Published in final edited form as:
Stat Med. 2009 October 15; 28(23): 2857–2875. doi:10.1002/sim.3669.

On the Use of Propensity Scores in Principal Causal Effect Estimation

Booil Jo and
Department of Psychiatry & Behavioral Sciences Stanford University School of Medicine Stanford, CA 94305-5795 booil@stanford.edu

Elizabeth A. Stuart
Department of Mental Health Department of Biostatistics Johns Hopkins Bloomberg School of Public Health 624 N Broadway, Baltimore, MD 21205 estuart@jhsph.edu

SUMMARY

We examine the practicality of propensity score methods for estimating causal treatment effects conditional on intermediate posttreatment outcomes (principal effects) in the context of randomized experiments. In particular, we focus on the sensitivity of principal causal effect estimates to violation of principal ignorability, which is the primary assumption that underlies the use of propensity score methods to estimate principal effects. Under principal ignorability, principal strata membership is conditionally independent of the potential outcome under control given the pre-treatment covariates; i.e., there are no differences in the potential outcomes under control across principal strata given the observed pretreatment covariates. Under this assumption, principal scores modeling principal strata membership can be estimated based solely on the observed covariates and used to predict strata membership and estimate principal effects. While this assumption underlies the use of propensity scores in this setting, sensitivity to violations of it has not been studied rigorously. In this paper, we explicitly define principal ignorability using the outcome model (although we do not actually use this outcome model in estimating principal scores) and systematically examine how deviations from the assumption affect estimates, including how the strength of association between principal stratum membership and covariates modifies the performance. We find that when principal ignorability is violated, very strong covariate predictors of stratum membership are needed to yield accurate estimates of principal effects.

Keywords

randomized experiments; intermediate outcomes; principal ignorability; principal scores; principal stratification; propensity scores

1 INTRODUCTION

Frangakis and Rubin [1] laid out the arguments regarding careful definition of causal effects in the presence of posttreatment intermediate outcome variables. This idea, known broadly as principal stratification, provides a framework for valid estimation of causal treatment effects conditional on intermediate outcomes. The crucial insight is that effects estimated in these settings must condition on the set of potential values of intermediate outcomes under all treatment conditions. Principal stratification refers to this classification of individuals based on sets of potential values of intermediate variables. The resulting categories, labeled as principal strata, are unaffected by treatment assignment, and therefore treatment effects calculated by conditioning on these categories, labeled principal effects, can be interpreted as causal. Treatment receipt behavior is one of the most frequently studied intermediate

outcome variables in this context. For example, in the two-arm experimental setting considered in Angrist, Imbens, and Rubin [2], according to principal stratification, individuals are categorized into compliance strata on the basis of their potential treatment receipt behavior under both the treatment and control conditions. In that setting, primary interest is in estimating the effect for compliers: those individuals who would take the treatment when in the treatment group and would take the control when in the control group.

Since the potential values of intermediate outcomes can be observed only under the condition to which each individual is actually assigned, estimating causal treatment effects conditional on principal strata remains challenging in practice. There has been relatively little investigation of the situations under which different approaches work well. One potentially intuitive and straightforward way of estimating principal effects is to use propensity score-based methods to classify individuals into principal strata. For example, in the Angrist et al. [2] setting, where the control group does not have access to the treatment, we may be interested in estimating causal treatment effects for those who would and would not receive the treatment under the treatment condition (i.e., for two principal strata, compliers and non-compliers). The challenge here is that stratum membership is known for individuals assigned to the treatment condition, but unknown for individuals assigned to control. If we could identify the stratum membership of individuals assigned to the control condition, inference would be straightforward. Given this goal, one potential way to estimate principal effects is to use propensity score methods to identify individuals in the control group who are likely to be compliers. We can then compare outcomes under the treatment and control conditions within each principal stratum. This intuitive idea has been used and discussed by previous researchers [3–6] in the context of principal stratification models, although little is known about the practicality or performance of the approach, and there has been relatively little formal discussion of the assumption that underlies the approach, which we term “principal ignorability.”

Follman [4] used the propensity score approach to estimate treatment effects accounting for levels of compliance. Follman [4] estimated a model of treatment receipt using the treatment group members (the propensity score model), and then used the predicted probabilities of treatment receipt in outcome models. In particular, he treated the propensity score as a baseline covariate and included an interaction of it and treatment assignment in the outcome model, essentially estimating a subgroup effect with the subgroup defined by predicted level of treatment receipt. Hill et al. [3] used a similar approach to look at the effects of high levels of participation in an early intervention for high-risk children, and found that higher-levels of participation led to stronger and longer-lasting effects. They labeled the individuals’ probability of belonging to particular principal strata “principal scores,” in connection with principal stratification. In the context of randomized trials with treatment noncompliance, Joffe et al. [5] used an approach similar to Follman [4], defining the compliance score as a measure of the effect of randomization on treatment received. They showed that validity of causal effect estimates depends on the link function when the compliance score is used as a regressor. Joffe et al. [6] also discussed identification of principal effects using the concept we refer to as principal ignorability in this paper. The key feature of this line of principal effect estimation methods is that estimation of principal scores and estimation of principal effects are separately conducted, which is possible based on its critical assumption that observed covariates are su cient for identifying principal stratum membership (principal ignorability). In contrast, in more commonly used principal effect estimation methods, which we term “joint estimation methods” and that do not necessarily require principal ignorability, principal stratum membership and the outcome are simultaneously modeled.

One potential benefit of the propensity score approaches to estimating principal effects is its separation of the estimation into two stages: first, a model relating the covariates to the intermediate outcome (e.g., treatment received), and second, a model relating the covariates to the potential outcomes, given the principal scores from stage 1. This two-stage approach leads to reduced reliance on a particular parametric model relating the covariates to the potential outcomes, and has been shown to be very beneficial in the observational studies context [7,8]. In particular, poor performance of parametric models in the principal stratification approach [9] leads us to consider whether the two-stage modeling approach could also be of use here. Another valuable advantage of the propensity score approaches is that they are methodologically simple and conceptually easy to understand. The trade-o is that propensity score approaches rely on being able to identify principal strata membership on the basis of only covariate information. However, in practice, there are often pretreatment covariates that are good predictors of principal stratum membership, and in some situations, it is also possible to actively and intentionally collect this information [10]. Given that, propensity score methods seem to be a promising strategy for estimating principal causal effects.

The goal of this paper is to assess when the covariate information is su cient to estimate principal effects well using propensity score methods and to examine how the strength of association between the principal strata membership and covariates affects the quality of the estimates. We believe that this is a critical step towards a better understanding of the practicality of propensity score methods in estimating principal causal effects. We also discuss commonly used joint estimation methods along with the two-stage methods. The main purpose of considering both the two-stage approach and the joint estimation approaches in the current paper is not to show superiority of one approach to another, but to better understand how the two-stage approach works in identifying principal effects. The next logical step would be to examine relative performance of the 2-stage and the joint estimation approaches under various conditions, which is not the focus of this paper.

