Published in final edited form as:
Prev Sci. 2008 December ; 9(4): 288–298. doi:10.1007/s11121-008-0104-y.

Estimating intervention effects of prevention programs: Accounting for noncompliance

Elizabeth A. Stuart[1,2], **Deborah F. Perry**[3], **Huynh-Nhu Le**[4], and **Nicholas S. Ialongo**[1]

Elizabeth A. Stuart: estuart@jhsph.edu

[1] Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, 624 N Broadway, 8th Floor, Baltimore, MD 21205; 410-502-6222; www.biostat.jhsph.edu/~estuart

[2] Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 N Wolfe St., Baltimore, MD 21205

[3] Department of Population, Family, and Reproductive Health, Johns Hopkins Bloomberg School of Public Health, 615 N Wolfe St., Baltimore, MD 21205

[4] Department of Psychology, The George Washington University, 2125 G St. NW, Washington, DC 20052

Abstract

Individuals not fully complying with their assigned treatments is a common problem encountered in randomized evaluations of behavioral interventions. Treatment group members rarely attend all sessions or do all “required” activities; control group members sometimes find ways to participate in aspects of the intervention. As a result, there is often interest in estimating both the effect of being assigned to participate in the intervention, as well as the impact of actually participating and doing all of the required activities. Methods known broadly as “complier average causal effects” (CACE) or “instrumental variables” (IV) methods have been developed to estimate this latter effect, but they are more commonly applied in medical and treatment research. Since the use of these statistical techniques in prevention trials has been less widespread, many prevention scientists may not be familiar with the underlying assumptions and limitations of CACE and IV approaches. This paper provides an introduction to these methods, described in the context of randomized controlled trials of two preventive interventions: one for perinatal depression among at-risk women and the other for aggressive disruptive behavior in children. Through these case studies, the underlying assumptions and limitations of these methods are highlighted.

Keywords

complier average causal effect; dosage effects; instrumental variables; randomized controlled trials

I. Introduction

Individuals not fully complying with (or adhering to) their assigned treatments is a common challenge encountered in experimental evaluations. Some individuals in the treatment group may not fully participate, while some assigned to the control group may find a way to participate in at least some aspects of the intervention of interest. This may be a particular problem in preventive interventions, where individuals identified to be at risk for a disorder may not feel the same motivation to comply as do individuals with an actual diagnosis who

may receive treatment for that condition (e.g., patients enrolled in a cancer treatment trial). One of the difficulties that results from individuals not fully participating is that, if the proportion of individuals who fully participate is relatively low, a large effect for them may be swamped by smaller (or no) effects for those who do not fully participate, leading to a small effect overall. Estimating the effect of fully participating (known as the complier average causal effect [CACE]) as well as the effect of being assigned to the treatment group (known as the intent to treat [ITT] effect) allows researchers to obtain a more complete view of the effects of their interventions.

Another reason researchers may want to estimate the effect of full participation (the CACE) is because it may be more generalizable than the ITT effect. Across different populations and studies, rates of compliance and patterns in the behavior of compliers may change. However, the effects of full participation in the program should not change in the same way, although it may still vary across individuals if there are moderators of the effects. This is similar to a randomized trial for a pharmaceutical drug treatment, where the compliance rates during drug testing phases may be very different from the adherence rates once the drug has been approved and stated to be effective. However, the biological effect of the drug in the body should stay constant. Likewise, if there is no effect moderation, the effects of complying with a behavioral intervention should be equivalent across different studies, even if the compliance behavior differs.

When there is noncompliance within treatment groups, standard statistical methods (e.g., t-tests or regression analysis) can contribute to our understanding of the effectiveness of an intervention. Specifically, these traditional methods generate unbiased estimates of the effect of assigning people to participate in the program of interest—what is commonly referred to as the ITT effect. What more advanced statistical methods add to this is an ability to separate out the effect of actually receiving the treatment—the CACE effect. More advanced methods are needed since individuals are not actually randomly assigned to their levels of participation. In fact, characteristics of these participants determine how much they comply with the demands of the treatment protocol. In this paper we describe these methods, known broadly as “instrumental variable” (IV) or CACE approaches (Angrist, Imbens, & Rubin, 1996; Bloom, 1984; Little & Rubin, 2000). There has been growing interest in these techniques as more sophisticated methods to handle noncompliance in randomized evaluations are becoming increasingly available. However many applied researchers still know little about the fundamentals of the CACE methods being used. This paper provides an overview of the main concepts behind and assumptions underlying CACE methods. An understanding of CACE methods, their assumptions, and their implications can help researchers learn more from their randomized evaluations and interpret the results appropriately.

Although there have been numerous advances in the statistical literature on methods for estimating CACE effects (e.g., Jo, 2002a; Little & Yau, 1998; Peng, Little, & Raghunathan, 2004), easy to understand descriptions of either the basic or more advanced methods are rare. Little and Rubin (2000) and Dunn, Maracy, and Tomenson (2005) both describe the fundamentals of these approaches, but these papers are primarily oriented towards statisticians, rather than applied researchers, and do not provide detailed discussion of the methods’ assumptions and their potential validity. In addition, neither paper is focused on the complexities encountered in evaluations of preventive interventions. Prevention scientists and psychologists have begun using CACE methods more frequently (e.g., Black et al., 2006; Connell, Dishion, Yasui, & Kavanagh, 2007), but these authors have tended to focus on substantive application of these methods to their scientific question and do not discuss the underlying methods and assumptions in depth. This paper aims to bridge this divide by providing an intuitive and non-technical description of CACE methods. To use

CACE methods properly, applied researchers need to understand what these methods can (and cannot) do, and what assumptions they rely on.

