



U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
Substance Abuse and Mental Health Services Administration
Center for Substance Abuse Prevention
www.samhsa.gov

ANALYZING DATA FROM NONRANDOMIZED GROUP STUDIES

Prepared by:
The Workplace Managed Care Cross Site Evaluation Team

Jeremy W. Bray, Ph.D.
William E. Schlenger, Ph.D.
Gary A. Zarkin, Ph.D.
RTI

Deborah Galvin, Ph.D.
Center for Substance Abuse Prevention
Substance Abuse and Mental Health Services Administration

Prepared under funding from
The Center for Substance Abuse Prevention

for the
Workplace Managed Care Steering Committee

Contact:
Jeremy W. Bray
RTI
3040 Cornwallis Road
Research Triangle Park, NC 27709-1294
(919) 541-7003
(919) 541-6683 (fax)
bray@rti.org

October 18, 2002

Revised for the
Youth Transition into the Workplace Initiative

February 18, 2004

ABSTRACT

Researchers evaluating prevention and early intervention programs must often rely on diverse study designs that assign groups to various study conditions (i.e., intervention versus control). Although the strongest designs randomly assign these groups to conditions, researchers frequently must use nonrandomized research designs in which assignments are made based on the characteristics of the groups. With nonrandomized group designs, there is little available guidance on how best to analyze the data. We provide guidance on which techniques work best under different data conditions and make recommendations to researchers as to how to choose among the various techniques. We use data from the Center for Substance Abuse Prevention's Workplace Managed Care initiative to compare the performance of the various methods commonly applied in quasi-experimental and group assignment designs.

INTRODUCTION

As policy makers demand more and better evidence on the effectiveness of specific policies or interventions that affect large numbers of individuals, researchers increasingly rely on study designs where groups of individuals are assigned to study conditions. Although these studies have been well documented in education research (e.g., 1, 2), where schools or classrooms are assigned to treatment or control conditions, they are becoming more prevalent in other areas (e.g., 3-5). Several authors have suggested analysis strategies for data where the groups are randomly assigned (1, 2, 6), but few address analysis issues related to nonrandom assignment, such as selection bias.

Frequently, policy makers and researchers want to investigate interventions in settings in which randomized samples are not practical and in which groups of individuals must be assigned to study conditions. Examples of such quasi-experimental designs, so called nonrandomized group designs (6), occur in a variety of prevention studies and workplace studies (e.g., 7-9). Although these studies are appropriately criticized for having increased threats to validity relative to experimental designs, demands by policy makers and practitioners for “best practices” and other information on how interventions work in settings that prohibit randomization frequently result in the need to use nonrandomized group designs.

Nonrandomized group designs pose two major data analysis challenges. First, they suffer from the same clustering problem that all group assignment studies face (6). If analysts do not appropriately address the clustering of individuals within groups, then standard errors may be underestimated, resulting in exaggerated statistical significance and false conclusions about the intervention’s effectiveness. Second, nonrandomized group designs suffer from the well-noted problem of bias created by nonrandom selection into the intervention and comparison conditions (10-13). By not randomly assigning groups to the study conditions, there is a greater chance of having systematic preexisting differences in background characteristics between the study and comparison groups. As with all quasi-experimental designs, failure to address the potential for selection bias can lead to misleading estimates of the intervention effect and, again, false conclusions about the intervention’s effectiveness.

In this paper, we consider the analysis of data from nonrandomized group designs. First, we provide a brief overview of the techniques commonly used to account for the clustering inherent in all group assignment designs. We also discuss the techniques used to address sample selection bias potentially created by nonrandom assignment. Next, we propose an adaptation of a method proposed by Heckman and Hotz (12) to address individual self-selection

for use in nonrandomized group designs and discuss its strengths and weaknesses and provide guidelines that researchers can use when deciding on an analysis strategy. We then demonstrate the application of these guidelines using data from a workplace substance abuse prevention/early intervention study.

NONRANDOMIZED GROUP STUDIES

The nonrandomized group design is a quasi-experimental design that assigns identifiable groups of individuals to the intervention or comparison condition in some nonrandom way (6). Study assignments are often made by the researcher based on characteristics of the groups for convenience (e.g., geographic location) or other pragmatic reasons, but perhaps equally as often the groups themselves (or some representative of the group, such as a principal or a worksite administrator) select their study condition. Regardless of the selection mechanism, the individuals within the groups are the analysis unit of interest. For example, many worksite programs are delivered to the entire worksite, and administrators at the worksite decide whether the program will be offered at their particular worksite. In such a situation, researchers attempting to assess the effect of the program on individual-level outcomes are faced with two key analysis issues: the clustering of individuals within groups and the potential for selection bias caused by the nonrandom assignment. Proper analysis of data from a nonrandomized group design requires awareness of and attention to both issues.

CLUSTERED DATA

The key feature that distinguishes a nonrandomized group design from other quasi-experimental designs is that identifiable groups of individuals, rather than the individuals themselves, are assigned to the study's treatment conditions, but the individual remains the unit of interest. Identifiable groups are groups that were not constituted at random. Examples include schools, classrooms within schools, worksites, clinics, or even whole communities. Because these groups are not constituted at random, their members usually share one or more traits in common. Typically, some of these traits, such as geographic location, socioeconomic status, or employee benefit structures, are measured in the study and therefore can be accounted for in the analysis. Because of pragmatic and other limitations, however, many more traits remain unmeasured, such as a common workplace culture or a shared work ethic, and therefore cannot be analyzed directly. The net effect of these shared traits is that an individual is more like other individuals within his or her group than individuals outside of his or her group. In other words, individuals are clustered within groups, and that clustering induces a correlation among the individuals within a group known as the intra-cluster correlation.

