Innovative Design and Analysis Strategies in the Evaluation of the National Youth Anti-Drug Media Campaign: Propensity Scores and Counterfactual Projection Weights in a National Probability Survey

Robert Orwin[1], Robert Hornik[2], David Judkins[1], Paul Zador[1], Sanjeev Sridharan[1], Robert Baskin[3]

[1] Westat, 1650 Research Blvd., Rockville, MD 20850/ robertorwin@westat.com, davidjudkins@westat.com, paulzador@westat.com, sanjeevsridharan@westat.com

[2] Annenberg School for Communication, University of Pennsylvania, 3620 Walnut Street, Philadelphia, PA 19104/ rhornik@asc.upenn.edu

[3] Administration on Health Care Quality, 2101 E. Jefferson, Rockville MD 20852/ RBaskin@ahrq.gov

Introduction

The National Youth Anti-Drug Media Campaign was funded by the Congress in 1998 to reduce and prevent drug use among young people by addressing youth directly, as well as indirectly, by encouraging their parents and other adults to take actions known to affect youth drug use. The major intervention components include television, radio, and other advertising, complemented by public relations efforts including community outreach and institutional partnerships. An evaluation of the Campaign is being conducted under contract to the National Institute on Drug Abuse (NIDA) by Westat and its subcontractor, the Annenberg School for Communication at the University of Pennsylvania. Funding of the evaluation is provided by ONDCP from the appropriation for the Media Campaign itself. The primary tool for the evaluation is the National Survey of Parents and Youth (NSPY). This survey is collecting initial and followup data from nationally representative samples of youth between 9 and 18 years of age and parents of these youth. The principal goal of the evaluation is to assess the effectiveness of the Campaign in reducing drug use initiation and its precursors (e.g., attitudes, intentions) among youth, particularly with respect to marijuana. The model of Campaign influence integrates propositions from the Theory of Reasoned Action and Social Cognitive Theory.

From an evaluation standpoint, it would have been desirable to have a) staged implementation in randomly selected test markets to provide a comparison group and b) an interview wave that preceded implementation (i.e., a true baseline). As is often the case with high-visibility federal programs, however, the program was implemented nationwide, and began full scale operations before a baseline survey could be fielded. With no designated comparison group and no controlled assignment to exposure conditions, the evaluation design had to rely on the exploitation of natural variation in exposure across respondents. Causal claims of Campaign effectiveness would rely on the ability to adjust associations between measures of Campaign exposure and measures of outcomes for competing explanations (i.e., confounders).

The analysis used propensity scoring to control for confounders. The method was first introduced by Rosenbaum and Rubin (1983) and is increasingly used to analyze date from observational studies (D'Agostino, 1998; Rubin, 1997). To our knowledge, this was the first attempt to apply propensity scoring to evaluate the effectiveness of a social intervention in the context of a complex, longitudinal national probability survey. An additional innovation was the creation of counterfactual projection (CFP) "weights" for each respondent. These weights were then used in the analysis as a way to remove the influence of confounders, in tandem with sampling weights, nonresponse weights, and post-stratification weights. Consequently, the effect estimates generated are simultaneously adjusted for all these factors.

This paper focuses on the methods for developing, testing, and using propensity scores in NSPY, the extension to CFP weights, and selected findings based on those efforts. For the full study design and findings to date, see the Fifth Semiannual Report: Evaluation of the National Youth Antidrug Media Campaign at http://www.nida.nih.gov/despr/westat/

Exposure-Outcome Associations as the Basis for Inferring Campaign Effects

The basis for inferring Campaign effect estimates was the association between exposure and outcome, adjusted for the influence of confounders and survey weights. Each association between exposure and outcome was estimated twice: cross-sectionally (exposure and outcome assessed in the same interview wave) and longitudinally (assessment of outcome lags assessment of exposure by one interview wave, with approximately 18 months between interviews). Cross-sectional

associations were indicative of short-term effects of Campaign exposure on outcomes, while longitudinal associations were indicative of longer-term effects. From a causal inference standpoint, the cross-sectional estimates have an inherent limitation. Suppose the adjusted association showed a favorable effect of exposure, e.g., parents reporting more exposure to ads report greater monitoring of their children. There remains an alternative explanation for the adjusted association, namely that outcome is the cause and (recall of) exposure is the effect (i.e., reverse causation). It would not be unexpected, for example, that a parent who is already engaged in monitoring their child's activities (one of the targeted outcomes for parents), is more predisposed to recall an ad that reinforces this behavior. We characterize this limitation as inherent because even if all confounders—observed and unobserved-- were successfully controlled for, the ambiguity of causal direction remains. This is not the case for the longitudinal association; if, after controlling for all confounders, exposure measured at time 1 is associated with outcome measured at time 2, then the causal direction must run from exposure to outcome since an effect cannot temporally precede its cause. Consequently, one can claim that a favorable cross-sectional association as *consistent* with a Campaign effect, while a longitudinal association is more persuasive as causal.