2 MOTIVATING EXAMPLE: JOBS II (THE JOB SEARCH INTERVENTION STUDY)

This paper was particularly motivated by the Job Search Intervention Study (JOBS II: [11]), which was a randomized experiment of an intervention for unemployed individuals. The control condition consisted of a booklet briefly describing job search methods and tips. The intervention condition consisted of five training sessions intended to prevent poor mental health and to promote high-quality reemployment. On the basis of the definition of treatment receipt as having attended at least one out of five total sessions, 55% of individuals who were assigned to the intervention condition are treatment receivers. This definition, which dichotomizes treatment receipt status, was adopted in previous studies that analyzed the data from JOBS II [10–13], and also will be used in the current paper.

Given that a substantial proportion of individuals in the intervention group did not receive the full intervention treatment, estimates of treatment efficacy vary depending on how we take treatment receipt into account. One strategy is to estimate causal treatment effects for individuals with different compliance types, which fits in the framework of principal stratification. Specifically, treatment receipt behavior is the posttreatment intermediate outcome of interest in this situation. In JOBS II, individuals assigned to the control condition were prohibited from attending intervention sessions. Then with binary treatment receipt and binary treatment assignment status, only two compliance types (principal strata) are possible. Following the convention of Angrist et al. [2], we will use compliers and never-takers to refer to these two types of individuals: compliers receive treatment if they are assigned to the treatment condition and never-takers who do not receive the treatment

even if they are assigned to the treatment condition. This setting provides an ideal situation to calculate propensity scores that reflect principal stratum membership (compliance type), which is fully observed among individuals assigned to the intervention condition.

In this paper, we focus on the high-risk group, based on previous studies [14,15] that indicated that the job search intervention had its primary impact on high-risk individuals. The risk score was computed based on a set of variables in the screening data that predict depressive symptoms at follow-up (depression, financial strain, and assertiveness). A total sample size of 410 was analyzed in this study after listwise deletion of cases that had missingness in covariates or outcomes. Of those 410 individuals, 273 are in the intervention condition and 137 are in the control condition. The true randomization probability was 0.7 for the intervention and 0.3 for the control (i.e., more individuals were assigned to the treatment condition). This unbalanced design was chosen in the JOBS II trial because considerable treatment noncompliance was expected. Table 1 shows the results from a logistic regression of treatment receipt (compliance) on pretreatment covariates, estimated among individuals assigned to the treatment condition. Among the seven variables included in the model, four variables were significant predictors of treatment receipt: sense of mastery, age, motivation to attend intervention seminars, and the level of education. In particular, as would be expected, motivation was very highly associated with actual treatment receipt (odds ratio = 3.782). The presence of good predictors of compliance in JOBS II provides a nice setting to investigate how the use of propensity scores works in principal causal effect estimation.

3 COMMON SETTING

More generally, we consider a simple two-arm experimental setting, where individuals are randomly assigned either to the treatment or to the control (absence of treatment) condition. The treatment assignment status $Z_i = 1$ if individual i is randomly assigned to the treatment condition, and $Z_i = 0$ if assigned to the control condition. The observed treatment receipt status $S_i = 1$ if individual i receives the treatment, and $S_i = 0$ otherwise. Let $S_i(1)$ and $S_i(0)$ denote the potential treatment receipt status for i when $Z_i = 1$ and $Z_i = 0$, respectively. For simplicity, we assume that individuals assigned to the control condition are not allowed to access the treatment (i.e., $S_i(0) = 0$ for all i), although this assumption is not essential for identifying principal effects using the propensity score approach.

In this setting, two principal strata (compliance types) are possible based on binary Z and binary S. The latent compliance type $C_i = 1$ if individual i would receive the treatment when the treatment is offered, and $C_i = 0$ if individual i would not receive the treatment regardless of the intervention assignment. According to Angrist et al. [2], these two types of individuals are compliers and never-takers. That is,

$$C_i = \begin{cases} 1 \text{ (complier)} & \text{if } S_i(1) = 1 \text{ and } S_i(0) = 0 \\ 0 \text{ (never-taker)} & \text{if } S_i(1) = 0 \text{ and } S_i(0) = 0. \end{cases}$$

Based on these two compliance types, we assume the following model for a continuous outcome Y for individual i and a single covariate X:

$$Y_i = \alpha_n + (\alpha_c - \alpha_n)\, C_i + \gamma_n Z_i + (\gamma_c - \gamma_n)\, C_i Z_i + \lambda_n X_i + (\lambda_c - \lambda_n)\, C_i X_i + \varepsilon_i. \quad (1)$$

The value of the potential outcome under treatment, $Y_i(1)$, as well as of the potential outcome under control, $Y_i(0)$, can both be obtained from the expression in Equation (1). In Equation (1), α_n and α_c are the mean potential outcomes under control for never-takers and

compliers, respectively, when $X = 0$. The difference, $\alpha_c - \alpha n$ can be interpreted as the difference in the potential outcome under control for compliers and never-takers with the same value of X, capturing additional differences between those groups not due to differences in the distributions of X. If the mean of X is not zero, α_c and α_n are intercepts, and we need to take into account the differential relationships of covariates with the outcome across strata to properly calculate the difference between compliers and never-takers. The relationship between the covariate (or, more generally, a vector of covariates) X and the outcome is expressed by λ_c for compliers and λ_n for never-takers. The average effect of treatment assignment for compliers is γ_c, known as the complier average causal effect (CACE). The average effect of treatment assignment for never-takers is $_n$, known as the never-taker average causal effect (NACE). The residual ε_i is assumed to be normally distributed with mean zero and variance σ_c^2 for compliers and σ_n^2 for never-takers. However, parametric assumptions such as normality are not essential for identifying principal effects using propensity scores.

In the presence of covariates that predict compliance, the proportions of compliers and never-takers can be expressed using logistic regression as

$$\begin{aligned} P(C_i{=}1 \mid X_i) &= \pi_i, \\ P(C_i{=}0 \mid X_i) &= 1 - \pi_i, \\ logit(\pi_i) &= \beta_0 + \beta_1 X_i, \end{aligned} \tag{2}$$

where π_i is the probability that individual i is a complier, β_0 is the logit intercept, and β_1 is a vector of logit coefficients that reflects the association between compliance and pre-treatment covariates. Let π_c denote the compliance rate.

We also assume that the following two common assumptions necessary to identify causal treatment effects hold.

- Ignorable treatment assignment: Treatment assignment is independent of the potential outcomes, given the observed covariates X [16–19]. In other words, there are no unobserved confounders that would lead to differences in the distributions of the pair of potential outcomes between the treatment and control groups, once we condition on the observed characteristics. This is automatically satisfied in randomized experiments.
- Stable unit treatment value (SUTVA): (1) the potential outcomes for each person are unaffected by the treatment assignment of other individuals, and (2) there is only one “version” of each treatment (e.g., the potential outcomes for someone under a given treatment do not depend on the person who delivers the treatment, and the control condition is the same for all individuals; [18–20]).

