Prev Sci (2013) 14:557–569
DOI 10.1007/s11121-012-0342-x

Selection Effects and Prevention Program Outcomes

Laura G. Hill · Robert Rosenman · Vidhura Tennekoon · Bidisha Mandal

Published online: 17 February 2013
© Society for Prevention Research 2013

Abstract A primary goal of the paper is to provide an example of an evaluation design and analytic method that can be used to strengthen causal inference in nonexperimental prevention research. We used this method in a nonexperimental multisite study to evaluate short-term outcomes of a preventive intervention, and we accounted for effects of two types of selection bias: self-selection into the program and differential dropout. To provide context for our analytic approach, we present an overview of the counterfactual model (also known as Rubin's causal model or the potential outcomes model) and several methods derived from that model, including propensity score matching, the Heckman two-step approach, and full information maximum likelihood based on a bivariate probit model and its trivariate generalization. We provide an example using evaluation data from a community-based family intervention and a nonexperimental control group constructed from the Washington State biennial Healthy Youth Survey (HYS) risk behavior data (HYS n=68,846; intervention n=1,502). We identified significant effects of participant, program, and community attributes in self-selection into the program and program completion. Identification of specific selection effects is useful for developing recruitment and retention strategies, and failure to identify selection may lead to inaccurate estimation of outcomes and their public health impact. Counterfactual models allow us to evaluate interventions in uncontrolled settings and still maintain some confidence in the internal validity of our inferences; their application holds great promise for the field of prevention science as we scale up to community dissemination of preventive interventions.

Keywords Selection effects · Translational research · Universal prevention · Family-focused interventions · Causal inference · Observational research · Nonexperimental research

Electronic supplementary material The online version of this article (doi:10.1007/s11121-012-0342-x) contains supplementary material, which is available to authorized users.

The first two authors are listed alphabetically and contributed equally to the conceptualization and writing of the manuscript.

L. G. Hill (✉)
Department of Human Development, Washington State University, PO Box 644852, Pullman, WA 99164-4852, USA
e-mail: laurahill@wsu.edu

R. Rosenman · B. Mandal
School of Economic Sciences, Washington State University, Pullman, WA, USA

R. Rosenman
e-mail: yamaka@wsu.edu

V. Tennekoon
Department of Economics, Eastern Washington University, Cheney, WA, USA

Introduction

Nonexperimental outcome evaluations of evidence-based prevention programs may be biased by selection effects at two phases: first, there may be systematic bias in who decides to attend a universal, community-based program, and second, of those who attend, there may be systematic bias in who completes a program. Because evidence-based interventions are increasingly being translated to community settings where experimental control is not possible, statistical correction methods represent an important tool in the evaluation of prevention programs. The overarching goal of this paper is to introduce an evaluation design and analytic method that can be used to address the problem of biased outcome estimates due to selection effects at these two stages in nonexperimental settings. In the service of that goal, we model a trivariate probit to identify selection effects in initial attendance and in program completion of a family-strengthening intervention and to determine whether, after correcting for bias due to those two types of selection effects, there was significant, positive, short-term change in participant outcomes that may be related to intervention goals and is unlikely to have occurred in the absence of that intervention. We also examine

which individual, program, and community variables were related to program participation, completion, and outcome and discuss the practical implications of this information.

Our introduction is organized as follows: we first describe the need for evaluation of prevention programs disseminated outside the context of research studies. Next, we discuss a major threat to causal inference in nonexperimental program evaluation—selection bias—and the difficulties of identifying and controlling for selection in community implementations. We then describe a family of methods designed to identify and control for selection bias in nonexperimental research. Finally, we describe the present study and its goals.

Evaluating the Public Health Effects of Large-Scale, Universal Prevention Efforts

Many prevention programs that have been demonstrated efficacious in clinical trials are now adopted by schools and communities and disseminated on a widespread basis. Formal effectiveness trials for many efficacious programs remain to be conducted, and even fewer programs have been examined outside the context of experimental studies. For example, in a brief review of intervention studies reported in *Prevention Science* between 2000 and 2010, we found 54 reports of efficacy trials but only 13 of effectiveness trials and only 3 of community disseminations. Thus, there remain many important questions about the effects of universal preventive interventions as those interventions are translated to real-world use. Does the program actually reach a broad spectrum of the population? Who decides to attend, who completes a program having decided to attend, and who benefits?

There are few studies that answer these questions, in part for historical reasons—systematic testing of preventive interventions in efficacy and effectiveness trials has occurred only over the past two decades, and translation to community settings is even more recent—and in part because there are few methods to deal with many of the problems that arise in evaluating effects of community-based programming. Although effectiveness trials can approximate community-based implementation, by virtue of being research trials, they are more controlled. A truly translational program of research must also incorporate evaluation of prevention efforts in "uncontrolled and uncontrollable" settings.

Selection Bias in Nonexperimental Program Evaluation

Selection effects pose problems for both internal and external validity of inferences made about program benefits, not only in nonexperimental research but also in clinical trials (Barnard et al. 2003; McGowan et al. 2010; Shadish et al. 2002). For example, parents who decide to attend a family skills program may be more motivated to change than parents who do not, and observed change may therefore be a result of motivation rather than of program participation. Similarly, an intervention designed to be universal might appear to have equivalent benefits across the entire range of participant attributes, but if variability among participants is restricted due to selection (e.g., the sample is predominantly high risk, and high-risk participants have the most ability to improve), inferences about a program's external validity may be incorrect. Unbiased determination of benefits is further complicated by additional selection effects, downstream from self-selection into the program. For example, there is evidence that higher-risk participants are less likely to complete a program and more likely to be lost to follow-up in longitudinal trials of all sorts (Biglan et al. 1991).

The randomized clinical trial (RCT) is held as the gold standard for causal inference because the random assignment of subjects to groups increases the likelihood of initial equivalence of treatment and control groups by minimizing self-selection effects, especially as sample size increases. Furthermore, RCTs allow for intent-to-treat outcome analyses, which, by including participants who enrolled in but did not complete a program, preserve randomization and increase internal validity of inferences when there is uneven attrition across groups. Finally, representativeness of the study sample in RCTs can be checked through comparison of program participants with the population at large. Determining the extent and nature of selection effects in program participation and completion is thus more straightforward in well-designed efficacy and effectiveness trials, and observed change can then be attributed to intervention effects with some confidence in the validity of that causal inference.