The primary effects of interest were the direct effects of youth exposure on youth behavior and parent exposure on parent behavior. We also examined the indirect effect of parent exposure on youth outcomes. This allowed us to explore an alternative mechanism for Campaign effects on youth that was consistent with the conceptual model of Campaign influence.[1]

Measures of Exposure and Outcome

The analysis focuses on five outcomes for youth: initiation of marijuana use; intentions to avoid initiating marijuana use; and three cognitive indices—attitudes and beliefs about marijuana use; perceptions of social norms about marijuana use; and self-efficacy to avoid marijuana use if it is available. There also are five outcome indices for parents: parent reports of talking with their children about drugs; an index of attitude and belief items concerning talk (talk cognitions); parent reports of monitoring their children; an index concerning monitoring (monitoring cognitions); and parent reports of engaging in fun activities with their children in and outside of the home. Parent and child responses are linked for some analysis (e.g., effect of parent exposure on child outcomes).

Ad exposure was measured in NSPY for both youth and parents by asking about recall of specific current or very recent TV and radio advertisements. The TV and radio advertisements were played for respondents on laptop computers in order to aid their recall. Youth were shown or listened to only youth-targeted ads, and parents were shown or listened to only parent-targeted ads. In addition, both youth and parents were asked some general questions about their recall of ads seen or heard on TV and radio, and in other media such as newspapers, magazines, movie theaters, billboards, and the Internet. NSPY used two measures of exposure; the first is based on general recall of anti-drug ads through all media, and the second is based on specific recall of currently broadcast ads on television and radio.

Development of the Confounder Pool

As part of the survey, a large number of cognitive and behavioral variables were obtained on each parent and child in addition to the exposure and outcome variables. These would form the basis for the confounder pool. Potential parent confounders included race, ethnicity, gender, age, income, marital status, strength of religious feelings, age of children, neighborhood characteristics, media consumption habits, language, and substance use history and current use (alcohol, tobacco, marijuana, and other illegal drugs). The youth confounder pool included all the parent characteristics plus youth school attendance, grade level, academic performance, participation in extra-curricular activities, plans for the future, family functioning, personal antisocial behavior, association with antisocial peers, use of marijuana by dose friends, personal tobacco and/or alcohol use of a long-standing nature, and sensation-seeking tendencies, among other factors.

Development of Propensity Scores

The propensity score for an individual is the probability of being in a particular group (in our case, having a particular exposure level) given the individual's values on a set of observed covariates. In the simplest and most common case, there are only two exposure levels, e.g., treatment vs. comparison. Formally, for subject i(i =1, …,N), the probability of assignment to the treatment group ($\boldsymbol{Z}_i$=1) versus comparison group ($\boldsymbol{Z}_i$=0) given the vector of covariates, $\boldsymbol{x}_i$, is $e(\boldsymbol{x})$=pr($\boldsymbol{Z}_i$=1| $\boldsymbol{X}_i$=$\boldsymbol{x}_i$), where it is assumed that given the $\boldsymbol{X}$'s, the $\boldsymbol{Z}_i$ are independent (D'Agostino, 1998). The propensity scores are created by regressing exposure on candidate confounders, typically with logistic regression, and outputting the distribution of

[1] For a complete exposition of the conceptual model of Campaign influence, see Chapter 2 at http://www.nida.nih.gov/despr/westat/Westat2003/Report.PDF

probabilities. While not addressing every concern with respect to causal attribution, propensity scoring methods bring tangible improvements over earlier methods like analysis of covariance/regression modeling. Specifically, it frees the regression modeling process from its usual limitation of reliance on a small number of covariates and simplistic functional forms (e.g., linear main effects only). Rather, a complex model with interactions and higher-order terms can be fit at the propensity scoring stage without great concern about overparameterization or multicollinearity. When subsequently included in the regression model, the propensity score carries all the information from the complex covariate model in a single variable, consuming only one degree of freedom. However, the most important advance may be that propensity scoring allows for direct diagnosis of the success with which confounder influence was removed, through tests of balance (described below). This is not possible with traditional ANCOVA models.[2]

