

Using Multilevel Modeling in the Evaluation of Community-Based Treatment Programs

David Livert, David Rindskopf
CUNY Graduate Center

Leonard Saxe
Brandeis University

Mike Stirratt
CUNY Graduate Center

Health and social intervention programs targeted at neighborhoods, cities, and counties are becoming increasingly common (cf. Winick & Larson, 1997) and there is a concomitant need for systematic evaluation (Connell, Kubisch, Schorr & Weiss, 1995; O'Connor, 1995). Interventions across multiple sites require a different – multilevel – approach to testing hypotheses. Although multilevel models are increasingly being utilized (cf. Rindskopf & Saxe, 1998), actual application of such models to program assessment is complex and there are few examples. The goal of this article is to describe the application of multilevel statistical models to the evaluation of a multi-site community-based intervention and to explicate key issues in the design, analysis, and presentation of results.

Since 1994, we have conducted the evaluation of *Fighting Back* (see Saxe, Reber, Hallfors, Kadushin, Jones, Rindskopf & Beveridge, 1997), a national demonstration program designed to test the feasibility of reducing substance abuse through coordination and expansion of community efforts (Jellinek & Hearn, 1991; Spickard, Dixon, & Sarver, 1994). Funded by the Robert Wood Johnson Foundation, *Fighting Back* has been implemented in 14 communities in 11 states. Although the program concept is shared in common across these communities, each *Fighting Back* program reflects the history and organizational structure unique to that site. The flexibility provided by

This study is supported by a grant from the Robert Wood Johnson Foundation. The authors would like to thank Ellie Buteau and Ann Cunningham for their editorial comments.

Correspondence concerning this article should be addressed to David Livert at the Department of Social/Personality Psychology, CUNY Graduate Center, 365 Fifth Avenue, New York, New York 10016.

multilevel models both to measure overall program outcomes and to explore differences in program implementation between communities seemed a natural fit to the research design (Kenny, 1996; Osgood & Smith, 1995).

Below, the *Fighting Back* program and the appropriateness of a multilevel model approach are reviewed. Then the evaluation design is described, focusing on the use of multiple comparison sites for each treatment community and issues of generalizability and statistical power. The analytic model used to test overall program effects is then described and the challenges of presenting results of this type of analysis are discussed.

Community-Based Drug Abuse Prevention Programs

Traditional interdiction and supply-reduction efforts, which seek to reduce drug abuse through incarceration and other law enforcement activity, are the primary forms of drug control policy in the United States (Saxe & Winick, in press). Community-based strategies – which provide alternatives to traditional interdiction approaches – have risen in popularity during the past 15 years. These strategies both focus on the reduction of illicit drug supplies as well as the reduction of demand for such drugs with the goal of reducing the harm to communities associated with substance abuse (Aguirre-Molina & Gorman, 1996; Kaftarian & Hansen, 1994; Pentz, Dwyer, MacKinnon, Flay, Hansen, Wang & Johnson, 1989; Perry, Williams, Veblen-Mortenson, Toomey, Komro, Anstine & McGovern, 1996).

Although community-based programs have become a key element of drug control policy (cf. ONCDP, 1999), the effectiveness of community-based programs has been difficult to establish. In part, this is due to the methodological challenges inherent in community research (Connell et al., 1995; Rindskopf & Saxe, 1998). The effective sample size is usually constrained by the number of communities, resulting in low power and precision. The program effect is usually small, because community-based interventions deliver their “dosage” broadly to a large number of potential recipients, instead of concentrating the “dosage” on a small cohort of participants as in clinical trials. Moreover, implementations are not exactly the same across treatment sites. Trade-offs and local contexts may require variation in implementation efforts which can reduce generalizability of outcome data (Rossi, Freeman, & Lipsey, 1999). For example, an evaluation may be designed to test an ostensibly community-wide intervention across all treatment sites over time, yet an individual site may narrow the focus of its efforts over the course of the program. Even with strict fidelity to program design, implementers may find that each community offers an idiosyncratic array of challenges that requires molding of the intervention to specific

needs. An accompanying challenge concerns outcomes: as the variation in implementation across sites increases, the number of indicators that can be used to evaluate program success which are common to all sites may be limited.

Fighting Back

In 1989 the Robert Wood Johnson Foundation (RWJF) initiated a program to support “intensive, community-wide initiatives to reduce demand for illegal drugs and alcohol” (Jellinek & Hearn, 1991, p. 79). The initiative was designed to demonstrate whether the use and associated harm from illicit substances could be reduced in a community through the consolidation of existing program, activities, and other resources. *Fighting Back’s* (FB) targets were mid-sized communities, both urban and rural, consisting of between 100,000 and 250,000 residents. RWJF funded 15 communities to receive planning grants. They included areas ranging from large portions of mid-sized cities (e.g., Milwaukee) to relatively small areas within a city (Washington, DC) to a large rural area (Northwest New Mexico). All but one of these communities later (in 1992) received implementation grants. Each FB community was charged with achieving the general goals of the program: to achieve a measurable reductions in alcohol and other drug (AOD) use, AOD use-related harm, and injuries, deaths and crime related to AOD use. Individual communities varied dramatically (i.e., Newark vs. Santa Barbara) in the amount of resources, the existing institutional structures, and treatment services available to pursue these goals. Individual FB programs were given considerable latitude with intervention activities (see Hallfors, Reber, Saxe, & Watson, 1998). Although the programs share organizational structure, institutional affiliations vary from site to site.