However, in community-driven disseminations there is no experimental control group to identify and control for possible selection effects and, because participants are lost to follow-up (or resources to conduct follow-up evaluation are scarce), intent-to-treat analyses to control for differential participation or attrition are often not feasible. When there is no control group or intent-to-treat sample available—that is, when control for selection is not possible through experimental design—it is common in some fields to use statistical control instead (see West and Thoemmes 2010 for an extended comparison of design versus statistical control methods). Statistical correction methods are widely used in economics (Heckman 1979) and in other fields that often do not have the option of conducting randomized trials, including political science (Berinsky 2004), sociology (Bushway et al. 2007), program evaluation (Cook et al. 2008), and epidemiology (Pearl 2000). Statistical control methods have also been discussed extensively in recent special issues of *Psychological Methods* (Maxwell 2010) and *Developmental Psychology* (Foster and Kalil 2008). In the next section, we provide a brief overview of the general approach to causal inference in nonexperimental program evaluation and of analytic methods derived from this approach.

Correcting for Bias Due to Selection: the Counterfactual Model

The logic of causal inference is rooted in the counterfactual model, also known as Rubin's Causal Model or the potential outcomes model (Neyman 1923/1990; Rubin 1974, 2004). In our description of the counterfactual model, we follow closely the descriptions of Shadish (2010) and Rubin (2004) and refer to program participation as "treatment" (T, where $T=1$ if treatment is received and $T=0$ otherwise). The other two elements of importance in the potential outcomes framework are units (program participants) and the outcome measure, Y. Before treatment all units have two potential, as-yet, unrealized outcomes: $Y(1)$ is the potential outcome of a person exposed to treatment, and $Y(0)$ is the potential outcome of that same person not exposed to treatment. In an ideal world, program participants would experience *both* outcomes, and the treatment effect would be measured as the difference between those two (potential) outcomes $[Y(1)-Y(0)]$ for each person; the average treatment effect would then be the mean of this difference across all participants. However, this is impossible, since experiencing the potential outcome $Y(1)$ means that the potential outcome $Y(0)$ cannot be observed for any given individual and vice versa. The observed outcome (Y_{obs}) is equal to $Y(1)$ for the individuals assigned to the treatment group and equal to $Y(0)$ for the individuals assigned to the control group, i.e., $Y_{obs}=T\times Y(1)+(1-T)\times Y(0)$.

The counterfactual framework allows us to conceptualize outcome estimation as a problem of missing data—$Y(0)$ is missing for individuals in the treatment group, and $Y(1)$ is missing for those in the control group. In true experiments, researchers draw inferences about those missing values by assuming that under random assignment, groups are completely equivalent, and the average observed outcome in the control group is thus equivalent to the (unobserved and unrealized) average potential outcome in the treatment group. Similarly, researchers infer that the average observed outcome in the treatment group is equivalent to the (unobserved and unrealized) potential average outcome in the control group. More formally, subject to the strong ignorability assumption, which states that the treatment assignment is independent of the potential outcomes after controlling for all the observables, $(T\perp Y(1),\ Y(0)|X)$, the inference is $P(Y_{obs}|X,T=1)=P(Y(1)|X,T=0)$ and $P(Y_{obs}|X,T=0)=P(Y(0)|X,T=1)$ where P refers to probability. This inference allows estimation of the average treatment effect as $E[(Y_{obs}|X,T=1)-(Y_{obs}|X,T=0)]$, where E refers to the statistical expectation.

However, inferences about the equivalence of observed with potential outcomes only hold true if treatment and control groups are truly equivalent. In a pristine RCT random assignment of units increases the likelihood that equivalence is achieved (but see Barnard et al. 2003 for a caution about the prevalence of "broken" RCTs). In nonexperimental studies, statistical approaches derived from the counterfactual model can be used to identify and control for nonequivalence of groups and to consistently estimate the treatment effect. Next we describe three such methods: propensity score matching, the Heckman two-step procedure, and bivariate and trivariate probit models with selection.

Propensity Score Matching In prevention science, the most widely known method of correcting for selection effects by creating equivalent groups in nonexperimental studies is propensity score matching (Rubin, 1997). In matching methods, generally, a nonexperimental control group is constructed, either from a supplemental dataset or from a nonrandomized control sample. This constructed control group is designed to be equivalent to the treatment group, which is accomplished through matching observations from the supplementary sample to those in the treatment sample on a set of explanatory variables presumed to be relevant to the outcome. Criteria used to create a matched group differ in various matching methods. In propensity score matching, propensity to select a treatment is calculated as a composite score of observed characteristics. Units in the treatment condition are then matched to units in the control condition with similar propensity scores to construct the nonexperimental control group. (See Technical Appendix A, part I, available online, for the assumptions underlying propensity score matching.)

In theory, the matching process ensures that units in treatment and control groups are equivalent subject to the ignorability assumption, which requires T to be independent of the potential outcomes after controlling for all the observables $(T\perp Y(1),Y(0)|X)$.[1] The ignorability assumption is believed to be satisfied when there are rich data, but this assumption is often not verifiable, and selection may thus be a function of unobserved variables that are related to both selection into treatment and to the observed outcome. For example, as noted earlier, families who are especially motivated to improve their family functioning may be both more likely to select into a program and more likely to find other means of improvement without a program. In this case, $E(Y(0)|T=1)$ is not equivalent to $E(Y_{obs}|T=0)$. So, when an unobserved variable (or set of variables) affects both selection into a treatment and the outcome of interest, the basic counterfactual assumption is violated, and the influence of systematic differences across groups cannot be ruled out as a threat to validity of causal inferences (Heckman et al. 1996).

The Heckman Approach In the late 1970s, Heckman (1979) pioneered a family of econometric methods to address the

[1] Note that this is a weaker assumption than the strong ignorability condition, which requires unconditional independence.

failure of the ignorability assumption by modeling selection as an omitted variable that may influence outcome. In Heckman's two-step correction procedure, as in propensity score matching, the researcher first creates a model of the selection process using observed variables, the results of which yield a predicted probability of selection for each individual (step 1). The error term from this selection equation represents both random error and the effect of unobserved variables driving selection—in other words, larger error terms could indicate that unobserved variables not included in the selection equation are more influential in predicting T (in the example above, large error terms could indicate families whose decision to attend a family program was driven by an especially large, unmeasured motivation to improve family functioning). A transformation of this error term is then modeled in the outcome equation as a proxy for the omitted variable (step 2).

The two-step procedure assumes a bivariate normal distribution of error terms, and estimates may be inconsistent if this assumption is violated (see Technical Appendix A, part II, available online, for a brief description of the two-step procedure and its assumptions and limitations; see also Bushway et al. 2007 for an accessible introduction to Heckman's approach).