The discussion is motivated and illustrated by two studies. The first is a randomized evaluation of the *Mamás y Bebés: Proyecto del Estado de Ánimo y la Salud/Mothers and Babies: Mood and Health* (MB project), a program to prevent perinatal depression among at-risk Latina women, carried out by a research team led by Dr. Le at the George Washington University (Le, Perry, Stuart, & Ortiz, 2008). The prevention course consists of eight weekly sessions of group psychoeducation during pregnancy, based on cognitive behavioral theory. However, most participants did not attend all eight sessions. Women and their babies were followed for a year postpartum to assess incidence of major depressive episodes. The second study is a randomized trial of a family school partnership (FSP) program aimed to improve academic achievement and reduce problem behaviors among schoolchildren (Ialongo, Werthamer, Brown, Kellam, & Wai, 1999). That trial was carried out by the Prevention Research Center (PRC) at Johns Hopkins University in the mid-1990's, and the participating students have been followed since. The FSP intervention included classroom-focused activities as well as 66 take-home activities for students to do with their families. However, almost none of the children did all of the activities. In both studies, there is interest in estimating not just the effect of being assigned to the intervention group, but also the effect of actually participating in most of the intervention activities.

This paper proceeds as follows. In Section II we describe the framework behind the estimation of the effects of full participation in an intervention of interest. This includes an intuitive discussion of the statistical methods used to estimate the CACE as well as the standard assumptions that underlie its estimation. Section III illustrates the use of these methods in the two preventive interventions described above, including discussion of the validity of the assumptions in those settings. Section IV concludes with areas for future work and advice for prevention researchers carrying out intervention studies.

II. Estimating the effect of full compliance

Current approaches and their limitations

What are the approaches researchers typically use currently to handle noncompliance in randomized trials? Perhaps the simplest approach is to ignore the noncompliance and focus only on ITT effects. However, that can limit the amount of information obtained from a study, as described above. A second approach is to use just the treatment group members to examine the relationship between different levels of participation and outcomes (e.g., Kam, Greenberg, & Walls, 2003). While this allows the determination of which levels of participation are associated with better outcomes, it does not estimate the causal effects of those levels of participation since there is no comparison with a comparable comparison group.

Perhaps the most common method for dealing with varying levels of compliance is what is known as an "as-treated" analysis, which compares the people who fully participate with those who didn't (regardless of their original treatment assignment). An "as-treated" analysis yields biased estimates of the effect of full participation because it is estimating the effect by comparing two different groups of people: those who fully participate in the treatment group with everyone else (which consists of the non-participants in the treatment group as well as the full control group—those who would have participated had they been in the treatment group, and those who wouldn't have. The people who fully participate in the program are likely different from those who do not, in both observed and unobserved ways. For example, in the MB intervention, there is some evidence that the women who fully participated had higher lifetime risk of having had a major depressive episode before entry

into the study. This means that the benefits of random assignment are lost when an as-treated analysis is done, since the groups being compared are no longer only randomly different from each other.

Related to an as-treated analysis is the approach of including the level of participation as a predictor in the model of the outcome, thus "controlling for" participation. Unfortunately this method also yields biased estimates of the effects, since it conditions on just the observed participation levels and implicitly compares individuals in the treatment and control groups with the same observed participation (Frangakis & Rubin, 2002). However, individuals in the treatment group with an observed level of participation (e.g., non-participation) may be different from individuals in the control group with that same level of participation. Different processes may lead to non-participation in the treatment and control groups, and in particular some of the apparent non-participants in the control group would likely have participated if they had been in the treatment group. An estimate of a causal effect needs to be a comparison of outcomes among similar individuals, which this approach does not do. This issue is more broadly known in the epidemiology literature as post-treatment selection bias (Robins & Greenland, 1992).

Instead, the intuitive idea behind CACE analyses is that we need to compare the participants in the treatment condition with a similar (sub)group of people from the control arm of the study—those who would have participated had they been given the opportunity to do so. If there was a straightforward way to identify those individuals directly, we could simply compare the outcomes of the (likely) participants in the treatment and control groups. However, we cannot identify those individuals directly—we do not know which control group members would have participated had they been in the treatment group. CACE methods instead use an indirect approach to estimate the effect of interest. The details of that approach are given below.

Compliance Types

Statisticians find notation is helpful to operationalize these ideas. We consider a situation where a set of N individuals have been randomly assigned to either a treatment condition (T_i =1) or a control condition (T_i = 0). However, not all of the treatment group members actually receive the treatment, and some of the control group members may find a way to do so. Denote the actual treatment ("dose") received by D, such that D_i (T_i =1) reflects the dose individual *i* receives if assigned to the treatment group and D_i (T_i = 0) is the dose individual *i* receives if assigned to control. We generally consider situations where the dose is either 0 or 1, indicating no treatment received or full treatment received, since the underlying assumptions are easier to conceptualize and discuss in this setting. We will discuss relaxation of that requirement below.

Each individual has two potential outcomes: Y_i (T_i =1, D_i (T_i =1)), the outcome we would observe if individual *i* is assigned to the treatment group, and Y_i (T_i = 0, D_i (T_i = 0)), the outcome we would observe if individual *i* is assigned to the control group. The causal effect of the treatment for individual *i* is defined as the difference in individual *i*'s outcomes if assigned to treatment versus control: $\tau_i = Y_i$ (T_i =1, D_i (T_i =1))− Y_i(T_i = 0, D_i (T_i = 0)). For each individual, however, we are able to observe only one of these two potential outcomes. For individuals in the treatment group we observe Y_i (T_i =1, D_i (T_i =1)) whereas for individuals in the control group we observe Y_i (T_i = 0, D_i (T_i = 0)). We can, however, estimate the overall average treatment effect (the ITT effect), which is a comparison of outcomes if everyone was assigned to the treatment group versus everyone assigned to the control group:

$$\tau=\frac{1}{N}\sum_{i=1}^{N} Y_i(T_i=1, D_i(T_i=1)) - Y_i(T_i=0, D_i(T_i=0)).$$

Because of randomization to the treatment and control groups, the difference in outcomes observed in the treatment and control groups gives us an unbiased estimate of τ, the ITT effect. However, unless there is full participation in the treatment group (and no one in the control group receives the treatment), τ does not necessarily provide information on the effects of the treatment itself—it merely tells us the effect of being assigned to receive the treatment. The CACE, defined precisely below, will provide us with an estimate of actually receiving the treatment.