4 ESTIMATION OF PRINCIPAL CAUSAL EFFECTS

4.1 PROPENSITY SCORE METHODS

Propensity scores are generally used in the context of non-experimental studies where there is interest in comparing the outcomes of treated (or exposed) and comparison (unexposed) groups. Use of them in the principal stratification context is similar, but will also take advantage of being in the context of a randomized experiment. The main idea behind propensity scores is to facilitate the comparison of subjects who are as similar as possible on background characteristics by collapsing the full set of covariates into their most important scalar summary, the propensity score. Formally, the propensity score is defined as the probability of being in the treatment group given the observed covariates, $P(Z = 1 \mid X)$. It is

often estimated using logistic regression. Properties of the propensity score laid out in Rosenbaum & Rubin [21] show that analyses that match, weight, or condition on the propensity score can yield unbiased estimates of treatment effects.

Whereas the conventional use of propensity scores is to model treatment group membership in observational studies (where treatment group membership is the same as treatment group receipt), the distinction in this paper is that we use propensity scores to model principal stratum membership. In particular, the propensity score will model treatment receipt (i.e., compliance) in the context of a randomized experiment, where treatment assignment does not necessarily imply treatment receipt. Following Hill et al. [3], we will use "principal scores" hereafter to refer to propensity scores that model principal stratum membership: $P(C = 1|X)$.

The principal score approach utilizes ideas similar to those of the conventional propensity score approach. The central assumption in standard propensity score settings is that the observed covariates are all that are needed to predict treatment assignment status: i.e., that treatment assignment is independent of the potential outcomes, given the observed covariates (known as ignorability or unconfounded treatment assignment, as defined earlier). This assumption is modified as follows in the principal score approach:

- Principal Ignorability (PI): Principal stratum membership is independent of the potential outcomes given observed information: $E(C_i|X_i,Y_i(0),Y_i(1)) = E(C_i|X_i)$. In the current setting we consider, where there are only two principal strata, stratum membership is known under the treatment condition. Therefore, the assumption applies only to the control condition. Another expression of this assumption is that $Y_i(0) \perp C_i|X_i$, meaning that there are no unobserved differences in the prognosis under control across compliance strata, given the observed covariates. On the basis of Equation (1), this assumption can be expressed as:

$$\alpha_n = \alpha_c, \tag{3}$$

$$\lambda_n = \lambda_c. \tag{4}$$

In the definition of PI described in equations (3) and (4), α_c and α_n are intercepts in the linear model shown in equation (1). In principle, the two assumptions in equations (3) and (4) can be combined into one assumption that implies that the potential outcomes under control are the same for never-takers and compliers, given the observed covariates. For example, if the covariates are centered at their means in equation (1), $\alpha_c - \alpha_n$ can be interpreted as the difference in outcomes for compliers and non-compliers when none of them take the treatment.

As defined above, the PI assumption clearly involves outcomes, although estimation of principal scores does not utilize any outcome information. The use of outcome information is restricted to the second stage, as described above. This assumes that the observed covariates are su cient for identifying stratum membership. Because of that, however, when principal scores are used to estimate principal effects, the results will be biased if the assumptions shown in equations (3) and (4) are not met, as demonstrated in Monte Carlo simulations below. While the PI assumption generally underlies the use of principal scores to estimate principal effects, it is certainly not the only assumption that can be used to identify principal effects. Below we discuss an alternative assumption, the exclusion restriction, and other possible assumptions are described by Joffe et al. [6].

The principal score method takes advantage of the fact that assignment to the treatment and control groups was randomized. Because of randomization, we can expect that the composition of principal strata (compliance types) and the relationship between principal stratum membership and covariates will be the same across randomized groups. Given this condition, principal scores can be predicted for the control group using the relationship between covariates and treatment receipt observed in the treatment group. The intuitive idea behind the principal score approach is to estimate the compliers' potential outcomes under control by finding the likely compliers in the control group: those individuals who look similar to the treatment group compliers in terms of their baseline covariates, as expressed via the principal score. A similar argument is used to estimate the NACE, by finding the likely never-takers in the control group.

The specific steps we employed to estimate principal scores and principal effects in the current paper are described below. In this paper, we focus on two uses of the principal score: matching [22] and weighting [23,24]. We describe these methods in the context of estimating the CACE and the NACE. Note that when estimating the CACE, the treatment condition never-takers are not used, and likewise, when estimating the NACE, the treatment condition compliers are not used.

Step 1: We first fit the principal score model using the treatment group members only, predicting compliance (i.e., treatment receipt) given the covariates. The logistic regression model shown in (2) is used for this purpose.

Step 2: From the model estimated in Step 1, predicted compliance probabilities (i.e., principal scores) are generated for treatment and control group members (P).

Once principal scores are obtained from Steps 1 and 2, principal effects are estimated for compliers and never-takers using either matching or weighting.

Step 3A. Matching: We match treatment group compliers to control group members with similar principal scores. We used full matching [25], which in our example uses all treatment group compliers and the full control group and forms matched sets of individuals with at least one treated and at least one control in each matched set. The matched sets are created in an optimal way, minimizing the average principal score distance within the matched sets. For example, in regions of the principal score distribution with many treated individuals and few control, the matched sets will have many treated individuals and few controls in each. In contrast, in regions of the principal score distribution with few treated and many controls, the matched sets will have few treated (e.g., 1) and many controls in each. In the JOBS II example discussed below, when estimating the CACE these matched sets ranged from having 1 treated and 13 controls to having 7 treated and 1 control. When estimating the NACE the matched sets ranged from having 1 treated and 9 controls to having 1 control and 6 treated. These matched sets are then used to form weights to be used in the subsequent analyses of the outcome. In particular, treated individuals are each given a weight of one. Control individuals are given weights that reflect the ratio of treated:control in each matched set. For example, in a matched set that contains 2 treatment group compliers and 3 control individuals, each treated individual receives a weight of 1 and the 3 controls each receive a weight of 2/3. In this way the treatment group compliers represent themselves, and the control group members are weighted to look like the treatment group compliers. Further details are provided in Hansen [26] and Stuart & Green [27]. Once the matched sets and resulting weights are obtained, the principal effects are estimated using regression adjustment, where the outcome values are regressed on a treatment indicator and the covariates, with the weights from full matching. This is the first time that the outcome values are used; doing this regression adjustment in the matched samples has been shown to

have very good performance [8,28,29]. After the matching and regression model is run for the compliers to estimate the CACE, the same process is used to match the non-compliers to the control group and estimate the NACE.