Design Issues

When community-based interventions are situated in more than one geographic location, evaluators must grapple with issues arising from a nested data structure: individual units (e.g., people, organizations) comprise one level and location comprises a second level. In the FB evaluation, multiple treatment communities are paired with two or more communities, analogous to case-control designs in medical research where there are multiple comparisons for each case. In contrast, however, communities are the units that receive a treatment or control, not individuals. Moreover, treatment and control conditions were not randomly assigned (see below). Control communities were selected from the same state (Northern and

Southern California and the District of Columbia were considered separate “states”); thus “state” became a third level of data (See Figure 1).

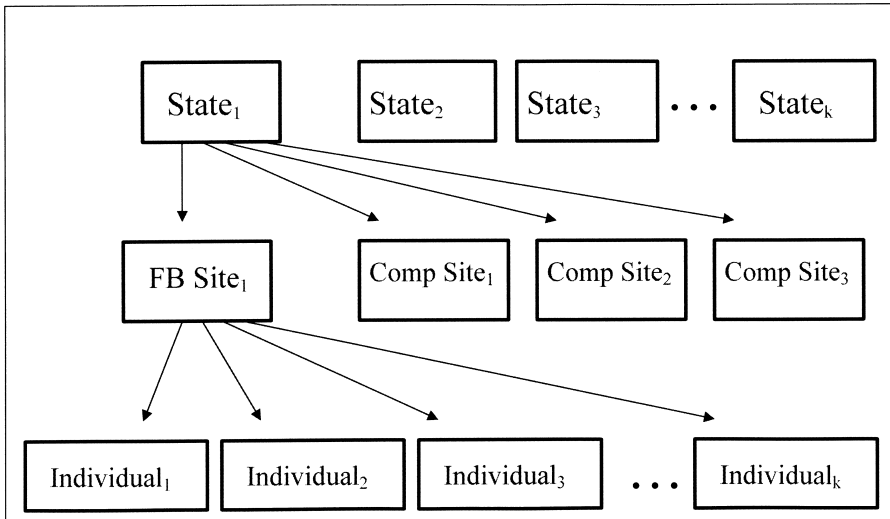
Fighting Back Survey Design

The data examined herein are based on three waves of a telephone survey of residents living in FB sites and comparison communities (see Kadushin et al., 1999; Saxe et al., 1997). Using random-digit-dialing (RDD) sampling procedures, the survey assessed AOD use and related attitudes, acquaintance substance use behaviors and attitudes, and neighborhood perceptions among residents in 12 FB sites (one rural and urban site were excluded) and 29 comparison communities, matched to the FB communities on demographic characteristics. The surveys were conducted in the Spring of 1995, 1997 and 1999 and targeted individuals aged 16-44. In 1995, the sample included approximately 500 respondents per treatment site and group of comparison communities ($n = 12,113$). The sample was expanded in 1997 to include up to 1000 respondents in each treatment community and 600 in each group of comparison communities ($n = 17,900$). The same method was employed in 1999 ($n = 17,469$). Because the most pertinent outcomes were AOD behavior, most of the evaluation’s outcomes are dichotomous and required the use of logistic regression to test program effects. Households without working telephones were excluded from the sampling frame, although since RDD procedures were used, the proportion is relatively small (Gfroerer & Hughes, 1992; Thornberry & Massey, 1988).¹

Pairing of Multiple Controls with a Treatment Community

For community-based prevention programs, the ultimate goal is an increase in desirable behaviors such as condom use or, in the case of FB, a reduction in an undesirable behavior, namely the use of illicit substances. A decrease in drug use in FB communities alone would not indicate program success: many competing explanations for a decrease would remain plausible (i.e., a regional or national decline in AOD use), so comparison communities are “matched” to the program sites in a quasi-experimental design (Cook & Campbell, 1979). Prior evaluations typically have used only matched pairs, in which a single control site is matched to one treatment site. To decrease the potential for competing explanations, the FB evaluation employed a multiple

¹ Telephone ownership is less likely in the South, in rural households, among blacks, among younger adults, and among those with lower incomes. Drug use rates for some of these populations are both lower (i.e., in rural areas) and higher (i.e., younger adults) so the impact of excluding these households from the evaluation are unclear (Gfroerer & Hughes, 1992).

**Figure 1**

Nested Data Structure: Fighting Back Treatment Sites and Matched Comparison Sites

matched control group design, in which each treatment site was matched to several comparison communities as discussed in Rindskopf and Saxe (1998).

Multiple matched controls provide several advantages. Additional control sites provide critical information about site-to-site variation within a state (in this case) which would otherwise not be available. General trends for control sites can be assessed, and the deviation of program sites from these trends can be measured. If control sites show similar trends in substance use outcomes, for example, then FB communities with desirable changes in such rates will be more easily detected. Nevertheless, if there is a considerable amount of variability among control communities, then differences between treatment and controls must be interpreted more carefully. Multilevel modeling can statistically address this issue. First, an overall treatment effect across treatment sites can be estimated. Second, the difference between treatment and controls within each matched pair can be estimated, yielding an effect size for each state. This is not possible with single matched controls: site-to-site variation within a pair of treatment-control sites is confounded with the estimate of treatment-control difference. As FB treatment and control communities were nested within a state, estimates of the program's effect size within that state could be developed. State-level characteristics can then be used to model differences

in treatment effects. It is thus possible to merge quantitative and qualitative traditions (Seltzer, 1994) by both answering “Did it work overall?” and determining what contextual factors may be associated with program success or failure in a particular state. However, the ability to determine the role of the contextual factors is limited: In our evaluation, there were only 12 Level 3 units (“states”).