Once the confounder pool was established, the propensity scores could be created. Standard propensity score methods assume that there are only two levels of exposure. However, in our set up, exposure is a three- or four-level variable. For this more complex problem, the method suggested by Joffe and Rosenbaum (1990) was used. With this method, an ordinal logit model is fit for each index. The structure of this model is

$$\ln\left(\frac{\sum_{j\le k} p_{ij}}{1-\sum_{j\le k} p_{ij}}\right)=a_k + X_i b\ .$$

Here p_{ij} is the propensity of the *i*-th subject for exposure level *j*, X_i denotes the vector of confounder scores for the same subject, a_k is a threshold parameter for the *k*-th exposure level, and b is a vector of slope parameters with one component for every confounder retained in the model. The point of the modeling is to identify which of the admissible potential confounders are actually predictive of exposure and then to estimate the vector of slope parameters for those predictors. The model was fit with a stepwise variable selection procedure in SAS. (The sampling weights were ignored in fitting the model.)

Four cross-sectional models were fitted, one for each type of parent exposure index and one for each type of youth exposure index. The cross-sectional models were fitted on the combined exposure data from Rounds 1 and 2. For the longitudinal analysis, a separate set of propensity models had to be fit that used only the Round 1 exposure data, concurrently with the exposure measure at that round, but prior to the Round 2 outcome measures. The confounder pool for the youth lagged model was identical to that of the cross-sectional model, while the lagged model for parents added initial Round 1 outcomes to the pool.[3] In all, there were four longitudinal propensity models as there were for cross-sectional: youth lagged general exposure, youth lagged specific exposure, parent lagged general exposure, and parent lagged specific exposure. After being created, scores were grouped into quintiles.[4]

Tests of Balance

Because propensity scoring is designed to remove the effects of confounding variables from the association between outcomes and exposures, the counterfactual projections of population means for the confounding variables should not vary across the exposure levels. This property is referred to as balance. If a confounder has been successfully balanced, then it will have the same counterfactual projection across all exposure levels. While the goals of balancing are clear, the optimal method of testing for balance is a matter of some debate. We have experimented with several approaches in NSPY. The approach used for the 5th semiannual report was to: 1) test all variables in the final model; 2) test variables that were eliminated from the final model by the stepwise regression but were still considered potentially important in predicting outcomes; and 3) test all variables for the full sample as well as within subgroups of race, gender, and age. WESVAR software was used to test linear trends and overall differences in the means of the variables across exposure levels within propensity quintiles for both general and specific exposure.[5] A variable was considered out of balance if either test was significant at $p<0.05$ within one or more quintiles, for the full sample or in one or more subgroups.

---

[2] With a single covariate (e.g., age), the same diagnosis could generally be made visually, but with many confounding covariates this is more difficult, and the issues of inadequate overlap and reliance on untrustworthy model-based extrapolations are more serious because small differences in many covariates can accumulate into a substantial overall difference (Rubin, 1997).

[3] The lagged model for youth *would* have added Round 1 outcomes to the confounder pool, except these were not measured on 9- to 11-year-olds.

[4] Simulations, studies of actual data, as well as formal proofs have shown that subclassification of the propensity score into about five strata or quintiles is generally sufficient to assess the quality of the adjustment for all the covariates that went into its estimation, no matter how many there are (Rubin, 1997).

[5] WesVar uses replication methods to ensure proper estimation of standard errors in complex survey designs. Weights are created for each replicate subsample, adjusted for nonresponse, post-stratified, and raked to control totals.

Following the initial tests of balance, the models were modified accordingly. Variables that had been eliminated by the stepwise procedure were added back in if they were significantly out of balance, as were interactions with age, race, and gender as needed to achieve balance within subgroups. This produced more complicated models than had been fit for previous reports, and required a large of amount of testing, refitting, and retesting, but resulted in the final models being in balance for all variables judged as logically related with the outcomes. For more details on the balance testing, see Appendix C of the Fifth Semiannual Report at http://www.nida.nih.gov/despr/westat/Westat2003/Appendix.PDF

Development of Counterfactual Projection Weights

Once the models had been fit and balanced, the next step was to use the models to remove the effects of the confounding variables from the causal analysis. This was done by following a suggestion by Imbens (2000) with some innovations. The basic suggestion of Imbens was to use the estimated propensities to calculate the expected response across the entire sample, which would be expected in the counterfactual event that everyone in the sample had received the same exposure level. This could be achieved with the estimator

$$\hat{y}_{Ck} = \sum_i \frac{d_{ik} y_i}{\hat{p}_{ik}},$$

where d_{ik} is an indicator variable for the *i*-th case having exposure level *k*, i.e.,

$$d_{ik} = \begin{cases} 1 & \text{if the } i\text{-th individual has observed exposure at level } k \\ 0 & \text{else} \end{cases}$$

and $\hat{p}_{ik}$ is the estimated propensity the *i*-th individual has for exposure level *k*. Note that, for each *i*, $\sum_k \hat{p}_{ik} = 1$ for every i.