To see this more clearly, consider estimating the effect of an intervention using the following regression equation for some outcome Y from a group randomized study:

$$Y_{ijt} = \alpha + \mathbf{X}_{ijt}\beta + \delta d_{jt} + U_{ijt}, \quad (1)$$

where i indexes individuals, j indexes groups, and t indexes pre-intervention ($t=1$) and post-intervention ($t=2$). α is the regression intercept (which is also the conditional mean of Y), \mathbf{X}_{ijt} is a vector of observed characteristics that influence Y , and β is a vector of slopes associated with the variables in \mathbf{X}_{ijt} . Not all of the variables in \mathbf{X}_{ijt} must necessarily vary at all three levels. Some may be time constant characteristics of the individual, such as race, while others may be characteristics of the group, such as geographic location. U_{ijt} is the error term. d_{jt} is an indicator variable that equals 1 if group j was exposed to the intervention in period t and 0 otherwise. The intervention effect is captured by the coefficient on d_{jt} , δ , and reflects the effect of the group-level intervention on the individual-level outcome Y . Equation 1 is the regression equation equivalent of the ANCOVA model suggested by Reichardt (14) and adapted for the nonrandomized group design.

If U_{ijt} is independently distributed across all individuals (i), groups (j), and time periods (t), then simple ordinary least squares (OLS) regression is the appropriate estimation method. However, there is likely to be some degree of intra-cluster correlation caused by the similarities among individuals within a group. Similarly, we can posit events and conditions that make individuals within a time period similar and therefore cause clustering within a time period. We can incorporate these and other levels of clustering into our model by decomposing U_{ijt} into various components. For example, consider the following decomposition:

$$U_{ijt} = \varepsilon_i + \zeta_j + \eta_t + \mu_{jt} + v_{ijt} \quad (2)$$

ε_i is a random variable that is specific to individual i and is constant over time. It reflects traits specific to an individual that induce a correlation within the observations on a specific individual over time. Similarly, ζ_j is a random variable that is specific to group j and reflects the shared traits of the individuals within group j . It therefore captures the correlation across individuals within a group. η_t is a random variable specific to time period t and captures the correlation across all observations occurring in time period t . μ_{jt} is a random variable that is specific to group j and time period t and captures the correlation among observations within a group-time period combination. v_{ijt} is a random variable that is unique to each person and time period and therefore represents an iid random error term.

One common method of dealing with intra-cluster correlation is the sandwich variance estimator (15-17). Sandwich variance estimators are an ex-post correction to the variance-covariance matrix and have a variety of names, including Huber, White, and generalized estimating equations (for a review of the use of sandwich variance estimators, see Norton et al. [18]). Sandwich estimators are most often used to correct for clustering at the group level (ζ_j) but are increasingly being used to handle clustering at other levels. The main advantage of sandwich variance estimators is that they are easily obtained in many statistical software packages (e.g., SAS, Stata, SUDAAN). One disadvantage of sandwich variance estimators is that, as most commonly implemented, they account for only one level of clustering at a time. Another limitation that arises in an ANOVA context is that the sandwich variance estimator does not alter traditional ANOVA sums of squares and so will not correct for clustering when used in a traditional ANOVA framework.

Another common method of dealing with intra-cluster correlation is the use of random effects or mixed models (see Murray [6]). In a random effects model, the various components of the error term are modeled as independently distributed random effects. By explicitly modeling the different error components, the random effects model efficiently handles many different levels of clustering. Identification of the random effects parameters, however, is achieved primarily through the assumption that the random effects are not correlated with each other or with any other variables in the model. As we will see later, this assumption is problematic under likely and plausible circumstances and, if violated, can bias the estimated intervention effect. Random effects or mixed models are becoming more widely implemented in statistical packages. Examples include SAS's proc mixed and glimmix macro, and Stata's gllamm ado.

SELECTION BIAS

Selection bias arises when there are underlying differences in the outcome between the comparison and intervention groups that are not caused by the intervention. As discussed by Heckman and Hotz (12), there are fundamentally two types of selection processes that can cause bias: selection on observables (or measured characteristics) and selection on unobservables (or unmeasured characteristics). Selection on observables occurs when there are differences in measured characteristics between the comparison and intervention groups that are correlated with the outcome of the intervention (e.g., age, race, or gender). Selection on unobservables occurs when there are differences in unmeasured characteristics between the comparison and intervention groups that are correlated with the outcome of the intervention (e.g., motivation or innate ability). The term "unobservables" is used by Heckman and Hotz (12)

to describe any characteristic that analysts cannot explicitly control for through some measured variable or proxy. It does not necessarily imply selection on a latent construct unless that construct remains unmeasured.

To see both types of selection, consider the following model of the study assignment process:

$$d_{jt} = 1 \text{ iff } I_j = \mathbf{Z}_j\boldsymbol{\gamma} + V_j > 0 \text{ \& } t = 2 \quad (3)$$

$$d_{jt} = 0 \text{ otherwise}$$

where \mathbf{Z}_j is a vector of measured group-level characteristics that determine the group's decision to participate in the intervention, V_j is an error term, and equations 1 and 2 still describe the study outcome and its error distribution.

Equation 3 describes the outcome of the decision process that led the group to participate in the intervention and therefore determines d_{jt} . Because the decision to participate in the intervention is determined in part by the characteristics of the group (e.g., the school, classroom, or worksite), it is possible and even likely that d_{jt} is correlated with the equation 1 error term, U_{ijt} . Selection bias occurs when a correlation between d_{jt} and U_{ijt} causes estimates of δ to be biased. If the correlation between U_{ijt} and d_{jt} arises because of a correlation between \mathbf{Z}_j and U_{ijt} , then the selection is said to be on observables. If the correlation between d_{jt} and U_{ijt} arises because of a correlation between V_j and U_{ijt} , then the selection is on unobservables.