Given binary treatment assignment and doses, there are four types of people, defined by their compliance behavior when assigned to the treatment and their compliance behavior when assigned to the control. These four types (Angrist et al., 1996) are:

- *Compliers (C)*, who fully participate when in the treatment group, and do not participate when in the control group: $D_i(T_i=1)=1$, $D_i(T_i=0)=0$ (e.g., in a standard clinical trial of a new drug, these are the people who would take the drug if in the treatment group but not if they are in the control group),
- *Always-takers (AT)*, who fully participate when in either the treatment or control group: $D_i(T_i=1)= D_i(T_i=0)=1$ (e.g., people who would take the drug if they are in the treatment group, and would also find a way to take it if they are in the control group),
- *Never-takers (NT)*, who do not participate when in either the treatment or control group: $D_i(T_i=1)= D_i(T_i=0)=0$ (e.g., people who would not take the drug if in either the treatment or control groups), and
- *Defiers (D)*, who do not participate when in the treatment group, but do participate when in the control group: $D_i(T_i=1)=0$, $D_i(T_i=0)=1$ (e.g., people who would not take the drug if in the treatment group, but would take the drug if in the control group).

Because the population can be broken up into these four types of people, we can also express the overall ITT effect as the average of the ITT effects for each of these four types, in the same way that we could, for example, calculate an overall effect by averaging the effects for males and females by the proportions of males and females in the population. In this way we can express the overall ITT effect τ as: $\tau=p_C\tau_C + p_{AT}\tau_{AT} + p_{NT}\tau_{NT} + p_D\tau_D$, where p_g is the proportion of the population of type "g" and τ_g 1 is the average treatment effect for group "g": $\tau_g=\frac{1}{N_g}\sum_{i=1}^{N_g} Y_i(T_i=1, D_i(T_i=1)) - Y_i(T_i=0, D_i(T_i=0))$. The key insight is that for the compliers, who do as they are told under each treatment condition, the effect of assignment (the ITT effect) is the same as the effect of fully participating. Thus, our interest is in the overall ITT effect as well as the effect of assignment for the compliers, the CACE.

Assumptions underlying CACE analyses

Because individuals are assigned to the treatment and control groups randomly in experimental evaluations, we can easily obtain an unbiased estimate of the ITT effect by comparing the average outcomes in the treatment and control groups. The challenge in estimating the CACE arises because participation is not randomly assigned and we observe each individual in either the treatment group or the control group—never both. People in the

control group who do not participate in the program may be either compliers or never-takers; those in the control group who do participate may be always-takers or defiers. People in the treatment group who do participate could be either compliers or always-takers; those in the treatment group who don't participate may be either never-takers or defiers. Standard IV methods impose a set of assumptions to help us estimate the CACE effect. Those assumptions are:

1. The outcomes of each individual are not affected by the treatment assignments of any other individuals. For example, in the MB study it assumes that the outcomes of one woman are not affected by the other women in her group. In the PRC study, it assumes that students' outcomes are not affected by which treatments other students receive. This assumption is made in nearly all studies estimating causal effects, and is known as "no interference" or the Stable Unit Treatment Value Assumption (SUTVA; Rubin, 1978).
2. Being given the opportunity to participate was assigned randomly.
3. Being given the opportunity to participate induces some individuals to actually participate. In other words, there are some compliers.
4. There are no defiers. This is sometimes called "monotonicity," to reflect the assumption that being assigned to the treatment group can only increase participation (that there is no one for whom assignment to the treatment group decreases participation). As detailed below, this assumption helps identify the compliers by ruling out particular behaviors.
5. There is no effect of assignment for the never-takers or for the always-takers. In other words, since it does not change their participation behavior, being assigned to the treatment group versus the control group also does not change their outcomes. Another way of saying this is that to benefit from the intervention you must actually participate in it. These assumptions (one for the always-takers and one for the never-takers) are known as the "exclusion restrictions."

Simple Description of Estimation

We can intuitively describe how these assumptions help us estimate the CACE. Our aim is to estimate the effect for compliers, τ_c, as expressed in the following formula: $\tau = p_C\tau_C + p_{AT}\tau_{AT} + p_{NT}\tau_{NT} + p_D\tau_D$. The assumptions detailed above enable us to obtain an estimate of τ_c. First, the no defiers assumption (Assumption 4) implies that $p_D = 0$. Second, the exclusion restrictions (Assumption 5) imply that $\tau_{AT} = 0$ and $\tau_{NT} = 0$. This means that the three final terms in the formula for τ go away, and the formula becomes: $ITT = \tau = p_C\tau_C$. So, under the assumptions detailed above, if we can estimate p_C we can obtain an estimate of τ_C, the estimand of interest, by simply dividing the ITT estimate by p_C.

So how do we estimate the proportion of compliers, p_C ? Consider the people in the treatment group who fully participate. They must be either compliers or always-takers. However, the people in the control group who fully participate must be always-takers, since we have assumed that there are no defiers. The proportion of people in the control group who fully participate thus gives us an estimate of the percentage of always-takers, p_{AT}. Random assignment (Assumption 2) implies that the proportion of always-takers should be the same in the treatment and control groups, and that the proportion of compliers should be the same in the treatment and control groups (just as the proportion of other subgroups, such as males, should be the same in the treatment and control groups). So from the control group we can obtain an estimate of the proportion of always-takers, p_{AT}, and from the treatment group we can obtain an estimate of the proportion of always-takers and compliers, $p_C + p_{AT}$. By subtracting one from the other we can obtain an estimate of the proportion of compliers,

p_C. By dividing the overall ITT effect by this proportion, p_C, we obtain our estimate of the CACE.