Step 3B. Weighting: Using a technique known as weighting by the odds, we assign weights to treated and control group individuals. In particular, treatment group compliers receive a weight of 1, so that the treatment group compliers represent themselves. As is the case with the matching approach, treatment group never-takers are not used when estimating the CACE, and thus essentially receive a weight of 0. The control group members each receive a weight of $W = P/(1 – P)$, where P is the principal score (the probability of being a complier). This weighting serves to make the control group look like the set of treatment group compliers [30,31]. These weights can be thought of as similar to survey sampling weights, used to weight a sample of subjects up to some broader population. In this case, the denominator, $1 – P$, weights the control group to look like the full population of treatment group compliers and controls. The numerator, P, is then used to weight the group to look like the sample of compliers [31]. Control group members with small values of P (low probabilities of being a complier) will receive small weights. Control group members with large values of P (and who thus have characteristics similar to the compliers in the treatment group) will receive larger weights. For example, a control group member with a predicted probability of being a complier of 0.4 ($P_i = 0.4$) will receive a weight of .4/.6 = 2/3. In contrast, a control group member with a higher predicted probability of being a complier, $P_i = 0.7$, will receive a larger weight, .7/.3 = 2.3, since their higher probability means that they look more similar to the compliers and thus should be upweighted in analyses. The CACE is then estimated using a weighted regression model with these weights. The same procedure is used to estimate the NACE, except that the treatment group never-takers receive a weight of 1 and the control group members receive a weight of $(1 – P)/P$. This weighting serves to make the control group members weight up to the group of never-takers.

In a connection to equation (1), this estimation procedure estimates the CACE considering the part relevant to compliers from equation (1). That is,

$$Y_i = \alpha_c + \gamma_c Z_i + \lambda_c X_i + \varepsilon_i, \quad (5)$$

Estimation of NACE is then separately conducted considering the part relevant to never-takers from equation (1).

$$Y_i = \alpha_n + \gamma_n Z_i + \lambda_n X_i + \varepsilon_i, \quad (6)$$

4.2 JOINT ESTIMATION METHODS

The propensity score methods have a clear two-step process of principal score estimation and weighting or matching, followed by effect estimation using regression adjustment using the matched samples or weights. At no point does the method require a large joint parametric model of covariates, compliance, and outcomes. In contrast, more commonly used principal effect estimation methods often simultaneously model compliance and outcomes to estimate principal causal effects. We use the term “joint estimation methods” to refer to these methods. These methods require explicit assumptions about the relationships between compliance, outcomes, and covariates. In this paper, we employ joint estimation methods to confirm our interpretation of PI and to better understand how propensity score methods work in identifying principal effects.

We will consider two joint estimation methods. The first uses the same PI assumption used in propensity score methods. However, in the joint estimation approach, PI is neither a required nor a widely used assumption in identifying principal effects. We considered the PI assumption in the joint estimation approach in order to help our understanding of the 2-stage approach. The second joint estimation method employs an alternative assumption, the exclusion restriction (ER) [2], which has been frequently used to estimate CACE in randomized experiments. Note that the ER assumption replaces PI in the second method, given that we do not need both assumptions to identify principal effects. For both joint estimation methods, we use maximum likelihood estimation using the EM algorithm (ML-EM), treating the unknown principal stratum membership (compliance status) among individuals assigned to the control condition as missing data. Parametric ML-EM methods are commonly used in estimating principal effects, and therefore we will not repeat the details of the procedures here. In particular, CACE analyses of the JOBS II data using the ML-EM method can be found in several previous studies [10,12,13]. Other estimation methods, such as Bayesian models, have also been used to estimate principal effects [32–34]. Further discussion of the ER and PI assumptions (although they do not explicitly use the term PI), as well as other possible assumptions used in estimating principal effects, can be found in Joffe et al. [6].

Joint PI Model: First, we consider a joint estimation method version of the propensity score methods. Focusing on the current setting, under PI, there are no differences in the potential outcome under control between principal strata given observed covariate information (i.e., $Y_i(0) \perp C_i|X_i$). When the outcomes are involved in identifying principal strata membership, this assumption needs to be explicitly applied to the relationship between the outcome and its predictors. In the outcome model shown in (1), PI is interpreted as that the relationship between the covariates and the outcome is constant across compliance strata ($\lambda_c = \lambda_n = \lambda$) and that there are no intercept differences between compliers and noncompliers ($\alpha_c = \alpha_n = \alpha$) under the control condition. Since the principal score approach and this joint PI method both rely on the same assumption of principal ignorability, we expect that they will yield similar estimated effects.

Under PI, the outcome model in (1) can be written as

$$Y_i=\alpha+\gamma_n Z_i+(\gamma_c-\gamma_n) C_i Z_i+\lambda X_i+\varepsilon_i. \quad (7)$$

Joint ER Model: In the second joint model we consider, we employ the ER, which is the most commonly used identifying assumption in estimating CACE and is an alternative to the PI assumption.

- Exclusion Restriction (ER): For those whose intermediate outcome (S) value does not change in response to treatment assignment, the distributions of the potential outcomes are independent of the treatment assignment (i.e., for units with $S_i(0) = S_i(1) = 0$ or $S_i(0) = S_i(1) = 1$, $Y_i(0,S_i(0)) = Y_i(1,S_i(1))$). In our setting where there are only two principal strata (compliers and never-takers), this restriction leads to $E(Y_i|Z_i,C_i = 0) = E(Y_i|C_i = 0)$. Another expression of this assumption is that $Y_i \perp Z_i|C_i = 0$, meaning that there is no effect of treatment assignment for never-takers. In other words, if treatment assignment does not affect the treatment an individual receives, it cannot affect their outcomes (also known sometimes as the assumption of "no direct effect" of assignment on outcomes).

In the outcome model shown in (1), ER means that $\gamma_n = 0$. Under the ER condition, the outcome model in (1) can be written as

$$Y_i=\alpha_n+(\alpha_c-\alpha_n)\ C_i+\gamma_c\ C_i\ Z_i+\lambda_n\ X_i+(\lambda_c-\lambda_n)\ C_i\ X_i+\varepsilon_i. \quad (8)$$

In clinical trials, where blind, double-blind, or placebo-control conditions are possible, the exclusion restriction is generally considered relatively benign. In other situations such as social-behavioral intervention studies (e.g., JOBS II), the exclusion restriction is often more questionable. The joint ER model will be used in analyzing the JOBS II data as a way of checking sensitivity to violation of PI in the propensity score methods.

5 MONTE CARLO SIMULATIONS

The Monte Carlo simulation results presented in this study are based on 1000 replications with a sample size of 500. We examine the bias, mean square error (MSE), and coverage of the estimators, where coverage is defined as the proportion of replications where the true parameter values are covered by the nominal 95% confidence interval of the parameter estimates. Data were generated according to the randomized trial setting described in equations (1) and (2). As in JOBS II, there are only two compliance strata, compliers and never-takers, and compliance stratum membership is observed in the treatment condition and unobserved in the control condition. For simplicity, one continuous covariate X is used in the simulation study, where $X_i \sim N(0, 1)$. The true ratio of the treatment and control groups is 50%:50%. The true ratio of compliers and never-takers is also 50%:50%. We also explored situations with unequal ratios of treated:control and compliers:non-compliers. With 70% or 30% treatment group members (i.e., randomization probabilities of either 0.7 or 0.3), the results are very similar to those presented here, but with slightly higher standard errors and coverage rates. With 70% or 30% compliers, we also found results similar to those presented here, but with slightly improved performance for the larger group and slightly worse performance for the smaller group. For example, when there are 30% compliers the CACE is estimated slightly worse while the NACE is estimated slightly better, but the overall results are very consistent with those reported here.