Correcting for selection on observed characteristics in a nonrandomized group study is relatively straightforward and relies on methods developed for more traditional quasi-experimental designs. The analyst simply includes controls for \mathbf{Z}_j in equation 1. Heckman and Hotz (12) and Heckman and Robb (11) discuss several ways of controlling for \mathbf{Z}_j , which they refer to as control functions. The simplest control function is to simply include \mathbf{Z}_j as a regressor in equation 1 (also referred to as the linear control function by Barnow, Cain, and Goldberger, [19]), but other commonly used control functions include the propensity score method (13), which is most useful when \mathbf{Z}_j contains a large number of variables.

Correcting for selection on unobserved characteristics is more complicated. There are two broad classes of methods designed to correct for selection on unobserved characteristics used in more traditional quasi-experimental designs: those that model the selection process and those that do not model the selection process. Approaches that model the selection

process correct for the correlation between V_j and U_{ijt} by estimating equation 3 in a way that allows or corrects for the correlation between V_j and U_{ijt} . For example, many Heckman sample selection techniques (20) assume that V_j and U_{ijt} are distributed jointly normal. Using this assumption, they correct for the selection bias by either jointly estimating equations 1 and 3 or by including an additional variable in equation 1 that captures the effects of the selection process. Instrumental variables (IV) approaches use equation 3 to predict d_{jt} as a function of variables that are not correlated with U_{ijt} and then use this predicted value in place of the actual d_{jt} when estimating equation 1 (21, 22). By construction, the predicted value is uncorrelated with U_{ijt} but highly correlated with d_{jt} and so gives an unbiased estimate of δ .

Although techniques that model the selection process can be quite effective, they also have two key limitations that often prevent analysts from using them with data from nonrandomized group studies. First, they almost universally rely on variables that appear in Z_j but not in X_{ijt} (i.e., variables that explain the selection into the intervention group but that do not influence the outcome, so called identifying instruments) to help identify the effect of the intervention. Without these variables, most techniques that model the selection process perform poorly. Unfortunately, these variables are often difficult to identify and measure. Second, techniques that estimate the selection process require enough groups (typically greater than 30 per study condition) to reliably estimate equation 3. Because few nonrandomized group studies meet these data requirements, we do not discuss these techniques further but refer interested readers to the previously referenced literature.

Techniques that do not model the selection process correct for selection bias by relying on assumptions about the nature of the unobserved factors causing the selection bias. Most commonly, they assume that U_{ijt} and V_j share a common component that causes a correlation between the two. For example, suppose U_{ijt} has the form given in equation 2, and V_j has the following form:

$$V_j = \zeta_j + v_j, \quad (4)$$

where v_j is a random variable that is unique to each group and therefore represents an iid random error term.

Under this assumption, the selection bias is caused by ζ_j and can be eliminated by controlling for ζ_j in equation 1. One way of controlling for ζ_j is by using a differences-in-differences (DD) estimator (also referred to as gain score analysis). DD estimators eliminate ζ_j from U_{ijt} by subtracting the baseline value of Y_{jt} from the follow-up value, creating a difference

value. Because ζ_j is assumed to be constant over time, it is the same in both the baseline and follow-up values and is therefore eliminated by the differencing. The average difference in the intervention group is then compared to the average difference in the comparison group to determine the intervention effect.

DD estimators are often implemented in a linear regression framework by including indicator variables for the study condition and for the post-treatment period, resulting in the following regression equation:

$$Y_{ijt} = \alpha_0 + \mathbf{X}_{ijt}\beta + \gamma_1\text{COND}_j + \gamma_2\text{POST}_t + \delta d_{jt} + U_{ijt}, \quad (5)$$

where COND_j is an indicator variable that equals 1 if group j is in the intervention condition and 0 otherwise, POST_t is an indicator variable that equals 1 if the measurement is from the follow-up and 0 otherwise, and γ_1 and γ_2 are coefficients to be estimated. The intervention effect is still captured by δ . When using DD estimators with nonlinear models, such as logistic regression, equation 5 is still appropriate. However, using the difference between the follow-up and baseline observations (the gain score) is not appropriate when using nonlinear models.

A similar non-model based approach is an adaptation of the individual-level fixed effects technique recommended by several authors (12, 23, 24)—the use of group-level fixed effects. This method estimates ζ_j by including a set of group-specific indicator variables, which allows a correlation between ζ_j and d_{jt} . The associated parameters are often called fixed effects and are identified by the variation across the groups (the between variation). Because all variation between the groups is captured by the fixed effects, this method relies solely on the variation within groups to identify the treatment effect. Because the study conditions are assigned at the group level (i.e., groups are nested within study conditions), the main study condition effect (COND_j in equation 5) cannot be separately identified from the group effects and so cannot be included in the model. The intervention effect is identified using the variation from the pre- to post-treatment observations within a group and so can be estimated if there are repeated observations on groups.

For linear models, fixed effects and DD methods are numerically identical and produce exactly the same estimate of the intervention effect, δ , if the data are balanced (i.e., the number of observations is the same in each time period) and if a post-treatment indicator is included in the fixed effects model. Fixed effects in nonlinear models, such as logistic regression, will produce similar but not identical treatment effect estimates as equation 5 and require special estimation methods if the number of observations per group is small (i.e., less than 30). For a

discussion of these issues, see Hsiao (23) or Baltagi (24). Other non-model based approaches include the random growth model (12), which assumes that U_{ijt} and V_j contain a shared component that changes linearly over time.