Multiple controls can be useful in dealing with several possible evaluation problems. A particular control site might turn out to be radically different from the treatment site on important dimensions, despite careful matching. Alternatively, a control community may experience significant change during the course of an evaluation, such as the closing of a military base, which may impact program outcomes. Or, a comparison community may adopt a program similar to the one being evaluated – the community now has a “treatment” underway – and can no longer be used as a control. In any of these cases with a single control design, evaluators might incorrectly attribute differences between treatment and control sites over time to the success of the intervention. With multiple controls, tests of program success can be conducted without the problematic control site.

Gathering data from multiple control communities also affords the opportunity to investigate the influence of site-level variables on programs. A single matched control design, with 12 treatment sites, would have allowed only very limited inferences about the effects of site-level variables as there would have been fewer than 24 degrees of freedom. Matched multiple controls provide greater degrees of freedom at the community level (about 40 in this study) with which the community-level influences on evaluation outcomes at the individual level can be measured. Moreover, a sufficient number of comparison communities permits answering questions that reach beyond the evaluation, such as the macro influences on individual AOD use and related attitudes.

There are some disadvantages to a multiple matched control design. First, matching communities may not be available and poor matches might be selected. For example, comparisons for the FB treatment site in Milwaukee were not ideal because the city is unique for its state. Zip codes in comparison communities were selected because they were demographically similar to those in Milwaukee. However, these communities – Racine and Madison – may yet differ from Milwaukee on many dimensions, some of which may be directly related to AOD use (e.g., proportion of college students). If the treatment site is radically different from all its controls, the entire state might have to be excluded from the analysis. Alternatively, the outlier could be dummy coded separately from other sites when testing program effects and retained.

Second, FB site boundaries did not necessarily follow political lines. Thus extra effort was required to identify comparable control areas within three or four other municipalities in the same state, rather than simply selecting the cities as a whole. To sample residents in comparison areas properly, screening questions based on zip codes or other landmarks had to be developed for each of the comparison sites. Any community indicators collected in the evaluation had to be collected for 29 control sites, as opposed to 12 if a single matched control design was employed. Although the multiple control sites design required considerably more effort, the advantages far outweighed the disadvantages.

Non-Random Sampling at Multiple Levels

The FB survey was designed to assess the program effects, rather than providing exact population estimates of illicit substance use. That is, the focus was on comparison of levels of use in communities where FB was and was not implemented. The survey design therefore diverges from a typical random sample survey in several ways.

An ideal evaluation design would have started with random selection of either states or communities and the random assignment of each selected community to be a treatment or control site. However, such a design was untenable for FB, given the participatory nature of the program (although in theory, one could have selected from among those communities willing to participate). Regardless of theoretical possibilities, the current evaluation (see Saxe et al., 1997) began after FB communities had been chosen and program development initiated. To avoid differences in laws and regulations regarding AOD use, the treatment site was matched with control sites within the same state (except in the District of Columbia). Further, selection was limited to those communities which were demographically most similar to the treatment sites, although in some cases the matching was better than in others. Thus, it cannot be assumed that the study's primary sampling units (PSUs) were randomly selected.

Another divergence concerns the sampling of respondents within PSUs: such sampling was not proportional to size, nor was it equal in all PSUs. Instead, sample sizes were chosen so that within each set of treatment and controls the treatment and group of control communities had approximately the same number of respondents. Therefore, if the FB site had 600 respondents, the matched control sites would contain 300 each (if there were two controls sites). These choices probably increased the power to detect the program effects, but are less than optimal if the survey is used to make inferences about the statewide population, because respondents within

treatment sites have been oversampled. Because the sites were not chosen randomly, it is more difficult to specify the population to which we could generalize, although “lower socioeconomic status urban areas” might be a reasonable general description.

Weights and Weighting

In the FB survey, telephone lines rather than people were randomly sampled in each community. If a household has two phone lines, for example, then a person living in that household is twice as likely to be selected as a person in a household with only one phone line.² Similarly, in a household with many people in the target age range (16-44 years old), each individual has a smaller chance of being surveyed than a person in a household with only one such person. Because individuals within a community do not have the same probability of being selected in the survey, statistical adjustments are required.

A variety of methods for handling these differing probabilities of selection are available (Lee, Forthofer, & Lorimor, 1989, Massey & Botman, 1989). The simplest method is to include these sampling variables – number of phones lines and household size in the target range – as predictors in each analysis. As long as there are no interactions between these variables and other predictors in the model, doing so does not bias the analyses. Another option is to calculate weights and analyze the data using a statistical package such as SUDAAN (Shah, Barnwell, & Bieler, 1997) which weights each respondent inversely proportional to their probability of being selected. In the FB design, the weights would be proportional to the household size divided by the number of phone lines in the household.

External Validity Issues

How well can the lessons learned from evaluating a community-based intervention be generalized to other communities and their populations? One set of external validity issues concerns generalizing evaluation results to the target population of interest (Cook & Campbell, 1979). FB’s target population was defined in terms of both age and location. Screening questions assured that only those in the target age range were included as respondents. The combination of an elaborate sampling process and an extensive geocoding effort after the data were collected ensured that

² In this article we ignore some problems of appropriately counting the number of phone lines in the household: Should cellular phones be counted? Beeper numbers? Computer modems or fax machines?

respondents to the telephone survey indeed lived within the target communities (Beveridge, Bucuvalas, Kadushin, Trippel, & Livert, 1997).

