to account for both the *observed* values of covariates $\mathbf{X}_i^{\mathbf{obs}}$ and the observed missingness patterns $\mathbf{R}_i$. Equation 1 is modified to Equation 4:

$$Y_t(i), Y_c(i) \perp Z_i \mid \mathbf{X}_i^{\mathbf{obs}}, \ \mathbf{R}_i. \tag{4}$$

Compared with Equation 1, this modified proposition states that the potential outcomes of t and c groups are conditionally independent of the actual group assignment Z_i, given both the *observed* values of covariates $\mathbf{X}_i^{\mathbf{obs}}$ *and* missingness patterns $\mathbf{R}_i$. When $\mathbf{X}_i^{\mathbf{obs}}$ and $\mathbf{R}_i$ are successfully equated in the t and c groups, the observational study will resemble the randomized experiment.

Equation 2 of the strong ignorability assumption is modified as well. Given incompletely observed $\mathbf{X}_i$, it is assumed that participant i has a nonzero probability of being in the t (or c) group, given $\mathbf{X}_i^{\mathbf{obs}}$ and $\mathbf{R}_i$. The propensity score of participant i, who has incompletely observed covariates, is modified to Equation 5 (D'Agostino & Rubin, 2000):

$$e\left(\mathbf{X}_i^{\mathbf{obs}}, \ \mathbf{R}_i\right) = \ Pr(Z_i = t \mid \mathbf{X}_i^{\mathbf{obs}}, \mathbf{R}_i). \tag{5}$$

This propensity score is the conditional probability that the participant i is assigned to the t group given the *observed* values of covariates $\mathbf{X}_i^{\mathbf{obs}}$, *and* the missingness patterns $\mathbf{R}_i$. $e(\mathbf{X}_i^{\mathbf{obs}}, \mathbf{R}_i)$ has similar properties to $e(\mathbf{X}_i)$. First, $e(\mathbf{X}_i^{\mathbf{obs}}, \mathbf{R}_i)$ is the function of $\mathbf{X}_i^{\mathbf{obs}}$ and $\mathbf{R}_i$, which can be used to equate the distributions of $\mathbf{X}_i^{\mathbf{obs}}$ and $\mathbf{R}_i$ between t and c groups. Second, if the strong ignorability assumption in Equations 4 and 5 holds for $\mathbf{X}_i^{\mathbf{obs}}$, the strong ignorability assumption also holds for $e(\mathbf{X}_i^{\mathbf{obs}}, \mathbf{R}_i)$ (Rosenbaum & Rubin, 1984, Appendix B). In order to produce an unbiased estimate of *ATE* or *ATT* based on sample data, the propensity score estimation model needs to be correctly specified (Drake, 1993) and the strong ignorability assumption must be met.

Pearl and colleagues have recently begun to develop a general approach to missing data problems (Mohan & Pearl, 2014; Mohan, Pearl, & Tian, 2013; Pearl, 2009; Pearl & Mohan, 2013; Thoemmes & Mohan, 2015). They extended Pearl's analysis of causal inference in terms of directed acyclic graphs (DAG) to formally identify two issues in missing data analysis. In this approach, researchers need to specify a DAG model for the missing data process. Two issues arise in identifying these models: (a) whether there are consistent estimates for the parameters in the model, and (b) whether the model has testable implications. These issues are termed recoverability and testability, respectively. This approach of Pearl and colleagues confirms that, when data are missing completely at random (MCAR) or MAR, there are consistent estimators (e.g., multiple imputation; discussed later) for the model parameters. MCAR implies that $\mathbf{X}_i^{\mathbf{mis}}$ does not depend on other measured or unmeasured variables (random). MAR implies that $\mathbf{X}_i^{\mathbf{obs}}$ accounts for the missingness in $\mathbf{X}_i^{\mathbf{mis}}$ (random, conditional on the measured variables). Tests of whether data are MCAR (e.g., Little, 1988) and a partial test of whether data are MAR (Potthoff, Tudor, Pieper, & Hasselblad, 2006) are available. This approach also shows that there are consistent estimators for some, but not all, models when data are missing not at random (MNAR). MNAR implies that $\mathbf{X}_i^{\mathbf{mis}}$ depends on the would-be unobserved values of the data that are missing (nonrandom). In applied research, this approach may present formidable challenges because researchers rarely have a correctly specified DAG model of the missing data process and the theoretical results are asymptotic. Therefore, we focus on approaches based on Rosenbaum and Rubin (1984), which have been far more extensively studied and for which information about their performance and guidelines can be offered for applied researchers.

In the following sections, propensity score estimation methods are reviewed. We begin with the traditional logistic regression method when all $\mathbf{X}_i$ are completely observed, followed by a review of approaches using logistic regression when $\mathbf{X}_i$ have missing values. We then very briefly consider general location modeling, a method whose use is limited to a very small number of covariates. Finally, we review newer propensity score estimation methods based on machine learning (classification trees, random forests, and generalized boosted modeling). Table 2 summarizes methods for estimating propensity scores with missing data in the covariates. Panel 1 presents approaches for addressing incomplete covariates within the logistic regression method: separate missingness patterns, imputation with constant plus missingness indicators, and multiple imputation. Panel 2 presents general location modeling. Panels 3 to 5 present the classification trees, random forests, and generalized boosted modeling methods. These methods adapt these approaches for incomplete covariates: separate missingness patterns, multiple imputation, and surrogate decision.

Propensity Score Estimation: Logistic Regression

Logistic regression was the first, and continues to be the most widely used, method to estimate propensity scores with complete data (Rosenbaum & Rubin, 1983, 1984; Thoemmes & Kim, 2011). When the row vector $\mathbf{X}_i$ is completely observed for all participants, the logistic regression equation to estimate the propensity score can be expressed as Equation 6,

$$ln\left(\frac{\hat{e}(\mathbf{X}_i)}{1 - \hat{e}(\mathbf{X}_i)}\right) = f(\mathbf{X}_i), \tag{6}$$

where $ln(\cdot)$ is the natural logarithm function; $ln\left(\frac{\hat{e}(\mathbf{X}_i)}{1-\hat{e}(\mathbf{X}_i)}\right)$ is known as the logit; and $f(\mathbf{X}_i)$ is the function describing the relationship between the logit and $\mathbf{X}_i$. The simplest form of $f(\mathbf{X_i})$, the linear regression model, is typically used. For $\mathbf{X}_i = (X_{1i}, X_{2i}, X_{3i}, X_{4i})$,

Table 2
Summary of Different Propensity Score Estimation Methods and Approaches to Handle Incomplete Covariates

Propensity score estimation method	Approaches to handle incomplete covariates
(1) Logistic regression • Linear functional form. • If linear model is believed to be misspecified, polynomial and interaction terms of covariates can be added.	(a) Separate missingness patterns • Estimate propensity scores separately in each missingness pattern. • Accounts for all missingness patterns by default. • Requires correct model specification for each missingness pattern. • Logistic regression is infeasible when number of participants is less than the number of observed covariates in any missingness pattern. • Produces unstable *ATE* or *ATT* estimates when few participants in missingness pattern. (b) Imputation with constant plus missingness indicators • Missing values are imputed using arbitrary constant values. • Polynomial/interaction terms of covariates are first computed, then are imputed. • Distinct missingness indicators are included in estimation model. • Requires correct model specification of the observed values of covariates *and* missingness indicators. • Produces stable *ATE* or *ATT* estimates given realistic sample size. (c) Multiple imputation • Multiple imputed data sets are created in the imputation phase. • Imputation procedure (data augmentation, MICE) accounts for all missingness patterns. • Imputation procedure requires incomplete covariates are missing at random. • Imputation procedure requires correct imputation model specifications. • Propensity score analysis is conducted in each imputed data set, then results are pooled. • Not necessary to include missingness indicators in propensity score estimation model. • Requires correct model specification for observed *and* missing values of covariates. • Produces stable *ATE* or *ATT* estimates given realistic sample size.
(2) General location modeling • Handles nominal and continuous covariates. • Often involves a huge number of parameters, resulting in inaccurate or unstable propensity scores.	(a) Expectation maximization (EM) algorithm • Accounts for all missingness patterns by default. • Requires that incomplete covariates are missing at random. • Requires correct model specification for observed *and* missing values of covariates.
(3) Classification trees • Recursively partitions participants into two nodes based on selected covariates. • Sets up nonlinear relationships between covariates and group assignment. • Significance test procedure for covariate selection is suggested to reduce selection bias. • Does not require model specification. • Requires specifying estimation stopping criteria. • Produces unstable trees model and propensity scores. • Often estimates incorrect propensity scores, when the true model has mainly linear effects of covariates. • Not as effective as random forests and generalized boosted modeling in reducing the bias of the *ATE* and *ATT* estimates.	(a) Separate missingness patterns (see above) • Classification trees is feasible when number of participants is less than the number of observed covariates in any missingness pattern. (b) Multiple imputation (see above) • Data augmentation and MICE are not suitable for the nonlinear relationships in classification trees. Alternative imputation procedures that handle nonlinear relationship in the imputation models (e.g., classification and regression trees) should be used. (c) Surrogate decision • Distinct missingness indicators are included in estimation model. • In estimation, observed values of covariates and the missingness indicators are used. • In each node, other covariates that resemble the original node are identified and used to estimate propensity cores, when participants have missing values in original node. • Requires correct model specification of the observed values of covariates *and* missingness indicators.
(4) Random forests • Multiple random subsamples are drawn from original data. Estimate propensity scores in each subsample using classification trees. Random forests propensity scores equal to the average of all classification trees propensity scores. • Believed to be less sensitive to the stopping criteria. • Produces more stable propensity scores. • More capable of handling linear effects of covariates.	(a) Separate missingness patterns (see above) • Same as that in classification trees. (b) Multiple imputation (see above) • Same as that in classification trees. (c) Surrogate decision • Same as that in classification trees.

Table 2 (*continued*)

Propensity score estimation method	Approaches to handle incomplete covariates
Suggested procedures: i. Sampling without replacement with the subsample size equals .632 times the original sample sizes. ii. Significance test procedure for covariate selection. iii. In each classification tree, propensity scores are estimated for the participants who are *not* in the random subsample for estimating the trees model. iv. Use with weighting methods for equating groups.	
(5) Generalized boosted modeling • Iterative process to estimate propensity scores using regression trees. Regression trees sets up nonlinear relationships by recursively partitioning participants into two nodes based on selected covariates. • Produce more stable propensity scores. • More capable of handling linear effects of covariates. • Increasing iterations can lead to model overfitting. • Determine the optimal iterations by summary of balance diagnostics of covariates. • Theoretically, similar performance to random forests in reducing the bias of *ATE* estimate.	(a) Separate missingness patterns (see above) • Same as that in classification trees. (b) Multiple imputation (see above) • Same as that in classification trees. (c) Surrogate decision • Same as that in classification trees.

