ELSEVIER

Evaluation and Program Planning 28 (2005) 209–220

EVALUATION and PROGRAM PLANNING

www.elsevier.com/locate/evalprogplan

Estimating causal effects of public health education campaigns using propensity score methodology☆

Itzhak Yanovitzky[a,*], Elaine Zanutto[b], Robert Hornik[b]

[a]*Department of Communication, School of Communication, Information and Library Studies, Rutgers University, 4 Huntington Street, New Brunswick, NJ 08901-1071, USA*

[b]*University of Pennsylvania, Philadelphia, PA, USA*

Abstract

Many evaluations of public health education campaigns attempt to draw conclusions regarding the effect of messages on audiences' attitudes, beliefs, and behaviors based on observational data. To make causal inferences in these instances, it is necessary to adjust estimated campaign effects for possible selection bias due to systematic differences between respondents exposed to the campaign and those that were not. In particular, it is necessary to adjust for the impact of confounding variables that are likely to be determinants of both campaign exposure and outcomes. In comparison to other available methods for adjusting for selection bias such as multiple regression and instrumental variable methods, propensity scores offer a particularly simple way of adjusting estimates of campaign exposure effects for selection bias. This paper discusses the logic of this approach and illustrates its application to the evaluation of the National Youth Anti-Drug Media Campaign.

In 1998, the US Office of National Drug Control Policy (ONDCP) launched the National Youth Anti-Drug Media Campaign (NYAMC) as part of an effort to reduce and prevent drug use among youth nationwide. The campaign has progressed through three phases of increasing complexity and intensity. Phase I (January–June 1998) involved pilot testing the intervention advertisements in 12 metropolitan areas. In Phase II (July 1998 through July 1999), these advertisements appeared in multiple media nationwide, not just in the test areas, largely using pre-existing advertisements. Phase III, which is currently in progress, marks the full implementation of the campaign with new advertisements and involves the dissemination of anti-drug messages to a national audience of youth and parents. The goals of the campaign are to prevent and reduce drug use among youth and to encourage parents to talk to their children about drug use and more closely monitor their children's activities. The message design strategy was developed with the help of many individuals and organizations with expertise in teen marketing, advertising and communication, behavior change, and drug prevention. A detailed description of the campaign appears elsewhere (Hornik et al., 2002).

Like many other public health education campaigns, NYAMC presents a challenge to evaluators. Since the campaign is implemented nationwide, any claim about the campaign's ability to reach its goals (i.e. providing evidence that exposure to the campaign's messages changed audiences' attitudes, beliefs, and behaviors) must rely on observational data. As a result, the issue of possible selection bias must be addressed when estimating average campaign exposure effects. For example, individuals who report a higher degree of exposure to the campaign and its messages may be different in some systematic way from those who report a lower degree of campaign exposure (e.g. they may be more educated), and these systematic differences (rather than the degree of exposure in itself) may explain why the two groups of individuals differ in terms of

☆ The data used in this paper were collected under contract with the National Institute on Drug Abuse for the Evaluation of the National Youth Anti-Drug Media Campaign (contract # N01DA-8-5063). The Westat Corporation holds the contract for the evaluation with the University of Pennsylvania as subcontractor. We are grateful to Co-Principal Investigator David Maklan (Westat) and his colleagues who have collaborated closely with us on the project, including on aspects of the study reported here. The views expressed in this paper are those of the authors alone and do not represent those of the evaluation team.

* Corresponding author. Tel.: +1 732 932 7500x8123; fax: +1 732 932 3756.

E-mail address: iy@scils.rutgers.edu (I. Yanovitzky).

0149-7189/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.evalprogplan.2005.01.004

their attitudes or behavior. Of particular concern in this respect is the impact of confounding variables (confounders), or covariates that are predictive of both a person's degree of exposure to campaign messages and this person's attitudes and behavior without being themselves caused by either campaign exposure or the outcomes. Without controlling for confounders, estimates of the campaign's effects on audiences' attitudes and behavior are likely to be biased.

This paper illustrates the use of propensity score methodology to adjust estimates of campaign effects (public health communication campaigns in general and NYAMC in particular) for selection bias. We begin by introducing readers to the logic of propensity score methods. Next, we provide a step-by-step account of the implementation of this approach in the evaluation of NYAMC. Finally, we discuss some of the complexities and potential limitations of using propensity scores in estimating effects of public health education campaigns.

1. Propensity scores methodology

Researchers interested in adjusting estimates of campaign effects for potential selection bias may use a number of available techniques. The most commonly used technique is to employ a standard multiple regression analysis in which the effect of all observed potential confounders on a certain outcome is controlled by including these variables as predictors along with level of exposure as an additional predictor. The basic strategy here is to find a set of control variables that, once included in a regression model, will remove any part of the correlation between the treatment variable and the outcome which is not due to the influence of the treatment on the outcome, namely, removing the correlation between the treatment variable and the error term (Winship & Morgan, 1999). This approach, however, has two important limitations (Perkins, Tu, Underhill, Zhou, & Murray, 2000; Rubin, 1997). The first is that standard regression methods assume that the relationship between the outcome and the covariates within each exposure group follows a particular functional form such as a linear or a logistic function (Rubin, 1997). This may not be a problem if there is a sufficient overlap on covariates across exposure groups and appropriate diagnostics are used to assess model fit. When there is insufficient overlap on covariates, however, these model-based methods rely on such pre-specified functional forms to extrapolate estimates of treatment effects, which may be biased. Second, in a multiple regression setting, the inclusion of many potential covariates in the regression model significantly reduces the number of degrees of freedom available for the analysis and increases the likelihood of multicollinearity.

Another available technique for bias reduction is the instrumental variable (IV) approach (Angrist, Imbens, & Rubin, 1996; Heckman, 1995; Winship & Morgan, 1999). The basic logic of this approach is to define a variable or a set of variables (the IV or IVs) that have two properties: they affect or cause variation in assignment to treatment (i.e. have a non-zero correlation with treatment) and they have no direct effect on the outcome of interest. In other words, these variables can only be assumed to impact the likelihood of assignment to treatment (e.g. level of exposure to the campaign) and thus should be used to adjust estimates of treatment effect on outcomes. The IV-adjusted estimate of treatment effect can be calculated in three steps (Winship & Morgan, 1999): (a) regress the observed outcome on the potential IV to calculate a predicted outcome, (b) regress the treatment variable on the IV to calculate a predicted treatment, and (c) regress the predicted outcome on the predicted treatment assignment. Put differently, both the dependent (Y) and independent variable (X) are expressed as functions of the IV that is uncorrelated with the error term for the effect of X on Y. Because of that, a consistent estimate of the treatment effect can be calculated by regressing the new predicted Y on the new predicted X. However, the IV approach has three important weaknesses. The first is that it is often difficult to identify an IV because the requirement that a variable be an important cause of treatment but not a potential cause of the outcome is rarely met (Perkins et al., 2000). Also, the standard errors produced by this procedure tend to be too large for generating precise estimates of effects if the instrument is weak or with a small sample size (Winship & Morgan, 1999). Finally, this bias reduction approach only consistently estimates the true average treatment effect when the treatment effect is constant for all individuals, an assumption that is often unreasonable (Newhouse & McClellan, 1998).