Because of this simple form of the estimate, the CACE estimates will generally be larger than the ITT effects. This is a consequence of Assumption 5. In its simplest form, the CACE estimate sets the effects for individuals who do not participate to 0, and redistributes the overall effect to the individuals who do participate. For example, consider a study with 4 people in the treatment group and an overall effect of 10 points, when comparing those 4 people with 4 people in a control group. However, we find that only three of the people in the treatment group (75%) actually took the treatment as assigned. The exclusion restriction (Assumption 5) says that the effect for the individual who did not take the treatment (did not participate) is 0. Therefore, the effect for the three who did participate must be larger, since the average still equals 10. In fact, it must be larger by a factor of 1/(75%). In other words, instead of assuming that the 4 individuals each had an effect of 10, we assume that one individual had an effect of 0 and the other 3 had an effect of 10/(3/4) = 10*(4/3) = 40/3 = 13.3, so that it still averages out to an overall effect of 10. Both estimates are consistent with the data that shows an overall effect of 10; one (the 13.3) takes into account the fact that not everyone in the treatment group participated, and thus that presumably not everyone in the treatment group actually benefited from the intervention. This procedure thus adjusts the effects for the participating individuals upward to account for the fact that some individuals did not participate.

More Complex Estimation

Estimation of the CACE is rarely done quite as simply as described in the previous section, but the main ideas remain the same. Estimation of the CACE is most often done using a "two-stage least squares" (TSLS) approach, which jointly models the two processes of participation and outcome (Angrist & Imbens, 1995). Other estimation approaches include maximum likelihood (Dunn et al., 2005) and Bayesian methods (Imbens & Rubin, 1997). TSLS essentially involves two models: a regression model of participation, and a regression model predicting the outcome, given participation. These models are estimated jointly, to calculate accurate standard errors that account for the uncertainty in the "first-stage" (participation) model. Using TSLS also allows the inclusion of covariates that predict participation and/or the outcome, which can increase the precision of the estimates. This framework also explains where the term "instrument" comes from (as in "instrumental variables" analysis). The "instrument" is the randomization to the treatment or control group, which is assumed to affect participation but not outcomes (except through its effect on participation). So in the two models, the treatment indicator T is used in the model of participation, but not in the outcome model. The specific models often look something like:

$$Y_i = \beta_0 + \beta_1 D_i + \theta X_i + e_i$$
$$D_i^* = \alpha_0 + \alpha_1 T_i + \pi W_i + \nu_i$$

where Y_i is the outcome of interest, T_i is the treatment indicator, and D_i^* is a latent continuous measure of the "dose" (level of treatment) received, related to the binary participation indicator D_i in the following way: $D_i = 1$ if $D_i^* > 0$ and $D_i = 0$ if $D_i^* \leq 0$. In this setting it is possible to consider a continuous dose by using D_i^* in the model for Y_i rather than the D_i shown above (e.g., Foster, 2000). Here we focus on the binary dose indicator D_i^* to clarify the underlying assumptions in these types of models. The matrix X represents covariates that predict the outcome; W consists of covariates that predict participation. There may be overlap between the predictors in X and W.

In the case studies described in this paper, where the control group does not have access to the intervention, the situation is somewhat simpler, but the same general ideas hold. In the simpler setting, always-takers and defiers cannot exist since no one in the control group can receive the intervention. There is thus only one exclusion restriction, for the never-takers. Finally, again since there are no always-takers, the estimation of the proportion of compliers can be done by simply estimating the proportion of people in the treatment group who fully participate. Below we discuss in more detail how well these assumptions hold in two trials of community-based preventive interventions.

III. Results: CACE analyses in the MB and PRC examples

The MB Study

As briefly described above, the MB Project aims to prevent perinatal depression among at-risk, low-income Hispanic women. Pregnant women receiving prenatal care at two sites in Washington, DC were approached to participate in the study. Women were eligible for enrollment if they were healthy pregnant women at 24 weeks or less gestational age, between 18–35 years old, and at high risk for depression, defined as scoring 16 or higher on the Center for Epidemiological Studies Depression Scale (CES-D; Radloff, 1977) and/or with a self-reported personal or family history of depression. Women were also screened for a current major depressive episode (MDE), necessary to meet criteria for major depression; three women exhibiting symptoms of an MDE at baseline were referred to community-based mental health treatment services and excluded from the study. A total of 217 women were randomized to either the MB course or usual care (UC), with 112 in the MB group and 105 in the UC group. The main ideas and theory behind the course are described in Muñoz et al. (2007). The version of the course investigated in the current study consists of eight weekly classes and three booster follow-up sessions. See Le et al. (2008) for more details on the version of the program investigated in the current study and on the study implementation and results.

The baseline characteristics and covariate values were obtained in a baseline, face-to-face interview, conducted in Spanish prior to random assignment. Data collected included demographics (native country and language use, number of years in the U.S., number of children, employment status), physical health status, receipt of prenatal care, and mental health status (including baseline measures of the outcome variables of primary interest, detailed below). The outcome reported here was measured directly after completion of the MB course (or three to four months after the baseline interview, for women in the UC group). In this paper we focus on one of the study's main outcome measures to illustrate the methods: the Beck Depression Inventory, Second Edition (BDI-II; Beck, Steer, & Brown, 1996). The BDI-II is a 21-item self-report instrument that measures severity of depressive symptoms in the previous two weeks. Higher BDI-II scores reflect greater levels of depressive symptomatology. A total score of 0–13 is considered within the "minimal" range, 14–19 "mild", 20–28 "moderate," and 29–63 "severe." See Le et al. (2008) for a more detailed description of the results of the intervention on a broader range of outcomes.