Bivariate Probit Model with Selection and the Trivariate Generalization The Heckman two-step estimation approach is used when the outcome of interest is an observed continuous variable. When the outcome is binary and the error term is normal, the efficient alternative is the bivariate probit with selection (Arendt and Holm 2006), a full information maximum likelihood method. The bivariate probit with selection is easily understood within the context of two binary variables, T and Y, each of which can take on a value of 0 or 1. T represents selection into treatment and takes on a value of 1 for individuals who attend a program and 0 otherwise; Y represents the outcome and takes on a value of 1 if there has been improvement in functioning and 0 otherwise. Hence, the table of possibilities is given by

T	Y
0	**0**
0	1
1	**0**
1	**1**

A standard bivariate probit model can be used when observations for all four rows are available. However, observations for the second row (not bolded) are eliminated as a possibility under the assumption that a person cannot choose to not have treatment (T=0) yet have the treatment be effective (Y=1). It should be noted that under the counterfactual model, it is necessary to frame this as an assumption, since a potential outcome is that a person could improve on the outcome of interest despite not receiving a treatment. The result is a bivariate probit model with selection, which accounts for the fact that the outcome T=0, Y=1 is not possible, and only the three remaining possibilities are represented in the likelihood function.

In the present study, we use a trivariate generalization of the bivariate probit model, with three equations instead of two, to account for a second level of selection—differential attrition—in our estimation of short-term results of a universal family intervention (the Strengthening Families Program for Parents and Youth 10–14, or SFP). Conceptually, the trivariate probit follows the same format as the bivariate probit with selection. Now there are three variables (T is for treatment, TC is for completion of treatment, and Y is for outcome) and eight possibilities:

***T*=0, *TC*=0, *Y*=0**	T=0, TC=1, Y=0
***T*=1, *TC*=0, *Y*=0**	T=0, TC=0, Y=1
***T*=1, *TC*=1, *Y*=0**	T=0, TC=1, Y=1
***T*=1, *TC*=1, *Y*=1**	T=1, TC=0, Y=1

However, we assume there is no improvement (for the time frame we evaluate) in those who are unexposed to the treatment or who fail to complete it. Under this assumption, there are only four potential outcomes, those in the left column (bolded). Hence, our likelihood function, based on a trivariate probit distribution, contains only the four bolded cases. To control for selection effects, the likelihood function is maximized over the correlation coefficients between the three residual terms, i.e., the correlation coefficients between the residual terms for T and TC (ρ_{12}), T and Y(ρ_{13}), and TC and Y (ρ_{23}), as well as over the parameters on the covariates explaining each of the three equations. (A more complete description of this model, including the underlying assumptions, is in Technical Appendix A, part III, available online.) In using the trivariate probit with selection, we propose a more integrated model of how selection, both for participation and program completion, might jointly affect outcome, and we derive a consistent statistical specification that adjusts the outcome equation for selection bias. If there is no selection bias and program participation and completion are completely random, we should expect $\rho_{12}=\rho_{13}=\rho_{23}=0$. Under this situation, the trivariate probit estimates are identical to the estimates from three separate univariate probit models. Estimates of non-zero correlation coefficients indicate selection, and ignoring selection may lead to biased inferences.

Increasing the Validity of Causal Inference with Nonexperimental Data Nonexperimental studies that use statistical

control may increase external validity of causal inferences at the possible cost of bias in estimation of outcomes. Analytic methods derived from the counterfactual model have been subject to the criticism that the influence of omitted variables and resulting selection bias can never be ruled out completely, even when error is modeled as an omitted variable driving selection. However, research has shown that these models enable accurate estimation of outcomes (that is, estimation of outcomes equivalent to those obtained from experiments), even in the absence of a randomized control group, under certain conditions. Within-study comparison designs (Cook and Steiner 2010), which compare estimates from nonexperimental studies to those from experimental benchmark conditions, have shown that bias in nonexperimental research is minimal under conditions designed to counter the problem of nonequivalence of groups in the absence of random assignment (Cook et al. 2008; Dehejia and Wahba 2002; Rubin 2008). In particular, estimates obtained from nonexperimental data are nearly identical to those obtained from experimental data when (1) the matched or control sample is drawn from a population equivalent to the program sample, (2) the choice of covariates adequately represents the mechanisms underlying selection, and (3) measurement instruments and methods are reliable (Cook and Steiner 2010).

The Present Study

In the present study, we examined questions about who attends, who completes, and who shows improvement in the context of a community-driven dissemination of an empirically validated family intervention. The original version of SFP was developed for substance-abusing parents with children aged 6–12 years, and SFP 10–14 was adapted into a universal intervention targeting families with young adolescents (Kumpfer et al. 1996). The program holds 2-hour sessions once weekly for 7 weeks; parents and youths meet separately for the first hour and come together for the second hour. In each 7-week cycle, families engage in activities that are designed to promote family strengths and to improve parents' clarity of communication about their expectations for youths' behavior and consequences for violating those expectations. SFP has been shown in its clinical trial to result in delayed initiation and reduced frequency of substance use (Spoth et al. 2008).

The dissemination described in the current study is community driven and the evaluation is a bottom–up process. Programs are initiated at the local level using state prevention funds or programming grants, there are no control groups, and providers make use of a free, centralized evaluation service voluntarily. Faculty from the extension system of a land-grant university have provided trainings on demand since 2000 and in 2005 initiated a series of trainings for the Spanish language version of the program as part of an outreach effort to the Latino communities of the state.

Because the ongoing dissemination is not a research trial there are no formal control groups. For this reason we designed our evaluation to include family-level risk and protective factor scales that are also used in the state's biennial Healthy Youth Survey and then used data from the school survey to construct a nonexperimental control group for identification of selection effects. In a previous paper (Hill et al. 2010), we described an analytic method for constructing a nonexperimental control group from existing data when there is contamination of the supplemental sample (i.e., observations from the intervention group are also represented in the supplemental sample). In the present paper, we focus on the analytic method of correcting for selection bias using a nonexperimental control group.