Note. ATE = average treatment effect; ATT = average treatment effect on the treated; MICE = multivariate imputation by chained equations.

$$ln\left(\frac{\hat{e}(\mathbf{X}_i)}{1-\hat{e}(\mathbf{X}_i)}\right) = b_0 + b_1X_{1i} + b_2X_{2i} + b_3X_{3i} + b_4X_4\,. \qquad (7)$$

Balance diagnostics (discussed later) can be used to detect misspecification of the propensity score estimation model. When the balance diagnostics indicate that the linear logistic regression model is misspecified, polynomial and interaction terms created from the covariates (e.g., X_{1i}^2, $X_{2i} \times X_{3i}$) should be added to the logistic regression model (for empirical examples, see Dehejia & Wahba, 2002; Diaz & Handa, 2006; Harder, Stuart, & Anthony, 2010).[3]

When the $\mathbf{X}_i$ include missing values, three different approaches described in the next three sections can be taken to estimate the propensity score $e(\mathbf{X}_i^{\mathbf{obs}}, \mathbf{R}_i)$: separate missingness patterns, imputation with a constant plus missingness indicators, and multiple imputation. A successful approach to estimate $e(\mathbf{X}_i^{\mathbf{obs}}, \mathbf{R}_i)$ needs to meet three criteria. First, all of the observed values of covariates $\mathbf{X}_i^{\mathbf{obs}}$ need to be used in the estimation. Second, all of the missingness patterns $\mathbf{R}_i$ need to be accounted for in the estimation. Third, the estimation model for $e(\mathbf{X}_i^{\mathbf{obs}}, \mathbf{R}_i)$ must be correctly specified (Drake, 1993). We now compare these approaches based on these three criteria.

Separate Missingness Patterns

In the separate missingness patterns approach (D'Agostino, Lang, Walkup, Morgan, & Karter, 2001; Rosenbaum & Rubin, 1984, Appendix B), participants' propensity scores are estimated *separately* for each distinct missingness pattern $\mathbf{R}_i$. Table 3 provides an illustration for four covariates $\mathbf{X}_i = (X_{1i}, X_{2i}, X_{3i}, X_{4i})$ and five different missingness patterns. The linear regression model in Equation 7 is used separately for each pattern. This approach meets the three evaluation criteria. In practice, this approach cannot be conducted when the number of participants is less than the number of observed covariates in any one missingness pattern—the logistic regression equation cannot be estimated. This problem commonly occurs when there are many distinct arbitrary missingness patterns in the sample. Even when the number of participants in the missingness pattern is greater than the number of covariates, but the subsample size is small (e.g., $n = 20$), *ATE* estimates will have very large standard errors (Qu & Lipkovich, 2009).

Imputation With Constant Plus Missingness Indicators

There are three steps in the imputation with constant plus missingness indicators approach (D'Agostino et al., 2001; Haviland, Nagin, & Rosenbaum, 2007; Rosenbaum, 2010, pp. 193–194, 240–242). First, indicators of the missingness patterns $\mathbf{R}_i$ for $\mathbf{X}_i$ are created. Second, the missing values of the covariates are imputed with arbitrary but constant values. Third, the values of the propensity scores $e(\mathbf{X}_i^{\mathbf{obs}}, \mathbf{R}_i)$ are estimated with logistic regression using the imputed covariates plus missingness indicators. Consider, again, the example in Table 3. There are four indicators of the missingness patterns, R_{1i}, R_{2i}, R_{3i}, and R_{4i}. We impute the missing values of X_{1i}, X_{2i}, X_{3i}, and X_{4i} using the arbitrary constant value 0. The propensity scores $e(\mathbf{X}_i^{\mathbf{obs}}, \mathbf{R}_i)$ are estimated using logistic regression, where covariates $(X_{1i}, X_{2i}, X_{3i}, X_{4i})$ and missingness indicators $(R_{1i}, R_{2i}, R_{3i}, R_{4i})$ are linearly related to the logit (Equation 8):

$$ln\left(\frac{\hat{e}(\mathbf{X}_i^{\mathbf{obs}},\ \mathbf{R}_i)}{1-\hat{e}(\mathbf{X}_i^{\mathbf{obs}},\ \mathbf{R}_i)}\right) = b_0 + b_1X_{1i} + b_2X_{2i} + b_3X_{3i} + b_4X_{4i} + b_5R_{1i} + b_6R_{2i} + b_7R_{3i} + b_8R_{4i}. \qquad (8)$$

This approach meets the three evaluation criteria. The first criterion is met because all $\mathbf{X}_i^{\mathbf{obs}}$ are utilized; the missing values are imputed. Identical estimated propensity scores will be obtained

[3] Some researchers have employed stepwise regression, which aims to prevent model overfit by discarding covariates with nonsignificant regression coefficients. Although model overfit leads to unfavorable outcomes (discussed later), stepwise regression is not recommended (Hill et al., 2011). Stepwise regression favors covariates that are strongly related to the actual group assignment, but may discard covariates more weakly related to group assignment that are strongly related to the outcome variable (Hill et al., 2011).

Table 3
Illustration of Separate Missingness Patterns Approach

	Missingness patterns $\mathbf{R}_i$				
	R_{1i}	R_{2i}	R_{3i}	R_{4i}	Regression model
1	1	1	1	1	$= b_0^{(1)} + b_1^{(1)}X_{1i} + b_2^{(1)}X_{2i} + b_3^{(1)}X_{3i} + b_4^{(1)}X_{4i}$
2	1	1	1	0	$= b_0^{(2)} + b_1^{(2)}X_1 + b_2^{(2)}X_2 + b_3^{(2)}X_3$
3	1	0	1	0	$= b_0^{(3)} + b_1^{(3)}X_1 + b_3^{(3)}X_3$
4	0	1	1	1	$= b_0^{(4)} + b_2^{(4)}X_2 + b_3^{(4)}X_3 + b_4^{(4)}X_4$
5	0	1	0	1	$= b_0^{(5)} + b_2^{(5)}X_2 + b_4^{(5)}X_4$

Note. When $\mathbf{R}_i = 1$, the covariate is observed. When $\mathbf{R}_i = 0$, the covariate is not observed. The superscript for the regression coefficients refer to each distinct missingness pattern.

regardless of the constant value chosen for imputation. A propensity score $e(\mathbf{X}_i^{\mathbf{obs}}, \mathbf{R}_i)$ estimated by "[logistic] regression is invariant under affine transformation of the [confounders]" (Rosenbaum, 2010, pp. 194, 241). An affine transformation is a linear transformation of the variable $X'_{1i} = a\,X_{1i} + b$, where X'_{1i} is participant i's affine transformed variable for X_{1i}, and a and b are any arbitrary numbers. In the imputation example, $a = 1.0$ and $b = 0.0$ when X_{1i} is observed, and $a = 0.0$ and $b = 0.0$ when X_{1i} is missing. D'Agostino et al. (2001) and Rosenbaum (2010) recommend choosing reasonable imputed values such as the means of the covariates. The second criterion can be met when the linear relationships of the missingness indicators (e.g., R_{1i} to R_{4i} in Equation 8) represent the full set of the missingness patterns $\mathbf{R}_i$. If the linear relationships of R_{1i} to R_{4i} are insufficient to (closely) meet the criterion, interaction effects of the missingness indicators can be included (Rosenbaum, 2010, p. 194). The third criterion is met when the relationships of the observed values of confounders $\mathbf{X}_i^{\mathbf{obs}}$ and missingness indicators are correctly specified.

Two potential technical issues are associated with this approach. First, different covariates can give rise to identical missingness indicators. Identical missingness indicators need to be deleted from the propensity score estimation model to avoid perfect collinearity. Second, for polynomial or interaction terms of the incomplete covariates, the polynomial or interaction term must be computed first. Then, the missing values for the polynomial or interaction terms are imputed with one arbitrary value.

Multiple Imputation

Multiple imputation can also be used to estimate propensity scores with missing data on the covariates (Crowe, Lipkovich, & Wang, 2010; Hill, 2004; Hughes, Chen, Thoemmes, & Kwok, 2010; Qu & Lipkovich, 2009). Multiple imputation is a procedure that creates m copies of the data set in which participants' missing values of the covariates $\mathbf{X}_i^{\mathbf{mis}}$ have been filled in with plausible values. Each data set is then analyzed and their results combined. Multiple imputation can produce unbiased results for $\mathbf{X}_i$ under MCAR or MAR situations. It can often reduce bias when data are MNAR. Enders (2010) and van Buuren (2012) provide full presentations.

Once the m imputed data sets are created, participants' propensity scores $e(\mathbf{X}_i^{\mathbf{obs}}, \mathbf{R}_i)$ are estimated using the same propensity score estimation model separately in each imputed data set (Hill, 2004). In the estimation model, the relationship between each full copy of the data set $\mathbf{X}_i$ (including $\mathbf{X}_i^{\mathbf{obs}}$ and imputed $\mathbf{X}_i^{\mathbf{mis}}$) and the actual group assignment is specified. Then *ATE* or *ATT* (as desired) and its corresponding standard error are estimated using the same propensity score equating method in each imputed data set. Finally, the *ATE* (or *ATT*) results from the m imputed data sets are combined using Rubin's (1987) rules, yielding a single summary that includes the *ATT* (or *ATE*) parameter estimate, corresponding standard error, and significance test result.

In order to correctly estimate the propensity scores $e(\mathbf{X}_i^{\mathbf{obs}}, \mathbf{R}_i)$, this approach needs to meet the assumptions of multiple imputation. First, missing values of covariates $\mathbf{X}_i^{\mathbf{mis}}$ are MAR or MCAR. If this assumption is not met (i.e., $\mathbf{X}_i^{\mathbf{mis}}$ are MNAR), incorrect propensity scores $e(\mathbf{X}_i^{\mathbf{obs}}, \mathbf{R}_i)$ will be estimated. One remedy is to include the missingness indicators of R_i in the propensity score estimation model. When the MAR assumption is not met, this remedy provides less biased *ATE* estimates than the propensity score estimation model without the missingness indicators (Qu & Lipkovich, 2009).

Second, the imputation models must be correctly specified. Even if the MAR assumption is met, misspecification of the imputation models can lead to incorrect imputed values for $\mathbf{X}_i^{\mathbf{mis}}$. As a result, the propensity scores $e(\mathbf{X}_i^{\mathbf{obs}}, \mathbf{R}_i)$ will not be properly estimated (White, Royston, & Wood, 2011). Two algorithms for multiple imputation are available: data augmentation (Schafer, 1997) and multivariate imputation by chained equations (MICE; Raghunathan, Lepkowksi, Van Hoewyk, & Solenbeger, 2001). Data augmentation assumes multivariate normality; it performs best when variables are normal or can be transformed to be normal. MICE offers imputation models for different types of incomplete covariates (e.g., nominal, ordered categorical, count, continuous). However, MICE has disadvantages relative to data augmentation in its computational burden and more limited convergence diagnostics (van Buuren, 2012, p. 117).

A final specification issue is that the imputation models offered by data augmentation and MICE assume linear relationships among covariates. If polynomial and/or interaction effects exist in the propensity score model that involve incomplete covariates, specialized imputation procedures must be used (Vink & van Buuren, 2013).

Comparisons and Summary

We have presented three approaches to estimate the propensity scores $e(\mathbf{X}_i^{\mathbf{obs}}, \mathbf{R}_i)$: separate missingness patterns, imputation with constant plus missingness indicators, and multiple imputation. Each of these three approaches can potentially meet the three evaluation criteria necessary to correctly estimate $e(\mathbf{X}_i^{\mathbf{obs}}, \mathbf{R}_i)$. However, in practice, difficulties may arise because the assumptions of the three approaches may not be met:

1. The separate missingness patterns approach can produce unbiased *ATE* (or *ATT*) estimates when $\mathbf{X}_i^{\mathbf{mis}}$ are MCAR or MAR.[4] It will not yield estimates using logistic regression when the number of participants is fewer than the number of observed covariates in any one missingness pattern. It will produce large standard errors of *ATE* (or *ATT*) estimates when the number of participants in a missingness pattern is small. The separate missingness patterns approach also requires the correct estimation model for *every* missingness pattern $\mathbf{R}_i$.