In comparison to these alternatives, propensity score methods (Rosenbaum, 2002; Rosenbaum & Rubin, 1983, 1984) offer researchers a particularly desirable way of adjusting estimates of campaign exposure effects for selection bias. First, in many cases treatment effects can be estimated without the need to explicitly model the relationship between the outcomes and the covariates. Second, propensity score modeling is more robust to model misspecification than linear regression because it is less vulnerable to bias from variables that are included but in the wrong functional form (Drake, 1993; Perkins et al., 2000; Rubin, 1997). Third, the diagnostics and fitting of a propensity score model are done independent of the outcome and, thus, approximate random assignment of subjects to treatment. Finally, a single propensity score can be used for adjusting several different outcomes for selection bias simultaneously (whereas, for example, several different regression models would need to be fit to complete a similar task).

Based on the counterfactual account of causality (for a detailed discussion see Holland, 1986; Winship & Morgan, 1999), propensity score methods seek to create comparison

groups which are similar (or balanced) on all confounders but different on their levels of treatment (campaign exposure in this case). The most attractive feature of this technique is the ability to replace a set of confounding covariates with a single function of these covariates, the estimated propensity score, which is a scalar variable obtained by modeling the probability of receiving treatment (level of campaign exposure) as a function of observed covariates (D'Agostino, 1998). Thus, an individual's propensity score is this person's probability of being assigned to a particular level of treatment (level of campaign exposure in this case) conditional on his or her covariates' values (Rosenbaum & Rubin, 1983). Since the estimated propensity score can be thought of as a single-number summary of the set of covariates from which the propensity score was estimated, differences in these covariates are expected to be minimal and random across treatment groups for subjects with similar estimated treatment propensities (Rosenbaum & Rubin, 1984). In this way, the estimated propensity score, to the extent it is adequately estimated, is a balancing score because within subclasses of respondents with similar estimated treatment propensities, the distribution of observed covariates is the same across treatment groups (Rosenbaum & Rubin, 1984, 1985). Therefore, assuming all relevant covariates have been controlled for, matching or subclassifying on the estimated propensity score is expected to remove selection bias due to the effect of confounders (Joffe & Rosenbaum, 1999). An adjusted estimate of campaign effects can then be calculated using various strategies such as matching (e.g. D'Agostino, 1998; Rubin & Thomas, 1996; Smith, 1997), stratification (D'Agostino, 1998; Perkins et al., 2000; Rubin, 1997), weighting (Hirano, Imbens, & Ridder, 2003; Imbens, 2000; Rosenbaum, 1987), and regression adjustment (D'Agostino & Rubin, 2000).

The assumption that all potential confounders have been measured and included in the propensity modeling process is part of the 'strongly ignorable treatment assumption' (Rosenbaum & Rubin, 1983). More precisely, we assume that assignment to treatment is unconfounded conditional on the observed set of covariates, that is, is based only on observable pre-treatment variables, and that the probability of being assigned to treatment is non-zero for all units (Rosenbaum & Rubin, 1983). In other words, a strong assumption is made that all potential confounders are measured. This is similar to the implicit assumption made in linear regression or instrumental variables analysis that there is no omitted variable bias. Only when the strongly ignorable treatment assumption is met does propensity score methodology produce approximately unbiased treatment effect estimates.

The following sections provide readers a more detailed account of the actual steps involved in the implementation of propensity score methodology. Propensity score adjustment is essentially a two-stage method. The first stage involves the balancing of confounders across treatment groups but for reasons of clarity we break down this stage into five successive steps (see steps 1–5 below). The second stage involves the actual estimation of propensity-score-adjusted campaign effects and is discussed below under the sixth step. Next, we describe the application of propensity score methodology to two specific cases: (a) a binary treatment (i.e. treatment and control) and (b) an ordinal categorical treatment (low, medium, and high campaign exposure).

1.1. Step 1: select a confounder pool

The first and perhaps the most critical step in employing a propensity score adjustment is to select a set of covariates (or potential confounders) from which to estimate the propensity score. This selection process should be made a-priori on theoretical grounds and based on previous available empirical evidence about relationships between variables of interest. It should not be based on patterns of association between variables that are found in the actual data used for the evaluation of campaign effects. Only by selecting confounders a-priori and blinding themselves to the actual data can researchers assure colleagues, stakeholders, and the public that the use of this method to adjust estimates of campaign effects is not geared toward generating findings that are consistent with the hypothesized effects of the campaign. Fig. 1 illustrates the logic of this process when measures of all variables are taken at the same time.

There are five types of variables in Fig. 1: variables measuring campaign exposure, variables measuring outcomes, confounders, mediators, and ambiguous variables. Only confounders should be included in the estimation of the propensity score. Confounders are those variables that can explain variations both in level of campaign exposure and in outcomes but themselves are not caused by campaign exposure or outcomes. This is illustrated in Fig. 1 by the direction of the arrows from Confounders to Campaign Exposure and Outcome variables. Thus, variables that are clearly causally prior to both campaign exposure and the outcome (for example, a person's age, gender, and race) should be included in the confounder pool. On the other hand, variables that are associated with the outcomes but themselves are caused

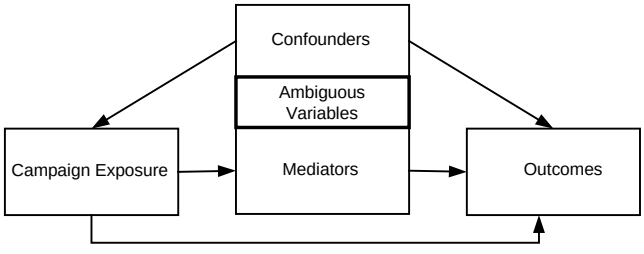

Fig. 1. Types of variables and the selection of potential confounders.