Method

Analytic Method

We used a trivariate probit analysis, described above, to jointly estimate three sequential equations: a first equation predicting *participation* (eligible members of the population attend a program, or not), a second predicting *completion* (having elected to attend, participants complete the program, or not), and a third predicting *outcome* (having completed the program, participants show positive short-term change on the binary outcome measure, or not). The first equation in the trivariate model estimated associations of demographics, individual risk/protective factor scores, and community-level risk factors on participation through comparison of SFP and the nonexperimental control group constructed from the Healthy Youth Survey, a school risk survey administered in schools throughout the state during the same years that the SFP data were gathered. The second equation, using only the SFP sample and accounting for the selection effects identified in the first equation, estimated effects of individual, community, and program attributes on completion, given participation. Finally, the third equation, also using only the SFP sample and accounting for selection in the first two equations, estimated effects of individual, community, and program attributes on outcome, given participation and completion. This trivariate extension of the bivariate probit with selection model (Lahiri and Song 2000; Lesaffre and Molenberghs 1991) allowed us to correct for potential bias in estimates of outcome that might result from the sequential selection effects in program participation and completion. Joint estimation also allowed for correlation of the equations' error terms, providing information about the interrelation of latent factors underlying participation, completion,

and outcome. We controlled for dependence of observations within the program with explanatory variables representing program-level average and standard deviation on a risk factor, as described below under "Selection of Explanatory Variables for Each Equation." We estimated robust standard errors to control for heteroscedasticity; results from this analysis were the same as without the correction. The trivariate estimates could also be sensitive to the model specification. Therefore, in order to verify the robustness of our estimates, we estimated the parameters of the model with several alternative specifications. The results from these runs, reported in Technical Appendix C (available online), show that our results are robust to changes in specifications.

We include a technical description of the model in the online Appendix, and the interested reader can find a more detailed description in Rosenman et al. (2010) (available at http://econofprevention.wsu.edu). We used the STATA cmp ("conditional mixed process") module (Roodman 2007, 2009) for the trivariate probit analysis. The Roodman (2009) paper discusses theory and implementation of the trivariate probit. Arendt and Holm (2006), Bhattacharya et al. (2006), Lahiri and Song (2000), and Lesaffre and Molenberghs (1991) describe applications of the method and simulation studies.

Participants

Strengthening Families Program Sample Our initial program sample (i.e., those who attended the program for at least one session) consisted of 1,502 youths (43 % female) who attended one of the 137 SFP 7-week cycles in 20 counties between 2006 and 2009 (see Table 1). Forty-five percent of the youth participants were White/European American, 20.2 % were Latino/a, 4.7 % American Indian/Alaska Native, 1.9 % Black/African American, 6.1 % other or multiple race/ethnicities, and 22 % did not report race/ethnicity. Of the 1,502 participants who attended the program, 785 (52.3 %) completed both pretest and posttest. We grouped participants into two cohorts (2006–2007 and 2008–2009) for purposes of comparison with the biennial Healthy Youth Survey data (see below).

Healthy Youth Survey Control Sample The Healthy Youth Survey, which assesses physical and emotional health, risk behaviors, and risk and protective factors, is a biennial school-based survey administered to students in grades 6, 8, 10, and 12 in the state of Washington (Washington State Department of Health, 2009). The primary sampling unit for HYS is the grade/school combination and is representative of the state population. We used data from two survey years (2006 and 2008), which included a total of 68,846 students and over 200 schools (see Table 1). School response rates for grades 6, 8, and 10 (matching the age range of SFP participants) ranged from 82–89 % in 2006 and from 83–88 % in 2008. The HYS sample was 47.9 % male, 50.8 % White/European American, 15.1 % Latino/a, 5.8 % American Indian/Alaska Native, 2.4 % Black/African American, and 23.9 % other or multiple ethnicity; 1.9 % did not report race/ethnicity.

Measures

Participant-Level Explanatory Variables We designed the SFP evaluation to include family risk and protective factor scales that are also in the HYS and that assess family practices targeted by the program. Psychometric properties of the scales from their original development and testing (Arthur, et al. 2007) and from HYS administrations (Washington State Department of Health 2006) have been reported elsewhere; here, we report on pretest scale properties for the SFP sample only. Protective factor scales used in the study were Opportunities for Prosocial Involvement (three items, α=0.65, 95 % CI=0.63–0.67), Rewards for Prosocial Involvement (three items, α=0.73, CI=0.71–0.74), and Family Attachment (two items, α=0.61, CI=0.58–0.64). Risk factor scales are Family Conflict (α=0.72, CI=0.70–0.74) and Poor Family Management (α=0.81, CI=0.80–83) (see Maydeu-Olivares et al. 2007 for use of confidence intervals in reporting coefficient alpha). All scales were scored such that higher values represented lower risk or greater protection. There is a Spanish translation of the evaluation, but all youth participants selected the English version.

Community-Level Explanatory Variables We used four community-level variables, averaged by county and cohort, to examine community effects on participation. First, we calculated the number of programs implemented in each county, by cohort, as a control for program availability during the 4 years of the study.[2] We used scores on the scale assessing Perceived Availability of Drugs (from the HYS dataset), averaged by county, as a community-level risk factor. Finally, we calculated average unemployment rates (US Department of Labor 2012) and median income (State of Washington Office of Financial Management 2010), also by county and by cohort, to control for potential economic effects on program participation.

Program-Level Explanatory Variables Pretest observations on participants within programs may not be independent of one another, because individual programs are likely to recruit or attract similar participants. To account for this

[2] Our number of program variable refers to frequency, not location within the county, and hence refers to opportunity for access, not ease of access.

Table 1 Comparison of Strengthening Families Program and healthy youth samples

	SFP (N=1,502)		HYS (N=68,846)		
	%		%		χ^2
Demographics					
Male	46.07		47.87		1.91
Female	43.14		51.86		44.71**
Gender missing	10.79		0.27		3,291.69**
White	45.21		50.84		18.68**
Latino/a	20.24		15.08		26.06**
American Indian	4.66		5.91		4.13*
African American	1.86		2.42		1.92
Other or multi	6.06		23.91		339.13**
Race missing	21.97		1.85		2,676.75**
	M	SD	M	SD	t
Family functioning					
Reinforcement	3.28	0.66	3.21	0.73	−3.37*
Involvement	2.85	0.67	3.04	0.77	9.80**
Family conflict	2.53	0.82	–	–	–
Attachment	2.97	0.78	–	–	–
Family Management	3.45	0.53	–	–	–

The t test represents comparison between SFP and HYS on the two variables (rewards for prosocial behavior and opportunities for prosocial behavior) included in the first selection equation (participation). The other three risk/protective factor variables were included only in the completion and improvement equations for SFP youths, in which there was no comparison with the HYS sample.