Threats to generalizability from attrition or non-response may occur at multiple levels. For example, in a survey of school systems attrition can occur at several levels: a randomly selected school district may refuse to participate; school principals within a cooperating district may refuse; a parent may not give permission for a child to participate; the child may be absent during the survey; or a child with parent's permission to participate may not answer certain questions. In the FB evaluation, all but two communities were included in the survey so non-response at this level was not a consideration. At the individual level, potential respondents may refuse to participate in the survey at the beginning of the interview, before it is possible to determine whether or not the household qualifies for inclusion in the sampling frame; a lower cooperation rate increases the potential for non-response bias. The average participation rate (completed interviews divided by those respondents actually contacted) across the three waves of the FB telephone survey was roughly 76%, an acceptable level for a survey of this type. Although item non-response can be a concern in telephone surveys that ask sensitive questions such as marijuana and cocaine use, efforts to properly "cushion" the items in the questionnaire and extensive pretesting kept rates relatively low.

Statistical Power Considerations

For the FB evaluation, several factors complicated the computation of power. The FB survey employs cluster sampling (individuals nested within communities) which may reduce power due to intraclass correlations (ICC) or design effects (Kish, 1965). Even a very low ICC (e.g., .01 or .02) can reduce power if there are a small number of sites with many individuals per site (Murray & Hannan, 1990; Siddiqui, Hedeker, Flay, & Hu, 1996). Another design feature that can increase statistical power, but complicates its estimate, is the inclusion of predictors at the individual level. In theory, additional covariates at the individual level could increase or decrease the ICC, in the FB data inclusion of such variables tended to decrease the ICC.

Our strategy was to start with a simple case then qualitatively, rather than quantitatively, adjust for complications. Power was first estimated for a simple random sample (SRS) design. Adjustments were then made for differences between the multilevel design and the SRS design. The effects of such differences are most easily understood as effects on the effective sample size (rather than actual sample size). For example, if there is an ICC in the data, the practical effect would be to make the sample behave as if it

were a smaller SRS, perhaps 2400 instead of 3000, thereby lowering the power. Based on these calculations, 1000 individuals per state were sampled for the first survey wave. To maximize the power to test the FB effect, cases were allocated evenly between the treatment site ($n = 500$) and the group of control sites in each state ($n = 500$): the sample size allotted to each comparison community depended on the number of such sites within a state.

Analytic Issues

Multilevel modeling was employed to test program outcomes measured in the three waves of our general population survey, as well as community indicators for which data existed prior to and throughout program implementation and evaluation. This technique is particularly useful to evaluators because the same statistical models can estimate overall program effects across all sets of treatment and control communities, as well as the degree to which the effect in a specific treatment site deviates from the overall average. As program outcomes were generally dichotomous, logistic regressions were used to test treatment effects. The following examples employ a dichotomous substance use behavior outcome: whether or not the respondent smoked marijuana during the last 30 days. Reflecting the structure of the survey data, a three-level logistic regression model was used. Several forms of the model are reviewed. A basic model is reviewed that would be employed after an intervention where data are collected at a single point in time; differences between treatment and controls sites constitute the program effect. Then, discussion focuses on models employing Level 1 covariates and how interpretation of certain parameters changed. Finally, models that test for program effects over time are explored.

Basic Model

Bryk and Raudenbush (1992) notation is used to describe the multilevel model. There are equations representing each level of the evaluation design.

Level 1

Level 1 consisted of individual respondents. The general form is:

$$(1) \quad y_{ijk} = \pi_{0jk} + \pi_{1jk} x_{1ijk} + \pi_{2jk} x_{2ijk} + \pi_{3jk} x_{3ijk} + \dots + e_{ijk},$$

where the dependent variable y_{ijk} is the response of individual i residing in community j in state k . The term, e_{ijk} , is the residual of individual i from community j in state k ; its variance, σ^2 is assumed to be constant across

communities. In addition to the constant term, independent variables are noted by x . For dichotomous dependent variables, the left-hand side of the equation would be replaced by the logit

$$(2) \quad \ln = \left(\frac{p_{ijk}}{1-p_{ijk}} \right) \pi_{0jk} + \pi_{1jk} x_{1ijk} + \pi_{2jk} x_{2ijk} + \pi_{3jk} x_{3ijk} + \dots$$

Most independent variables were categorical; these were coded into dummy variables. Where used, continuous variables were centered so that values on the transformed variables represented deviations from the average. Interpretability of the constant term was crucial to evaluation objectives, therefore all predictor variables were coded so that a zero value on each represented a “typical” respondent in our survey: a White 19 year old male with a high school education, who was currently employed in the labor force. Dummy coding rather than effects coding, was used so that the intercept in the Level 1 equation would represent an actual (possible) person in the sample. Although this approach provides an easily interpretable constant, estimates may be less precise in communities in which the “typical” respondent is atypical (e.g., Washington, DC, which is mainly non-White).