(or hypothesized to be caused) by campaign exposure (i.e. mediators), as illustrated by the direction of the arrow linking Campaign Exposure to Mediators in Fig. 1, should be excluded from the confounder pool because including them would lead to biased estimation of campaign effects. For example, if exposure to the messages of the campaign leads youth to change their belief about the percentage of their peers that use marijuana, and these belief changes lead, in turn, to changes in marijuana use by youth, then it would be a mistake to include this belief in the confounder pool. Therefore, all variables that are likely to be an outcome of campaign exposure should be excluded from the confounder pool.

Unfortunately, some variables may play both confounding and mediating roles. These are termed Ambiguous Variables in Fig. 1. There are conflicting imperatives for these variables: their effects must be both removed and not removed at the same time. If they are excluded, there is some risk of bias in estimates of treatment effects because some of the observed association may be due to the excluded variable (Drake, 1993). If they are included, there is some risk that the effects will be biased because the effects of treatment on outcome are captured by the presence of the potential consequence variable.

As an example of an ambiguous variable, consider the potential role of an adolescent's frequency of association with marijuana-using peers. Adolescents that associate frequently with marijuana-using peers are likely to use marijuana themselves and may also be more exposed to the messages of the campaign due to their interest in marijuana use sparked by their social interactions with these peers. Thus, the frequency of association with marijuana-using peers is a confounding variable. At the same time, there may be adolescents who reduced their frequency of interaction with marijuana-using peers following exposure to the messages of the campaign—so the frequency of association with marijuana-using peers is also a mediating variable. The solution is to include ambiguous variables in the confounder pool only when one is confident that they are not mediators. In this case, we had no a-priori reason to believe that the frequency of association with marijuana-using peers would be affected by campaign exposure, and so this variable was included in the confounder pool. This particular decision and similar others were made following deliberations among members of the evaluation team and independent of actual relationships between variables in the data. Such deliberations are needed to reduce the potential for mistakenly excluding potential confounders or including potential mediators.

1.2. Step 2: determine the initial imbalance on confounders

The initial imbalance in each of the covariates with an interval level of measurement can be estimated through a two-sample t-test (for treatments with only two levels) or a one-way analysis of variance (for treatments with more than two levels) in which campaign exposure is the independent variable. For dichotomous confounders, a two-sample test of differences in proportions (for treatments with only two levels) or a logistic regression with the confounder as the outcome and dummy variables for the treatment levels as the predictors (for treatment with more than two levels) can be used to assess initial imbalance. There are two straightforward reasons for beginning propensity score analysis with estimating the initial imbalance on confounders across treatment (or exposure) groups. If it turns out that the distribution of covariates is adequately balanced, there is no need to employ the propensity score technique.[1] Simple bivariate associations of exposure and outcomes will suffice to reliably estimate campaign effects (assuming there are no important unobserved covariates). On the other hand, if an imbalanced distribution of confounders exists, the information on imbalanced covariates may be used later as a benchmark against which to verify that the propensity score adjustment methodology has, in fact, increased the balance in the covariates.

1.3. Step 3: estimate the propensity score

The propensity score for each subject can be estimated with different methods including discriminant analysis and probit models (D'Agostino, 1998) or even classification trees (Stone, Obrosky, Singer, Kapoor, & Fine, 1995). The most common method in use, however, is a logistic regression model with campaign exposure as the outcome and all the candidate confounders as predictors. If the exposure variable is dichotomous, the equation takes the following form:

$$\mathrm{Log}\left(\frac{P_i}{1-P_i}\right) = \alpha + \beta X_i$$

Where, P_i is the probability of respondent i to be exposed to the campaign and X_i is a vector of confounders from which the propensity score is estimated. The estimated propensity score is then calculated as the estimated probability of being exposed to the campaign, conditional on X_i. In cases where exposure is an ordinal variable, McCullagh's (1980) ordinal logistic regression model (also called the proportional odds model) can be used to generate an estimated propensity score per individual for each level of exposure. This ordinal logistic regression model for an exposure variable with K

[1] Adequate balance can be defined as the amount of balance expected in a completely randomized experiment. In other words, approximately 5% of the significance tests for balance would be expected to be statistically significant at the 0.05 level purely by chance in a randomized experiment. If this is achieved in the propensity score analysis this is as much balance as could be expected in a randomized experiment and is often the benchmark against which we decide if we have adequate balance (Rosenbaum & Rubin, 1984; Zanutto, et al., in press).

levels has the following form:

$$\mathrm{Log}\left(\frac{P(Y_i \le k)}{1 - P(Y_i \le k)}\right) = \alpha_k + \beta X_i \quad k = 1, 2, \ldots K - 1$$

where Y_i is the exposure level received by respondent i. In this case, the estimated linear predictor βX_i (or the linear combination of all individual scores on confounders weighted by their estimated slope parameters,) is a single scalar balancing score.[2] This estimated linear predictor (that many available data analysis computer programs can save as a new variable in the dataset) is not a propensity score because it does not represent the probability of being exposed to the campaign at a certain dose conditional on covariates, but it is a balancing score that can be used to construct matched pairs or strata that balance the covariates (Joffe & Rosenbaum, 1999; Lu, Zanutto, Hornik & Rosenbaum, 2001; Zanutto, Lu, & Hornik, in press). However, to simplify terminology, we will refer to it as a propensity score in the rest of this paper.

To verify that the data can support a comparison of treatment and control groups that are balanced on all covariates, the distribution of the estimated propensity scores for the treated and control groups should be checked for adequate overlap. This can be accomplished by creating overlapping histograms, or by comparing quintiles of the estimated propensity scores for the treatment and control groups. If there is no overlap in the distribution of propensity scores across exposure groups, then no estimates of the treatment effect can be made without relying on unverifiable modeling assumptions. Moreover, there needs to be enough cases in each propensity score quintile in each exposure group (i.e. in each quintile by exposure cell) so that it is possible to make comparisons across exposure groups within each propensity score quintile. If the covariates are very powerful predictors of exposure, then this condition may not be met particularly for the distribution of propensity scores across the lowest and highest quintile. Thus, even if the distribution of propensity scores across exposure groups does overlap considerably, treatment effects can only be estimated for the respondents who have propensity scores that overlap that of others. For example, if the estimated propensity scores for the control group range between 0.05 and 0.7 and the estimated propensity scores for the treatment group range between 0.4 and 0.95, then we can only estimate treatment effects for the portion of the population with propensity scores between 0.4 and 0.7. In the range 0.05–0.4, there are no treated units to compare to the observed control units, and in the range 0.7–0.95, there are no control units to compare to the observed treated units. This ability to highlight the range over which reliable treatment estimates can be made is an advantage of the propensity score methods. In this simple example, further analysis (e.g. Steps 5 and 6) should be restricted to those treated and control units with estimated propensity scores between 0.4 and 0.7 (for an illustration of this approach see Dehejia & Wahba, 1999; Zanutto et al., in press). In some cases there may only be a few treated and control units in the non-overlapping tail areas and including them in the analysis may not adversely affect the balance of the covariates in which case there is no need to exclude these observations from the analysis.[3]