2. The imputation with constant plus missingness indicators approach can produce unbiased *ATE* (or *ATT*) estimates when $\mathbf{X}_i^{\mathbf{mis}}$ are MCAR or MAR. It does not account for all possible missingness patterns by default, potentially leading to incorrect estimates. The imputation with constant plus missingness indicators approach also requires correct specification for observed values $\mathbf{X}_i^{\mathbf{obs}}$ and missingness indicators.[5]

3. The multiple imputation approach can produce unbiased *ATE* (or *ATT*) estimates when $\mathbf{X}_i^{\mathbf{mis}}$ are MCAR or MAR. It requires correct specification for $\mathbf{X}_i$ (both $\mathbf{X}_i^{\mathbf{obs}}$ and $\mathbf{X}_i^{\mathbf{mis}}$). If the assumptions underlying multiple imputation are not met, the imputed values can potentially affect the estimated propensity scores.

In conclusion, we recommend the separate missingness pattern approach when it is feasible and when there are a sufficiently large number of participants in each missingness pattern. This approach is the least prone to incorrectly estimate $e(\mathbf{X}_i^{\mathbf{obs}}, \mathbf{R}_i)$ among the three approaches (D'Agostino et al., 2001). However, when the number of participants in each separate missingness pattern is small, the imputation with constant plus missingness indicators and multiple imputation approaches are preferred. The two approaches require different assumptions to properly estimate the propensity scores. Another related approach in the literature (D'Agostino, 2004; D'Agostino & Rubin, 2000), known as *general location modeling*, is *not* recommended in typical propensity score applications; it produces unstable propensity scores when $\mathbf{X}_i$ includes more than a few covariates (Belin, Hu, Young, & Grusky, 1999).

Machine Learning Propensity Score Estimation Methods

We now present three machine learning methods to the estimation of propensity scores with missing data: classification trees, random forests, and generalized boosted modeling. Previously, we discussed that polynomial and interaction terms of the covariates can be added to the logistic regression model. Nevertheless, the logistic regression method still assumes linearity between the covariates and their polynomial and interaction terms and the logit of the propensity scores. On the other hand, the three machine learning methods estimate arbitrary nonlinear relationships between predictors (here, the covariates) and a binary dependent variable (here, assignment to the *t* or *c* group).

Classification Trees

Classification trees is an early machine learning procedure that approximates the relationships between predictors and a binary dependent variable (Breiman, Friedman, Stone, & Olshen, 1984; Morgan & Sonquist, 1963). We initially describe the application of classification trees when all covariates are completely observed, and then describe the complications in the procedures when covariates have missing values. Classification trees provide the foundation for the two newer methods that have superseded it; this approach is presented for pedagogic reasons.

Procedure. Figure 1 illustrates a propensity score analysis based on the classification trees method using a real data example. Classification trees use a recursive process to estimate the tree model from top to bottom in Figure 1. At each step of the process, a covariate (potential confounder) and its cutoff value are selected to classify the participants into two nodes. At the top of Figure 1, the first recursive process selects the covariate AVGZF1 and its cutoff value ($= -0.69$). The participants who score ≤ -0.69 are classified to the left node below, whereas the participants who score > -0.69 are classified to the right node below. The process repeats in each node to build the trees model further down, until the stopping criteria are met (discussed in the next paragraphs). Consider the left node immediately below AVGZF1. At this node, the participants who score $\leq$ 3.67 on the covariate TACH1 are classified to the leftmost bottom Node A (terminal node) that includes 108 participants. The bar charts of the bottom nodes reflect the composition of the terminal node in terms of the *c* group participants (marked as 0, light bar) and *t* group participants (marked as 1, dark bar). In the leftmost bottom node (Terminal Node A), there are slightly fewer *c* group participants (43.5%) than *t* group participants (56.5%). Participants in the group who are classified into terminal node A (AVGZF1 ≤ -0.69 and TACH1 $>$ 3.67) are assigned a propensity score of .565, equal to the proportion of *t* group participants in terminal node A. As illustrated in Figure 1, the relationships of the covariates and actual group assignment are nonlinear in the classification trees method. Further, interaction effects between covariates are not represented in the same manner in the classification trees and logistic regression methods (Strobl, Malley, & Tutz, 2009, pp. 328–330). Figure 1 illustrates one type of interaction effect in the classification trees method. Below the first recursive process of node AVGZF1, the

[4] The separate missingness and imputation with constant plus missingness indicators approaches fully meet the strong ignorability condition defined by Equations 4 and 5. Recall that under this condition, the potential outcomes are independent of the treatment condition, conditional on both the observed covariates and missing data patterns. The multiple imputation approach requires slightly stronger assumptions that the imputation model needs to be correctly specified.

[5] D'Agostino (2004, p. 166) and D'Agostino and Rubin (2000, p. 753) show that the missingness patterns $\mathbf{R}_i$ can be omitted from propensity score estimation, when the actual group assignment Z_i is independent of the missingness patterns $\mathbf{R}_i$. Researchers may examine this condition using a contingency table test, examining the effect size. We recommend that the missingness patterns $\mathbf{R}_i$ be included in propensity score estimation unless there is strong support that this additional condition is met.

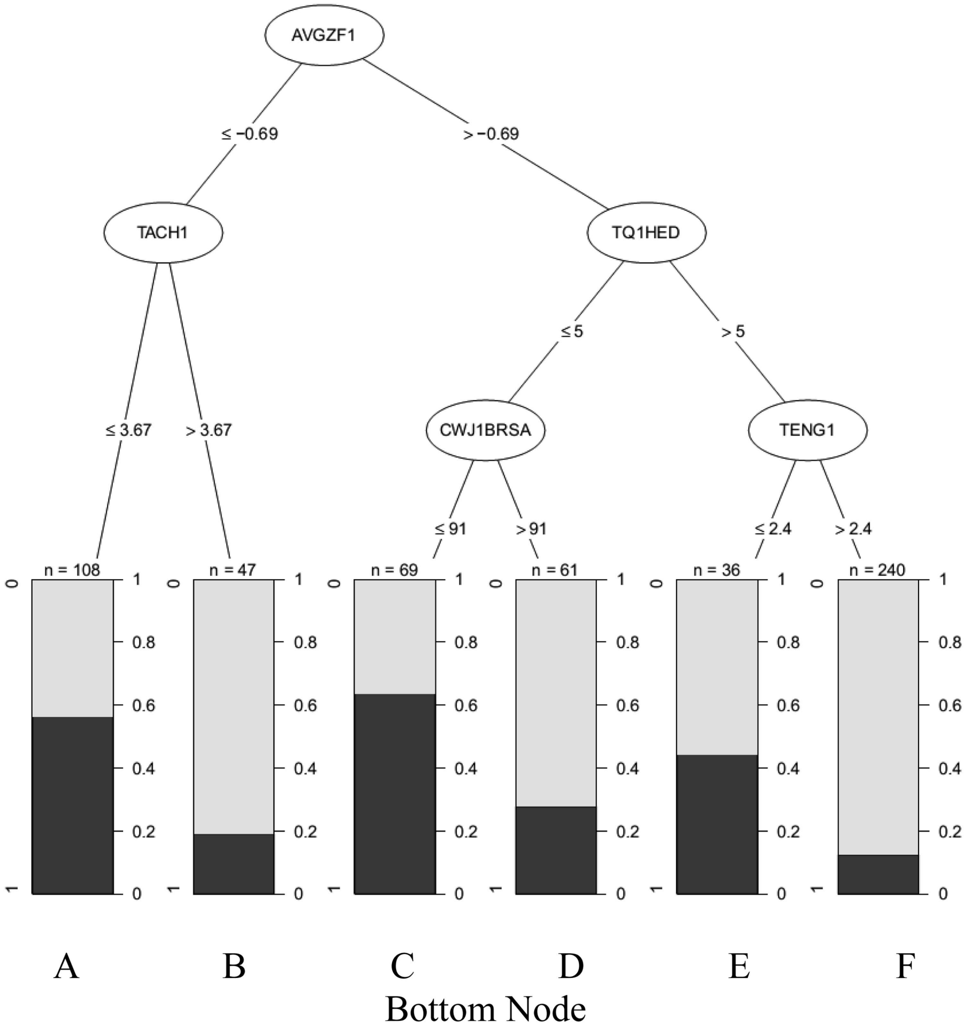

Figure 1. An example of a classification trees model. The variables in ellipses are the nodes to classify the participants into two groups. At each bottom node (A to F) of the classification trees model, a bar chart is displayed. The number of participants at each bottom node is shown at the top of the bar. The bar chart represents the proportion of control group participants (= 0) with the light bar and the proportion of treatment group participants (= 1) with the dark bar at each bottom node.

second recursive process selects the covariate TACH1 in the left branch, and a different covariate TQ1HED in the right branch.

The most important decision in the classification trees model is how to select the covariate and its cutoff value in the recursive process. The traditional procedure is to search for the partition (covariate and its cutoff value) that maximizes the reduction of an impurity measure such as the Gini index after the partitioning. However, this procedure is biased toward selecting predictors with many categories, continuous predictors, and predictors with many missing values (e.g., Berk, 2008, pp. 137, 150; Hastie, Tibshirani, & Friedman, 2009, p. 310). In propensity score estimation, when covariates have missing values, it is not recommended.

The newer, recommended procedure is a three-phase conditional significance test, which, in the context of propensity score analysis, reduces the covariate selection bias (Hothorn, Hornik, & Zeileis, 2006; Strobl, Boulesteix, Zeileis, & Hothorn, 2007; see Berk, 2008, p. 137). In the first phase of the procedure, separate significance tests of the contribution of each covariate to group assignment, controlling for all other covariates, are conducted. In the second phase, an omnibus significance test of the association of the set of all covariates with group assignment is conducted. If the null hypothesis of no association is rejected, the covariate with the smallest p value in the first phase is selected for partitioning; otherwise, no partitioning is conducted. In the third phase, the cutoff value of the selected covariate is chosen using another significance test. The null hypothesis is that, at the tested cutoff value, the proportions of the t group participants who are partitioned between the two nodes are identical (available in R software package party; Hothorn, Hornik, et al., 2006).

Approaches for incompletely observed covariates. When the covariates are not completely observed, three approaches can be used in conjunction with the classification trees method that carry over to the newer random forests and generalized boosting modeling: separate missingness patterns, multiple imputation, and surrogate decision. We previously described and evaluated the separate missingness patterns and multiple imputation approaches in the context of estimation of propensity scores using logistic regression. However, distinctions become apparent when these first two approaches are used in conjunction with the classification trees method:

1. For the separate missingness patterns approach, the classification trees method *is* feasible when there are fewer participants than the number of observed covariates in any one missingness pattern.

2. For multiple imputation, appropriate imputation procedures have been developed (Burgette & Reiter, 2010; Doove, Van Buuren, & Dusseldorp, 2014; van Buuren, 2012, p. 84) that use classification trees and random forests (discussed later) as imputation models. The standard data augmentation and MICE algorithms do not properly capture the nonlinear relationships among the covariates. In addition, these imputation procedures are applicable to big data sets having a large samples size and large number of covariates.