Examination of the scatter plot of age and marijuana revealed a non-linear relationship with a gradual increase in marijuana use up to the age of 19, followed by a gradual decline. To model this relationship accurately, age was centered with 19 as the reference category. Two variables were then created: the first representing ages 16 to 18 (-3, -2, -1) with 0 for ages 19 and older and the second, representing ages 20 and older (1, 2, 3 and so on) with 0 for ages 20 and younger. Bryk and Raudenbush (1992, p.148ff) discuss similar models. This specification permitted the increased drug use in late adolescence and gradual decline thereafter to be more accurately modeled than if a single age term were used.

Because many independent variables were categorical, the number of predictor items in the actual Level 1 equations is often much larger than the number of independent variables. For example, an individual’s race/ethnicity required three dummy variables in the equation: Black, Hispanic, and other (White was the reference category). The task of testing interactions was simplified by considering only interactions between the FB variable (at Level 2) and the covariates. This strategy would permit differences in the relationship between gender and marijuana use, for example, to vary between FB and control sites and, thus, be partialled out of the program effect term. It is certainly possible to include other interactions between predictors, but these are not explored in the example given below.

Level 1 Sampling Variables. The standard procedure for adjusting for differences in probability of selection among respondents is to calculate the reciprocal of the probability of selection. For evaluation purposes, the estimation of population usage rates is less relevant. Household size and number of phone lines were, however, weakly correlated with substance use and were included in the model as covariates. One phone line per household – the modal category – was the contrast category and households with more than one phone line were coded “1.” Although it would be simplest to treat household size as a continuous variable, examination of the bivariate relationship between household size and substance use outcomes revealed a curvilinear relationship. As a result dummy codes for 1, 3, and 4 or more people of qualifying age in the household were used (the modal two-person household became the reference category).

Level 2. Community-level effects are modeled at Level 2. For the purposes of evaluation, the most important Level 2 characteristic is whether the community was a FB implementation or comparison site. Thus, the Level 2 model for the intercept is

$$(3) \quad \pi_{0jk} = \beta_{00k} + \beta_{01k} \text{FB}_{jk} + r_{0jk},$$

where β_{00k} is the mean response across comparison communities in state k ; and r_{0jk} is the deviation of community j from state k . The term β_{01k} is the difference between the FB community and the comparison community mean in state k . The covariances between effects at different levels are assumed to be zero; the variance-covariance matrix of r_{0jk} is denoted τ_{π} . This model permits the community mean for substance use to vary within a state; each community’s level of drug use can be estimated from the Level 2 intercept and residual term. Other individual level terms are considered fixed. One justification for this decision is a lack of research suggesting that geographic location conditions the relationship between personal characteristics and illicit substance use. A more pragmatic justification arises from the task of interpreting treatment effects, which become quite complicated when they are tested with multilevel models employing a large number of cross-level interactions and random effects terms.

With 41 Level 2 units, enough degrees of freedom were available to include several Level 2 predictors in addition to the FB dummy variable. A number of variables could have been included at the community level, but have not yet been. Variables can be aggregated from individual level data in the survey, such as the percentage of respondents employed full-time. Census data and other surveys of the community can also provide useful

Level 2 variables that can reflect aggregate or holistic characteristics of the community: population size, population density, average temperature, annual days of bad weather, or number of police per 1,000 population.

Level 3. Level 3 effects are at the “state” level. In the case of FB, there are 12 “states” at Level 3. The basic Level 3 model is

$$(4) \quad \beta_{00k} = \gamma_{000} + u_{00k}$$

$$(5) \quad \beta_{01k} = \gamma_{010} + u_{01k}$$

where γ_{000} is the grand mean among all comparison communities, and u_{00k} is the deviation of the state k comparison community mean from the grand comparison community mean. The average FB program effect is represented by γ_{010} , and u_{01k} is the deviation of the program effect in each state from the average program effect. The variance of u_{01k} ($\tau_{\pi 010}$) indicates how much the FB program effect varies among the states in which it was implemented.

Combining the models for each level, the resulting model of the FB program effect with one time point was

$$(6) \quad y_{ijk} = \gamma_{000k} + \gamma_{010k} \text{FB}_{jk} + u_{00k} + u_{01k} \text{FB}_{jk} + r_{0jk} + e_{0ijk}.$$

It is conceivable that multi-community programs may demonstrate overall significant outcomes in the desired direction, while effects in an individual community may even be in the opposite direction (Seltzer, 1994). Thus, an examination of Level 3 residual terms (u_{01k}) is particularly relevant to evaluators. Examination of the Level 3 residuals for the intercept (u_{00k}) may also be of interest. For example, in the case of the FB evaluation, state-to-state differences in outcome variables such as monthly use of alcohol or marijuana may be due to state-level characteristics such as the severity of drug possession laws or the minimum legal age for purchasing alcohol. Of course, with only 12 states the opportunity for extensive testing of such effects is limited.

A Simple Example

The following equations demonstrate the potential flexibility of the multilevel statistical models employed in the FB evaluation. Subscripts for individuals, cities and states are omitted; the dependent variable is a dichotomous variable (monthly marijuana use).