SFP Strengthening Families Program, *HYS* Healthy Youth Survey, *Reinforcement* Reinforcement for Prosocial Behavior, *Involvement* Opportunities for Prosocial Behavior, *Family Management* Poor Family Management: Poor Family Management and Family Conflict are reverse scored so that higher scores indicate better family management practices

* $p<0.05$; ** $p<0.001$

interdependence of observations, we calculated a program-level average of risk and protective factor scale pretest scores and their average standard deviation and included these variables in the equations predicting completion and outcome. We used dummy variables for each county to control for within-program similarity related to geographic location.

Outcome Measure We created a binary outcome measure from scores on the five risk and protective factor scales (described above under participant-level explanatory variables) using the following algorithm: observations received a "1" for each dichotomized risk/protective factor variable on which their posttest score was more than one half a standard deviation higher than their pretest score. Observations who improved on at least three of the five risk factors, and who did not show a decrease of one half a standard deviation or more on any risk factor, received a "1" on the dichotomous improvement variable and "0" otherwise. The outcome measure is child reported but describes parent behaviors and family functioning. Therefore, throughout, we consider the outcome variable as an indicator of family functioning.

Selection of Explanatory Variables for Each Equation

We used youths' demographic characteristics (age, race, and gender) and pretest scores from two of the five risk and protective factor scores, Opportunities for Prosocial Involvement and Rewards for Prosocial Involvement, as explanatory variables in the participation equation. Items in these scales assess the degree to which parents and youths are engaged with one another and youths feel reinforced by that engagement; we chose these protective factor scales as the most likely to represent a family's positivity and willingness to attend SFP (and therefore as most likely to model selection into the program), given the interactive nature of the program. As explanatory variables[3] in the selection equation, we also used county-level variables, which we considered likely to be related to participation but not to a family's completion or outcome: the number of programs offered in a county

[3] The county-level variables are used as instrumental variables (explanatory variables that are correlated with unobserved predictors but not with outcome) in our estimation. Instrumental variable methods are explained in Technical Appendix B.

during the period of data collection for program availability[4] (from our records), the Perceived Availability of Drugs in a community (from the HYS, as an indication of environmental risk that might cause families concern) for a family's perception of drug problems in the community, and the county-level unemployment rates and median income, as economic hardship may be a barrier to participation.

As explanatory variables in the completion and outcome equations, we added pretest scores from two risk factors (Family Conflict and Poor Family Management Skills) and a protective factor (Attachment). We also used program-level averages of the Family Management scale score and its standard deviation to control for non-independence of observations within programs in both completion and outcome equations.[5]

Missing Data

Because this is a voluntary community evaluation, data quality in some programs was poor—for example, 22 % of SFP participant evaluations were missing data for race and 10.8 % for gender. We created two dummy variables representing missing gender and missing race and included them in analyses to examine whether missingness on these variables was related to outcome, either because program quality was poorer in programs where data quality was also poor, or because participant attributes were related to non-report of demographics. In 14 programs, representing 9.8 % of the participant sample, facilitators did not collect risk and protective factor data from adolescent participants. A further 10 % of participants had posttest but not pretest data. These participants were not concentrated in any particular programs, but family functioning at posttest as measured by risk and protective scores was significantly lower for this group. Presumably, those participants were not present on the first night of the program and thus did not complete the pretest evaluation; lower program dosage might explain their lower scores. It might also be that lower-functioning families were less likely to make it to the program's first night.

Using dummy variables for missing gender or race provides some evidence of whether or not the missing values are missing at random and, in extreme cases, if the people who do not report these variables are different from those that do. Suppose, for example, the estimates show that males and females who report their genders are different. If gender is missing at random, then the missingness should be in proportion to the population, and the coefficient on gender missing should be the weighted average of the two groups. Our results indicate that those not reporting gender or race are somehow different from those that do, and missing gender or race cannot be considered missing at random (see Technical Appendix C).

Results

Initial Participation in the Program

In the participation equation, we compared participant and community characteristics of those who attended SFP with those who did not (Table 2). Of those who reported their children's gender, families with male children were more likely to attend the program than those with female children (β=0.07, p=0.008). An even stronger positive effect was found for those families for whom the child's gender was missing from the data (β=1.59, $p<0.001$). Relative to white youths, Latino youths were more likely to participate (β=0.13, $p<0.001$), and relative to older adolescents, youths in the targeted age range (10–14) were more likely to participate ($p<0.001$). American Indian families (β=−0.17, p=0.004) and those who reported "other" or multiple race/ethnicity (β=−0.61, $p<0.001$) were significantly less likely to attend, but those who lacked race/ethnicity data were more likely to attend (β=0.83, $p<0.001$). Rewards for Prosocial Activities was positively associated with participation (β=0.25, $p<$ 0.001), but Opportunities for Prosocial Involvement was negatively associated with attendance (β=−0.35, $p<0.001$).

At the community level, likelihood of participation increased when there were more programs offered in a county (β=0.07, $p<0.001$) and in counties where Perceived Availability of Drugs was higher (β=2.22, $p<0.001$). Also at the community level, unemployment rates (β=−0.08, $p<$ 0.001) and median income (β=−0.14, $p<0.001$) were negatively associated with participation.

Completion of the Program

After correcting for selection effects due to participation, the only significant individual-level predictors of program completion were male gender and Family Management pretest scores: Families who reported male children were more likely to complete (β=0.17, p=0.04), as were families with youths who rated Family Management skills higher (β= 0.21, p=0.03) (see Table 2). Families of youths who did

[4] As indicated in footnote 2, this variable refers to opportunity for access not ease of access, hence is unlikely to be related to completion.

[5] One alternative would have been to include dummy variables for each program; however, the program-specific risk factor averages and standard deviations contain more information. (The county-level dummies provide an additional control for non-independence of observations.) Another approach would have been to treat program as a random effect, but we are more interested in the fixed effects of programs than in the programs as a random sample of programs from which we wish to generalize to the population of programs at large (cf. Serlin et al. 2003).