The PRC study

The second example is drawn from an evaluation assessing the Family School Partnership program, carried out by the Johns Hopkins University Prevention Research Center (PRC) in the mid-1990's (Ialongo et al., 1999). The evaluation compared two first-grade classroom interventions with a control condition (i.e., their usual classroom activities). All first grade students in nine Baltimore schools were randomly assigned to intervention or control conditions. The intervention we focus on here is the Family School Partnership, which included both classroom activities and 66 activities for the students to do at home with their

families. The aim was to decrease disruptive behavior and improve academic achievement. The sample we use here is the same as that used in Jo (2002a), and includes 284 children with complete data on the covariates and outcome. The covariates were measured in the fall of first grade while the outcome considered was measured approximately 18 months later, during the spring of second grade. The covariates available include demographics, parent characteristics (including whether the student was eligible for free lunch, a measure of socioeconomic status), and teacher ratings of classroom behavior. The illustrative outcome we examine is the teachers' ratings of how often the child engages their teacher and classmates in appropriate social interaction, measured during the spring of second grade. All teacher ratings (both baseline and outcome) were obtained using the Teacher Observation of Classroom Adaptation-Revised score (TOCA-R; Werthamer-Larsson, Kellam, & Wheeler, 1991). The analysis here complements those in Jo (2002a) by providing a simpler description of the methods and providing estimates under slightly different assumptions, as detailed below.

Defining Full Participation

In both the MB and PRC examples there is interest in the effects of "full" participation in the interventions. The first question, then, is how to define "full" participation. In the MB study, participation is defined as going to at least four of the eight classes. Four classes was believed to be the minimum number of classes a woman would need to attend in order to benefit from the program (Le et al., 2008). In the PRC study, completing 45 of the 66 take-home activities was defined as full participation, as in Jo (2002a).

Validity of Assumptions in MB and PRC examples

Next, we will examine the degree to which the five key assumptions of CACE analyses are met in the context of each of these preventive interventions. These assumptions are briefly restated: 1) The outcomes of one individual are not affected by the treatment assignments of any other individuals (SUTVA); 2) The opportunity to participate was assigned randomly; 3) There are some compliers; 4) There are no defiers (monotonicity); and, 5) There is no effect of assignment for the always-takers or the never-taker (the exclusion restrictions).

SUTVA (Assumption 1) is potentially problematic in both interventions, as it may be in many prevention programs. The MB program involves group sessions; the composition of the group may have an effect on women's outcomes. Similarly, in the PRC intervention, which combined classroom and family activities, the classroom environment and classmates' behavior may affect individuals' outcomes. Unfortunately, there has been very little progress in developing methods to account for the interaction of individuals in either ITT or CACE estimation. Rarely do researchers even acknowledge that this assumption has been violated and what the consequences of this might be for their findings. One exception is Jo, Asparouhov, Muthén, Ialongo, and Brown (2008), who discuss some of the dangers in ignoring clustering of compliance behavior and propose some modeling options. For the purposes of this paper, we will assume that SUTVA holds well enough for these analyses.

Both studies involved random assignment of individuals to the treatment and control groups, and so we know that random assignment (Assumption 2) is satisfied in both studies. It is also easy to see that there are some compliers in both studies. In the MB study, approximately 55% of the women in the treatment group completed four or more classes. In the PRC study, 50% of the students in the treatment group completed over 45 of the take-home activities. Since control group members did not have access to the intervention in either study, there are no always-takers and thus these treatment group members who participated must be compliers, satisfying Assumption 3. The fact that individuals in the control group did not have access to the treatment in either study also implies that there

cannot be any defiers (or always-takers), satisfying Assumption 4. Although in other studies of the same interventions it is possible that the control group would have access to the treatment (and thus some individuals may be always-takers or defiers in those studies, changing the inferences), the relevant classification for these analyses is behavior in the current studies, which did not allow control group members access to the interventions.

The exclusion restriction (Assumption 5) is perhaps the most problematic for both studies. The exclusion restriction states that there is no effect on the people who do not fully participate. Since participation is defined in the MB study as 4 or more classes and in the PRC study as 45 or more take-home activities, this assumption implies that women in the MB study who attend 3 classes receive no benefits of those 3 classes. Similarly, there is no effect of doing 44 take-home activities in the PRC study. In this paper we assume that the exclusion restriction (Assumption 5) holds well enough for both studies; whether or not it does is an important, substantive question that can only be answered through increased dialogue and collaboration between statisticians and applied researchers. Hirano, Imbens, Rubin, and Zhou (2000) and Jo (2002a) provide alternative estimation techniques that allow relaxation of the exclusion restrictions.

CACE Results: MB Program

To put these methods into practice, we first compare the characteristics of women who fully participated in the MB program (i.e., attended four or more classes) with those who did not (Table 1). This analysis relies on just the women in the treatment group, since we do not know the compliance status of the women in the control group—we do not know if they would have participated had they been in the treatment group. In the treatment group, 62 of the 112 women attended four or more classes, which implies 55% compliers.

Although there are differences between the women who participated and those who did not (e.g., the percentage receiving previous psychiatric treatment and the percentage in the U.S. for less than one year), few of these differences are statistically significant (Table 1). However, while Table 1 indicates few observed differences between these two groups of women in the treatment condition, it is likely that there are also unobserved differences between those who did and did not participate. Therefore, we should not simply compare the outcomes of the participants with those of non-participants. We instead use the CACE methods described above to estimate the effect of participation. This involves using the two-stage least squares approach to jointly model participation in the program and outcomes, conditional on participation. To increase the precision of the CACE estimates, we use the variables that were statistically significant predictors of baseline depressive symptoms (the baseline measure of our primary outcome of interest) as predictors in the CACE model; in this sample, these were: married or cohabitating, single or widowed, husband's job status, and self-rated health status excellent or very good (see Le et al., 2008 for more details). All analyses were run in R Version 2.5.1 (R Core Development Team, 2007). ITT estimates were calculated using t-tests and the CACE estimates were obtained using the two-stage least squares function "tsls."[1] Similar functions are available in other packages, including the "ivregress" command in Stata 10 (StataCorp, 2007)[2] and the "2-stage least squares" regression commands in SPSS (SPSS Inc., 2007).