The third approach, surrogate decision (Berk, 2008, pp. 132–135; Hapfelmeier, Hothorn, Ulm, & Strobl, 2014; Hastie et al., 2009, p. 311; Hothorn, Hornik, et al., 2006, p. 658; Strobl et al., 2009, p. 342), includes both the covariates and missingness indicators for $\mathbf{R}_i$ as variables in the regression trees model (Cham, Hughes, West, & Im, 2015). In the significance test procedure to estimate the trees model, all of the observed values of each covariate and its missingness indicator are utilized. The challenge is to estimate participants' propensity scores when they have missing values on some nodes that determine partitioning in the trees model. In the surrogate decision procedure, for each node in the trees model, other covariates and their cutoff values are identified that approximate the decision rule of the original node to partition the participants. These proxy covariates are ranked according to their ability to make decisions resembling those of the original node. Participants with missing values on the original node are classified by the first-ranked proxy covariate if this covariate is observed. If not, the second-ranked proxy covariate is used if this covariate is observed, and so forth.

Summary. Compared with logistic regression, the classification trees method is attractive because it does not require propensity score model specification. However, simulation studies show that it provides less bias reduction in the *ATE* and *ATT* estimates than random forests and generalized boosted modeling (Lee, Lessler, & Stuart, 2010; Setoguchi, Schneeweiss, Brookhart, Glynn, & Cook, 2008). Hence, we turn to random forests and generalized boosted modeling, which partially overcome this limitation.

Random Forests

Random forests is a machine learning method that builds on the foundation of the classification trees method, addressing a number of its potential disadvantages (Breiman, 2001). It involves several steps. The first step is to draw B random subsamples from the data. We suggest a large value of B (e.g., 1,000), because increasing the value of B does not lead to the disadvantage of model overfitting (discussed later; Breiman, 2001, Theorem 2.1; see also Berk, 2008, pp. 201–202; Strobl et al., 2009, p. 344). Based on a simulation study, Strobl et al. (2007) showed that sampling *without replacement* to create each subsample with size equal to 0.632 times the original sample size minimized covariate selection bias.

The second step is to estimate a classification trees model separately in each random subsample using the identical model specification. The three-phase significance test procedure is used to select a covariate and its cutoff value for a node; this minimizes the bias of the *ATE* estimate (Cham, 2013). The random forests method offers an extra option in the recursive partitioning process. In every recursive partitioning process, a random subsample of m covariates is drawn out of all k covariates. This option can reduce the variability of the estimated propensity scores across repeated sampling (Berk, 2008, pp. 194–198; Breiman, 2001; Hastie et al., 2009, p. 588; Strobl et al., 2009, p. 333). How to choose the optimal value of m (number of covariates out of k available covariates) in propensity score estimation is an open question. In other random forest applications, Hastie et al. (2009, p. 592) and Strobl, Boulesteix, Kneib, Augustin, and Zeileis (2008, p. 10) suggest $m = \sqrt{k}$ is the optimal choice. For propensity score analysis, applied research by Cham et al. (2015) found that using all covariates (i.e., $m = k$) led to the greatest reduction of bias in the *ATT* estimate. A simulation study (Cham, 2013) found that the optimal choice of m to maximize the bias reduction in the *ATE* estimate depends on the original sample size. Future research is needed to clarify the optimal basis for this choice.

The third step is to estimate participants' propensity scores. Each participant's propensity score is estimated in each tree model in each sample. Two options are available:

1. Propensity scores can be estimated for all participants in the original data (including those who were not in the random subsample used to estimate the classification trees model).

2. Propensity scores can be estimated only for the participants who are in the "holdout sample" that is not used to estimate the classification trees model.

There are two advantages of Option 2 relative to Option 1. The variability of the propensity scores is less biased (Berk, 2008, p. 189; Strobl et al., 2009, p. 335), and bias in the estimate of the *ATE* is reduced (Cham, 2013). Once propensity scores have been estimated, the random forests propensity score for each participant i is set equal to the unweighted average of the estimated propensity scores for participant i in the B random subsamples (Hastie et al., 2009, p. 283; Strobl et al., 2009, p. 334). When the covariates have missing values, random forests can utilize the same three approaches as the classification trees method to estimate propensity scores $e(\mathbf{X}_i^{\mathbf{obs}}, \mathbf{R}_i)$: separate missingness patterns, multiple imputation, and surrogate decision.

Random forests can address some of the potential limitations of classification trees approach. First, random forests is believed to be less sensitive to the stopping criteria (e.g., minimum node size, maximum depth of trees model) because multiple tree models are estimated (Berk, 2008, p. 235; Hastie et al., 2009, pp. 596–598). Second, the random forests model is far more stable than a single classification trees model (Berk, 2008, pp. 180–182; Hastie et al., 2009, Chapter 15; Strobl et al., 2009, p. 332), so that the variability of the estimated propensity scores is reduced. Third, the compiling of multiple tree models into one solution smooths the nonlinear relationships, which helps to better approximate smooth functional (e.g., linear) relationships of covariates (Bühlmann & Yu, 2002). In addition, the two extra options for random forests (draw random subsample of *m* out of all *k* covariates in every recursive partitioning process; estimate propensity scores for the participants who are not in the random subsample for the trees model) have the potential to further reduce the bias of the *ATE* estimate.

Simulation studies have supported the superiority of the random forests relative to the classification trees method for propensity score estimation. When all the covariates are completely observed and when the weighting method (e.g., Schafer & Kang, 2008; West et al., 2014) is used to equate the groups, Lee et al. (2010) found that random forests is more successful than classification trees in reducing the bias of the *ATE* estimate. Cham (2013) found that random forests propensity scores are more successful in reducing the bias of *ATT* estimate when the weighting equating method rather than the nearest neighbor matching equating method is used (discussed later). We recommend the use of random forests with the weighting method than the classification trees method to estimate propensity scores. In the random forests method, careful consideration should be given to the number of covariates to be randomly selected in each recursive partitioning process. Random forests is available in the cforest command in R software package party (Hothorn, Bühlmann, Dudoit, Molinaro, & Van Der Laan, 2006; Strobl et al., 2007, 2008).

Generalized Boosted Modeling

Generalized boosted modeling is another machine learning method based on the foundation of the classification trees method (Berk, 2008, Chapter 6; Bühlmann & Hothorn, 2007; Hastie et al., 2009, Chapter 10). McCaffrey, Ridgeway, and Morral (2004) proposed the application of this method to propensity score estimation and developed the R software package twang (Ridgeway, McCaffrey, Morral, Ann, & Burgette, 2014). The first step of generalized boosted modeling is to define the loss function for actual group assignment predicted by the covariates. When there are two groups (*t* and *c*), the negative binomial log-likelihood function is used (Hastie et al., 2009, Equation 10.18; McCaffrey et al., 2004, Equation B1),

$$\text{loss function} = -l(Z_i\,,\ e(\mathbf{X}_i)) = -Z_i\, ln(e(\mathbf{X}_i)) + (1 - Z_i)(1 - ln(e(\mathbf{X}_i))), \qquad (9)$$

where $l(Z_i, e(\mathbf{X}_i))$ is participant *i*'s binomial log-likelihood function of group assignment Z_i (= 1 for *t*, = 0 for *c*), given the propensity score $e(\mathbf{X}_i)$. Equation 9 evaluates the propensity score $e(\mathbf{X}_i)$ by how well it is assigned large probabilities when $Z_i = 1$, and small probabilities when $Z_i = 0$, on average (McCaffrey et al., 2004, p. 424). To facilitate the computational algorithm, the loss function can be reexpressed as Equation 10,

$$\text{loss function} = ln(1 + exp(-2\, Z_i' g(\mathbf{X}_i)), \qquad (10)$$

where $g(\mathbf{X}_i) = 0.5 \times ln(e(\mathbf{X}_i)/(1 - e(\mathbf{X}_i)))$, which is one half of the logit of the propensity score; $exp(\cdot)$ is the exponential function; and Z_i' is the effect coded Z_i (= 1 for *t*, = −1 for *c*). The second step is to calculate the participants' negative gradient, which is defined as the negative first-order partial derivative of the loss function (Equation 10) with respect to $g(\mathbf{X}_i)$, given the value of the estimated $g(\mathbf{X}_i)$. For the first iteration, the start value for the participants' estimated $g(\mathbf{X}_i)$ is the proportion of *t* group participants in the sample.

The third step is to use the regression trees analysis to estimate the relationships between the covariates $\mathbf{X}_i$ and the participants' negative gradient (dependent variable). Regression trees analysis can be viewed as a "continuous dependent variable" version of the classification trees method. In the R software package twang, the regression trees analysis selects the covariate and its cutoff value that will serve as nodes by minimizing the mean squared residuals of the negative gradient in the two nodes (McCaffrey et al., 2004, p. 424). The participants' estimated negative gradient is equal to the means of participants' negative gradient in the corresponding bottom nodes to which they are classified. The analyst must specify a stopping criterion for the regression trees, such as the maximum depth of the trees model. McCaffrey et al. (2004, p. 409) suggest using depth equal to 4. A random subsample of the original data can be used to estimate the regression trees in each iteration. Sampling without replacement to create each subsample with the size equal to 50% of the original data size is recommended to reduce model bias and variance (Hastie et al., 2009, pp. 365–367; McCaffrey et al., 2004, pp. 408–409).

The fourth step is to update the estimate of $g(\mathbf{X}_i)$ (half of the logit of the propensity score). The current estimate of $g(\mathbf{X}_i)$ and the current estimate of the negative gradient value $h(\mathbf{X}_i)$ from the regression trees are combined: $g(\mathbf{X}_i)^{new} = g(\mathbf{X}_i) + \nu\, h(\mathbf{X}_i)$, where ν is the user-specified shrinkage parameter that may range from 0.0 to 1.0. McCaffrey et al. (2004, p. 409) suggested $\nu = 0.0005$ for propensity score estimation. Smaller values of ν increases the accuracy of the estimated propensity scores at a cost of requiring a larger number of iterations (Bühlmann & Hothorn, 2007, p. 480; Hastie et al., 2009, p. 365). We suggest trying multiple values to investigate the sensitivity of the choice of ν for propensity score estimation.

Generalized boosted modeling iterates until the algorithm reaches the designated maximum number of iterations. The estimated propensity score is obtained based on the estimate of $g(\mathbf{X}_i)$ in the last iteration. Through the iteration process, generalized boosted modeling produces more stable models and propensity score estimates become better approximations to smooth (e.g., linear) functions. Choosing the number of iterations is a challenge. Increasing the number of iterations can result in overfitting the generalized boosted model, with the negative consequence that estimated propensity scores are biased toward 0 or 1 (Berk, 2008, pp. 262, 272–273; Mease, Wyner, & Buja, 2007). This bias leads to undesirable properties in propensity score analysis (discussed later). McCaffrey et al. (2004) proposed stopping criteria to determine the optimal number of iterations; however, these criteria have clear limitations:

1. The summary of the balance diagnostic statistics only represents in part the balance of covariates' multivariate distributions and may not represent the bias reduction of the *ATE* or *ATT* estimate.

2. The R package twang currently offers only weighting methods to equate the propensity scores between the *t* and *c* groups; other equating methods (e.g., matching) are not available.

Lee et al. (2010) found in a simulation study that the default setting of twang successfully reduces the bias of the *ATT* estimate. When covariates have missing values, generalized boosted modeling can utilize the same three approaches (separate missingness patterns, multiple imputation, surrogate decision) as classification trees to estimate $e(\mathbf{X}_i^{\mathbf{obs}}, \mathbf{R}_i)$.