Table 2 Results of trivariate probit model of participation, completion, and improvement

	Participation		Completion		Improvement	
	β	SE	β	SE	β	SE
Intercept	−5.93**	0.39	0.55	0.90	3.20**	0.96
Individual						
Male	0.07**	0.03	0.17*	0.08	−0.05	0.10
Gender missing	1.59**	0.09	0.10	0.18	−0.54*	0.23
African American	−0.08	0.09	−0.47	0.28	−0.06	0.39
American Indian	−0.17**	0.06	0.21	0.19	−0.03	0.19
Latino	0.13**	0.04	0.15	0.12	0.17	0.11
Other or mixed	−0.61**	0.05	0.01	0.17	−0.04	0.20
Race missing	0.83**	0.05	−0.53**	0.14	−0.06	0.19
Age 10–11	0.70**	0.08	0.56***	0.30	−0.13	0.38
Age 12	0.75**	0.08	0.56***	0.31	−0.22	0.38
Age 13	0.53**	0.08	0.45	0.31	−0.29	0.38
Age 14	0.75**	0.09	0.46	0.31	−0.52	0.40
Age 15	−0.09	0.09	0.12	0.34	−0.69	0.45
Reinforcement[a]	0.25**	0.03	−0.09	0.09	−0.18***	0.10
Involvement	−0.35**	0.02	0.11	0.09	−0.10	0.10
Attachment			−0.01	0.07	−0.10	0.08
Family Conflict[a]			0.02	0.05	−0.11***	0.06
Family Management[a]			0.21*	0.03	−0.12	0.12
Community						
Number of programs	0.08**	0.00				
Perceived availability of drugs	2.22**	0.22				
Unemployment rate cohort	−0.08**	0.01				
Median income	−0.14**	0.02				
Program						
Family Management average			−0.48*	0.24	−0.29	0.25
Family Management SD			−0.45	0.34	−0.33	0.36
ρ_{12} (participation–completion)			−0.05	0.08		
ρ_{13} (participation–improvement)					−0.10	0.09
ρ_{23} (completion–improvement)					0.59*	0.23

* $p<0.05$; ** $p<0.01$; *** $p<0.10$

Reinforcement Positive Reinforcement for Prosocial Behavior, *Involvement* Opportunities for Prosocial Behavior, *Attachment* Attachment to Parents/Caregivers, *Family Management* Poor Family Management

[a] Poor family management and family conflict risk factors are reverse scored so that higher scores indicate better familymanagement practices. County dummies for 19 counties were included in the Completion model but are not presented here. Community-level variables were calculated by cohort (2004–2006 and 2006–2008) and by county

not report their race/ethnicity were less likely to complete the program ($\beta=-0.53$, $p<0.001$).

At the community level, coefficients for 9 of 20 counties were significant and positive, indicating that program location was related to retention. The most sustained implementations of SFP in the state, and therefore the most experienced facilitators, are located in six of those nine counties. Although we originally included county dummies as a proxy for geography and population characteristics, the higher completion rates in some counties revealed that the county dummies may have served instead as a proxy for program quality, since in most cases, the programs delivered in each county were provided by a single coordinating agency and by the same cadre of facilitators.

Finally, there was a significant effect of program-level family functioning: in programs with higher average risk at pretest, participants were less likely to complete ($\beta=-0.48$, $p=0.05$). Since individual risk and protective factors were also included in the equation, this estimate represents an effect of program composition rather than of individual attributes.

Program Outcome

There was significant improvement from pretest to posttest on the outcome variable (β=3.20, p=0.001, CI=1.33–5.07) (see Table 2). Youths who did not report their gender were less likely to show improvement in outcome from baseline to posttest (β=−0.54, p=0.02). Rewards for Prosocial Involvement (β=−0.18, p=0.09) and Family Conflict were also associated with outcome (β=0.11, p=0.06). Other risk and protective factor scores were not significantly related to outcome, though all estimates were in the same direction and of about the same magnitude. We interpret these results as qualified support for the universality of SFP effects; it seems that those who completed the program were likely to show improvement irrespective of demographic and preprogram risk factors, though there is some evidence of greater improvement in families who began the program with lower functioning on Rewards for Prosocial Involvement and Family Conflict.

Interrelations of Latent Factors of Participation, Completion, and Outcome

The latent factor explaining participation, representing selection effects of those who initially chose to attend the program, was not significantly correlated with the latent factors explaining completion (ρ_{12}=−0.05, ns) and outcome (ρ_{13}=−0.10, ns). In other words, self-selection into the program did not likely introduce substantial bias in estimates of the outcome equation; nonetheless, estimates in the participation equation provide valuable information about factors significantly related to initial program attendance. The latent variables underlying observed program completion and outcome were significantly correlated (ρ_{23}=0.59), indicating that failure to correct for differential attrition would result in biased estimates in the equation for outcome.

Discussion

Our evaluation design allowed us to identify selection effects in program participation despite lack of a formal control group, and we estimated and corrected for joint effects of participation and completion using a trivariate extension of the bivariate probit model with selection. In the introduction, we noted that all causal claims are inferential and discussed the importance of counterfactual models as they relate to the validity of causal inferences in nonexperimental program evaluation. Experimental trials of preventive interventions increase our confidence in the internal validity of causal inferences, but they do so at the expense of external validity (Shadish et al. 2002). Counterfactual models allow us to evaluate interventions in uncontrolled settings and still maintain some confidence in the internal validity of our inferences about program effects. Benchmark studies show that, when carefully designed, studies using methods derived from the counterfactual model control for a primary threat to internal validity (selection) while strengthening the external validity of causal inferences. Thus, the application of these methods holds great promise for the field of prevention science as we move to large-scale community dissemination of evidence-based programs.

Below, we illustrate the practical value of the trivariate probit model with selection by showing how identification of selection effects can be used to inform recruitment and retention strategies and to interpret results of short-term evaluation.

Participation: Who Decides to Attend?

Estimates from the first equation of the trivariate analysis using the HYS control sample showed significant selection effects related to both participant and community attributes. The finding that Latino families were more likely to attend reflects a concerted outreach effort to Latino families across the state during the study period. American Indian families were significantly less likely to attend, which reflects both lower numbers of trained American Indian facilitators and less availability of the program on reservations. Families with children in the targeted age range were also significantly more likely to attend than families with older children, which indicates that the program is attracting its intended audience. Youths whose parents used more positive reinforcement and frequently expressed pride in them were more likely to attend, but those whose parents regularly involved them in family activities and decisions were less likely to attend. Thus, it appears that families with positive interactions, as well as those families in which children are less likely to be involved in decision making, are those most likely to attend the program.

There were also significant selection effects related to community attributes. After controlling for availability of programs, we found that families in counties with greater perceived access to harmful substances were significantly more likely to attend. Families in communities with higher median income were less likely to attend; this finding is inconsistent with previous research on the clinical trial of SFP (Redmond et al. 2002). On the other hand, families in communities with higher unemployment were also less likely to attend. In counties hit hard by the economic crisis, there was most likely a higher percentage of families experiencing upheaval and who were therefore less likely to engage in new activities.