Because fixed effects methods control for all variation between the groups in the outcome, they correct for clustering within groups as well as for selection that results from time-invariant group-level characteristics. When used to correct for clustering, some authors have criticized the fixed effect method as overstating the true statistical significance of the intervention effect (i.e., inflated Type I error rates), which leads to invalid inferences about the true effect of the intervention (25, 26), but this problem only occurs in certain situations.

The first situation occurs in a group randomized design without repeated measures. Because the fixed effect identifies the intervention effect from the within-group variation only, it cannot identify a unique intervention effect if there is no within-group variation in the intervention condition. Some estimation techniques (e.g., traditional ANOVA) will provide estimates of an intervention effect in this case (26), but these estimates are fundamentally unidentified as an intervention effect and therefore yield invalid inferences about the true intervention effect. The second situation is when fixed effects are used to adjust for only one level of clustering, leaving other levels of clustering unaddressed. This problem arises when investigators use fixed effects to control for group-level clustering without addressing other possible sources of clustering, such as group by time clustering (i.e., μ_{jt} in equation 2).

Techniques such as fixed effects models that do not model the selection process have at least two limitations worth noting. First, these techniques may limit the external validity of the parameter estimates. Because most techniques that do not model the selection process condition on the analysis sample in some way, they potentially limit the researcher's ability to generalize the results beyond the analysis sample. When these techniques are used, the results can in principle be generalized to the full population only to the extent that the theory or logic model relating the intervention to the outcome is correct. If δ is a true theoretical parameter, then any unbiased estimate of it is generalizable, but if equation 1 is only loosely based on theory, then the external validity of estimates from techniques that do not model the selection process may be greatly limited.

Another limitation is that these techniques may only partially handle selection bias. For example, group-level fixed effects techniques only control for selection bias that is caused by unobserved group-level factors that do not vary over time. Although other techniques are available that relax this constraint, all non-model based approaches to dealing with selection

bias rely on assumptions about the cause of the selection bias. If these assumptions are incorrect, or if they only capture some of the factors that may cause selection bias, then techniques that do not model the selection process may yield misleading results.

SELECTING THE APPROPRIATE TECHNIQUE

We have presented several estimation techniques that analysts might use to estimate intervention effects using data from a nonrandomized group study. These range from the naïve linear model represented by equation 1, to equation 1 with clustering corrections, to the DD model presented in equation 5, to the use of group-level fixed effects. Given such an array of possible analysis techniques, how should an analyst choose among them?

First, the analyst should write down the regression equation that arises from the theory or logic model that relates the intervention to the outcome (i.e., equation 1). Next, the analyst should add an error term for every level of identifiable clustering that occurs in the data (i.e., equation 2). The analyst should then estimate this model using a mixed or random effects model to control for each of the clustering terms. This model, which assumes no correlation between clustering terms and the intervention indicator, serves as the base model against which to compare estimates from models that control for selection bias.

Next, the analyst should allow for a correlation between the error terms and the intervention indicator by estimating a DD model (i.e., equation 5) using a mixed model to control for the error terms previously identified. The results of this model can then be compared to those from the random effects base model previously estimated using a Hausman test (27). The Hausman test statistic for a significant difference between the two estimates is calculated as

$$z = (\delta_{DD} - \delta_{RE}) / [\sqrt{(\text{SE}(\delta_{DD})^2 - \text{SE}(\delta_{RE})^2)}], \quad (6)$$

where δ_{RE} is the estimate of the intervention effect from the base random effects model, δ_{DD} is the estimate of the intervention effect from the DD model, $\text{SE}(\delta_{RE})$ is the standard error of the intervention effect from the base random effects model, and $\text{SE}(\delta_{DD})$ is the standard error of the intervention effect from the DD model. The test statistic z is distributed standard normal, and so the difference in the estimates is significant if z exceeds standard critical values (i.e., 1.96 for a two-tailed significance level of 0.05). The Hausman test is a low power test, however, and so significance levels of 0.10 or even 0.15 should be considered by researchers when making inferences based on the results of the test. To test multiple coefficients simultaneously, as in a

study with more than one intervention condition, use a vector of coefficients and the variance-covariance matrix to compute a χ^2 test statistic (see Greene [27]).

If there is no significant difference between the two estimates, then the base random effects model is preferred because it yields more precise estimates of the intervention effect. If the DD estimate is significantly different from the base random effects model estimate, then the random effects assumption of no correlation between the error terms and the intervention indicator is probably violated. Thus, the DD estimate is preferred.

Next, the analyst should estimate a group-level fixed effects model by including indicator variables for the groups and a pre-post indicator in equation 1, while still estimating equation 1 with a mixed model to account for all clustering other than group-level clustering. The resulting estimate of the intervention effect should then be compared to the DD model estimate using a Hausman test calculated as follows:

$$z = (\delta_{FE} - \delta_{DD}) / [\sqrt{(SE(\delta_{FE})^2 - SE(\delta_{DD})^2)}], \quad (7)$$

where δ_{FE} is the estimate of the intervention effect from the fixed effects model, $SE(\delta_{FE})$ is the standard error of the intervention effect from the fixed effects model, and all other terms are as defined previously. If there is no significant difference between the estimates, then the equation DD is preferred, again because it yields more a precise estimate of the intervention effect. If there is a significant difference, then the DD model may not have fully corrected the selection bias and the fixed effects estimate is preferred.

Importantly, the analyst should estimate and test all models before deciding on a final estimate of the intervention effect. All models provide information and all make assumptions that may be violated. The analyst should estimate all models and consider all available information when making inferences about the effectiveness of the intervention.