We first calculate the effect on BDI-II scores of being given the opportunity to participate in the MB program—the ITT effect, which is −2.30 (p=0.03). Women who were given the opportunity to participate in the program had lower BDI–II scores at the end of the

[1]The exact line of code for R is "tsls(outcome ~ D + x1 + x2, ~ T + x1 + x2, data=dataset)".
[2]The exact line of code for Stata is "ivregress 2sls outcome x1 x2 (D=T)".

intervention than women who were not given the opportunity to participate. However, the ITT estimate does not take into account how much the women actually participated.

As expected, the CACE estimates are somewhat larger than the ITT estimates, but with similar significance levels (CACE estimate=−3.70, p=0.04). Consistent with our intuitive description of CACE estimation above, the CACE estimates are in fact approximately the ITT estimates divided by 0.55 (the proportion of compliers in the population). (They are not exactly the ITT estimates divided by 0.55 because of the use of predictors in the models of participation and the outcome).

CACE Results: PRC intervention

We now turn to the second example, the PRC evaluation of the Family School Partnership intervention. Using 45 take-home activities as the minimum for full participation, approximately 50% of the students in the treatment group fully participated in the intervention. Table 2 compares the baseline characteristics of the participants and non-participants in the treatment group. As with the MB program, we do not see large differences in the observed characteristics of the participants and non-participants. However, there are some indications of differences in the parent's age and in the student's race, with younger and white parents more likely to fully participate.

For the FSP intervention, the ITT estimate of the effect on shyness in second grade is −0.33 (p=0.01) and the CACE estimate is −0.65 (p=0.01). The adjustment models for both estimates include all of the covariates shown in Table 2. Again we see that the CACE estimates are larger than the ITT estimates, but with similar significance levels. Since the participation rate is 50%, the CACE estimate is approximately double the ITT estimate. This is a consequence of the exclusion restriction, which assumes that there was no effect of the program for the 50% of children who completed fewer than 45 take-home activities; it is important to understand that these estimates rely heavily on the assumptions described above, such as the exclusion restriction. Without those assumptions these estimators are invalid. These results are very similar to those reported in Jo (2002a). That paper also uses alternative model restrictions to relax the exclusion restriction; when that adjustment is made, Jo (2002a) finds that the effects on both compliers and non-compliers are not statistically significant, indicating that the results are sensitive to the exclusion restriction.

IV. Discussion

Researchers are often interested in gaining more insights into how to transport their findings from a randomized controlled trial to real-world applications, for example by estimating the effects of different levels of participation. CACE models provide one way of examining this issue, by allowing researchers to estimate the effect of full participation. As CACE methods become more popular it is important that researchers understand their underlying assumptions as well as their limitations. Thus far in this paper we have attempted to clarify the underlying assumptions of CACE methods; in this discussion we expand on some of the limitations and unresolved issues.

One limitation of the most basic CACE methods described above is the need to define "full" participation as a binary variable. Most preventive interventions are not one-time events with a clear "yes" or "no" for participation. This leads to two particular complications. First, there is a trade-off in selecting the cut-off defining "full" participation. On one hand, defining a high cut-off will imply a larger estimated CACE. On the other hand, the exclusion restriction (Assumption 5) may be harder to justify with a high cut-off. This can be particularly problematic for many behavioral interventions, since the exclusion restriction assumes that any participation less than full participation yields no effect; this means that, in

the MB project example, attending 3 MB classes does not led to no benefits, while attending 4 classes does. Some research has addressed this by running analyses with different cutpoints (e.g., Black et al., 2006), however when doing so it is also important to carefully consider and discuss which cutpoints make the exclusion restrictions reasonable. Relaxing the exclusion restriction through alternative model restrictions (e.g., Jo, 2002a) and sensitivity analyses that assess the robustness of the results to violation of the exclusion restrictions (e.g., Dunn et al., 2003) are important directions for further methodological research and applied use. Understanding the theory underlying the intervention, and what level of participation would be expected to be required to see effects, can also greatly inform these analyses, as discussed further below.

The second complication with binary indicators of participation is that, when continuous measures of participation are available, researchers may be reluctant to create a single dichotomous variable. If this is the case, dose-response or partial compliance models that allow for varying levels of participation (e.g., Foster, 2000, 2003; Jin & Rubin, 2008; Rosenbaum, 2002) can be useful. Unfortunately, however, those approaches require even more modeling assumptions than described here, and it is another area in which the methods and clear statements of their underlying assumptions are still developing.

A related unresolved issue is how to combine the multiple measures of implementation that are often available. For example, for studies like the PRC trial, there may be both individual-level participation, as used here, as well as measures of how well the teacher implemented the classroom-based aspect of the intervention. Researchers generally focus on one measure, or a very simple summary measure of overall participation. A promising approach could be to use latent class analysis among treatment group members to identify different participation classes (e.g., as in Lin, Ten Have, & Elliott, in press), and then use those as the levels of compliance when estimating the CACE, but that approach has not been systematically examined.

Another common complication not addressed in detail in this paper is the effect of missing data, both in covariates and outcomes. The two most common approaches for handling missing data within the CACE framework are maximum likelihood methods and multiple imputation. Software such as Mplus (Muthén & Muthén, 1998–2007) can handle both approaches, with the maximum likelihood methods implemented using full information maximum likelihood under a missing at random assumption. For the MB example described above we did a sensitivity analysis where missing values were imputed using multiple imputation and found that the results were virtually identical to those reported above (Le et al., 2008). One particular complication encountered when there is both noncompliance and missing data is that the missingness may depend on the compliance status, which makes the missingness not missing at random (NMAR). In this case even the ITT estimates may be biased, and more sophisticated modeling approaches are needed (Baker & Kramer, 2005; Frangakis & Rubin, 1999; O'Malley & Normand, 2005). Another limitation of currently available methods is a lack of power analyses that can account for noncompliance; development of methods that estimate power (for both ITT and CACE), given certain levels of noncompliance, would help design future intervention trials in the most appropriate ways (e.g., Jo, 2002b).