Implications Identification of selection effects using a nonexperimental control group allowed us to verify the effectiveness of our outreach to Latino families, to identify training needs for American Indian facilitators, and to learn

that programs were more likely to successfully recruit in higher-risk counties. This information is useful to strategic planning of large-scale training and recruitment efforts.

Completion: Who Keeps Coming?

Completion of the program was related to fewer individual and county/program effects than initial participation. Families who reported male children were more likely to complete the program, even after we accounted for the fact that they were also more likely to select into the program. The higher dropout rate of those who did not report race/ethnicity could be a result of attendees dropping out of programs that had less-experienced or less-skilled providers; these providers may also have been less skilled at conducting the evaluation, resulting in higher rates of missing data from the demographics form collected at the beginning of the program. Participants who were reluctant to submit personal information (e.g., immigrants without legal status) may also have been more likely to drop out of the program. Families with higher levels of Family Management skills were more likely to complete the program. Since the Family Management scale assesses monitoring and clear communication of rules and expectations, higher scores may reflect greater structure generally in family organization; greater organization would facilitate regular program attendance. The average level of baseline family functioning within the program was marginally negatively related to completion, but within-program heterogeneity of family functioning was not significantly related to completion. Thus, there was only weak evidence that the composition of a program affected program retention, but programs in counties with more experienced facilitators tended to have higher retention rates.

Implications Lochman and van den Steenhoven (2002) posit that low program attendance presents the greatest barrier to public-health-level effects of prevention, and in our study attrition analysis demonstrated that some kinds of families were more likely to drop out, some program attributes were associated with higher dropout, and providers in some counties were especially effective in retaining participants. Identification of selection effects at this stage allowed us to determine that facilitators should target extra retention efforts toward families who appear to have less structure initially, and the significant county effects provide signposts to facilitators who might need more technical assistance. Many facilitators have questions about whether it is acceptable to include higher-risk families in a program with others; our results show that within-program variation in risk factor levels did not appear to be a problem. The finding that there were lower completion rates among those with incomplete demographics is a puzzle and provides direction for discussion with facilitators of those programs.

Outcome: Who Benefits?

Despite the conservative algorithm used to calculate outcome (scores improved by at least one half a standard deviation on at least three of five risk factors and no decrease on any score), the estimate for the intercept in the third equation was significant, showing that there was a positive probability of significant short-term improvement in risk/protective factor outcome scores. Program outcome was not explained by race, gender, age, community factors, or program composition, after accounting for effects of participation and completion. There was weak evidence that two of the family risk and protective factors were related to outcome.

Implications We interpret these results as providing qualified support that SFP was universally beneficial, at least in the short term, to families who chose to participate in and then completed the program. However, because there were significant selection effects in both participation and completion, we are unable to determine the extent to which improvement in outcome would have been universal had the case mix been more representative of the population at large. For example, families from counties with lower median incomes and higher perceived availability of drugs were more likely to attend; we are unable to determine, from this study, whether the pattern of results would differ if families from higher-income, lower-risk communities were equally likely to attend.

Limitations

As recommended from comparison of results from nonexperimental studies with benchmark experiments (cf. Cook and Steiner 2010), we constructed a comparison group drawn from an equivalent population to that of our sample; our measures were reliable, and our models for selection into the program and completion go beyond simple demographics with the inclusion of individual, program, and county variables. However, as noted earlier, a primary criticism of matching and Heckman-type methods is that unobserved variables influencing selection are unlikely to be adequately modeled and that causal inferences drawn from nonexperimental research may therefore be invalid. Additionally, research has consistently shown that quality of implementation is related to program retention and outcomes (Durlak and DuPre 2008), and though we had indirect indicators of program quality in the county dummy variables, we did not directly assess the specifics of implementation quality (e.g., program fidelity and participant/provider fit and rapport). We speculate that addition of such information would be especially influential in the second equation, explaining program completion. Recent research

by McGowan and colleagues has shown that variations in attendance may be systematically related to factors that are also related to outcomes, and even small selection effects in their clinical trial data caused substantial bias in estimation of outcomes (McGowan, et al. 2010). In the current study, we have not accounted for potential bias related to attendance and dosage effects, nor are we able to determine the effects of missing pretest data. Our findings about race and gender are limited by the number of observations missing those data. Hence, what we were able to infer is limited only to those who reported race and gender. In fact, our results indicate that those not reporting race and/or gender may be different from those who do.

One of the purposes of the paper was to present a design and analytic approach that decreases bias in the estimation of outcomes by controlling for selection in community-based disseminations. We cannot claim with certainty that a change over time equals a causal effect; however, modeling the error terms to control for selection increases the internal validity of our inferences about program outcomes.

Conclusion

Evaluation of programs in uncontrolled settings is complicated but critical to our accurate understanding of the public health benefits of preventive interventions. The development of new methodologies to address problems of internal and external validity inherent in nonexperimental studies, and the importing of such methodologies from other fields, represent an important next step in the translational research agenda for prevention science. In the current paper, we described a family of methods based on the counterfactual model and demonstrated the use of one, a trivariate probit model with selection, that can be used to estimate and correct for joint effects of selection and attrition in nonexperimental studies.

Besides demonstrating how statistical methods can improve both internal and external validity of statistical analysis and inferences, our results, while applicable only to the program we explored, have important implications more generally as prevention programs demonstrated efficacious in clinical trials are implemented in communities. Two stand out. The first is that program participation varied greatly by gender, race, availability, and other factors. This is not surprising, as effectiveness trials have shown similar results. However, the study design allowed us to examine this variability in a naturalistic context without experimental control and to verify that we were successfully reaching some targeted groups but not others. Second, we found no correlation between selection into the program and short-term improvement at its end. If the intervention is causal in this improvement, these findings indicate that who joins is not predetermining the outcome, supporting the universality of SFP benefits. However, we did find a correlation between our second level of selection—completing the program—and short-term improvement. This is somewhat disconcerting because it indicates that factors related to who reaches the end of the program are also related to improvement. If this fact is not controlled for, parameter estimates are biased and inconsistent, which could lead to incorrect inferences from the results. Our results and inferences are more robust because we have controlled for this possibility.

Acknowledgments This study was supported in part by the National Institute of Drug Abuse (grants R21 DA025139-01Al and R21 DA19758-01). We thank the Washington State Department of Health for providing the supplementary data sample, and we thank the program providers and families who participated in the program evaluation.

References

Arendt, J.N. & Holm, A. (2006). Probit models with binary endogenous regressors (working paper 4/2006). Retrieved from Department of Business and Economics at the University of Southern Denmark website: http://static.sdu.dk/mediafiles/Files/Om_SDU/Institutter/Ivoe/Disc_papers/Disc_2006/dpbe4%202006%20pdf.pdf. Accessed 23 Oct 2012.