Level 1 (individual)

$$(7) \quad \ln\left(\frac{p}{1-p}\right) = \pi_0 + \pi_1 \text{FEMALE}$$

Level 2 (city)

$$(8) \quad \pi_0 = \beta_{00} + \beta_{01} \text{FB} + r_0$$

$$(9) \quad \pi_1 = \beta_{10} + \beta_{11} \text{FB} + r_1$$

Level 3 (state)

$$(10) \quad \beta_{00} = \gamma_{000} + u_{00}$$

$$(11) \quad \beta_{01} = \gamma_{010} + u_{01}$$

$$(12) \quad \beta_{10} = \gamma_{100} + u_{10}$$

$$(13) \quad \beta_{11} = \gamma_{110} + u_{11}$$

Substituting Equations 8 to 13 into Equation 7 provides an equation representing the whole model

$$(14) \quad y = \gamma_{000} + \gamma_{010} \text{FB} + \gamma_{100} \text{FEMALE} + \gamma_{110} \text{FEMALE} \times \text{FB} \\ + u_{00} + u_{01} \text{FB} + u_{10} \text{FEMALE} + u_{11} \text{FEMALE} \times \text{FB} + r_0 + r_1 + e$$

In this model, the log-odds (logit) of a person smoking marijuana is influenced by the sex of the respondent. The term π_0 represents the logit for males, and π_1 represents the difference between females and males.

Equation 8 shows that the logit for males in a particular city is a function of the overall rate for males in control cities (β_{00}) and the difference between males in FB and comparison communities (β_{01}). The term r_0 indicates that there is variation among the cities, not accounted for by FB (the only Level 2 variable in the model). The term β_{10} in Equation 9 represents the difference in marijuana use between males and females in the average comparison community in a state; β_{11} allows the program effect (FB) to vary for females (compared to males) in that state. This would be useful if there were reason to expect differential treatment effects on males and females (in the actual analysis such terms were not included).

The Level 3 equations each have an overall effect (γ), and a residual term (u). The presence of a residual term indicates that the corresponding effect (β) varies from state to state. A particular advantage of this formulation is that the average effect of FB, γ_{010} , may differ across states, as indicated by $\tau_{\pi 01}$, the variance of u_{01} . Likewise, the average female-male differences in treatment effect, γ_{110} , may differ across states, as indicated by $\tau_{\pi 11}$, the variance of u_{11} .

One feature of the multilevel model is that the Level 2 and Level 3 effects each represent differences among cities and states *after controlling for individual-level variables*. Thus, any differences between cities or states in racial composition, education level, labor force participation, and other measured individual characteristics can be controlled (they were omitted for this simple example). This would lower the intra-site correlation to the degree that differences in these factors across sites account for differential drug use across sites. As noted before, this will increase statistical power.

Modeling Program Effects Over Time

With several waves of data collected through the evaluation's general population survey (1995, 1997 and 1999), examination of FB program effects over time was possible. Our analytic goal was to determine whether substance use trends for respondents living in FB sites diverged in the desired direction from those of control site respondents. This required establishing the time trends for control sites in each state and then testing whether the time trends for treatment sites diverged from the control site trend. This was not a panel study: We did not survey the same respondents in each wave, but different samples of the same communities. Time was coded as a continuous variable (Wave 1 = 0, Wave 2 = 1, Wave 3 = 2) and treated as an individual characteristic at Level 1. The program effect test consisted of the difference in time trends between FB and comparison sites, which was represented by the cross-product of a Level 1 variable (time) and a Level 2 variable (FB):

$$(15) \quad y = \gamma_{000} + \gamma_{010} \text{FB} + \gamma_{100k} \text{TIME} + \gamma_{100} \text{FB} \times \text{TIME} \\ + u_{00} + u_{01} \text{FB} \times \text{TIME} + r_0 + e.$$

γ_{110} is an interaction term representing the program effect: the difference in time trend between FB and control sites. For simplicity of interpretation, the Level 1 variation (time) and Level 2 variation (FB) are fixed, they do not have an accompanying residual term. The FB \times TIME effect, however, can vary between states (u_{01}).

Additional models examined whether the time effect fit a quadratic function (not shown) and whether the program effect exhibited resembled a “step function.” The step function modeled differences in the two time intervals. For this model, two steps were coded separately with STEP1 representing the difference between Time 1 and Time 2 and STEP 2 representing the difference between Time 2 and Time 3. The step function variables are defined as

$$\begin{aligned} \text{STEP1} &= 0 \text{ if Time} = 0; 1 \text{ if Time} = 1 \text{ or } 2 \\ \text{STEP2} &= 0 \text{ if Time} = 0 \text{ or } 1; 1 \text{ if Time} = 2. \end{aligned}$$

Step functions can be useful for coding any ordered predictor. The parameter associated with each such variable estimates the difference in response between one category of the ordered predictor and the next category. If these estimates are all similar, then a straight line (linear function of the original order variable) is appropriate. If any of the estimates is not significantly different from zero, then the response does not change as the predictor changes from one category to the next. Step functions are also quite useful when analyzing time trends of three or more data points.

Interpreting Multilevel Results

Analytic issues notwithstanding, a key set of challenges associated with a three level multilevel analysis concern presentation of results. A key issue is how to present both the technical details and interpretations for audiences with different level of expertise; in particular, the challenge is how to present data for stakeholder audiences who may be less sophisticated statistically.

The verbal interpretation of parameters given above is used when describing the results to relatively sophisticated policy makers. Additionally, in the case of explaining random effects, the square root of variances is used because standard deviations are usually more easily interpreted, and key covariances are translated into correlations. Furthermore, if examination of the residuals indicates approximate normality is reasonable, confidence intervals for standard deviations are also included.