A final complication encountered in many preventive interventions is clustering. Unfortunately this is another topic with limited methodological development. Thus, for this paper we ignored the clustering of individuals within groups in the MB study and within classrooms and schools in the PRC study, as is often commonly done in applications of IV methods (e.g., Connell et al., 2007). Nevertheless, clustering is an important issue that can impact the results of a study (Jo et al., 2008). Jo et al. describe advanced statistical methods

to account for the correlation of both compliance behavior and outcomes when randomization is conducted at the group level. Further work needs to be done to consider settings where individuals are randomized but interact with each other in groups, as in the MB and PRC studies considered here.

As an alternative to the CACE models described here, another approach sometimes used to estimate the effects of levels of participation is propensity score matching, where individuals in the treatment group with some participation level are matched to individuals in the control group with similar baseline characteristics (Foster, 2003; Hill, Brooks-Gunn, & Waldfogel, 2003). This procedure assumes that there are no unobserved differences between the participants and non-participants, given the observed characteristics. In other words, we assume that we can identify the individuals in the control group who would have fully participated, just on the basis of their observed covariates. That assumption may be more or less reasonable than the exclusion restrictions in CACE models, depending on the extent of covariates available. Future work should investigate the situations in which each approach is most appropriate.

Given that these techniques offer new insights into data collected by a wide range of prevention researchers, there are several strategies for mitigating the challenges encountered in CACE analyses and facilitating the estimation of the effects of full participation in preventive interventions. These include:

- Encouraging high rates of participation, and identifying (and trying to eliminate) the barriers that individuals face to participation. The higher the participation rate is in the treatment group, the easier it will be to estimate the effects of full participation.
- Limiting access to the program for individuals in the control group. Estimation of the effects of full participation is facilitated when individuals only have access to the treatment if they are assigned to the treatment group, ruling out the existence of defiers and always-takers.
- Measuring baseline characteristics that are believed to affect participation rates. This may include variables not normally considered for standard data collection, such as whether the individual has a car or how long it takes them to get to the program site. The better the baseline variables predict participation, the better the CACE estimation will be.
- Drawing on the theory behind the intervention to determine what constitutes “full” participation. In this way, effective CACE estimation relies on close collaboration between substantive researchers and statisticians, so that the validity of the CACE assumptions can be carefully considered.

There are also important policy implications that follow from a broader application of these methods in prevention science. As evidence-based interventions are transported from randomized controlled trials (that are often well funded and staffed by trained researchers) to real-world settings, a more nuanced understanding of the predictors of participation will be very useful. Knowing what effects can reasonably be expected to result from specific levels of participation can help program managers implement the necessary supports and strategies to achieve this end. This type of information can also help avoid promising policy makers population-based outcomes that are not generalizable on the basis of traditional ITT analyses. In this current climate of expediting the broad scale adoption of evidence-based interventions in communities across the U.S., it is vital that policy makers and program managers also understand the assumptions and limitations of estimating the effects of these strategies.

In conclusion, CACE methods for estimating the effects of actual participation in a program of interest are important and useful. When using them, however, researchers need to carefully consider the underlying assumptions and their appropriateness. Further methodological research is also needed to develop methods that account for the complexities encountered in real-life studies, such as missing data and clustering. Through collaborative partnerships between applied prevention scientists and statisticians, better estimates of the true effect of interventions can be calculated and used to improve the public's health and well-being.

Acknowledgments

This research supported in part by grant R40 MC 02497 from the Maternal and Child Health Bureau (Title V, Social Security Act), Health Resources and Services Administration, Department of Health and Human Services (PI: Le) as well as by the Center for Prevention and Early Intervention, jointly funded by the National Institute of Mental Health and the National Institute of Drug Abuse (MH066247; PI: Ialongo).

References

Angrist J, Imbens GW. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. Journal of the American Statistical Association 1995;90(430):431–442.

Angrist J, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. Journal of the American Statistical Association 1996;91:444–472.

Baker SG, Kramer BS. Simple maximum likelihood estimates of efficacy in randomized trials and before-and-after studies, with implications for meta-analysis. Statistical Methods in Medical Research 2005;14(4):349–367. [PubMed: 16178137]

Beck, AT.; Steer, RA.; Brown, GK. Manual for the Beck Depression Inventory. 2. San Antonio, TX: The Psychological Corporation; 1996.

Black MM, Bentley ME, Papas MA, Oberlander S, Teti LO, McNary S, Le K, O'Connell M. Delaying second births among adolescent mothers: A randomized, controlled trial of a home-based mentoring program. Pediatrics 2006;118:1087–1099. [PubMed: 16951002]

Bloom HS. Accounting for no-shows in experimental evaluation designs. Evaluation Review 1984;8:225–246.

Connell AM, Dishion TJ, Yasui M, Kavanagh K. An Adaptive Approach to Family Intervention: Linking Engagement in Family-Centered Intervention to Reductions in Adolescent Problem Behavior. Journal of Consulting and Clinical Psychology 2007;75(4):568–579. [PubMed: 17663611]

Dunn G, Maracy M, Dowrick C, Ayuso-Mateos JL, Dalgard OS, Page H, Lehtinen V, Casey P, Wilkinson C, Vazquez-Barquero JL, Wikinson G. Estimating psychological treatment effects from a randomised controlled trial with both non-compliance and loss to follow-up. British Journal of Psychiatry 2003;183:323–331. [PubMed: 14519610]

Dunn G, Maracy M, Tomenson B. Estimating treatment effects from randomized controlled trials with noncompliance and loss to follow-up: the role of instrumental variables methods. Statistical Methods in Medical Research 2005;14:369–395. [PubMed: 16178138]

Foster EM. Is more better than less? An analysis of children's mental health services. Health Services Research 2000;35(5):1135–1158. [PubMed: 11130814]

Foster EM. Propensity score matching: An illustrative analysis of dose response. Medical Care 2003;41:1183–1192. [PubMed: 14515114]

Frangakis CE, Rubin DB. Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. Biometrika 1999;86(2):365–379.