In equation (1), the true γ_c (CACE) is 0.5 and the true γ_n (NACE) is 0.0, which means that the exclusion restriction holds. Readers may notice that the estimated CACE and NACE in Tables 2-4 always sum to approximately 0.5 for each simulation setting and method. For example, in Table 2, when the outcome means are 0.5 SD's apart and effects are estimated using matching, the CACE is 0.281 and the NACE is 0.224. This is not a coincidence. In this setting, the true average effect across the population is 0.25 (half of the population are compliers, with an effect of 0.5 and the other half are non-compliers, with an effect of 0). Thus, when estimating the effects for compliers and non-compliers, the average will still be 0.25, just as it would be 0.25 if, for example we estimated the effects separately for males and females and then averaged them. This point, and the intuition behind it, is discussed further in Stuart et al. [35].

The true residual variance is 1.0 for both compliers and never-takers ($\sigma_c^2=\sigma_n^2=1$). The true value for both γ_n and γ_c is 0.0. The covariate effect is set at the same value so that the difference between α_c and α_n is the only source of deviation from PI. The true α_c is always 1.0, but the true α_n takes values of 1.0, 1.5, 2.0, and 3.0, reflecting differences between α_c and α_n of 0.0, 0.5, 1.0, and 2.0 SD's. In equation (2), the true association between C and X varies from 0.1 to 0.5, expressed as an odds ratio.

Tables 2, 3, and 4 present simulation results using the two propensity score methods described earlier. The principal scores are estimated using the compliance model described in equation (2), without involving any outcome values. Once principal scores are estimated

for everyone, the principal effects are estimated using weighting or matching. Restrictions such as those imposed in the joint models described in equations (7) and (8) are not necessary in this procedure. From the joint estimation approach's point of view, this is like estimating compliance stratum membership assuming the outcome model described in equation (7) and then estimating principal effects using the model described in equation (1). In other words, we expect some bias in principal effect estimates using propensity score methods due to the discrepancy between the two models (i.e., deviation from PI). Little is known about how large the impact of the deviation from PI is and how the strength of association between *C* and *X* affects the method's performance.

Table 2 shows the sensitivity of the CACE and NACE estimates to the violation of PI when the odds ratio for the association between compliance and the covariate is 0.5. The results are slightly better in terms of bias when weighting is used instead of matching. In both methods, when PI holds, principal effect estimates show good coverage rates with small MSE. If similar results were obtained from real data analyses under the PI condition, we are likely to conclude that treatment assignment had a significant positive effect on compliers and had no effect on never-takers, which is the correct conclusion. As the deviation from PI increases, the quality of principal effect estimates rapidly deteriorates. With a moderate deviation from PI (0.5 SD), the coverage rates of the principal effect estimates decrease substantially, to between 50% and 60%. However, these results do not change our inference regarding the CACE in the sense that we would still conclude that the treatment assignment effects are positive and significant. When α_n and α_c are 1 SD apart, our inference regarding the CACE and NACE can be incorrect in that we may conclude that treatment assignment had no effect on compliers and had a significant positive effect on never-takers. When PI is severely violated (2 SD), the results may lead to conclusions that fully contradict the truth: that treatment assignment had a negative effect on compliers and a positive effect on never-takers.

Table 3 shows the sensitivity of the CACE and NACE estimates to the violation of PI when the odds ratio of the association between compliance and the covariate is 0.3. Again the results are slightly better in terms of bias when weighting is used instead of matching. Again, when PI holds, principal effect estimates show good coverage rates with small MSE. However, as the deviation from PI increases, the quality of principal effect estimates rapidly deteriorates, but at a slower rate than when the odds ratio is 0.5.

Table 4 shows the sensitivity of the CACE and NACE estimates to the violation of PI when the association between compliance and the covariate is extremely strong (odds ratio = 0.1). With both matching and weighting, the results show much improved quality of principal effect estimates than when the odds ratio was 0.5 or 0.3. This occurs because *X* is a stronger predictor of compliance status and thus the principal score better predicts which control individuals are compliers and which are never-takers. The improved performance of weighting as compared to matching is also somewhat larger than when the odds ratio is 0.5 or 0.3. Coverage rates still quickly deteriorate as the deviation from PI increases. However, even with a substantial deviation from PI (1 SD), we are still likely to reach conclusions that are consistent with the truth: that treatment assignment had a positive effect on compliers and no effect on never-takers. With a severe deviation from PI (2 SD), both methods are likely to fail to detect the positive effect of treatment assignment on compliers. However, we are at least unlikely to reach the opposite conclusion, that treatment assignment had a negative effect on compliers. The results reported in Tables 2, 3, and 4 imply that principal effect estimates obtained using propensity score methods need to be interpreted with caution unless the association between compliance and the covariate is very strong, or we have good reasons to believe that a substantial deviation from PI is unlikely.

6 APPLICATION TO JOBS II

As implied in the simulation study, propensity score methods may yield biased principal effect estimates if PI does not hold, especially if the association between compliance and the covariate is not very strong. In practice, however, it is hard to know whether PI is very likely or very unlikely to hold, and it is also unrealistic to expect almost perfect predictors of principal stratum membership. Given this situation, it is not straightforward to evaluate the quality of principal effect estimates obtained using propensity score methods. In this section, we present principal effect estimation results focusing on two outcomes (sense of mastery and depression) from JOBS II. With both outcome measures, ER approximately holds according to propensity score methods, which leads to an interesting situation where we can examine the deviation from PI assuming that ER holds. We utilize the fact that, under the ER condition, the principal effect for compliers (CACE) can be identified using the joint ER model.

Tables 5 and 6 show the principal effect estimation results using propensity score methods as well as the two joint estimation methods described earlier. For the propensity score methods we used the MatchIt package [36] for the R statistical software package [37]. For ML-EM estimation of principal effects using the JOBS II data, we used the M*plus* program version 5.1 [38]. The baseline covariates presented in Table 1 are used in estimating the principal effects: depression, mastery, economic hardship (econ), age, motivation, grade, and gender. All 7 covariates are used in propensity score methods to estimate the principal scores and in estimating the outcome model. All 7 covariates are also used in modeling compliance and the outcome in the joint estimation methods. However, in practice, compliance and the outcome often have different predictors. The 2-stage methods and the joint estimation methods both work well with different sets of covariate predictors for compliance and the outcome. Note that having good predictive covariates of compliance is more critical when modeling both compliance and the outcome than when modeling the outcome only.

For the outcome model, we report only the covariates that are significant predictors of the outcome according to any of the 4 analysis methods. Unlike in the Monte Carlo simulations presented in the previous section, the effect of covariates may vary across principal strata in real data analyses. As an approximate way of maintaining the interpretation of α_c and α_n as outcome means instead of as intercepts, we centered covariates at their observed means.