EMPIRICAL EXAMPLE

Data for this example come from the Workplace Managed Care (WMC) Program. The WMC Program, funded by SAMHSA's Center for Substance Abuse Prevention (CSAP), was a 3-year, multiprotocol, multipopulation cooperative agreement program designed to generate a broad understanding of the nature and scope of substance abuse prevention and early intervention efforts of workplaces in collaboration with their health care providers, employee assistance programs, health/wellness programs, human resources, unions, and security. The WMC Program also intended to increase understanding of how these programs function for a

variety of populations of employees and their families within a variety of contexts. The WMC Program began in September 1997, with the award of nine cooperative agreements and a Coordinating Center contract. The participating grantees and their collaborating worksites studied a variety of existing prevention/early intervention strategies targeted toward reducing the incidence and prevalence of alcohol and drug use among employees and their families. The prevention/early intervention strategies included health risk assessments, enhanced drug-free workplace programs, drug testing, employee assistance programs, health wellness/promotion, peer interventions, and parent training.

Because the interventions studied were within existing workplace environments, randomization of study groups was impractical for most of the grantees. Furthermore, many of the prevention programs were implemented at the worksite level, not at the individual level. Thus, most of the nine grantees had nonrandomized group trial designs.

To illustrate the analysis methods described above, we use data from one of the nine WMC grantees and its participating corporate partner. The analyses presented in this paper were performed to illustrate the methods described above and should not be interpreted as definitive estimates of the effect of the intervention being examined. In particular, we posit no specific logic model linking the intervention to the outcome. Interested readers are referred to Blank et al. (28) for a more detailed analysis of the example intervention.

The selected firm is a manufacturing company specializing in the production of a wide variety of engineered products. The company employs approximately 1,300 individuals in sites located in seven states. The WMC grantee evaluated the effects of varying rates of random drug testing on a variety of substance abuse and workplace outcomes. The grantee planned to implement random drug testing at the various intervention sites at annual rates of 100, 200, and 400 percent (i.e., annual, semiannual, or quarterly testing of all employees), but due to business and environmental factors beyond the grantee's control, the actual rate of drug testing in each of the participating worksites was determined by worksite administrators and was lower than intended.

To evaluate the effects of drug testing on various outcomes, surveys were conducted in eight study worksites. All employees in each worksite were asked to complete a short survey that collected data on basic demographics and perceptions about drug and alcohol use. These anonymous employee surveys were administered in two waves approximately 1 year apart, and the annual rate of drug testing in the year prior to the survey was obtained from administrative records. Worksite-level survey response rates ranged between 80 percent and 95 percent.

The workplace outcome analyzed in this study is whether or not the respondent thinks drug use is a problem in his/her plant or office. The demographic covariates included in the analysis are age, gender, and race. For the purposes of this analysis, we have created two measures of the intervention. The first is a worksite-level intervention indicator that equals 1 if the site increased the rate of drug testing from the first survey wave to the second. The second intervention measure is the actual continuous annual drug testing rate in each site. Because the surveys were anonymous, individual employees cannot be tracked from one survey wave to the next, and the two survey waves are treated as independent cross sections. The analysis data file contains 1,039 observations.

METHODS

We begin by estimating the following model with no clustering or selection corrections:

$$\text{Prob}(Y_{ijt} = 1) = f(\alpha + \mathbf{X}_{ijt}\beta + \delta d_{jt} + U_{ijt}), \quad (8)$$

where Y_{ijt} is an indicator variable that equals 1 if respondent i in worksite j reported believing that drug use was a problem in his or her worksite at wave t . \mathbf{X}_{ijt} is a vector of demographic variables that includes age, gender, and race. To demonstrate the various methods described above, we estimate two variants of equation 8: one with the dichotomous treatment indicator and one with the continuous dose variable described above.

In all models, equation 8 is estimated as a logit model. For each intervention measure, we estimate equation 8 five times (for a total of 10 estimations). First, we estimate equation 8 as an ordinary logit model with no corrections for either clustering or sample selection—the logit analog of equation 1. Second, we estimate it using sandwich standard errors to correct for clustering at the worksite level. Third, we estimate it as a random effects logit model in which we include a worksite and a worksite by time random effect to control for both levels of clustering simultaneously. Fourth, we estimate it as a DD logit model including the condition and time main effects as fixed effects and including a worksite and a worksite by time random effect—the logit analog of equation 5. Finally, we estimate it as a logit model with worksite and time fixed effects and with a worksite by time random effect.

RESULTS

Table 1 presents the means and standard deviations of the variables used in the analysis by survey wave. The dependent variable is an indicator for whether the individual thinks illegal drug use is a problem at the worksite. In wave 1, just over 17 percent of

respondents thought drug use was a problem. In wave 2, that percentage dropped slightly to just over 16 percent of respondents. In wave 2, approximately 44 percent of respondents worked in a worksite that increased the drug testing rate from wave 1 to wave 2. The average annual rate of drug testing faced by respondents was 92 percent in wave 1 and 129 percent in wave 2. An annual rate of greater than 100 percent indicates that, on average, every employee was tested at least once in the year and some employees were tested more than once.

The demographic characteristics of the worksites remained relatively stable over the two survey waves. Not surprisingly, the worksite populations became slightly older, with the percentage of the population in the 18 to 25 and 26 to 35 year old categories declining from wave 1 to wave 2, and the percentage in the 36 to 50 and over 50 categories increasing. The education level of the population dropped slightly, with a lower percentage of respondents having completed college or a trade or technical school in wave 2 than in wave 1. Finally, the prevalence of union membership increased substantially, from just under 40 percent in wave 1 to just over 50 percent in wave 2.