Frangakis CE, Rubin DB. Principal stratification in causal inference. Biometrics 2002;58:21–29. [PubMed: 11890317]

Hill JL, Brooks-Gunn J, Waldfogel J. Sustained effects of high participation in an early intervention for low-birth-weight premature infants. Developmental Psychology 2003;39:730–744. [PubMed: 12859126]

Hirano K, Imbens GW, Rubin DB, Zhou X. Assessing the effect of influenza vaccine in an encouragement design with covariates. Biostatistics 2000;1:69–88. [PubMed: 12933526]

Ialongo N, Werthamer L, Brown CH, Kellam S, Wai SB. The proximal impact of two first grade preventive interventions on the early risk behaviors for later substance abuse, depression and antisocial behavior. American Journal of Community Psychology 1999;27:599–642. [PubMed: 10676542]

Imbens GW, Rubin DB. Bayesian inference for causal effects in randomized experiments with noncompliance. The Annals of Statistics 1997;25(1):305–327.

Jin H, Rubin DB. Principal Stratification for Causal Inference with Extended Partial Compliance. Journal of the American Statistical Association 2008;103(481):101–111.

Jo B. Estimation of intervention effects with noncompliance: Alternative model specifications. Journal of Educational and Behavioral Statistics 2002a;27(4):385–409.

Jo B. Statistical power in randomized intervention studies with noncompliance. Psychological Methods 2002b;7:178–193. [PubMed: 12090409]

Jo B, Asparouhov T, Muthen BO, Ialongo NS, Brown CH. Cluster randomized trials with treatment noncompliance. Psychological Methods 2008;13(1):1–18. [PubMed: 18331150]

Kam CM, Greenberg MT, Walls CT. Examining the role of implementation quality in school-based prevention using the PATHS curriculum. Prevention Science 2003;4(1):55–63. [PubMed: 12611419]

Le, HN.; Perry, DF.; Stuart, EA.; Ortiz, G. Preventing perinatal depression in pregnant at-risk Latinas: A randomized trial. 2008. Manuscript under review

Lin JY, Ten Have TR, Elliott MR. Longitudinal Nested Compliance Class Model in the Presence of Time-Varying Noncompliance. Journal of the American Statistical Association. (in press).

Little RJ, Rubin DB. Causal effects in clinical and epidemiological studies via potential outcomes: Concepts and analytical approaches. Annual Review of Public Health 2000;21:121–145.

Little RJA, Yau L. Statistical techniques for analyzing data from prevention trials: Treatment of no-shows using Rubin's causal model. Psychological Methods 1998;3:147–159.

Muñoz RF, Le HN, Ghosh Ippen C, Diaz MA, Urizar GG Jr, Soto J, Mendelson T, Delucchi K, Lieberman AF. Prevention of postpartum depression in low-income women: Development of the Mamás y Bebés/Mothers and Babies Course. Cognitive and Behavioral Practice 2007;14:70–83.

Muthén, LK.; Muthén, BO. Mplus User's Guide. Fifth Edition. Los Angeles, CA: Muthén & Muthén; 1998–2007.

O'Malley AJ, Normand ST. Likelihood methods for treatment noncompliance and subsequent nonresponse in randomized trials. Biometrics 2005;61(2):325–334. [PubMed: 16011678]

Peng Y, Little RJ, Raghunathan TE. An extended general location model for causal inferences from data subject to noncompliance and missing values. Biometrics 2004;60:598–607. [PubMed: 15339281]

R Core Development Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2007. www.r-project.org

Radloff LS. The CES-D Scale: A self-report depression scale for research in the general population. Applied Psychological Measurement 1977;1:385–401.

Robins JM, Greenland S. Identifiability and exchangeability of direct and indirect effects. Epidemiology 1992;3:143–155. [PubMed: 1576220]

Rosenbaum, PR. Observational Studies. 2. New York: Springer-Verlag; 2002.

Rubin DB. Bayesian inference for causal effects: The role of randomization. The Annals of Statistics 1978;6:34–58.

SPSS Inc. SPSS Base 16.0 for Windows. Chicago, IL: SPSS Inc; 2007.

StataCorp. Stata Statistical Software: Release 10. College Station, TX: StataCorp LP; 2007.

Werthamer-Larsson L, Kellam SG, Wheeler L. Effect of first-grade classroom environment on child shy behavior, aggressive behavior, and concentration problems. American Journal of Community Psychology 1991;19:585–602. [PubMed: 1755437]

Table 1

Comparison of baseline characteristics of participants and non-participants randomized to the MB Intervention

Baseline Characteristic	Mean Among Participants	Mean Among non-Participants	p-value of difference
Age	25.3	26.3	.24
From Central America	81%	78%	.52
Less than 1 year in US	37%	21%	.27
Married or cohabitating	73%	66%	.11
Number of children	0.94	1.12	.35
Employed	29%	44%	.20
Husband employed	64%	63%	.59
Previous psychiatric treatment	15%	2%	.35
BDI-II at Baseline	15.9	15.4	.80
Number of women	62	50	

* Significant at 10% level

Note: Participation defined as attending four or more classes. Comparison done using treatment group only.

Table 2

Comparison of baseline characteristics of participants and non-participants randomized to the Family School Partnership intervention

Baseline Characteristic	Mean Among Participants	Mean Among non-Participants	p-value of difference
Male	52%	46%	.63
Qualify for free lunch	59%	54%	.67
Parent limited by health problems	4%	10%	.15
Parent's age	32.4	34.7	.09 *
Primary caregiver male	6%	11%	.69
Non-white	80%	92%	.06 *
TOCA-R aggression rating at baseline	1.50	1.49	.96
TOCA-R shyness rating at baseline	2.17	2.40	.15
Number of students	71	71	

* $p < .10$

Note: Participation defined as completing 45 or more take-home activities. Comparison done using treatment group only.