Arthur, M. W., Briney, J. S., Hawkins, J. D., Abbott, R. D., Brooke-Weiss, B. L., & Catalano, R. F. (2007). Measuring risk and protection in communities using the Communities that Care Youth Survey. *Evaluation and Program Planning, 30*, 197–211.

Barnard, J., Frangakis, C. E., Hill, J. L., & Rubin, D. R. (2003). Principal stratification approach to broken randomized experiments. *Journal of the American Statistical Association, 98*, 299–323. doi:10.1198/016214503000071.

Berinsky, A. (2004). *Silent voices: Opinion polls and political representation in America*. Princeton: Princeton University Press.

Bhattacharya, J., Goldman, D., & McCaffrey, D. (2006). Estimating probit models with self-selected treatments. *Statistics in Medicine, 25*, 389–413. doi:10.1002/sim.2226.

Biglan, A., Hood, D., Brozovsky, P., Ochs, L., Ary, D., & Black, C. (1991). Subject attrition in prevention research. *NIDA Research Monograph, 107*, 213–234.

Bushway, S., Johnson, B. D., & Slocum, L. A. (2007). Is the magic still there? The use of the Heckman two-step correction for selection bias in criminology. *Journal of Quantitative Criminology, 23*, 151–178. doi:10.1007/s10940-007-9024-4.

Cook, T. D., & Steiner, P. M. (2010). Case matching and the reduction of selection bias in quasi experiments: The relative importance of pretest measures of outcome, of unreliable measurement, and of mode of data analysis. *Psychological Methods, 15*, 56–68. doi:10.1037/a0018536.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management, 27*, 724–750.

Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics, 84*, 151–161.

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology, 41*, 327–350.

Foster, E. M. (2010). Casual inference and developmental psychology. *Developmental Psychology, 46*, 1454–1480.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica, 47*, 153–161.

Heckman, J. J., Ichimura, H., Smith, J., & Todd, P. (1996). Sources of selection bias in evaluating social programs: An interpretation of conventional measures and evidence on the effectiveness of matching as a program evaluation method. *Proceedings of the National Acadamy of Science, 93*, 13416–13420.

Hill, L. G., Goates, S. G., & Rosenman, R. (2010). Detecting selection effects in community implementations of family-based substance abuse prevention programs. *American Journal of Public Health, 100*, 623–630.

Kumpfer, K. L., Molgaard, V., & Spoth, R. (1996). The Strengthening Families Program for the prevention of delinquency and drug use. In R. D. V. Peters & R. J. McMahon (Eds.), *Preventing childhood disorders, substance abuse, and delinquency. Banff International Behavioral Science Series (Vol. 3)* (pp. 241–267). Thousand Oaks: Sage Publications.

Lahiri, K., & Song, J. G. (2000). The effect of smoking on health using a sequential self-selection model. *Health Economics, 9*, 491–511.

Lesaffre, E., & Molenberghs, G. (1991). Multivariate probit analysis: A neglected procedure in medical statistics. *Statistics in Medicine, 10*, 1391–1403.

Lochman, J. E., & van den Steenhoven, A. (2002). Family-based approaches to substance abuse prevention. *The Journal of Primary Prevention, 23*, 49–114.

Maxwell, S. E. (2010). Introduction to the special section on Campbell's and Rubin's conceptualizations of causality. *Psychological Methods, 15*, 1–2.

Maydeu-Olivares, A., Coffman, D. L., & Hartmann, W. M. (2007). Asymptotically distribution-free (ADF) interval estimation of coefficient alpha. *Psychological Methods, 12*, 157–176.

McGowan, H. M., Nix, R. L., Murphy, S. A., & Bierman, K. L. (2010). Investigating the impact of selection bias in dose–response analyses of preventive interventions. *Prevention Science, 11*, 239–251.

Neyman, J. (1923/1990). On the application of probability theory to agricultural experiments: Essay on principles. Section 9. *Statistical Science, 5*, 465–480.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge: Cambridge University Press.

Redmond, C., Spoth, R., & Trudeau, L. (2002). Family- and community-level predictors of parent support seeking. *Journal of Community Psychology, 30*, 153–171.

Roodman, D.M. (2007). CMP: Stata module to implement conditional (recursive) mixed process estimator. *Statistical software components*. http://ideas.repec.org/c/boc/bocode/s456882.html. Accessed 23 Oct 2012.

Roodman, D. (2009). Estimating fully observed recursive mixed-process models with cmp. http://www.cgdev.org/content/publications/detail/1421516. Accessed 23 Oct 2012.

Rosenman, R., Mandal, B., Tennekoon, V., & Hill, L.G. (2010). Estimating treatment effectiveness with sample selection (working paper 2010-05). Retrieved from School of Economic Sciences at Washington State University website: http://faculty.ses.wsu.edu/WorkingPapers/Rosenman/WP2010-5.pdf. Accessed 23 Oct 2012.

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

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

Rubin, D. B. (2004). Teaching statistical inference for causal effects in experiments and observational studies. *Journal of Educational and Behavioral Statistics, 29*, 343–367.

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *Annals of Applied Statistics, 2*, 808–840.

Serlin, R. C., Wampold, B. E., & Levin, J. R. (2003). Should providers of treatment be regarded as a random factor? If it ain't broke, don't "fix" it: A comment on Siemer and Joormann (2003). *Psychological Methods, 8*, 524–534.

Shadish, W. R. (2010). Campbell and Rubin: A primer and comparison of their approaches to causal inference in field settings. *Psychological Methods, 15*-, 3–17.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton, Mifflin and Company.

Spoth, R., Randall, G. K., & Shin, C. (2008). Increasing school success through partnership-based family competency training: Experimental study of long-term outcomes. *School Psychology Quarterly, 23*, 70–89.

US Department of Labor, Bureau of Labor Statistics (2012). http://www.bls.gov/data/#unemployment. Accessed 23 Oct 2012.

Washington Office of Financial Management (2010). http://www.ofm.wa.gov/localdata/default.asp. Accessed 23 Oct 2012.

Washington State Department of Health (2006). Washington State Healthy Youth Survey. http://www.doh.wa.gov/Portals/1/Documents/Pubs/WashingtonStateHYS2006.pdf. Accessed 23 Oct 2012.

West, S. G., & Thoemmes, F. (2010). Campbell's and Rubin's perspectives on causal inference. *Psychological Methods, 15*, 18–37