1.4. Step 4: stratify on the propensity score

Once the propensity score has been estimated and the analysis has been restricted to the region of propensity score overlap, observations are stratified on the propensity score. Typically observations are divided into five equally-sized groups (strata) based on the distribution of the estimated propensity score (this is also called subclassification on the propensity score). Studies have shown that creating five strata based on propensity scores can remove approximately 90% of the initial imbalance in each of the confounders between treatment groups (Cochran, 1968; Rosenbaum & Rubin, 1984). This stratification strategy will produce an estimate of the treatment effect in the population (or, more specifically, the population represented by the region of overlap). Alternatively, in the binary treatment case, observations can be divided into five groups where strata boundaries are defined by the quintiles of the estimated propensity scores for the treated units (rather than for all units). This approach will produce estimates of the treatment effect on the treated units that may be a useful estimate when only a select group of people tend to get the treatment and it is desired to estimate the treatment effect on those type of people who are likely to get the treatment, rather than the treatment effect on the population as a whole. The definition of the 'treatment effect on the treated' is less well-defined in the case of multiple treatment levels, however, if desired, it is possible to stratify so as to obtain an estimate of the treatment effect on a particular subgroup (e.g. those with high exposure) rather than for the whole population. With regard to the evaluation of the campaign discussed below, the basic question was whether the campaign was effective on average for a population in which the majority of people (over 70%) had some exposure to the campaign. In this particular case it made sense to estimate treatment effect in the population. In another

[2] Note that each individual has a set of estimated propensities, one corresponding to each level of treatment, and so unlike the binary treatment case they do not provide us with a single scalar balancing score. However, when McCullagh's ordinal logistic regression models is used, or a similar model such as an ordinal probit model, the linear predictor is a single scalar balancing score (Joffe & Rosenbaum, 1999; Lu et al., 2001).

[3] As an alternative to restricting the analysis to the region of propensity score overlap, the entire data set can be analyzed by combining propensity score subclassification with additional model-based adjustments, especially outside the region of propensity score overlap (Rosenbaum & Rubin, 1985). This allows inferences to be made for the entire population, rather than for a restricted subpopulation, but is also more model-dependent.

situation (e.g. a campaign tailored to a specific target audience), a researcher might choose to restrict the comparison samples so to gauge effects on those who would be exposed, and ignore those who were not likely to be in the audience for the program.

1.5. Step 5: check balance on confounders across treatment groups

The goal of this step is to verify that subclassification on the estimated propensity score removes any initial bias on confounders (see Step 2). Using the subclassification method described in the previous step, one may check for balance on covariates using multiple two-way analysis of variance (ANOVA) analyses, where campaign exposure is one factor, the propensity score quintile to which the individual was assigned is a second factor (coded as a categorical variable with four degrees of freedom), and each one of the covariates (or confounders) is a dependent variable.[4] If balance is achieved, there should not be a statistically significant main effect of campaign exposure on the covariate. There should also not be a statistically significant interaction effect of exposure by quintile. If these two conditions are not met, the propensity score should be re-estimated by adding interaction terms and/or non-linear functions (e.g. quadratic or cubic) of imbalanced covariates to the propensity score model. At minimum, balance should be achieved on key covariates (i.e. covariates that are clearly causally prior to campaign exposure). As indicated above, the usual standard of 'adequate balance' is the amount of balance that would be achieved on average in a completely randomized experiment. In this way, Steps 3–5 are repeated until balance is achieved or until no further improvement in balance can be made.

1.6. Step 6: calculating confounders-adjusted estimates of effects

There are several ways of using the propensity score to generate estimates of campaign effects (for more details see D'Agostino, 1998; Perkins et al., 2000; Rosenbaum & Rubin, 1983; Rubin, 1997; Rubin & Thomas, 1996; Smith, 1997; Winship & Morgan, 1999). However, subclassification (or stratification) on the propensity score and the calculation of the average treatment effect is probably the most prevalent strategy of estimating treatment effects (Rubin, 1997). First, average treatment effects are estimated separately within each propensity score quintile by subtracting the average outcome among treated individuals from that of untreated individuals. Next, an overall estimate of treatment effect is calculated by averaging the differences between treatment and control groups across all 5 quintiles. This procedure is summarized in the following expression:

$$\hat{\delta} = \sum_{k=1}^{5} \frac{n_k}{N} (\bar{Y}_{tk} - \bar{Y}_{ck})$$

Where $\hat{\delta}$ is the estimated overall treatment effect, k indexes the propensity score quintile (quintile 1 through 5), N is the total number of units, n_k is the number of units in the propensity score quintile k, and $\bar{Y}_{tk}$ and $\bar{Y}_{ck}$, respectively, are the average level of the outcome recorded for exposed (treatment) and unexposed (control) units within a specific propensity score quintile.[5] The estimated standard error of this estimated treatment effect is commonly calculated as:

$$\hat{s}(\hat{\delta}) = \sqrt{\sum_{k=1}^{5} \frac{n_k^2}{N^2} \left(\frac{s_{tk}^2}{n_{tk}} + \frac{s_{ck}^2}{n_{ck}} \right)}$$

Where n_{tk} and n_{ck} are the number of treated and control units, respectively, in quintile k, and s_{tk}^2 and s_{ck}^2 are the sample variances of the treated and control units, respectively, in quintile k. A similar strategy can be used to estimate the difference in treatment effect between two levels of a multi-level treatment variable (e.g. the difference between high and low exposure).