Table 5 shows the estimated effects on sense of mastery six months after the intervention. Sense of mastery was one of the outcomes hypothesized to be affected by the intervention, and it may also be related to later outcomes such as reemployment [39]. Among the seven covariates, baseline sense of mastery was a significant predictor of later sense of mastery in all 4 methods. All 4 methods indicate a positive impact of treatment assignment for compliers and little impact for never-takers on sense of mastery. The estimates of CACE and NACE based on the joint PI method are very similar to those from the propensity score methods, confirming that the joint PI method is comparable to propensity score methods under the setting we consider in Equations (1) and (2). The differences between the two approaches may increase as we deviate from this setting (e.g., deviation from normality, nonlinear relationship between Y and X).

A particularly interesting feature is how the methods can be used to inform each other, and in particular the validity of the underlying assumptions. The principal score approaches assume PI and allow us to estimate the NACE, which informs how well the ER holds. The joint estimation ER approach, on the other hand, assumes ER, and allows us to examine how likely PI is to hold. According to the propensity score methods, ER approximately holds,

with NACE estimates not statistically significant and close to zero. According to the joint ER model, there is a minor deviation from PI. That is, α_n (3.606) and α_c (3.543) are about 0.13 SD apart (based on SD pooled across the treatment and control conditions). This likely explains why similar principal effect estimates are obtained across methods assuming PI and ER: both assumptions seem reasonable for this outcome.

Table 6 shows the principal effect estimation results when the level of depression six months after the intervention is the outcome. Among the mental health problems associated with job loss, depressive symptoms are the most commonly reported [40]. Among the 7 covariates, baseline sense of mastery, economic hardship, and motivation were each significant predictors of depression in at least one of the methods. In all 4 methods, the results indicate a positive impact (i.e., lower depression rates) of treatment assignment for compliers and little impact for never-takers on depression. However, the size of the CACE is somewhat different depending on whether PI or ER is assumed. Again, according to propensity score methods (which assumes PI), ER approximately holds when depression is the outcome. We utilized this information in the joint ER model. However, unlike for sense of mastery, the results for depression show a larger deviation from PI. According to the joint ER model, there is a moderate deviation from PI. That is, α_n (1.961) and α_c (2.360) are about 0.55 SD apart, which explains the noticeable, although not dramatic, differences in the principal effect estimates across the methods assuming PI and ER. In JOBS II, depression was measured with a subscale of 11 items based on the Hopkins Symptom Checklist [41] such as crying easily and feeling no interest in things. The 11-item scale required respondents to indicate on a five-point scale how much they experienced each item (1=not at all, 2=a little bit, 3=moderately, 4=quite a bit, 5=extremely). The overall mean of the depression score at the 6 month followup is 1.99 and SD is 0.73. Given this distribution, a 1 SD difference is less than a one unit difference in the 5-point scale. In other words, a 0.5 SD to 1.0 SD difference in the depression score across compliers and never-takers seems to be within the practically possible range of deviation from PI.

The impact of violating PI we find here is consistent with the results we obtained in our simulation study. In Table 6, the size of CACE when using the joint ER method (–0.485) is about 1.5 times of the size of CACE when using the propensity weighting method (–0.331). In Table 3 (which is comparable to JOBS II in terms of the association between C and X), the true CACE (0.500) is about 1.6 times of the average size of CACE (0.322) when estimated using the propensity weighting method with a moderate deviation from PI (0.5 SD).

7 CONCLUSIONS

The Monte Carlo simulations presented here showed that, when principal ignorability holds, propensity score methods provide valid principal causal effect estimates regardless of the level of association between principal stratum membership and covariates. However as the deviation from principal ignorability increases, the quality of principal effect estimates rapidly deteriorates. Depending on how principal ignorability is violated, principal effects can not only be underestimated, but also can be overestimated. In general, principal effect estimates are quite sensitive to deviation from principal ignorability. A stronger association between principal stratum membership and covariates somewhat alleviates this sensitivity. However, noticeable benefit does not show until association between principal stratum membership and covariates is extremely strong. These results suggest that principal effect estimates obtained using propensity score methods need to be interpreted with caution unless the association between compliance and the covariate is very strong, or we have good reasons to believe that a substantial deviation from principal ignorability is unlikely.

In analyzing the JOBS II data, we employed both propensity score methods and more commonly used joint estimation methods. The two approaches are closely related, although how they handle missing principal stratum information is different. One benefit of considering both types of methods is that the two approaches can be used together to improve sensitivity analysis strategies in the sense of testing each other's assumptions. According to the combined information from propensity score methods and joint estimation methods, principal ignorability approximately holds with the sense of mastery outcome, and is moderately violated with the depression outcome. The ER appears to hold for both outcomes. In this example, the association between principal stratum membership and covariates is fairly strong. Under these conditions, applying propensity score methods seems reasonable, although the results should be cautiously interpreted considering the possible ranges of bias due to deviation from principal ignorability. However, JOBS II may not be a typical example in terms of the degree of deviation from principal ignorability. In some real data applications, principal ignorability may be more severely violated. We also utilized the fact that the exclusion restriction approximately holds in JOBS II according to the propensity score methods. This made it easy in our analyses to connect the propensity score and joint estimation methods. In other real data applications, this may not be the case. Further research is warranted to guide efficient methods of sensitivity analysis for propensity score methods in the principal stratification context.

This work prompts discussion of areas for future work. One is the complication of missing data. For simplicity we used simple case-wise deletion to illustrate these methods. Unfortunately, incorporating missing data into propensity score methods is not straightforward and is an area with relatively little work (the few papers in this area include D'Agostino et al. [42] and Song et al. [43]). It is more straightforward to take missing data into account in the joint estimation methods. Another important direction for future work is investigation of other matching methods. A wide variety of matching methods exist [22], and should be investigated to determine if some approaches may work better (or worse) in particular settings. For example, 1:1 matching without replacement was not possible in the JOBS II data because the number of control individuals was smaller than the number of treatment group compliers (and thus one control match could not be found for each treatment group complier). We used full matching as an alternative. Full matching has the benefit that the full control group is used and no controls are discarded, but it is possible that in other data analyses 1:1 matching may in fact be preferable. More work is needed to identify when different matching approaches are most effective.

This work can also inform the design of future trials, in particular highlighting the need for good predictors of compliance behavior when there is interest in estimating the CACE. This may involve collecting data on variables not normally thought of in study design, such as the motivation variable collected in JOBS II, or other variables that may capture how di cult it is for individuals to participate in the treatment (such as whether they own a car, live on a bus route, have child care, etc.). In the early stages of the study, researchers should consider the types of effects that they are interested in estimating, and determine which variables may help them predict the necessary behaviors, such as compliance. For example, the simulations shown here show the importance of having strong predictors of compliance when estimating the CACE.

Acknowledgments

The research of the first author was supported by NIMH (MH066319, MH066247) and the research of the second author was also supported by NIMH (MH066247, MH083846). We thank participants of the Prevention Science Methodology Group for useful feedback. We also appreciate valuable input provided by Amiram Vinokur.