Table 2 presents results for the models using a dichotomous intervention indicator. The first column presents results from the ordinary logit model; the second column presents results from a logit with sandwich estimator standard errors; the third column presents results from a model that includes worksite and worksite by time random effects; the fourth column presents results from the DD mixed logit model; and the last column presents results from a model that includes an indicator for the survey wave, worksite-level fixed effects, and worksite by time random effects. Results for model 1 were obtained using SAS proc logistic, and all other results were obtained using the SAS glimmix macro.

Looking first at the estimated intervention effect from column 1, we see that the ordinary logit model finds a significant intervention effect of 0.575 (OR = 1.78), suggesting that increasing the annual drug testing rate increases employees' likelihood of perceiving a drug problem at the worksite. Accounting for clustering within worksites using the sandwich variance estimator makes this same effect insignificant by increasing its standard error. Recall that the sandwich variance estimator is an ex-post correction that does not affect point estimates. When worksite and worksite by time random effects are included, the estimated intervention effect is -0.927 (OR = 0.40) and insignificant. Note that the standard error of the intervention effect in column 3 is larger than that in column 2. This is because in column 3 we have accounted for clustering at two levels (worksite and worksite by time), but in column 2 we account for only one level of clustering (worksite). Also note the change in sign of the estimated intervention effect.

If the assumption of no correlation between the random effects and the intervention indicator in column 3 was correct, then the estimated intervention effect from columns 1 and 2 should be approximately the same as the effect in column 3. Thus, the sign change suggests that the assumptions of the random effect model are violated.

Column 4 includes the condition and time main effects found in the DD model. Here we find a significant intervention effect of -1.692 (OR = 0.18). A Hausman test shows that the estimated intervention effect from column 3 is marginally significantly different from that in column 4 ($z = -1.35$, $p = 0.09$). Finally, column 5 replaces the intervention site main effect with a set of worksite fixed effects, which are not presented in Table 2 but are available upon request. The model in column 5 yields a significant intervention effect of -1.793 (OR = 0.17), which is approximately the same as that found in column 4. Note, however, that the standard error of the intervention effect in column 5 is larger than that in column 4. A Hausman test reveals that the difference in the two estimates is not significant ($z = -0.31$, $p = 0.38$), and so the column 4 estimate is preferred.

Table 3 presents results from the models that use the continuous drug testing rate as the measure of the intervention. Columns 1 through 5 use the same corrections for clustering and selection on unobservables as their counterparts in Table 2. We see that the ordinary logit yields a significant effect of the drug testing rate of 0.593 (OR = 1.81). Correcting for clustering on the worksite using the sandwich variance estimator increases the standard error somewhat, but does not make the estimated effect insignificant. As in Table 2, including worksite and worksite by time random effects causes our point estimate to become negative, but it is now insignificant. Including the design main effects to control for selection on unobservables via a DD model increases the magnitude of the estimated effect (i.e., it becomes more negative) to -0.777 (OR = 0.46), but the effect remains insignificant. A Hausman test shows that the difference between the column 3 and the column 4 estimates is significant ($z = -2.09$, $p = 0.02$), suggesting that the column 4 estimate is preferred over that in column 3. Including worksite-level fixed effects as in column 5, however, causes the estimated effect to become significant at -1.308 (OR = 0.27). As in Table 2, estimates for the worksite-level fixed effects are not presented but are available upon request. A Hausman test shows that the difference between the column 4 and the column 5 estimates is insignificant at the 0.10 level but is significant at the 0.15 level ($z = -1.23$, $p = 0.11$). Although not a definitive rejection of the column 4 estimate, the low power of the Hausman test and the relatively substantial change in the magnitude of the

estimated effect suggest that the column 5 estimates should be considered when making inferences about the estimated intervention effect.

DISCUSSION

The demand for the study of interventions that are implemented in “real world” settings has resulted in more nonrandomized group studies being performed. A nonrandomized group study is a quasi-experimental study in which identifiable groups of individuals are assigned to the intervention and comparison groups in a systematic way. These designs have two major analysis challenges: clustered data and potential selection bias. Previous literature has identified a variety of methods for dealing with clustered data, including ex post corrections to the estimated variances, random effects models, and fixed effects models. Several options are also available to correct for sample selection bias. If the selection is on observed characteristics, then researchers can simply include measures of the observed characteristics in their analyses. If selection is on unobserved characteristics, then more complicated corrections are needed, such as the Heckman model or IV techniques that estimate the selection process, or DD or fixed effects methods that do not estimate the selection process. Unfortunately, there is little guidance as to analysis methods for researchers who are analyzing data from nonrandomized group trials.

We examined various methods for the analysis of data from nonrandomized group studies. Many of the methods for addressing clustered data can be readily applied to nonrandomized group studies. As with any group assignment study, however, researchers should use methods that address all levels of clustering in the data. Similarly, many of the methods for correcting for selection on unobserved characteristics can also be applied to nonrandomized group trial data. Techniques that estimate the selection process, however, require a sufficient number of groups (typically greater than 30) to model the group-level decision to participate in the intervention. Because many nonrandomized group trials have relatively few groups, these approaches may not be appropriate in many cases. Techniques that do not estimate the selection process, however, can be used whenever the researcher has access to both pre- and post-treatment data.

We used data from SAMHSA’s WMC Program to explore the estimated intervention effect using various estimation methods. We found that both clustering and sample selection corrections have substantial impacts on quantitative and qualitative conclusions about the effects of an intervention. In particular, we found that the estimated intervention effect can

switch from positive and significant to negative and significant when both clustering and sample selection are addressed.