An interpretive challenge is that non-technical consumers do not understand the implications of a logistic model being a nonlinear transformation of a probability. They want to know how big each effect is on a probability scale, not on a logit scale; they are not happy being told “it depends where you are on the scale.” Our goal was to present the results in a way that would be meaningful to consumers of our results, but with

maximal (though not always total) technical correctness. The logistic scale can easily be converted back into the scale of probabilities, using the formula

$$(16) \quad p = \left[\frac{\exp(\textit{logit})}{1 + \exp(\textit{logit})} \right].$$

Transforming the intercept, which represents the log odds of our typical person using some drug, into a proportion is relative easy. To find the effect of each predictor (i.e., Time, FB, Female), we would change their values one at a time, recompute the predicted logit, and then translate back into proportions. Of course, this only gives the effect size in proportions for the “typical” person; because of the nonlinearity of the logit, that size will differ for other individuals. Confidence intervals are also derived by multiplying the SE of a coefficient by 1.96 and calculating the upper and lower limits by adding it and subtracting it from the coefficient; these logits are then converted into percentages.

Presentation of Multilevel Model Results

We illustrate some of the issues encountered in presenting multilevel logistic regression data for evaluating community-based programs. We have simplified the model used here: Not all Level 1 predictors are used, and no Level 2 or Level 3 predictors other than those required to define program effects are used. Furthermore, as discussed above, most interactions were not considered in this model. The important random effects that were kept in these models are those that allowed both the intercept (the general level in the control populations within a state) and the treatment effect (FB \times TIME – the difference between FB and control communities) to vary across states.

Level 1 (individual)

$$(17) \quad \ln\left(\frac{p}{1-p}\right) = \pi_0 + \pi_1 \text{TIME} + \pi_2 \text{FEMALE} + \dots$$

Level 2 (city)

$$(18) \quad \pi_0 = \beta_{00} + \beta_{01} \text{FB} + r_0$$

$$(19) \quad \pi_1 = \beta_{10} + \beta_{11} \text{FB}$$

D. Livert, D. Rindskopf, L. Saxe, and M. Stirratt

$$(20) \quad \pi_2 = \beta_{20}$$

$$(21) \quad \pi_j = \beta_{j0} \text{ for all other predictors}$$

Level 3 (state)

$$(22) \quad \beta_{00} = \gamma_{000} + u_{00}(\text{CONSTANT})$$

$$(23) \quad \beta_{01} = \gamma_{010}(\text{FB})$$

$$(24) \quad \beta_{10} = \gamma_{100}(\text{TIME})$$

$$(25) \quad \beta_{11} = \gamma_{110} + u_{11}(\text{FB} \times \text{TIME})$$

$$(26) \quad \beta_{20} = \gamma_{200}$$

$$(27) \quad \beta_{lm} = \gamma_{lm0} \text{ for all other predictors}$$

(Combined)

$$(28) \quad y = \gamma_{000} + \gamma_{010} \text{FB} + \gamma_{100} \text{TIME} + \gamma_{100} \text{FB} \times \text{TIME} + \gamma_{200} \text{FEMALE} + \dots + u_{00} + u_{01} \text{FB} \times \text{TIME} + r_0 + e$$

In contrast to the model explicated in Equations 10 to 13, there are omitted terms in this model. Residual terms for main effects were omitted for FB at Level 2 (r_1), FB at Level 3 (u_{01}), and Time at Level 3 (u_{10}). Earlier tests of the multilevel model revealed these terms to be zero or near zero. In addition, interpretation of the program effects parameters would have become considerably more complex had they been included in the model.

The interpretation of each parameter is displayed in Table 1. The dependent variable in this example is whether or not the respondent had smoked marijuana in the last 12 months. The individual-level predictor variables are described in Table 2. Data consist of three waves of the evaluation's general population survey and provide a test of the program effect: do respondents in FB sites smoke marijuana less over time, relative to trends in control sites?

Interpreting Program Effects

Presentation of a program effect that represents differences in trends over time can be challenging for even sophisticated audiences. We have found that it is useful to review the meaning of the estimated parameters

Table 1

Interpretation of Important Terms in a Three-Level Logistic Regression Model

Term	Interpretation
p	Probability that an individual used a drug, drank more than five drinks, etc.
$p/(1 - p)$	Odds that an individual used a drug, drank more than five drinks, etc.
$\ln[p/(1 - p)]$	Logarithm of the odds; usually called the logit
π_0	Log-odds that a White (i.e., non-Black, non-Hispanic, ...) male will use a drug
π_1	Difference in log-odds of drug use between Time 1 and Time 2
π_2	Difference in log-odds of drug use between females and males
π_3	Difference in log-odds of using a drug between Blacks and Whites (and so on for other individual-level predictors)
β_{00}	Predicted log-odds of drug use for a "typical person" (White, male, etc.) in control sites within a particular state
β_{01}	Predicted difference in the log-odds of drug use between individuals in FB and control sites within a particular state
β_{10}	Predicted difference in the log-odds of drug use between individuals in control sites within a particular state between Time 1 and Time 2
β_{11}	Predicted difference in time trend between individuals in FB sites and control sites within a particular state
TIME	Indicator variable, with the value 0 for Time 1, 1 for Time 2, and 2 for Time 3
FB	Indicator variable, with the value 1 for FB sites, and 0 for control sites
r	Difference between each control site and the average control site for a particular state.
$(\tau_\beta)^{1/2}$	Standard deviation of r , the variation in control sites around the average control site within a state
γ_{000}	Average log-odds of drug use in control sites, averaged across all states
γ_{010}	Average difference in log-odds of drug use between FB sites and control sites, averaged across all states
γ_{100}	Average difference in log-odds of drug use between Time 1 and Time 2, averaged across all states
γ_{110}	Average difference in time trend between FB and control sites, averaged across all states
u_{00}	Amount by which log-odds of drug use differs between control sites in this state and the overall average control site
Term	Interpretation
u_{11}	Amount by which the program effect (FBxTIME) differs between this state and the average across all states
$(\tau_{\pi00})^{1/2}$	Standard deviation of the log-odds of drug use across states (controlling for individual-level variables)
$(\tau_{\pi11})^{1/2}$	Standard deviation of the amount by which the program effect (FB \times TIME) differs across states (i.e., the standard deviation of the program effect across the 12 states)