An innovation of the NSPY analysis was to project the expected response to the entire eligible population by using the sampling weights. This is important in this study given the differential probabilities of selection for youth and parents, depending on family composition. Youth aged 14 to 18 had a higher probability of selection if they had siblings in the 12 to 13 or 9 to 11 brackets, all youth had a lower probability of selection if they had a sibling in the same age bracket, and married parents had lower probabilities of selection than single parents. Also, there is variation in the probability of response to the survey that is reflected in the sampling weights. Using the sampling weights, the counterfactual estimator of response on variable *y* to exposure *k* would be

$$\hat{Y}_{Ck} = \sum_i \frac{d_{ik} y_i w_i}{\hat{p}_{ik}},$$

where w_i is the sampling weight for the *i*-th respondent, adjusted for nonresponse and poststratified to population controls.

Summary of Findings

Several approaches were employed to detect the presence of Campaign effects. With each, the influence of the complex sample design, nonresponse adjustment, and CFP weights were reflected as fully as possible. However, most of the interpretation was based on a test of the Gamma statistic of significance for monotone relationships. The monotone dose-response test assessed the overall association between exposure and outcome.

Effects of Parent Exposure

As seen in Table 1, parents who reported more exposure to Campaign messages scored better on four out of five outcomes after applying the statistical controls described above (The association criterion is whether or not the gamma estimate was significant at $p<.05$). In addition, parents who had more exposure the first time they were measured were more likely to talk with their children and do fun activities with their children subsequently. However, there was little evidence for Campaign effects on parents' monitoring behavior, a major focus of the parent Campaign and the one parent behavior most associated with youth nonuse of marijuana. In addition, there was no evidence for favorable indirect effects on youth behavior as the result of parent exposure to the Campaign.

Table 1. Parent exposure effects on parent and youth outcomes for parents of 12- to 18-year-olds

	Cross-sectional effects association		Delayed-effects Association	
	General	Specific	General	Specific
Parent Outcomes				
Talking behavior	Favorable	Favorable	Favorable	No
Talking cognitions	Favorable	Favorable	No	No
Monitoring behavior	No	No	No	No
Monitoring cognitions	Favorable	No	No	No
Doing fun activities	Favorable	Favorable	Favorable	No
Youth MJ Outcomes				
Past year use	No	No	No	No
Intentions to use	No	No	No	No
Attitudes & beliefs	No	No	No	No
Social norms	No	No	No	No
Self efficacy	No	No	No	No

Effects of Youth Exposure

To date, the youth cross-sectional analysis shows no tendency for those reporting more exposure to Campaign messages to hold more desirable beliefs. Moreover, the longitudinal analysis suggested an unfavorable delayed effect of Campaign exposure on subsequent intentions to use marijuana and on other beliefs. Table 2 shows the exposure levels and associated gammas for each outcome from the delayed-effects analysis. The exposure columns represent the level of exposure reported by these youth at Round 1 to Campaign television advertising. The rows represent average scores on the five outcomes of interest at Round 2 for the same youth. All estimates in the cells are adjusted, as described above, as well as being survey weighted to represent the U.S. population. Note that for the eight cognitive outcome effects, all of the gammas are negative, with four of the eight statistically significant. These outcomes involve intentions, social norms, and self-efficacy. While intentions are strong predictors of subsequent initiation of marijuana use, the evidence for an unfavorable effect on initiation was not statistically significant overall or for any subgroup.

Table 2. Exposure per month at Round 1 and outcomes at Round 2 among 12- to 18-year-olds who were nonusers of marijuana at Round 1

		Round 1 Exposure				
Round 2 Outcome	Exposure type	<1 exposure	1 to 3 exposures	4 to 11 exposures	12+ exposures	Gamma (95%CI)
Percent not intending to use marijuana	General	84.0%		78.4%	77.4%	**-.14* (-.25 to -.03)**
	Specific	82.3%	78.2%	76.5%		**-.12* (-.21 to -.02)**
Attitudes/Beliefs Index (Mean score)	General	99.6		87.4	90.5	-.03 (-.08 to .01)
	Specific	92.3	93.4	86.0		-.03 (-.08 to .02)
Social Norms Index(Mean score)	General	99.2		79.5	83.0	**-.07* (-.12 to -.02)**
	Specific	90.2	85.9	77.8		-.05 (-.11 to .00)
Self-Efficacy Index (Mean score)	General	105.8		105.8	106.7	-.01 (-.07 to .05)
	Specific	120.0	102.2	104.3		**-.08* (-.15 to -.02)**
Percent Initiation of Use	General	12.0%	11.8%	13.2%		.04 (-10 to .18)
	Specific	12.8%	13.2%	12.8%		-.00 (-.11 to .11))

* $p < .05$.