2. Estimating the effect of NYAMC on target audiences using propensity scores

We will now provide two illustrations of the use of propensity score methodology to adjust estimates of campaign effects on outcomes in the context of NYAMC. The data used in the illustrations below were collected through the National Survey of Parents and Youth (NSPY) that was developed for the purposes of evaluating Phase III of the campaign. The first three waves (November 1999 to May 2000, July to December 2000, and January to June 2001) included 8133 youth aged 9–18 and 5606 of their parents. These respondents represent the approximately 40 million youth and 43 million of their parents who are the target audience for NYAMC. The interviewers for NSPY achieved a response rate of 65% for youth and 63% for parents across waves. Questionnaires were administered in respondents' homes on touch-screen laptop computers and included extensive measurement of youth exposure to the campaign messages and other anti-drug messages, their beliefs, attitudes, intentions, and behaviors with regard to drugs and a wide variety of other factors either known to be related to drug use or likely to make youth more or less susceptible to NYAMC messages. Questionnaires for

[4] Since our sample sizes are so large, we believe these tests are acceptable even for our binary covariates. Alternatively, logistic regressions may be used to achieve the same goal.

[5] When estimating the treatment effect on the treated units, using strata boundaries defined by the quintiles of the estimated propensity scores for the treated units, N is the total number of treated units, and n_k is the number of treated units in the propensity score quintile k.

parents included questions about exposure to the campaign messages and other anti-drug messages as well as questions about their beliefs, attitudes, intentions, and behaviors with regard to their interactions with their children. These behaviors included talking with their children about drugs, parental monitoring of children's lives, and involvement in activities with their children.

It is worth noting that the estimates of campaign effects generated by these examples were not adjusted for non-response, and the complex sample design, and some possible confounders were not included (although these are all taken into account in all evaluation reports). For this reason, the estimates of campaign effects reported here cannot be generalized to the population of youth from which the actual sample was drawn and, therefore, are likely to differ from those reported by the Evaluation (see Hornik et al., 2002).

2.1. Illustration 1: campaign exposure as two treatment conditions

The most basic premise of NYAMC is that, all other things being equal, those exposed to the campaign should hold, on average, stronger anti-drug attitudes, beliefs, and intentions than those who were not exposed to the campaign. Two measures of campaign exposure were employed. The first was aimed at capturing exposure to anti-drug advertisements in general, not only to those generated by the campaign. Respondents were asked about the frequency of seeing or hearing anti-drug advertisements in each of the major media channels (television and radio, newspapers and magazines, outdoor venues, and movies) that were utilized by the campaign. Based on the answers each respondent gave to these questions, he or she received a score on a general exposure index ranging from low exposure (less than four times a month) to high exposure (12 or more times per month) with a medium level of exposure (4 to less than 12 times a month) in between. The second measure of exposure was designed to assess frequency of exposure to campaign-specific anti-drug ads. Here respondents were asked to accurately recall a specific ad played to them from a sample of ads, some that were actually aired and some that were not. By summing answers across ads, each respondent received a score on a three-point recall-aided exposure index: low exposure (less than four times a month), medium exposure (four to less than 12 times a month), and high exposure (more than 12 times a month).

The specific exposure index is used in Illustration 2 with regard to parents' discussions about marijuana use with their kids. In this illustration we examine the association between perceived marijuana trial by others (measured as a binary variable where '1' represents a belief that none or only few other kids of the same age tried marijuana) and the level of general exposure to anti-drug messages among 12–18-year-olds who have never tried marijuana ($N=4{,}282$). Information on the true prevalence of marijuana trial among teens nationwide is expected to reduce the perception that this behavior is normative for individuals this age. Of course, there is always some risk that such efforts would have an opposite effect. That is, a large number of messages on the topic of marijuana trial by teens, whether generated by the campaign itself or as part of a more general 'buzz' in the media, may actually lead to increased perception of marijuana use by others.

For the purposes of estimating the effect of general exposure to anti-drug messages on perceived marijuana trial by others (as well as demonstrating the use of propensity scores when campaign exposure is coded as two treatment conditions), we recoded the original exposure index into a binary variable wherein respondents who could not recall being exposed to anti-drug messages were assigned a value of '0' (not exposed) and those who could recall being exposed to one or more such messages were assigned the value of '1' (exposed). We used a propensity score technique to generate an estimate of the causal effect of exposure on this belief. Following the procedure described above (see Step 1), we identified over 50 potential confounders including demographic and family characteristics, previous drug experience, personality traits, and media consumption patterns reported by youth and their parents (for a complete list see Hornik et al., 2002). Moving to Step 2 in the analysis, we examined the initial balance on covariates between exposure and control groups by employing two-sample tests of differences in proportions (for binary variables) and two-sample t-tests (for continuous variables). For reasons of clarity, Table 1 lists only those confounders for which a statistically significant initial bias between exposure and control groups was found.

Table 1 suggests that while the majority of youth was exposed to general anti-drug messages, the number of non-exposed individuals was substantial enough to allow a valid comparison of exposure effects across groups. Overall, there were 19 covariates with initial bias: 14 relating to youth characteristics and five pertaining to their parents. The most notable confounders are those measuring habitual media consumption. In general, youth (and households, for that matter) who are heavy consumers of media are more likely to be exposed to general anti-drug messages. Other significant confounders include youth and parent demographics, relationships with parents, association with delinquent peers, and parental drug use.

To ensure that estimated effects of exposure are not confounded with the effects of these covariates, a propensity score adjustment was employed. First, each respondent's probability of exposure to anti-drug messages was estimated from the combination of all original covariates and all their first-order interactions. Next, to verify that the data can support a comparison of treatment and control groups (see Step 4), the distribution of the estimated propensity scores within the exposure and control groups was checked for adequate overlap. Thirty-one respondents in the control (non-exposure) group and 20 respondents in the treatment

Table 1
Comparison of differences between exposure and control groups on significant confounders before and after propensity score adjustment among 12–18 year-olds who have never tried marijuana, United States, 1999–2001