Comparisons and Summary for Incompletely Observed Covariates

There are three approaches for classification trees, random forests, and generalized boosted models with incompletely observed covariates to estimate propensity scores. We recommend the separate missingness pattern approach because it is the least prone to incorrectly estimate $e(\mathbf{X}_i^{\mathbf{obs}}, \mathbf{R}_i)$ (D'Agostino et al., 2001). It is suitable for MCAR or MAR covariates. For these three machine learning methods, the separate missingness pattern approach is feasible when the number of covariates is more than the number of participants in a missingness pattern. Surrogate decision and multiple imputation approaches require different assumptions to properly estimate the propensity scores. The surrogate decision approach is suitable for MCAR or MAR covariates. It does not account for all possible missingness patterns by default, potentially leading to incorrect estimates. Multiple imputation is suitable for MCAR or MAR covariates. If the assumptions are not met, the imputed values can potentially affect the estimated propensity scores.

Balance Diagnostics, Equating Methods, Sensitivity Analysis

Balance diagnostics are used to examine how well the estimated propensity score model has performed in equating the *t* and *c* groups. Austin (2009, 2011) provides a comprehensive review of balance diagnostics when all covariates are completely observed. Here, we (a) briefly review balance diagnostics to detect overfitting of the propensity score estimation model, and then (b) consider balance diagnostics for incomplete covariates and missingness patterns.

When a propensity score estimation model is overfitted, an important negative consequence is that the estimated propensity scores will be biased toward the minimum value (0) or the maximum value (1). As a result, the common support region (overlap) of the estimated propensity scores between the *t* and *c* groups is reduced (Harder et al., 2010; Hill, Weiss, & Zhai, 2011). As the size of the common support region is reduced, the generalization of *ATE* or *ATT* results becomes increasingly limited (Caliendo & Kopeinig, 2008; Harder et al., 2010; Thoemmes & Kim, 2011). With extreme overfitting, all *c* group participants can have estimated propensity scores near 0.0, and all *t* group participants can have estimated propensity scores near 1. This situation violates the strong ignorability assumption that participant must have a nonzero probability of being in the *t* (or *c*) group (Equation 2; Rosenbaum & Rubin, 1983; West et al., 2014). The balance diagnostic used to examine model overfitting is to compare the common support region (overlap) of estimated propensity scores between groups before and after propensity score equating. A variety of graphical displays can be used for this purpose: box plots, empirical cumulative density functions, histograms, kernel density plots, and quantile-quantile plots (see Austin, 2009, 2011). If the graphical displays show an insufficient common support region after equating, model overfitting may have occurred.

When covariates are incompletely observed, the propensity scores $e(\mathbf{X}_i^{\mathbf{obs}}, \mathbf{R}_i)$ are estimated based on the observed values of covariates $\mathbf{X}_i^{\mathbf{obs}}$ and the missingness patterns $\mathbf{R}_i$. If the strong ignorability assumption is met and the propensity score estimation model is correctly estimated, $\mathbf{X}_i^{\mathbf{obs}}$ and $\mathbf{R}_i$ will be equal between the *t* and *c* groups, on average, across repeated samples (Equation 4; Rosenbaum & Rubin, 1984; West et al., 2014). For single univariate covariates, Austin (2009, 2011) provides a comprehensive review of the common balance diagnostic statistics and graphical displays, such as the standardized mean difference, variance ratio, and quantile-quantile plot. These balance diagnostics utilize all of the observed values of the single covariate by default. Researchers can directly apply these existing balance diagnostics to incomplete covariates. In addition, these balance diagnostics can be applied to examine the polynomial and interaction terms of the covariates (complete or incomplete; Austin, 2009; Hill et al., 2011; Stuart, 2010; West et al., 2014). To date, the use of balance diagnostics to examine multivariate distributions of the covariates is still uncommon.

Balance diagnostics for the missingness patterns depend on the approach used to handle the missingness patterns. The separate missingness and multiple imputation approaches address all missingness patterns by default. Therefore, balance diagnostics are not needed for missingness patterns with these approaches. In contrast, the imputation with constant plus missingness indicators approach (for logistic regression) and surrogate decision approach (for classification trees, random forests, generalized boosted modeling) require correct specification of the missingness indicators. We recommend that researchers examine the balance between the *t* and *c* groups of the missingness indicators and their two-way interaction terms using the balance diagnostics described earlier in this section for univariate covariates. However, note that when the missing data rate is low (e.g., 10%), these diagnostics may not be sensitive enough to detect improvement or degradation in the balance of the missingness patterns (Cham et al., 2015).

Given satisfactory results of the checks of the balance diagnostics for the estimated propensity scores, the final procedure is to equate the propensity scores between the *t* and *c* groups. Equating methods, such as matching, subclassification, weighting, and analysis of covariance, are described in the review articles mentioned at the beginning of this article. For example, the nearest neighbor matching procedure pairs one *t* group participant with one (or more) *c* group participant(s) with equal or the closest propensity scores. Note that in the multiple imputation approach, each imputed data set may have a different number of *t* and *c* group participants following matching (Hughes et al., 2010). This may add an additional layer of complexity in applied research. Once the *t* and *c* groups have been equated, the *ATE* or *ATT* is estimated. Weighting assigns each *t* and *c* group participant a weight based on the estimated propensity scores. When estimating the *ATE* or *ATT* and its standard error, participants' weights are taken account of by the estimator. Common choices of the estimators are related to those used in survey sampling procedures—here, weighting by a function of the inverse of the propensity scores (e.g., Horvitz-Thompson estimation; Schafer & Kang, 2008; Stuart, 2010; West et al., 2014).

A final procedure is to conduct sensitivity analysis to probe how the *ATE* or *ATT* results might change if there are unmeasured confounder(s). Sensitivity analysis is not affected by whether measured putative confounders have missing values. Approaches to conducting sensitivity analyses are described by Rosenbaum (2010) and Stuart (2010).

Illustrative Example

Our illustrative example provides a comparison of how different propensity score estimation methods for incomplete covariates perform in reducing the bias of the average treatment effect (*ATE*) estimate. We created a simulated data set based on a study of the effect of grade retention (holding students back in grade a year) in elementary school on students' motivation to complete high school assessed in Grade 9 (Cham et al., 2015). The "treatment" (*t*) group was the students who were retained in grade in elementary school. The control (*c*) group was the students who were continuously promoted to the next grade.

Participants and Method

In the original research, participants were recruited from a larger sample of academically at-risk students from three school districts. Students were defined as being academically at risk if they scored below the median score on a state-approved district-administered measure of literacy at entrance to elementary school. A comprehensive set of covariates (potential confounders) were measured in Grade 1 before retention, chosen based on prior research on grade retention and academic achievement. Moser et al. (2012) present the details of the original study.

To create the data set for our illustrative example, we used the original data of the actual group assignment (retained or promoted) and 10 of the measured covariates, selected because they were highly correlated with motivation to complete high school (Cham et al., 2015). Using only participants with complete data (98 retained; 305 promoted), we generated the outcome variable by setting up the treatment-outcome model (see Appendix for details). We used a linear regression model in which the actual group assignment and covariates are linearly related to the generated outcome variable with moderate correlations ($r \sim 0.3$). This constituted our complete data set. We then generated missing values for five of the covariates. The Appendix shows the details of the generated missingness patterns and missing data mechanisms of the covariates. The simulated data had five distinct missingness patterns, each with similar number of participants ($n \sim 80$). Four covariates were generated to be MAR; their missingness was accounted for by the five remaining complete covariates. One covariate was generated to be MCAR.

Several properties of this illustrative example should be noted. First, the true propensity score is unknown, because the covariates and actual group assignment were from the original data. Second, the imputation with constant plus missingness indicators and the surrogate decision approaches can account for all missingness patterns by specifying linear relationships of the distinct missingness indicators in the propensity score estimation model. Third, the missing data mechanisms of the covariates (MCAR and MAR) are known, fulfilling one assumption of multiple imputation. But the true imputation models of the covariates are unknown.

Analysis

The following propensity score estimation methods were investigated: logistic regression, random forests, and generalized boosted modeling. For each method, three approaches were used to address the incomplete covariates: (a) separate missingness patterns, (b) imputation with constant plus missingness indicators (for logistic regression) or surrogate decision (for random forests and generalized boosted modeling), and (c) multiple imputation.

In the logistic regression method, all covariates are linearly related to the logit of actual group assignment. All logistic regression analyses were conducted using R glm (Version 3.1.1; R Core Team, 2014). For this example, the separate missingness patterns approach was feasible with logistic regression. In the imputation with constant plus missingness indicators approach, all distinct missingness indicators were linearly related to the logit of actual group assignment. This specification accounts for all missingness patterns (i.e., no misspecification). In the multiple imputation approach, the MICE algorithm implemented in the R package mice (Version 2.22; van Buuren & Groothuis-Oudshoorn, 2011) was used to impute missing values. Linear regression was used to impute continuous incomplete covariates, and linear logistic regression was used to impute binary incomplete covariates. Twenty imputed data sets were generated. No remedies were applied; all covariates met the MAR assumption in the simulated data.

In the random forests method, we adapted the model specification by Cham et al. (2015). Table 4 lists the details of the model specification. All random forests analyses were conducted using the cforest command in R software package party (Version 1.0–15; Hothorn, Bühlmann, et al., 2006; Strobl et al., 2008). In the multiple imputation approach, we used the imputation algorithm by Doove et al. (2014) in the R software package mice. The imputation algorithm used classification trees to impute incomplete categorical covariates and regression trees to impute incom-

Table 4
Specification of Random Forests Propensity Score Estimation Method in Illustrative Example

Random forests model specification
1. Number of classification trees = 1,000.
2. Sampling without replacement to draw random subsamples.
3. Each subsample has size equal to .632 times the original sample size.
4. Details of the significance test procedure for covariate selection:
(a) Permutation test statistic in quadratic form was used.
(b) Minimum *p* value as stopping criterion for the estimation process was not set.
5. All covariates are used in each covariate selection process.
6. Minimum number of participants in each terminal node = 5.
7. Maximum depth of classification trees model = No set value.
8. In each classification tree, propensity scores are estimated for the participants who are *not* in the random subsample for estimating the tree model.

plete continuous covariates. In the generalized boosted modeling method, the R software package twang was used (Version 1.4–0; Ridgeway et al., 2014). The recommended model specifications in the software manuals were used. We used the mean Kolmogorov–Smirnov test statistics of the covariates comparing the retained and promoted students as balance diagnostics to select the optimal number of iterations. In the multiple imputation approach, we used the same imputation algorithm as we did for random forests.

We used the inverse probability of treatment weights (IPTW; e.g., Schafer & Kang, 2008; West et al., 2014) to equate each set of estimated propensity scores between the retained and promoted students. In this procedure, each retained student was given a weight of $1/\hat{e}(\mathbf{X}_i^{\mathbf{obs}}, \mathbf{R}_i)$, whereas each promoted student was given a weight of $1/(1 - \hat{e}(\mathbf{X}_i^{\mathbf{obs}}, \mathbf{R}_i))$. Participants' weights were incorporated using survey sampling weighting procedures (R software package survey, Version 3.30–3; Lumley, 2004, 2014). To investigate the overfitting of the propensity scores, we produced box plots of each set of propensity scores between the retained and promoted students, before and after IPTW equating (Austin, 2009, 2011). To investigate the balance of the observed values of the covariates, we examined standardized mean differences (*SMD*s) of the covariates between the retained and promoted students, both before and after IPTW equating using each set of propensity scores (Austin, 2009, 2011). In the multiple imputation approach, we present the diagnostics from the first imputed data set. We recommend that applied researchers examine the diagnostics for each imputed data set to ensure satisfactory balance. We did not investigate the balance of the missingness patterns, because all approaches for incomplete covariates account for all missingness patterns for the example. Finally, we estimated the average treatment effect (*ATE*) and its standard error, 95% confidence interval, and performed the hypothesis test.