Discussion

Two innovations were successfully implemented in this study. First, propensity scoring was applied to evaluate the effectiveness of a social intervention in the context of a complex, longitudinal national probability survey. Second, the propensity scoring was extended to create CFP weights for each respondent. The CPF weights were then used in tandem with sampling weights, nonresponse weights, and post-stratification weights in a single, integrated analysis. Consequently, the effect estimates generated are simultaneously adjusted for all these factors.

The adjustments for confounders were based in statistical theory, and the operational procedures underwent extensive quality

control. Nonetheless, and particularly in light of the counterintuitive findings from the youth lagged-effect analysis, we thought it important to rule out the possibility that the observed results might be an artifact of these complex adjustment procedures. To ensure that the procedures did not *create* the counterintuitive findings, we compared the estimates in Table 2 to an alternate set that incorporated the survey weights but were not adjusted for confounder control. As shown in Table 3, these results make it clear that the unfavorable associations did not result from the procedures used to adjust for confounders. In almost every case, the original association was *more* unfavorable to the Campaign before the confounder controls were introduced. The same patterns were apparent in the raw (unweighted), ruling out the possibility that the unfavorable findings were caused by an undetected error in the application of the survey weights. Finally, we regressed the outcomes directly onto the propensity scores as an alternative to using the CFP weights. This too yielded similar results.

Table 3. Exposure per month at Round 1 and outcomes at Round 2 among 12- to 18-year-olds who were nonusers of marijuana at Round 1- (data not corrected for confounders)

		Round 1 Exposure				
Round 2 Outcome	Exposure type	<1 exposure	1 to 3 exposures	4 to 11 exposures	12+ exposures	Gamma (95% CI)
Percent (Not) intending to use	General	85.4%		80.1%	75.1%	**-.22* (-.31,-.14)**
	Specific	85.7%	78.8%	74.9%		**-.20* (-.27,-.13)**
Attitudes/Beliefs Index (Mean score)	General	106.5		91.2	83.6	**-.08* (-.11,-.05)**
	Specific	102.3	94.7	81.3		**-.08* (-.11,-.04)**
Social Norms Index (Mean score)	General	106.2		84.8	74.7	**-.13* (-.17,-.09)**
	Specific	103.4	88.7	70.8		**-.12* (-.16,-.09)**
Self-Efficacy Index (Mean score)	General	109.5		110.5	105.8	**-.05* (-0.10,-0.0)**
	Specific	123.8	104.1	102.7		**-.09* (-0.14,-.04)**
Percent Initiation of Marijuana use	General	10.6%		11.6%	14.1%	**.12* (.01,.23)**
	Specific	10.4%	12.9%	13.8%		.09 (-.01,.19)

* $p < .05$

Propensity scoring has clear advantages over traditional ANCOVA/regression modeling for confounder control in observational studies. Like the earlier methods, however, propensity scoring can adjust only for confounders that are observed and measured. This is always a limitation of nonrandomized studies compared with randomized studies, where the randomization tends to balance the distribution of all covariates, observed *and* unobserved. In theory, this "omitted covariate" problem, a type of misspecification bias, can lead to false conclusions about the relationship between exposure and outcome. To minimize the risk, we considered a wide range of background variables that might affect exposure, and included as many as possible as part of the questionnaire design and the acquisition of geographic information. The questionnaires can be found at: http://www.nida.nih.gov/DESPR/Westat/index.html. Researchers can view the questions and decide for themselves if important variables might have been left out.

A larger concern in the present study was that the emphasis and priority placed on achieving balance in the confounders, in conjunction with the stepwise modeling approach, may have resulted in overfitting. We very cautiously included all variables that appeared to be out of balance for the full sample or for any subgroup. However, it was not clear this improved the results, and may have reduced sensitivity to effects. For the next report, which will include a 3rd round of interviews, we are being more judicious in selecting variables for inclusion, focusing more on sample-wide balance and less on subgroups, and allowing for up to 5 percent of variables to remain unbalanced. The last criterion mirrors what would happen with random assignment, in that some variables would be out of balance merely by chance.

References

D'Agostino, R.B., Jr. (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.

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

Joffe, M. M., and Rosenbaum, P. R. (1990). Propensity scores. *American Journal of Epidemiology*, 150, 327-333.

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

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine*, 127(8 Pt 2), 757-63.