References

1. Frangakis CE, Rubin DB. Principal stratification in causal inference. Biometrics. 2002; 58:21–29. [PubMed: 11890317]
2. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. Journal of the American Statistical Association. 1996; 91:444–455.
3. Hill JL, Brooks-Gunn J, Waldfogel JBB. Sustained effects of high participation in an early intervention for low-birthweight premature infants. Developmental Psychology. 2003; 39:730–744. [PubMed: 12859126]
4. Follman DA. On the effect of treatment among would-be treatment compliers: An analysis of the Multiple Risk Factor Intervention Trial. Journal of the American Statistical Association. 2000; 95:1101–1109.
5. Joffe MM, Ten Have TR, Brensinger C. The compliance score as a regressor in randomized trials. Biostatistics. 2003; 4:327–340. [PubMed: 12925501]
6. Joffe MM, Small D, Hsu C-Y. Defining and estimating intervention effects for groups that will develop an auxiliary outcome. Statistical Science. 2007; 22:74–97.
7. Rubin DB. The use of matched sampling and regression adjustment to remove bias in observational studies. Biometrics. 1973; 29:185–203.
8. Ho DE, Imai K, King G, Stuart EA. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Political Analysis. 2007; 15:199–236.
9. Griffin BA, McCaffrey DF, Morral AR. An application of principal stratification to control for institutionalization at follow-up in studies of substance abuse treatment programs. Annals of Applied Statistics. 2008; 2:1034–1055. [PubMed: 19779599]
10. Jo B. Model misspecification sensitivity analysis in estimating causal effects of interventions with noncompliance. Statistics in Medicine. 2002; 21:3161–3181. [PubMed: 12375297]
11. Vinokur AD, Price RH, Caplan RD. From field experiments to program implementation: Assessing the potential outcomes of an experimental intervention program for unemployed persons. American Journal of Community Psychology. 1991; 19:543–562. [PubMed: 1755435]
12. Jo B. Estimating intervention effects with noncompliance: Alternative model specifications. Journal of Educational and Behavioral Statistics. 2002; 27:385–420.
13. Little RJA, Yau L. Statistical techniques for analyzing data from prevention trials: Treatment of no-shows using Rubin's causal model. Psychological Methods. 1998; 3:147–159.
14. Vinokur AD, Price RH, Schul Y. Impact of the JOBS intervention on unemployed workers varying in risk for depression. American Journal of Community Psychology. 1995; 23:39–74. [PubMed: 7572826]
15. Price RH, van Ryn M, Vinokur AD. Impact of a preventive job search intervention on the likelihood of depression among the unemployed. Journal of Health and Social Behavior. 1992; 33:158–167. [PubMed: 1619263]
16. Holland PW. Statistics and causal inference. Journal of the American Statistical Association. 1986; 81:945–970.
17. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology. 1974; 66:688–701.
18. Rubin DB. Bayesian inference for causal effects: the role of randomization. Annals of Statistics. 1978; 6:34–58.
19. Rubin DB. Discussion of “Randomization analysis of experimental data in the Fisher randomization test” by D. Basu. Journal of the American Statistical Association. 1980; 75:591–593.
20. Rubin DB. Comment on Neyman (1923) and causal inference in experiments and observational studies. Statistical Science. 1990; 5:472–480.
21. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983; 70:41–55.
22. Stuart, EA.; Rubin, DB. Best Practices in Quasi-Experimental Designs: Matching methods for causal inference.. In: Osborne, J., editor. Best Practices in Quantitative Social Science. Sage Publications; Thousand Oaks, CA: 2007. p. 155-176.

23. Lunceford JK, Davidian M. (2004) Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. Statistics in Medicine. 2004; 23:2937–2960. [PubMed: 15351954]
24. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association. 1994; 89:846–866.
25. Rosenbaum PR. A Characterization of Optimal Designs for Observational Studies. Journal of the Royal Statistical Society. 1991; 53:597–610.
26. Hansen BB. Full matching in an observational study of coaching for the SAT. Journal of the American Statistical Association. 2004; 99:609–618.
27. Stuart EA, Green KM. Using Full Matching to Estimate Causal Effects in Non-Experimental Studies: Examining the Relationship between Adolescent Marijuana Use and Adult Outcomes. Developmental Psychology. 2008; 44:395–406. [PubMed: 18331131]
28. Abadie A, Imbens GW. Large sample properties of matching estimators for average treatment effects. Econometrica. 2006; 74:235–267.
29. Robins JM, Rotnitzky A, Bickel PJ, Kwon J. Inference for semiparametric models: Some questions and an answer. Statistica Sinica. 2001; 11:920–936. Comment on.
30. Hirano K, Imbens GW, Ridder G. Efficient estimation of average treatment effects using the estimated propensity score. Econometrica. 2003; 71:1161–1189.
31. McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. Psychological Methods. 2004; 9:403–425. [PubMed: 15598095]
32. Barnard J, Frangakis CE, Hill JL, Rubin DB. A Principal Stratification approach to broken randomized experiments: a case study of School Choice vouchers in New York City. Journal of the American Statistical Association. 2003; 98:299–323.
33. Hirano K, Imbens GW, Rubin DB, Zhou XH. Assessing the Effect of an Influenza Vaccine in an Encouragement Design. Biostatistics. 2000; 1:69–88. [PubMed: 12933526]
34. Jin H, Rubin DB. Principal stratification for causal inference with extended partial compliance. Journal of the American Statistical Association. 2008; 103:101–111.
35. Stuart EA, Perry DF, Le H-N, Ialongo NS. Estimating intervention effects of prevention programs: Accounting for noncompliance. Prevention Science. 2008; 9:288–298. [PubMed: 18843535]
36. Ho, DE.; Imai, K.; King, G.; Stuart, EA. MatchIt: Nonparametric Preprocessing for Parametric Causal Inference (Version 2.211) [Software].. The Journal of Statistical Software. Forthcomingin press. Available at http://gking.harvard.edu/matchit/
37. R Development Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria: 2008. http://www.cran.r-project.org
38. Muthén, LK.; Muthén, BO. Muthén & Muthén; Los Angeles, CA: 1998-2008. Mplus user's guide..
39. Vinokur AD, Schul Y. Mastery and inoculation against setbacks as active ingredients in intervention for the unemployed. Journal of Consulting and Clinical Psychology. 1997; 65:867–877. [PubMed: 9337505]
40. Fryer D, Payne R. Being unemployed: A review of the literature on the psychological experience of unemployment. International Review of Industrial and Organizational Psychology. 1986:235–278.
41. Derogatis, LR.; Lipman, RS.; Rickles, K.; Uhlenuth, EH.; Covi, L. The Hopkins Symptom Checklist (HSCL).. In: Pichot, P., editor. Psychological measurements in psychopharmacology: Modern problems in pharmacopsychiatry. Karger, Basel; New York: 1974. p. 79-110.
42. D'Agostino RB Jr. Lang W, Walkup M, Morgan T. Examining the impact of missing data on propensity score estimation in determining the effectiveness of self-monitoring of blood glucose (SMBG). Health Services & Outcomes Research Methodology. 2001; 2:291–315.
43. Song J, Belin TR, Lee MB, Gao X, Rotheram-Borus MJ. Handling baseline differences and missing items in a longitudinal study of HIV risk among runaway youths. Health Services & Outcomes Research Methodology. 2001; 2:317–329.