Confounder	Exposure ($N=3,184$)		Non-exposure ($N=807$)		Pre adjustment ($N=3,991$)	Post adjustment ($N=3,940$)[a]	
	Mean	(SD)	Mean	(SD)	F-value[b]	F-value (main effect)[c]	F-value (interaction effect)[d]
Teen variables							
Gender (1=male, 2=female)	1.50	(0.50)	1.44	(0.49)	7.45*	0.010	0.94
Child's average grade in school	6.4	(2.1)	6.64	(2.1)	7.8*	0.028	1.45
TV viewing hours on weekdays	4.86	(1.95)	4.14	(2.0)	86.1**	0.79	0.78
TV viewing hours on weekends	4.92	(1.77)	4.36	(1.8)	62.8**	0.001	1.77
Radio listening hours on weekdays	3.75	(2.18)	3.50	(2.22)	8.24*	0.03	1.02
Radio listening hours on weekends	3.56	(1.88)	3.24	(1.87)	13.4**	0.34	1.17
Internet use (yes/no)	0.86	(0.34)	0.82	(0.37)	6.86*	0.001	0.98
Magazine reading frequency	3.1	(1.1)	2.93	(1.1)	13.9**	0.47	1.12
Negative peer influence	1.54	(0.76)	1.41	(0.65)	18.9**	0.084	0.75
Child reports on family fighting (yes/no)	0.23	(0.42)	0.19	(0.39)	5.2*	0.026	0.28
No unsupervised time with friends (yes/no)	0.12	(0.33)	0.16	(0.37)	10.3*	0.096	0.31
Sensation-seeking tendencies (1=high, 2=low)	1.51	(0.50)	1.55	(0.49)	4.84*	0.067	0.66
White (dummy)	0.66	(0.48)	0.73	(0.44)	15.7**	0.608	1.3
Black (dummy)	0.16	(0.37)	0.10	(0.30)	15.9**	0.48	1.5
Parent variables							
Education (0=less than college; 1=college+)	0.84	(0.98)	0.76	(0.97)	4.62*	0.021	1.5
Time watching TV	1.38	(0.59)	1.31	(0.63)	10.04*	0.14	2.17
Parent-child participation in outdoor activities	3.3	(1.73)	3.5	(1.8)	10.36*	0.12	0.74
Prior illicit drug use (yes/no)	0.77	(0.47)	0.81	(0.50)	4.37*	0.17	1.62
Prior smoking behavior (yes/no)	0.92	(0.73)	0.86	(0.73)	4.32*	0.01	1.69

*$p<0.05$, **$p<0.001$.
[a] 51 Observations were dropped due to insufficient overlap in the propensity score.
[b] F-statistics was calculated by squaring two-sample t-statistics scores.
[c] F-statistics for main effect of general exposure after propensity score adjustment.
[d] F-statistics for the interaction effect of general exposure and propensity score quintile.

(exposure) group had propensity scores that were outside the region of overlap (i.e. could not be matched with similar scores in the corresponding group). These 51 individuals were excluded from the subsequent analysis and the remaining ones were classified according to 5 quintiles of estimated propensity score.

Before moving to estimate the effect of general exposure to anti-drug messages on youth perception of marijuana trial by other kids their age, we checked again for imbalanced covariates (see Step 5). To this end, we employed multiple two-way analysis of variance (ANOVA) analyses, where general exposure was one factor, the propensity score quintile was a second factor, and each one of the covariates (or confounders) served as the dependent variable. The results of this analysis are summarized in the two right-hand columns in Table 1. Clearly, the propensity score adjustment reduced the initial bias on confounders (as indicated by the lower F-statistics scores), turning all initial differences between exposure and control group into non-significant ones. Similarly there were no significant effects of the exposure-by-quintile interaction on these covariates. It is worth noting that, occasionally, the propensity score adjustment can create imbalance on covariates that were initially balanced. Fortunately, that was not the case here.

We were now ready to estimate the causal effect of general exposure to anti-drug messages on youth perception of marijuana trial by other kids their age. We chose to do so by employing the sub-classification strategy (see Step 6). That is, we calculated the estimated campaign effects for each propensity score quintile and then averaged them into an overall estimate of campaign effect. As Table 2 demonstrates, there was no evidence of general exposure effects on this particular outcome. More specifically, while across quintiles and overall the estimated exposure effects were in the hypothesized direction (namely that exposure to general anti-drug messages will result in lower estimates of peers who use marijuana, there were no significant differences in the average perception score between the exposure and control group (as indicated by the non-significant t-test values).

2.2. Illustration 2: campaign exposure as more than two treatment conditions

An important intermediate goal of the campaign is to encourage parents to have more discussions about drugs with their children. Hence, if the campaign is successful, there should be an association between parental exposure to the campaign and the frequency of parent-child

Table 2
Estimated effects of general exposure to anti-drug messages on perceptions of marijuana use by others among 12–18 year-olds who have never tried marijuana using sub-classification on the propensity score

Propensity score quintile	Exposure level	Group size	Average perception score (SD)	Estimated effect	*t*-Value
Quintile 1	Non-exposed	263	0.62 (0.48)	−0.03	−0.81
	Exposed	517	0.59 (0.49)		
Quintile 2	Non-exposed	178	0.58 (0.49)	−0.02	−0.48
	Exposed	609	0.56 (0.49)		
Quintile 3	Non-exposed	144	0.57 (0.49)	−0.03	−0.66
	Exposed	652	0.54 (0.49)		
Quintile 4	Non-exposed	111	0.61 (0.49)	−0.07	−1.4
	Exposed	673	0.54 (0.49)		
Quintile 5	Non-exposed	87	0.63 (0.48)	−0.11	−1.86
	Exposed	707	0.52 (0.50)		
Overall	Non-exposed	783	0.60 (0.002)[a]	−0.05	−1.9
	Exposed	3,158	0.55 (0.008)[a]		

[a] Overall estimates averaged over propensity score quintiles (with corresponding standard error).

discussions about drugs. As noted above, the measure of exposure used here is self-reported exposure to campaign-specific anti-drug ads. The specific exposure index was measured as a categorical variable ranging from low exposure (less than four times a month) to high exposure (12 or more times per month) with a medium level of exposure (4 to less than 12 times a month) in between. The dependent variable in this case is the extent of parent-child conversations about drug use reported by parents of youth 12–18-year-old who have never used marijuana ($N=4{,}018$). Parents could earn up to three points (on a scale ranging from 0 to 3) if they reported talking about drugs at least twice in the past 6 months as well as talking about family rules about drugs, and about specific things a child could do to avoid drugs. Our expectation here was that parents who report a greater dose of exposure to the campaign's anti-drug messages will also report more parent-child conversations about drug use.

As in the previous example, to use a propensity score approach to adjust for confounders, we first identified a list of 28 potential confounders including demographic characteristics (race, gender, age, education, income, marital status, religiousness, number of children under 18, and urbanity), media consumption habits (e.g. television, radio, newspapers, magazines, and internet), and prior alcohol, tobacco, and drug use (for a complete list see Hornik et al., 2002). We then examined the initial imbalance of the covariates across the three exposure groups using one-way ANOVAs with campaign exposure as the factor and each covariate, separately, as the dependent variable.[6] Table 3 lists only those confounders with a statistically significant initial bias across exposure groups.