Results and Discussion

Figure 2 shows the box plots of each set of propensity scores of the students who were retained in grade (*t*) and promoted to the next grade (*c*), before and after IPTW equating. In the multiple imputation approach, we used the first imputed data set to produce the plots. The box plots before IPTW equating show the size of the common support region (overlap) of the propensity scores between two groups, which determines the generalizability of the results of the propensity score analysis. Among the three estimation methods, random forests propensity scores had the largest common support region. Among the missingness approaches, the separate missingness patterns approach produced the largest common support region, followed by surrogate decision and multiple imputation. Logistic regression produced a smaller common support

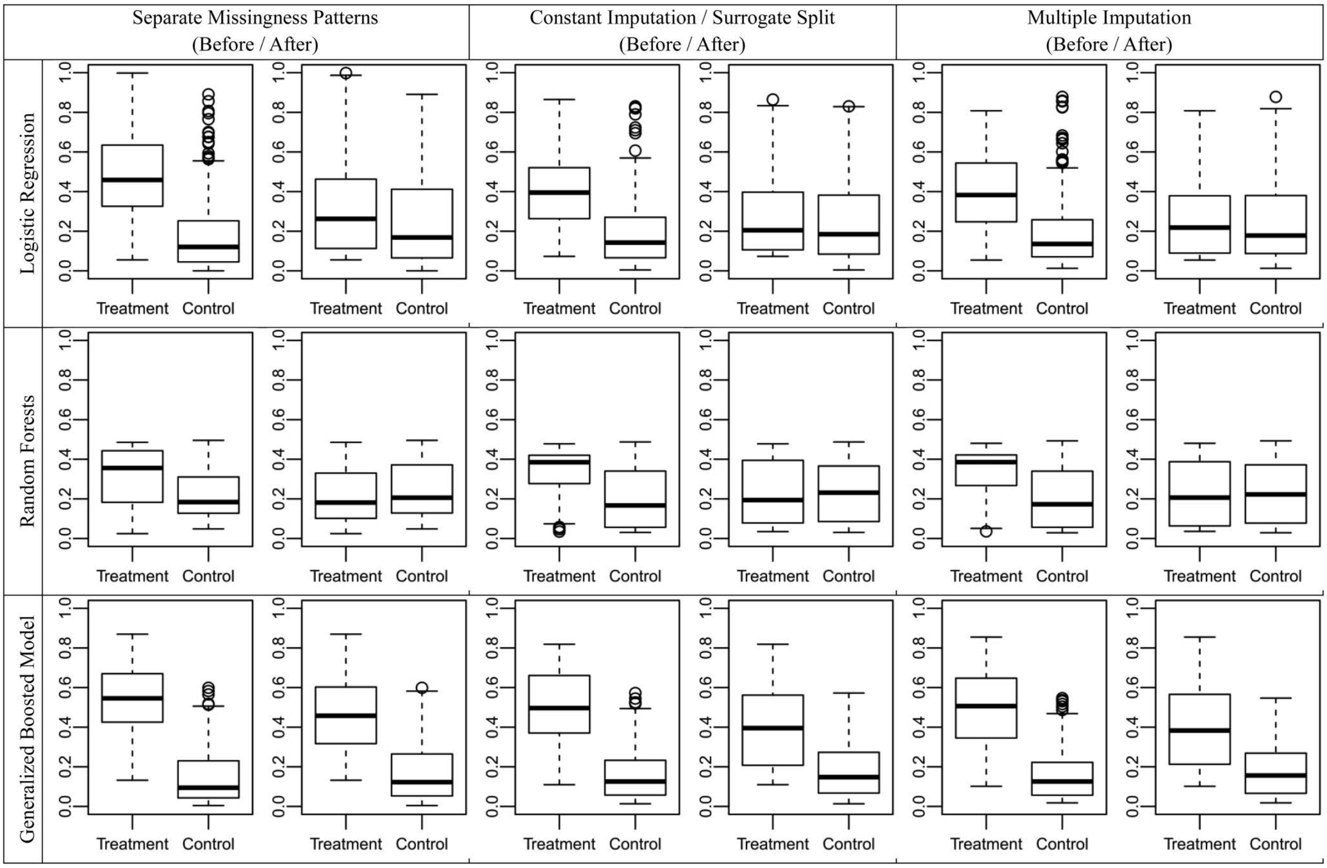

Figure 2. Box plots of the estimated propensity scores between retained (treatment) and promoted (control) students, before (left) and after (right) inverse probability of treatment weights equating. Constant imputation means imputation with constant plus missingness indicators.

region (with the order of the approaches being imputation with constant ≈ multiple imputation > separate missingness patterns). Generalized boosted modeling produced the smallest common support region (all three approaches to missing data performed similarly), possibly because model overfitting might have occurred. After IPTW equating, successful equating should produce box plots with similar propensity score distributions between the two groups. The logistic regression and random forests methods using any of the three approaches led to successful equating of the two groups. However, generalized boosted modeling using any of the three approaches did not successfully equate the retained and promoted groups. Therefore, it is expected that generalized boosted modeling propensity scores would produce the least balance between the covariates as well as the least reduction of bias of the *ATE* estimate.

Figure 3 shows the dot plots of the *SMD* of the observed values of covariates between retained and promoted students, before and after IPTW equating by each set of propensity scores. *SMD* = 0.0 indicates perfect balance. Rubin (2001) suggested a guideline that *SMD*s within −0.25 and 0.25 are acceptable (marked by two vertical lines in Figure 3). Before IPTW equating, all but one of the 10 covariates had *SMD*s outside the suggested range. For logistic regression, all three approaches reduced the covariates' *SMD*s into the suggested range. For random forests, the surrogate decision and multiple imputation approaches reduced the covariates' *SMD*s into suggested range; their performance was similar to that of logistic regression. The separate missingness patterns approach failed to reduce half of the covariates with problematic *SMD*s into the suggested range. For generalized boosted modeling, all three approaches failed to reduce the *SMD*s into the suggested range.

Table 5 shows the *ATE* results. We compared the *ATE* estimate from each set of propensity scores with the "true" complete sample *ATE* estimate. The complete sample *ATE* estimate is obtained by estimating the correct treatment-outcome model in the generated data set (= 8.75). The complete sample *ATE* estimate is close to the true population value (= 9.00; see Appendix). We evaluated the *ATE* estimate by calculating the percent bias of the *ATE* estimate compared with the complete sample *ATE* estimate. Per-

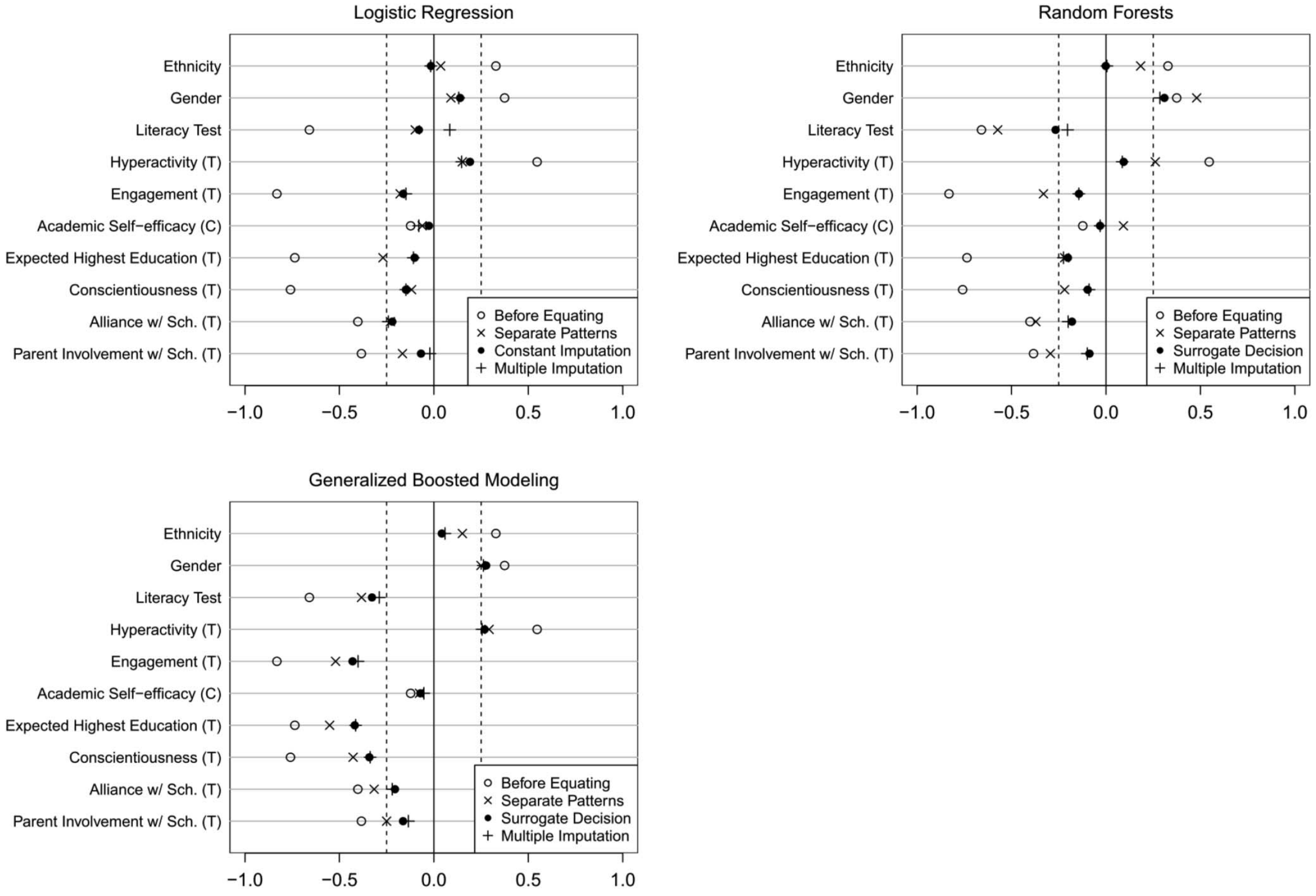

Figure 3. Dot plots of the standardized mean differences (*SMD*s) of the covariates between retained and promoted students, before and after inverse probability of treatment weights equating by each set of propensity scores. SMD = 0.0 indicates no differences. The two dashed vertical lines at −0.25 and 0.25 indicate Rubin's (2001) suggested range for *SMD*. On the vertical axis, "(C)" means the covariate is child-reported, and "(T)" means the cofounder is teacher-reported. In the box in bottom-right, "Separate Patterns" mean separate missingness patterns approach, and "Constant Imputation" means imputation with constant plus missingness indicators approach. Sch. = school.