Based on these analyses, we propose the following recommendations to researchers analyzing nonrandomized group trial data. First, use random effects to account for all levels of identifiable clustering. Next, estimate DD models with random effects to account for clustering, and compare the results to the simple random effects model using a Hausman test. If the DD model does not yield significantly different estimates, then the simple random effects model is preferred. If the DD model does yield significantly different results, then it is preferred. Next, estimate a group-level fixed effects model (controlling for clustering at any level other than the group with random effects) and compare the results to the DD results using a Hausman test. If the fixed effects model does not yield significantly different estimates from the DD model, then the DD model is preferred. Otherwise, the fixed effects model is preferred. Based on our empirical results, group-level fixed effects appear to be especially important in the presence of a continuous measure of the intervention, such as a continuous dose variable. For both continuous and discrete outcomes, all of our proposed analyses can be easily performed in standard statistical software packages, such as SAS or Stata.

Although the methods proposed here will greatly improve researchers' ability to draw inferences from nonrandomized group trial data, as with any quasi-experimental design, a causal interpretation must ultimately depend on the validity of the underlying theory or logic model that links the intervention to the outcome. No empirical strategy can alleviate concerns about the plausibility of an estimated intervention effect that arise from doubts about the underlying theory. However, strong empirical methods can help to eliminate competing explanations for why an effect might be found and therefore bolster theoretical arguments about the causal nature of any such intervention effect.

REFERENCES

1. Bryk AS, Raudenbush SW. Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model. *American Journal of Education* 1988; 97(1):65-108.
2. Goldstein H. Multilevel covariance component models. *Biometrika* 1987; 74(2):430-431.
3. Farquhar, JW, Fortmann SP, Flora JA, et al. Effects of communitywide education on cardiovascular disease risk factors: The Stanford five-city project. *JAMA* 1990; 264(3):359-365.
4. Luepker RV. Community trials. *Prev Med* 1994; 23:602-605.
5. Carleton RA, Lasater TM, Assaf AR, et al. The Pawtucket heart health program: Community Changes in cardiovascular risk factors and projected disease risk. *AJPH* 1995; 85(6):777-785.
6. Murray DM. Design and analysis of group randomized trials. New York (NY): Oxford University Press; 1998.
7. Zarkin GA, Bray JW, Karuntzos GT, Demiralp B. The effect of an enhanced employee assistance program (EAP) intervention on EAP utilization. *Journal of Studies on Alcohol* 2001; 62(3):351-358.
8. Lapham SC, Chang I, Gregory C. Substance abuse intervention for health care workers: A preliminary report. *J Behav Health Serv Res* 2000; 27:131-143.
9. Ames GM, Grube JW, Moore RS. Social control and workplace drinking norms: A comparison of two organizational cultures. *J Stud Alcohol* 2000; 61:203-219.
10. Cook TD, Campbell DT. Quasi-experimentation: Design & analysis issues for field settings. Boston (MA): Houghton Mifflin Company; 1979.
11. Heckman JJ, Robb R. Alternative methods for evaluating the impact of interventions: An overview. *Journal of Economics* 1985; 30:239-267.

12. Heckman JJ, Hotz VJ. Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of manpower training. *Journal of the American Statistical Association* 1989; 84(408):862-880.
13. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 1984; 79:516-524
14. Reichardt CS. The statistical analysis of data from nonequivalent group designs. In: Cook TD, Campbell DT, editors. *Quasi-experimentation: Design & analysis issues for field settings*. Boston: Houghton Mifflin Company; 1979. p. 147-205.
15. Huber PJ. The behavior of maximum likelihood estimates under nonstandard conditions. In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* 1967; 1:221-233.
16. White H. A heteroskedasticity-consistent covariance matrix and a direct test for heteroskedasticity. *Econometrica* 1980; 48:817-838.
17. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; 73:13-22.
18. Norton EC, Bieler GS, Ennett ST, Zarkin GA. Analysis of prevention program effectiveness with clustered data using generalized estimating equations. *J Consult Clin Psychol* 1996; 64(5):919-926.
19. Barnow BS, Cain GG, Goldberger AS. Issues in the analysis of selectivity bias. In: Stromsdorfer E, Farkas G, editors. *Evaluation studies*, vol. 5. San Francisco: Sage; 1980.
20. Heckman J. Sample selection bias as a specification error. *Econometrica* 1979; 47:153-161.
21. Newhouse JP, McClellan M. Econometrics in outcomes research: The use of instrumental variables. *Annu Rev Public Health* 1998; 19:17-34.

22. Heckman J. Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations. *The Journal of Human Resources* 1997; 32(3):441-462.
23. Hsiao C. *Analysis of panel data*. Cambridge (MA): Cambridge University Press; 1986.
24. Baltagi BH. *Econometric analysis of panel data*. Chichester (UK): John Wiley & Sons; 1995.
25. Murray DM. Design and analysis of group-randomized trials: A review of recent developments. *AEP* 1997; 7(57):S69-S77.
26. Zucker DM. An analysis of variance pitfall: The fixed effects analysis in a nested design. *Educa Psychol Meas* 1990; 50:731-738.
27. Greene WH. *Econometric analysis*, third edition. New Jersey: Prentice-Hall; 1997.
28. Blank D, Walsh JM, Cangianelli L. Effect of drug-testing and prevention education strategies on employee behavior and attitudes at a manufacturing company. Working paper. Bethesda, MD: The Walsh Group, P.A.; June 11, 2002.