Table 1

JOBS II: Logistic Regression of Compliance on Baseline Covariates in the Treatment Condition (compliers vs. never-takers)

Parameter	Estimate	*SE*
Intercept	–1.756	1.846
Depression	–0.384	0.455
Mastery	–0.572	0.271
Economic hardship	–0.315	0.163
Age	0.064	0.015
Motivation	1.330	0.314
Grade	0.278	0.070
Female	–0.471	0.283

Table 2

Simulation: Principal causal effect estimation using propensity scores when there is a predictor of compliance with OR = 0.5

Deviation from PI ($\alpha_c - \alpha_n$)	Propensity Matching				Propensity Weighting			
	Raw Bias	SE	MSE	Coverage	Raw Bias	SE	MSE	Coverage
CACE (γ_c), true value=0.5								
0.0 SD	0.006	0.109	0.017	0.896	–0.003	0.113	0.013	0.944
0.5 SD	–0.219	0.111	0.065	0.515	–0.215	0.114	0.059	0.536
1.0 SD	–0.444	0.117	0.216	0.059	–0.432	0.118	0.201	0.047
2.0 SD	–0.895	0.139	0.825	0.000	–0.867	0.132	0.769	0.000
NACE (γ_n), true value=0								
0.0 SD	–0.002	0.109	0.017	0.895	–0.003	0.113	0.014	0.939
0.5 SD	0.224	0.111	0.068	0.470	0.215	0.114	0.061	0.517
1.0 SD	0.449	0.117	0.221	0.068	0.434	0.118	0.203	0.055
2.0 SD	0.900	0.139	0.836	0.000	0.870	0.132	0.776	0.000

Stat Med. Author manuscript; available in PMC 2010 October 15.

Table 3

Simulation: Principal causal effect estimation using propensity scores when there is a predictor of compliance with OR = 0.3

Deviation from PI ($\alpha_c - \alpha_n$)	Propensity Matching				Propensity Weighting			
	Raw Bias	SE	MSE	Coverage	Raw Bias	SE	MSE	Coverage
CACE (γ_c), true value=0.5								
0.0 SD	0.005	0.109	0.018	0.888	0.000	0.118	0.015	0.944
0.5 SD	–0.190	0.111	0.054	0.589	–0.178	0.120	0.047	0.659
1.0 SD	–0.384	0.117	0.167	0.135	–0.355	0.123	0.142	0.207
2.0 SD	–0.772	0.136	0.622	0.000	–0.710	0.137	0.524	0.000
NACE (γ_n), true value=0								
0.0 SD	–0.002	0.109	0.018	0.895	–0.002	0.118	0.015	0.939
0.5 SD	0.193	0.111	0.056	0.559	0.176	0.120	0.046	0.689
1.0 SD	0.387	0.117	0.170	0.133	0.354	0.123	0.142	0.201
2.0 SD	0.775	0.136	0.628	0.001	0.710	0.137	0.525	0.004

Stat Med. Author manuscript; available in PMC 2010 October 15.

Table 4

Simulation: Principal causal effect estimation using propensity scores when there is a predictor of compliance with OR = 0.1

Deviation from PI ($\alpha_c - \alpha_n$)	Propensity Matching				Propensity Weighting			
	Raw Bias	SE	MSE	Coverage	Raw Bias	SE	MSE	Coverage
CACE (γ_c), true value=0.5								
0.0 SD	0.002	0.109	0.020	0.872	–0.001	0.128	0.018	0.933
0.5 SD	–0.136	0.111	0.039	0.718	–0.109	0.129	0.030	0.851
1.0 SD	–0.275	0.115	0.097	0.452	–0.217	0.132	0.067	0.617
2.0 SD	–0.552	0.131	0.331	0.032	–0.432	0.144	0.210	0.171
NACE (γ_n), true value=0								
0.0 SD	–0.001	0.109	0.020	0.879	–0.001	0.128	0.017	0.947
0.5 SD	0.138	0.111	0.040	0.695	0.106	0.129	0.029	0.862
1.0 SD	0.276	0.115	0.099	0.349	0.214	0.132	0.064	0.617
2.0 SD	0.554	0.131	0.335	0.048	0.430	0.144	0.207	0.162

Stat Med. Author manuscript; available in PMC 2010 October 15.

Table 5

JOBS II: Estimation of Principal Causal Effects on Sense of Mastery

	Matching		Weighting		Joint PI		Joint ER	
Parameter	**Estimate**	**SE**	**Estimate**	**SE**	**Estimate**	**SE**	**Estimate**	**SE**
γ_c (CACE)	0.243	0.048	0.224	0.057	0.203	0.047	0.238	0.072
γ_n (NACE)	0.056	0.052	0.036	0.053	0.054	0.051	0.000	.
α_c	3.538	0.036	3.557	0.048	3.578	0.036	3.543	0.066
α_n	3.612	0.036	3.633	0.038	3.578	0.036	3.606	0.068
λ_c (Mastery)	0.575	0.048	0.607	0.086	0.551	0.035	0.537	0.050
λ_n (Mastery)	0.559	0.052	0.530	0.050	0.551	0.035	0.559	0.056
σ_c^2	0.162		0.171		0.145	0.016	0.143	0.015
σ_n^2	0.160		0.136		0.151	0.020	0.149	0.021

Stat Med. Author manuscript; available in PMC 2010 October 15.

Table 6

JOBS II: Estimation of Principal Causal Effects on Depression

	Matching		Weighting		Joint PI		Joint ER	
Parameter	**Estimate**	**SE**	**Estimate**	**SE**	**Estimate**	**SE**	**Estimate**	**SE**
γ_c (CACE)	−0.285	0.082	−0.331	0.099	−0.275	0.084	−0.485	0.153
γ_n (NACE)	−0.045	0.090	−0.087	0.099	−0.150	0.091	0.000	.
α_c	2.161	0.061	2.206	0.083	2.150	0.067	2.360	0.144
α_n	2.050	0.063	2.088	0.079	2.150	0.067	1.961	0.063
λ_c (Mastery)	−0.363	0.082	−0.404	0.136	−0.303	0.078	−0.403	0.100
λ_n (Mastery)	−0.189	0.090	−0.255	0.100	−0.303	0.078	−0.138	0.113
λ_c (Econ)	0.131	0.048	0.161	0.065	0.156	0.043	0.185	0.066
λ_n (Econ)	0.090	0.051	0.146	0.062	0.156	0.043	0.132	0.061
λ_c (Motiv)	0.143	0.084	0.289	0.111	0.111	0.080	0.066	0.102
λ_n (Motiv)	0.221	0.116	0.268	0.158	0.111	0.080	0.056	0.152
σ_c^2	0.478		0.488		0.475	0.075	0.498	0.054
σ_n^2	0.479		0.469		0.491	0.092	0.397	0.062

Stat Med. Author manuscript; available in PMC 2010 October 15.