Table 5
Average Treatment Effect (ATE) Results of the Illustrative Example

Estimation method	Approach for incomplete covariates	*ATE* estimate	Standard error	95% confidence interval	Percent bias to true sample *ATE* estimate
Logistic regression	Separate patterns	7.29*	.95	[5.42, 9.16]	−16.7%
	Constant imputation	8.15*	1.17	[5.86, 10.43]	–6.9%
	Multiple imputation	8.08*	1.05	[5.89, 10.27]	–7.7%
Random forests	Separate patterns	5.85*	1.41	[3.08, 8.62]	–33.1%
	Surrogate decision	8.03*	1.25	[5.58, 10.49]	–8.2%
	Multiple imputation	8.10*	1.23	[5.52, 10.67]	–7.5%
Generalized boosted modeling	Separate patterns	6.50*	.95	[4.64, 8.37]	–25.7%
	Surrogate decision	7.27*	1.04	[5.24, 9.30]	–16.9%
	Multiple imputation	7.55*	1.07	[5.33, 9.76]	–13.7%

Note. "Separate patterns" means separate missingness patterns approach, and "Constant imputation" means imputation with constant plus missingness indicators approach. True sample *ATE* estimate is obtained by estimating the true treatment-outcome model in the generated data set (= 8.75; true population value = 9.00).
* $p < .01$.

cent bias less than 10% was defined as being acceptable. For logistic regression, the imputation with constant plus missingness indicators and multiple imputation approaches produced unbiased *ATE* estimates (bias = −6.9% and −7.7%, respectively). Both approaches also yielded similar standard errors of the *ATE* estimate. For the separate missingness patterns approach, the *ATE* was underestimated (bias = −16.7%), possibly because the number of participants (~80) in each missingness pattern was insufficient to produce stable *ATE* estimate. This approach produced the smallest standard error of the *ATE* estimate, a problematic finding given the magnitude of bias. For random forests, the results were similar to those of the logistic regression approach. The surrogate decision and multiple imputation approaches produced unbiased *ATE* estimates (bias = −8.2% and −7.5%, respectively). Both approaches produced similar standard errors for the *ATE* estimates; these quantities were slightly larger than those produced by logistic regression. The separate missingness patterns approach greatly underestimated the *ATE* (bias = −33.1%). Note that this bias was greater than the same approach using logistic regression (bias = −16.7%). This result suggests that random forests method is less stable than the logistic regression method for propensity score analysis. For generalized boosted modeling, all three approaches underestimated *ATE* (bias < −10%). The results likely occurred because the estimated propensity scores and the observed values of covariates were unsuccessfully equated between the retained and promoted students.

The results of this illustrative example provide several potential insights for applied research. First, the separate missingness patterns approach requires a large number of participants in each missingness pattern. In this example with 10 covariates, 80 participants in each missingness pattern was *not* sufficient to produce stable *ATE* estimates. When this approach was used in conjunction with the random forests method, the results were even less stable than with the logistic regression method. When the missingness patterns can be correctly specified, imputation with constant plus missingness indicators (using logistic regression) and surrogate decision (using random forests) approaches produce unbiased and stable *ATE* estimates. When the covariates are MCAR or MAR, multiple imputation approaches using logistic regression and random forests produce unbiased and stable *ATE* estimates as well. In this example, generalized boosted modeling using the recommended model specifications failed to produce unbiased *ATE* estimates. More research is needed to understand the specifications under which generalized boosted modeling could produce unbiased *ATE* estimates. Austin (2012) has shown that increasing the depth of the regression trees models can improve the bias reduction. Last but not least, the sample size and number of covariates of the illustrative sample are far smaller than in typical big data analyses that may utilize classification trees, random forests, and generalized boosted modeling. These three approaches are easily utilized and may be more successful in propensity score analyses with big data sets.

Missing Values in the Posttest Outcome Variable

In practice, the posttest outcome variable can also have missing values. When the outcome variable is MCAR or MAR due to the observed values of covariates $\mathbf{X}_i^{\mathbf{obs}}$, two approaches may be used to produce unbiased *ATE* or *ATT* estimates. The first approach is to impute both the incompletely observed outcome variable and the incompletely observed covariates using the multiple imputation approach (Hill, 2004; Qu & Lipkovich, 2009). With a small sample size, this approach may not be optimal when only a few covariates in a comprehensive set of $\mathbf{X}_i^{\mathbf{obs}}$ are related both to the outcome variable and to the missingness of the outcome variable (Hardt, Herke, & Leonhart, 2012).

A second approach can be used with the separate missingness patterns, imputation with constant plus missingness indicators, and the surrogate decision approaches. After the participants are successfully equated by the estimated propensity scores, the incompletely observed outcome variable is accounted for by multiple imputation or the expectation-maximization (EM) algorithm. In this approach, we can include only those covariates that are related to the outcome variable and to the missingness of the MAR outcome variable in the imputation model or the EM algorithm. We prefer this second approach to the first approach, because of flexibility in specifying the variables to account for the MAR outcome variable.

When an MNAR outcome variable is measured at a single time point, there are some approaches to handle the MNAR outcome (e.g., selection model and pattern mixture model; Enders, 2010). When the MNAR outcome variable is measured longitudinally and is analyzed using latent growth models, longitudinal extensions of these approaches are also available (Enders, 2011; Yang & Maxwell, 2014). However, these analyses "rely heavily on untestable assumptions, and even relatively minor violations of these assumptions can introduce substantial bias" (Enders, 2011, p. 1).

Conclusion

The primary goal of this article was to provide a comprehensive review of propensity score analysis with incompletely observed covariates. We reviewed the theory of propensity score analysis when all the covariates are completely observed and when the covariates have missing values. We presented different propensity score estimation methods for both completely observed and incompletely observed covariates, including machine learning techniques (classification trees, random forests, and generalized boosted modeling). These estimation methods and approaches were evaluated according to criteria based on the theory of propensity score analysis, and were compared based on the findings in the literature (see Table 2 for a summary). We have also considered these methods and approaches when the covariates are MCAR and MAR.

We used an illustrative example with simulated data based on an actual data set to compare these methods and approaches. Our results offer some insights and suggestions for applied researchers for making choices among the estimation methods and approaches to handle incomplete covariates. Nonetheless, we recommend that applied researchers use different estimation methods and compare their results, looking for convergence that can help bracket the size of the true effect (e.g., Harder et al., 2010). More extensive simulation studies are needed to provide thorough evidence on which to base recommendations for applied researchers.

We presented balance diagnostics to help detect model overfitting in the propensity score estimation models and applied these diagnostics in the illustrative example. We also noted that balance diagnostics for incompletely observed covariates and missingness patterns are the same as those for completely observed covariates. Propensity score equating methods and sensitivity analyses do not depend on whether the covariates have missing values or not. Given space limitations, we refer readers to other comprehensive review articles for the details of these topics. In the illustrative example, we used one equating method (IPTW) to estimate *ATE*. In applied research, we suggest that researchers consider carefully the choice of the combination of the estimation method and equating method, because different combinations may lead to different *ATE* or *ATT* results (Cham, 2013; Harder et al., 2010; Hill et al., 2011). We have also discussed treatments for an incompletely observed posttest outcome variable in the MCAR, MAR, and MNAR situations.

References

Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine, 28,* 3083–3107. http://dx.doi.org/10.1002/sim.3697

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research, 46,* 399–424. http://dx.doi.org/10.1080/00273171.2011.568786

Austin, P. C. (2012). Using ensemble-based methods for directly estimating causal effects: An investigation of trees-based G-computation. *Multivariate Behavioral Research, 47,* 115–135. http://dx.doi.org/10.1080/00273171.2012.640600

Belin, T. R., Hu, M. Y., Young, A. S., & Grusky, O. (1999). Performance of a general location model with an ignorable missing-data assumption in a multivariate mental health services study. *Statistics in Medicine, 18,* 3123–3135. http://dx.doi.org/10.1002/(SICI)1097-0258(19991130)18:22<3123::AID-SIM277>3.0.CO;2-2

Berk, R. A. (2008). *Statistical learning from a regression perspective*. New York, NY: Springer.

Breiman, L. (2001). Random forests. *Machine Learning, 45,* 5–32. http://dx.doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. Boca Raton, FL: Chapman and Hall.

Bühlmann, P., & Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science, 22,* 477–505. http://dx.doi.org/10.1214/07-STS242

Bühlmann, P., & Yu, B. (2002). Analyzing bagging. *Annals of Statistics, 30,* 927–961. http://dx.doi.org/10.1214/aos/1031689014

Burgette, L. F., & Reiter, J. P. (2010). Multiple imputation for missing data via sequential regression trees. *American Journal of Epidemiology, 172,* 1070–1076. http://dx.doi.org/10.1093/aje/kwq260

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys, 22,* 31–72. http://dx.doi.org/10.1111/j.1467-6419.2007.00527.x

Cham, H. (2013). *Propensity score estimation with random forests* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (Accession Order No. UMI 3567836)

Cham, H., Hughes, J. N., West, S. G., & Im, M. H. (2015). Effect of retention in elementary grades on grade 9 motivation for educational attainment. *Journal of School Psychology, 53,* 7–24. http://dx.doi.org/10.1016/j.jsp.2014.10.001

Chapin, F. (1947). *Experimental designs in sociological research*. New York, NY: Harper.

Crowe, B. J., Lipkovich, I. A., & Wang, O. (2010). Comparison of several imputation methods for missing baseline data in propensity scores analysis of binary outcome. *Pharmaceutical Statistics, 9,* 269–279. http://dx.doi.org/10.1002/pst.389

D'Agostino, R., Jr. (2004). Propensity scores with missing data. In A. Gelman & X. L. Meng (Eds.), *Applied Bayesian modeling and causal inference from incomplete-data perspectives* (pp. 163–174). Hoboken, NJ: Wiley.

D'Agostino, R., Jr., Lang, W., Walkup, M., Morgan, T., & Karter, A. (2001). Examining the impact of missing data on propensity score estimation in determining the effectiveness of self-monitoring of blood glucose (SMBG). *Health Services and Outcomes Research Methodology, 2,* 291–315. http://dx.doi.org/10.1023/A:1020375413191

D'Agostino, R., Jr., & Rubin, D. B. (2000). Estimating and using propensity scores with partially missing data. *Journal of the American Statistical Association, 95,* 749–759. http://dx.doi.org/10.1080/01621459.2000.10474263

Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *The Review of Economics and Statistics, 84,* 151–161. http://dx.doi.org/10.1162/003465302317331982

Diaz, J. J., & Handa, S. (2006). An assessment of propensity score matching as a nonexperimental impact estimator: Evidence from Mexico's PROGRESA program. *The Journal of Human Resources, 41,* 319–345. http://dx.doi.org/10.3368/jhr.XLI.2.319

Doove, L. L., Van Buuren, S., & Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis, 72,* 92–104. http://dx.doi.org/10.1016/j.csda.2013.10.025

Drake, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics, 49,* 1231–1236. http://dx.doi.org/10.2307/2532266

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.