As shown in Table 3, there were 13 covariates with initial imbalance. As in the previous example, there were significant imbalances in the confounders measuring habitual media consumption, with more media consumption generally associated with more exposure to the campaign. Other imbalanced confounders included education, income, Internet use, tobacco and marijuana use, and a few demographics. We then fit an ordinal logistic regression predicting the propensity of exposure to campaign-specific anti-drug information from all confounders. Following Steps 3–5 we decided on a final propensity score model consisting of main effects for all the confounders, all their first-order interactions, and quadratic terms for all continuous confounders. Simpler models were estimated during this process (e.g. a model with only the main effects for all potential confounders), but this particular model balanced the most confounders. A propensity score was calculated for each individual and the region of propensity score overlap was assessed. Ninety-nine observations fell outside the region of overlap and were dropped from further analysis. The remaining observations were classified according to 5 quintiles of the estimated propensity score.

To verify that stratification on the propensity score created groups of individuals with balanced covariates, we used two-way ANOVAs with campaign exposure and propensity score quintile as factors and each covariate as the dependent variable. The results of this analysis are summarized in the two right-hand columns of Table 3. Stratification on the propensity score successfully eliminated all the initial imbalances (except for tobacco use) and no new imbalances were created. With only a single F-statistic significant at the 0.05 level, this level of balance is greater than that expected by chance in a randomized experiment and, therefore, we proceeded to estimate the effect of campaign exposure on parents' talking behavior controlling for confounders. We did so

[6] Since our sample sizes are so large, we believe these tests are acceptable even for our binary covariates. Alternatively logistic regressions may be used to achieve the same goal.

Table 3
Comparison of differences between three campaign exposure groups on significant confounders before and after propensity score adjustment among parents of 12–18 year-olds who have never tried marijuana, United States, 1999–2001

Confounder	Campaign exposure level						Pre adjustment ($N=3,807$) F-value[a]	Post adjustment ($N=3,708$)[b]	
	Low ($N=1,102$)		Medium ($N=1,331$)		High ($N=1,374$)			F-value (main)[c]	F-value (interaction)[d]
	Mean	(SD)	Mean	(SD)	Mean	(SD)			
Education	2.84	(1.02)	2.71	(1.00)	2.50	(1.00)	36.05**	0.455	1.336
Income	5.32	(1.88)	5.16	(1.90)	4.69	(1.90)	38.32**	0.918	0.818
TV consumption	1.29	(0.63)	1.39	(0.59)	1.44	(0.57)	20.71**	1.215	1.189
Radio consumption	0.98	(0.72)	1.01	(0.72)	1.13	(0.73)	14.16**	0.801	0.745
Newspaper consumption	1.64	(0.62)	1.69	(0.61)	1.62	(0.65)	3.81*	1.432	1.624
Ethnic TV channels consumption...	0.34	(0.59)	0.40	(0.63)	0.56	(0.72)	40.36**	0.130	1.289
Tobacco use	0.88	(0.71)	0.94	(0.72)	0.94	(0.76)	3.15*	1.096	2.105*
Marijuana use	0.57	(0.56)	0.57	(0.56)	0.51	(0.57)	5.37**	1.413	1.128
Internet use	0.77	(0.89)	0.85	(0.89)	0.95	(0.91)	11.67**	0.165	1.163
White (dummy)	0.73	(0.45)	0.70	(0.46)	0.63	(0.48)	15.51**	0.171	1.056
Urban (dummy)	0.27	(0.44)	0.30	(0.46)	0.36	(0.48)	12.92**	0.080	0.912
Suburban (dummy)	0.30	(0.46)	0.29	(0.45)	0.19	(0.39)	23.86**	2.565	1.469
Northeast (dummy)	0.18	(0.38)	0.16	(0.37)	0.13	(0.34)	5.01*	0.112	1.339

*$p<0.05$, **$p<0.001$.
[a] F-statistics was calculated using a one-way ANOVA (Type III sum of squares).
[b] 99 Observations were dropped due to insufficient overlap in the propensity score.
[c] F-statistics for main effect of campaign exposure after propensity score adjustment.
[d] F-statistics for the interaction effect of campaign exposure and propensity score quintile.

by calculating the average parent-child drug discussions score (ranging from 0 to 3) for each campaign exposure level in each propensity score quintile (shown in the column labeled 'Average talking score' in Table 4) and then averaged them into an overall estimate of this outcome using the equation:

$$\bar{Y}_{\mathrm{adj},t} = \sum_{k=1}^{5} \frac{n_k}{N} \bar{Y}_{t,k}$$

where $\bar{Y}_{\mathrm{adj},t}$ is the propensity-adjusted average outcome corresponding to exposure (treatment) level t, and $\bar{Y}_{t,k}$ is

Table 4
Estimated campaign effects on parents' discussion about drugs with children 12–18 year-olds who have never tried marijuana using sub-classification on the propensity score

Propensity score quintile	Exposure level	Group size	Average talking score (SD)	Estimated effect (high-low)	F-value
Quintile 1	Low	335	2.34 (0.87)	0.15	1.34
	Medium	240	2.36 (0.83)		
	High	104	2.49 (0.76)		
Quintile 2	Low	254	2.39 (0.84)	0.16	1.88
	Medium	272	2.45 (0.78)		
	High	147	2.54 (0.71)		
Quintile 3	Low	168	2.42 (0.84)	0.10	0.97
	Medium	271	2.51 (0.72)		
	High	242	2.52 (0.75)		
Quintile 4	Low	115	2.49 (0.74)	0.06	0.51
	Medium	273	2.48 (0.77)		
	High	286	2.54 (0.77)		
Quintile 5	Low	87	2.46 (0.83)	0.17	2.00
	Medium	171	2.63 (0.69)		
	High	422	2.63 (0.71)		
Overall	Low	959[a]	2.42 (0.03)[b]	0.13	5.54*
	Medium	1227[a]	2.49 (0.02)[b]		
	High	1201[a]	2.54 (0.02)[b]		

*$p<0.05$.
[a] 3387 Observations had observed values of the outcome variable and all the covariates.
[b] Overall estimates averaged over propensity score quintiles (with corresponding standard error).

the average outcome corresponding to exposure level t in propensity score quintile k. This overall propensity-adjusted estimate and the corresponding standard error are shown in the 'Propensity-Adjusted Overall Estimate' row of Table 4.

To test for the significance of estimated campaign effects we ran a two-way ANOVA with campaign exposure level and propensity score quintile as factors and parental talking behavior as the dependent variable (main effect for exposure: $F(2,3372)=5.54$, $p<0.001$). This analysis indicates that there is evidence of statistically significant campaign effects on parental talking behavior (as we hypothesized), although the effect size is small.