Table 1: Means of Analysis Variables

	Wave 1 (N=508)	Wave2 (N=496)
Thinks illegal drug use is a problem at the worksite	0.171 (0.377)	0.161 (0.368)
Worksite increased rate of drug testing	—	0.435 (0.496)
Continuous drug testing rate	0.919 (0.744)	1.292 (1.084)
Age		
18 to 25	0.079 (0.270)	0.071 (0.256)
26 to 35	0.226 (0.419)	0.183 (0.387)
36 to 50	0.392 (0.489)	0.419 (0.494)
Over 50	0.303 (0.460)	0.327 (0.469)
Education		
Less than high school	0.063 (0.243)	0.071 (0.256)
Completed high school	0.579 (0.494)	0.655 (0.476)
Completed college	0.232 (0.423)	0.192 (0.394)
Trade or technical school	0.126 (0.332)	0.083 (0.276)
Male	0.762 (0.426)	0.748 (0.435)
Married	0.659 (0.474)	0.653 (0.476)
Minority	0.142 (0.349)	0.157 (0.364)
Union member	0.398 (0.490)	0.506 (0.500)

Note: Standard deviations in parentheses.

Table 2: Logit Estimates for Dichotomous Intervention Measure

	Ordinary Logit	Clustering Adjustment		Selection Corrections	
		Sandwich Variance	Random Effects	DD	Fixed Effects
Increased drug testing rate	0.575*** (0.199)	0.575 (0.389)	-0.927 (0.572)	-1.692** (0.806)	-1.793** (0.868)
Age (18 to 25 age group is reference category)					
26 to 35	-0.078 (0.343)	-0.078 (0.416)	0.287 (0.355)	0.277 (0.358)	0.316 (0.363)
36 to 50	-0.507 (0.336)	-0.507 (0.454)	0.163 (0.348)	0.172 (0.352)	0.211 (0.356)
Over 50	-0.207 (0.343)	-0.207 (0.310)	0.658* (0.360)	0.689* (0.364)	0.721* (0.368)
Education (Completed high school is reference category)					
Less than high school	-0.190 (0.351)	-0.190 (0.278)	-0.079 (0.368)	-0.092 (0.370)	-0.079 (0.376)
Completed college	-0.377 (0.249)	-0.377 (0.353)	-0.106 (0.267)	-0.072 (0.268)	-0.062 (0.274)
Trade or technical school	-0.519 (0.330)	-0.519 (0.401)	-0.279 (0.339)	-0.272 (0.343)	-0.266 (0.346)
Male	0.265 (0.220)	0.265 (0.436)	0.083 (0.234)	0.068 (0.236)	0.051 (0.240)
Married	0.375* (0.199)	0.375* (0.193)	0.364* (0.203)	0.374* (0.205)	0.369* (0.207)
Minority	0.476** (0.230)	0.476** (0.241)	0.947*** (0.271)	0.981*** (0.271)	0.988*** (0.281)
Union member	-0.137 (0.190)	-0.137 (0.331)	-0.546** (0.239)	-0.504** (0.239)	-0.565** (0.246)
Intervention worksite	—	—	—	2.345*** (0.783)	—
Wave 2 survey	—	—	—	0.474 (0.532)	0.487 (0.576)
Intercept	-1.832*** (0.353)	-1.832** (0.749)	-2.601*** (0.613)	-3.588*** (0.637)	-0.366 (0.675)

Note: Standard errors in parentheses.

* $p < 0.10$

** $p < 0.05$

*** $p < 0.01$

Table 3: Logit Estimates for Continuous Intervention Measure

	Ordinary Logit	Clustering Adjustment		Selection Corrections	
		Sandwich Variance	Random Effects	DD	Fixed Effects
Continuous drug testing rate	0.593*** (0.100)	0.593*** (0.158)	-0.411 (0.442)	-0.777 (0.476)	-1.308** (0.641)
Age (18 to 25 age group is reference category)					
26 to 35	-0.043 (0.346)	-0.043 (0.379)	0.290 (0.355)	0.293 (0.358)	0.315 (0.363)
36 to 50	-0.343 (0.339)	-0.343 (0.397)	0.162 (0.348)	0.185 (0.351)	0.209 (0.356)
Over 50	0.025 (0.347)	0.025 (0.221)	0.652* (0.359)	0.693* (0.364)	0.718* (0.368)
Education (Completed high school is reference category)					
Less than high school	-0.306 (0.354)	-0.306 (0.263)	-0.065 (0.367)	-0.063 (0.370)	-0.062 (0.375)
Completed college	-0.196 (0.253)	-0.196 (0.391)	-0.102 (0.267)	-0.078 (0.269)	-0.057 (0.274)
Trade or technical school	-0.380 (0.335)	-0.380 (0.442)	-0.268 (0.339)	-0.262 (0.342)	-0.259 (0.345)
Male	0.260 (0.223)	0.260 (0.428)	0.088 (0.234)	0.071 (0.236)	0.053 (0.239)
Married	0.368* (0.201)	0.368* (0.197)	0.363* (0.203)	0.370* (0.204)	0.370* (0.207)
Minority	0.419* (0.232)	0.420*** (0.147)	0.947*** (0.271)	0.991*** (0.273)	0.984*** (0.279)
Union member	-0.080 (0.191)	-0.080 (0.213)	-0.561** (0.239)	-0.537** (0.240)	-0.564** (0.245)
Intervention worksite	—	—	—	2.466** (0.987)	—
Wave 2 survey	—	—	—	-0.064 (0.402)	0.039 (0.446)
Intercept	-2.640*** (0.394)	-2.640*** (0.550)	-2.346*** (0.782)	-2.892*** (0.697)	1.798 (1.437)

Note: Standard errors in parentheses.

* $p < 0.10$

** $p < 0.05$

*** $p < 0.01$

ACKNOWLEDGMENTS

Funding for this work was provided by the Office of Workplace Programs, Center for Substance Abuse Prevention (CSAP) through a subcontract with the CDM Group. The authors would like to thank the Workplace Managed Care Steering Committee and Publications Subcommittee for helpful suggestions. The authors also thank Susan Murchie for editorial assistance and Erica Brody and Christian Evensen for research assistance. The opinions expressed are not necessarily those of RTI or of CSAP.