Enders, C. K. (2011). Missing not at random models for latent growth curve analyses. *Psychological Methods, 16,* 1–16. http://dx.doi.org/10.1037/a0022640

Hapfelmeier, A., Hothorn, T., Ulm, K., & Strobl, C. (2014). A new variable importance measure for random forests with missing data. *Statistics and Computing, 24,* 21–34. http://dx.doi.org/10.1007/s11222-012-9349-1

Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods, 15,* 234–249. http://dx.doi.org/10.1037/a0019623

Hardt, J., Herke, M., & Leonhart, R. (2012). Auxiliary variables in multiple imputation in regression with missing X: A warning against including too many in small sample research. *BMC Medical Research Methodology, 12,* 184.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York, NY: Springer. http://dx.doi.org/10.1007/978-0-387-84858-7

Haviland, A., Nagin, D. S., & Rosenbaum, P. R. (2007). Combining propensity score matching and group-based trajectory analysis in an observational study. *Psychological Methods, 12,* 247–267. http://dx.doi.org/10.1037/1082-989X.12.3.247

Hill, J. (2004). *Reducing bias in treatment effect estimation in observational studies suffering from missing data*. Institute for Social and Economic Research and Policy Working Paper 04–01, Columbia University, New York, NY. Retrieved from http://academiccommons.columbia.edu/download/fedora_content/download/ac:129152/CONTENT/2004_01.pdf

Hill, J., Weiss, C., & Zhai, F. (2011). Challenges with propensity score approaches in a high-dimensional setting and a potential alternative. *Multivariate Behavioral Research, 46,* 477–513. http://dx.doi.org/10.1080/00273171.2011.570161

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81,* 945–960. http://dx.doi.org/10.1080/01621459.1986.10478354

Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A., & van der Laan, M. J. (2006). Survival ensembles. *Biostatistics (Oxford, England), 7,* 355–373. http://dx.doi.org/10.1093/biostatistics/kxj011

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics, 15,* 651–674. http://dx.doi.org/10.1198/106186006X133933

Hughes, J. N., Chen, Q., Thoemmes, F., & Kwok, O. M. (2010). An investigation of the relationship between retention in first grade and performance on high stakes tests in third grade. *Educational Evaluation and Policy Analysis, 32,* 166–182. http://dx.doi.org/10.3102/0162373710367682

Kelcey, B. (2011). Covariate selection in propensity scores using outcome proxies. *Multivariate Behavioral Research, 46,* 453–476. http://dx.doi.org/10.1080/00273171.2011.570164

Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine, 29,* 337–346.

Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association, 83,* 1198–1202. http://dx.doi.org/10.1080/01621459.1988.10478722

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software, 9,* 1–19. Retrieved from https://www.jstatsoft.org/article/view/v009i08/

Lumley, T. (2014). Survey: Analysis of complex survey samples (Version 3.30-3) [Software]. Retrieved from https://cran.r-project.org/web/packages/survey/index.html

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. http://dx.doi.org/10.1037/1082-989X.9.4.403

Mease, D., Wyner, A. J., & Buja, A. (2007). Boosted classification trees and class probability/quantile estimation. *Journal of Machine Learning Research, 8,* 409–439. Retrieved from http://dl.acm.org/citation.cfm?id=1248675

Mohan, K., & Pearl, J. (2014). On the testability of models with missing data. In S. Kaski & J. Corander (Eds.), *JMLR Workshop and Conference Proceedings (Vol. 33): Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics* (pp. 643–650). Retrieved from http://jmlr.csail.mit.edu/proceedings/papers/v33/

Mohan, K., Pearl, J., & Tian, J. (2013). Graphical models for inference with missing data. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing System 26 (NIPS-2013)* (pp. 1277–1285). Retrieved from http://papers.nips.cc/paper/4899-graphical-models-for-inference-with-missing-data

Morgan, J. N., & Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association, 58,* 415–434. http://dx.doi.org/10.1080/01621459.1963.10500855

Moser, S. E., West, S. G., & Hughes, J. N. (2012). Trajectories of math and reading achievement in low-achieving children in elementary school: Effects of early and later retention in grade. *Journal of Educational Psychology, 104,* 603–621. http://dx.doi.org/10.1037/a0027571

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511803161

Pearl, J., & Mohan, K. (2013). *Recoverability and testability of missing data: Introduction and summary of results (R-417)*. Los Angeles, CA: University of California, Los Angeles, Computer Science Department.

Potthoff, R. F., Tudor, G. E., Pieper, K. S., & Hasselblad, V. (2006). Can one assess whether missing data are missing at random in medical studies? *Statistical Methods in Medical Research, 15,* 213–234. http://dx.doi.org/10.1191/0962280206sm448oa

Qu, Y., & Lipkovich, I. (2009). Propensity score estimation with missing values using a multiple imputation missingness pattern (MIMP) approach. *Statistics in Medicine, 28,* 1402–1414. http://dx.doi.org/10.1002/sim.3549

Raghunathan, T. W., Lepkowksi, J. M., Van Hoewyk, J., & Solenbeger, P. A. (2001). Multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology, 27,* 85–95.

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Ridgeway, G., McCaffrey, D., Morral, A., Ann, B., & Burgette, L. (2014). twang: Toolkit for weighting and analysis of nonequivalent groups (Version 1.4–0) [Software]. Retrieved from http://cran.r-project.org/web/packages/twang/index.html

Rosenbaum, P. R. (2010). *Design of observational studies*. New York, NY: Springer-Verlag. http://dx.doi.org/10.1007/978-1-4419-1213-8

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70,* 41–55. http://dx.doi.org/10.1093/biomet/70.1.41

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. http://dx.doi.org/10.1080/01621459.1984.10478078

Rosenbaum, P. R., & Rubin, D. B. (1985). The bias due to incomplete matching. *Biometrics, 41,* 103–116. http://dx.doi.org/10.2307/2530647

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66,* 688–701. http://dx.doi.org/10.1037/h0037350

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley. http://dx.doi.org/10.1002/9780470316696

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology, 2,* 169–188. http://dx.doi.org/10.1023/A:1020363010465

Sagarin, B. J., West, S. G., Ratnikov, A., Homan, W. K., Ritchie, T. D., & Hansen, E. J. (2014). Treatment noncompliance in randomized experiments: Statistical approaches and design issues. *Psychological Methods, 19,* 317–333. http://dx.doi.org/10.1037/met0000013

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. London, UK: Chapman & Hall/CRC. http://dx.doi.org/10.1201/9781439821862

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7,* 147–177.

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods, 13,* 279–313. http://dx.doi.org/10.1037/a0014268

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. http://dx.doi.org/10.1002/pds.1555

Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics, 9,* 307. http://dx.doi.org/10.1186/1471-2105-9-307

Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics, 8,* 25. http://dx.doi.org/10.1186/1471-2105-8-25

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods, 14,* 323–348. http://dx.doi.org/10.1037/a0016973

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science, 25,* 1–21. http://dx.doi.org/10.1214/09-STS313

Thoemmes, F. J., & Kim, E. S. (2011). A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research, 46,* 90–118. http://dx.doi.org/10.1080/00273171.2011.540475

Thoemmes, F., & Mohan, K. (2015). Graphical representation of missing data problems. *Structural Equation Modeling, 22,* 631–642. http://dx.doi.org/10.1080/10705511.2014.937378

van Buuren, S. (2012). *Flexible imputation of missing data*. Boca Raton, FL: Chapman & Hall/CRC. http://dx.doi.org/10.1201/b11826

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45,* 1–67. Retrieved from http://www.jstatsoft.org/article/view/v045i03

Vink, G., & van Buuren, S. (2013). Multiple imputation of squared terms. *Sociological Methods & Research, 42,* 598–607. http://dx.doi.org/10.1177/0049124113502943

West, S. G., Cham, H., Thoemmes, F., Renneberg, B., Schulze, J., & Weiler, M. (2014). Propensity scores as a basis for equating groups: Basic principles and application in clinical treatment outcome research. *Journal of Consulting and Clinical Psychology, 82,* 906–919. http://dx.doi.org/10.1037/a0036387

West, S. G., & Thoemmes, F. (2010). Campbell's and Rubin's perspectives on causal inference. *Psychological Methods, 15,* 18–37. http://dx.doi.org/10.1037/a0015917

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine, 30,* 377–399. http://dx.doi.org/10.1002/sim.4067

Yang, M., & Maxwell, S. E. (2014). Treatment effects in randomized longitudinal trials with different types of nonignorable dropout. *Psychological Methods, 19,* 188–210. http://dx.doi.org/10.1037/a0033804

Appendix

Data Generation for the Illustrative Example

Correlations, Means, and Standard Deviations of Retention Status (Actual Group Assignment) and Covariates in Original Research

Variables	1	2	3	4	5	6	7	8	9	10	11
1. Actual group assignment	1										
2. Ethnicity	.15	1									
3. Gender	.16	.01	1								
4. Literacy test	−.30	−.27	−.06	1							
5. Hyperactivity (T)	.23	.17	.28	−.22	1						
6. Engagement (T)	−.34	−.21	−.26	.24	−.79	1					
7. Academic self-efficacy (C)	−.05	.06	.04	−.04	−.12	.15	1				
8. Expected highest education (T)	−.30	−.32	−.06	.32	−.36	.45	.12	1			
9. Conscientiousness (T)	−.31	−.19	−.28	.22	−.79	.99	.14	.41	1		
10. Alliance w/ sch. (T)	−.18	−.23	−.10	.14	−.38	.44	−.01	.35	.40	1	
11. Parent involvement w/ sch. (T)	−.16	−.21	−.05	.14	−.07	.20	.05	.30	.16	.45	1
Mean	.24	.24	.53	−.54	.86	3.27	3.45	5.85	3.09	4.11	2.18
Standard deviation	.43	.43	.50	.60	.65	1.06	.55	2.37	1.05	.74	.48

Note. For actual group assignment, retained = 1, promoted = .0. For ethnicity, 1.0 = African American, .0 = other ethnicity. For gender, 1.0 = male, .0 = female. C = covariate is child-reported; T = covariate is teacher-reported; sch. = school.

Treatment-Outcome Model

(Outcome) = 9.0 × (Group Assignment) – 4.0 × (Ethnicity) – 4.0 × (Gender) + 1.8 × (Literacy Test) – 3.5 × (Hyperactivity) – 7.5 × (Engagement) + 4.0 × (Academic Self-efficacy) + .3 × (Expected Highest Education) + 7.0 × (Conscientiousness) + .3 × (Alliance w/ Sch.) + 2.5 × (Parent Involvement w/ Sch.) + (Residual)

The residual is a random normal variable with $M = 0.0$ and $SD = 5.0$.

Missingness Patterns and Missing Data Mechanisms of the Covariates

Missingness pattern	Number of participants	Covariates with missing values	Missing data mechanism
1	80	None	Nil
2	80	Ethnicity	Delete 40 participants in this missingness pattern who have the lowest value of $\frac{exp[0.01 \times (Gender) + 0.005 \times (Expected\ Highest\ Education)]}{\{1 + exp[0.01 \times (Gender) + 0.005 \times (Expected\ Highest\ Education)]\}}$
3	80	Hyperactivity	Delete 40 participants in this missingness pattern who have the lowest value of $\frac{exp[0.01 \times (\text{Engagement}) + 0.005 \times (\text{Academic Self-efficacy})]}{\{1 + ;exp[0.01 \times (\text{Engagement}) + 0.005 \times (\text{Academic Self-efficacy})]\}}$
4	80	Conscientiousness, Alliance w/ sch.	Delete 40 participants in this missingness pattern who have the lowest value of (parent involvement w/ sch.)
5	83	Literacy test	Randomly delete 40 participants

Note. The participants are randomly sorted into five distinct missingness patterns. sch. = school.

Received January 2, 2015
Revision received December 17, 2015
Accepted January 7, 2016 ■