3. Discussion

The two basic illustrations we provide above demonstrate the use of the propensity score methodology to reduce bias in estimates of campaign effects when exposure and outcomes are measured simultaneously. In both illustrations, the propensity score approach appears to have removed much of the selection bias of covariates for which imbalances had been demonstrated. While observational data were used here, this method may also be useful in the experimental domain, particularly when the random assignment of subjects to conditions fails due to the premature drop out of subjects or other circumstances over which the researcher has no control. In this case propensity scores may be used to allow a close approximation of the controlled trial (D'Agostino, 1998).

Like any other method, the propensity score approach has a number of limitations that we discuss here briefly. First, it is important to remember that propensity score methods can only adjust for selection bias and not for other types of biases such as those associated with measurement errors. Moreover, they only adjust for selection bias that is attributed to observed confounders and not for unobserved ones (unless they are strongly correlated with observed covariates) under the assumption of strong ignorability. Hence, this methodology cannot serve as a replacement for randomized studies which can balance both observed and unobserved covariates (Rosenbaum & Rubin, 1983; Rubin, 1997). However, estimates can be devised to determine the robustness of the conclusions to potential influences of unobserved covariates. Such sensitivity analyses suppose that a relevant but unobserved covariate has been left out of the propensity score model. By explicating how this hypothetical unmeasured covariate is related to treatment assignment and outcome, one can estimate how the treatment effect that adjusts for it might change if such a covariate was available for adjustment (Rosenbaum, 1986, 2002).

Second, this approach assumes that the investigator can distinguish between confounders (i.e. variables causally prior to treatment and outcome) and mediators (variables that are caused by treatment and are associated with an outcome). If mediators are mistakenly included in the confounder pool, the effects of treatment may be biased. Simliarly, if important confounders are left out of the propensity score pool (and, therefore, are not balanced), the analysis may bias treatment effects. As discussed above, this problem may be avoided by a careful selection of a confounder pool and the inclusion of all potential confounders in the estimation of the propensity score. Based on our experience, we recommend that the selection of a confounder pool will be made following extensive deliberations among members of the evaluation team and after seeking expert advice from colleagues.

It is also worth noting that propensity score methodology tends to work better in larger samples because imbalances of some covariates after sub-classification on the propensity score are likely to be minor in large samples but substantial enough to bias effect estimates in smaller samples (Rubin, 1997). In addition, like other applications, the effectiveness of propensity score methodology is conditional on the quality of data available for the analysis. In particular, a large number of missing values on covariates may present a problem when estimating the propensity score (D'Agostino & Rubin, 2000). In the illustrations above, the number of missing values was too small to be consequential and a listwise deletion procedure was used. When the problem is more severe, researchers may want to consider a different strategy of handling missing values such as multiple imputation for missing data (Allison, 2001; Rubin, 1987; Schafer, 1997). Finally, in the case of a treatment variable with more than two levels, it is possible that McCullagh's ordinal logistic regression model does not provide an adequate fit to the data. This is obviously true for non-ordinal treatment levels. For example, a recent study (Ta-Seale, Croghan, & Obenchain, 2000) used a propensity score methodology in a case where treatment consisted of a nominal variable with 3 categories (assignment of patients to three different drugs). While approaches for dealing with this complexity are currently being developed (see Imbens, 2000), further research is necessary to extend the propensity score methodology to such cases.

The bottom line is that propensity score methodology may be a particularly useful tool for researchers who are interested in improving causal inference from observational data. This methodology is widely used in several fields, particularly in the area of applied medicine and biostatistics (D'Agostino, 1998). We therefore, encourage program planners and evaluators to explore this methodology further and implement it when evaluating the effectiveness of programs based on observational data.

References

Allison, P. D. (2001). *Missing data*. Thousand Oaks, CA: Sage.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*, 444–472.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, *24*, 205–213.

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

D'Agostino, R. B., & Rubin, D. B. (2000). Estimating and using propensity scores with partially missing data. *Journal of the American Statistical Association*, *95*(451), 749–759.

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

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

Heckman, J. J. (1995). *Instrumental variables: A cautionary tale (No. Technical Working Paper No. 185)*. Cambridge, MA: National Bureau of Economic Research.

Hirano, K., Imbens, G., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, *71*, 1161–1189.

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

Hornik, R., Maklan, D., Cadell, D., Barmada, C. H., Jacobson, L., Prado, A., et al. (2002). *Evaluation of the National Youth Anti-Drug Media Campaign: Fifth semi-annual report of findings*. Rockville, MD: Westat.

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

Joffe, M. M., & Rosenbaum, P. R. (1999). Invited commentary: Propensity scores. *American Journal of Epidemiology*, *150*(4), 327–333.

Lu, B., Zanutto, E., Hornik, R., & Rosenbaum, P. R. (2001). Matching with doses in an observational study of a media campaign against drug abuse. *Journal of the American Statistical Association*, *96*(456), 1245–1253.

McCullagh, P. (1980). Regression model for ordinal data. *Journal of the Royal Statistical Society*, *42*, 109–142.

Newhouse, J., & McClellan, M. (1998). Econometrics in outcomes research: The use of instrumental variables. *Annual Review of Public Health*, *19*, 17–34.

Perkins, S. M., Tu, W., Underhill, M. G., Zhou, X.-H., & Murray, M. D. (2000). The use of propensity scores in pharmacoepidemiologic research. *Pharmacoepidemiology and drug safety*, *9*, 93–101.

Rosenbaum, P. R. (1986). Dropping out of high school in the United States: An observational study. *Journal of Educational Statistics*, *11*, 207–224.

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

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

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

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

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, *39*, 33–38.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine*, *127*(Part 2), 757–763.

Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics*, *52*, 249–264.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. New York: Chapman and Hall.

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

Stone, R. A., Obrosky, D. S., Singer, D. E., Kapoor, W. N., & Fine, M. J. (1995). Propensity score adjustment for pretreatment differences between hospitalized and ambulatory patients with community-acquired pneumonia. *Medical Care*, *33*, AS56–AS66.

Ta-Seale, M., Croghan, T. W., & Obenchain, R. (2000). Determinants of antidepressant treatment compliance: Implications for policy. *Medical Care Research and Review*, *57*, 491–512.

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

Zanutto, E. L., Lu, B., & Hornik, R. (2005). Using propensity score subclassification for multiple treatment doses to evaluate a National Anti-Drug Media Campaign. *Journal of Educational and Behavioral Statistics*.