Table 2
Description of Individual-Level Predictors Used in the Example

Variable	Description and Coding
FB	1 if treatment site, 0 if control
Time	Survey wave: 0 = 1995, 1 = 1997, 2 = 1999
HH1,HH3,HH4	Number of residents age 16-44 living in household, 2 is the reference category
Phones	1 = 2 or more phone lines in household, 0 = 1
Female	1 if female, 0 if male
Age16-18	Respondent age under 19, -3 = 16, -2 = 17, -1 = 18, 0 otherwise
Age20-44	Respondent age over 19, 1 = 20, 2 = 21, 3 = 22 for those through age 44, otherwise 0
Black	1 if Black, 0 otherwise
Hispanic	1 if Hispanic, 0 otherwise
Other Race	1 if non-White, non-Black, non-Hispanic, 0 otherwise
Neweduc	Educational attainment, centered at High School graduate (-2 = Grade 0-8, -1 = Grade 9-11, 0 = HS Grad/GED, 1 = College 1-3 yrs, 2 = College Grad, 3 = Grad School/Degree)
No Labor	1 if not currently employed full- or part-time, 0 if otherwise

before presenting the actual data. Table 3 shows straightforward descriptions for each of the parameters in a simple model. Figure 2 conveys the meaning of the parameters in a simple graphic form. The horizontal line (1) represents no time effect. A main effect for time in control sites is shown by line (2). The difference between lines (2) and (3) represents the difference between FB sites and controls at Time 1. If time effects are greater in FB sites than control sites, then this FB × Time interaction will be represented by the vertical difference between lines (3) and (4). This term represents the program effect.

The parameter estimates and standard errors are presented in Table 4 as they come from the computer program MLWin (Rasbash, Healy, Browne, & Cameron, 1998) with some minor editing. Some of the fixed effect results are of substantive interest but are incidental to the evaluation findings. For example, females are less likely to report having smoked marijuana than males; and Blacks, Hispanics, and those of other races all are less likely to

Table 3
Verbal Explanation of Trend Model

Verbal Description	Parameter	Model Term
What is the level of drug use in the comparison sites?	γ_{00}	Constant
What was the change in drug use in the from 1995 to 1997 in the comparison sites? From 1997 to 1999?	γ_{10}	Time
What was the FB vs. comparison site difference in 1995?	γ_{01}	FB at baseline
What is the effect of FB?	γ_{11}	FB \times Time Interaction
How much variation in the FB effect is there across sites?	$\tau_{\pi 11}$	Level 3 FB \times Time Residual Variance

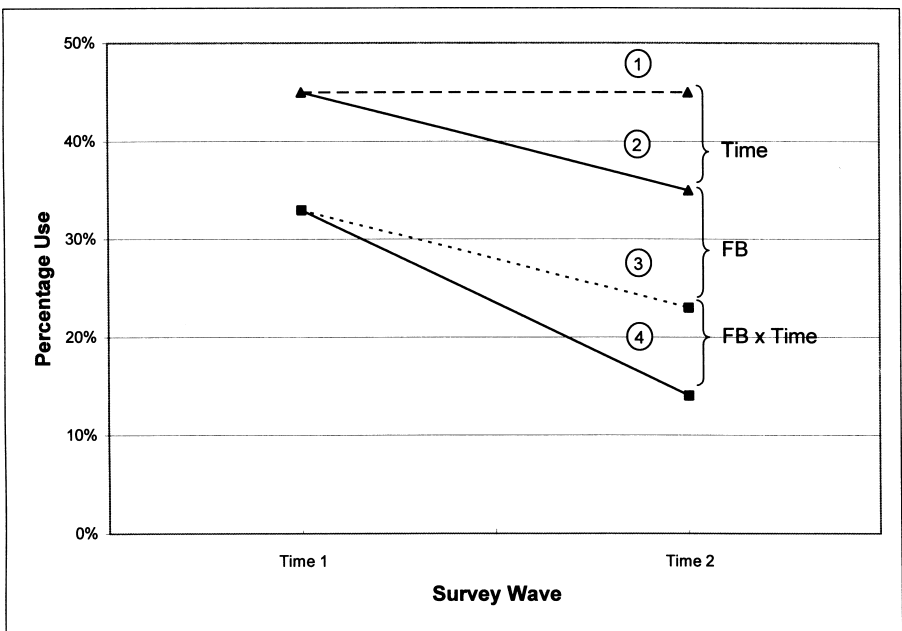


Figure 2
Differences in Slopes Over Time

report having smoked than Whites. For respondents between the ages of 16 and 18 each year brings a slightly higher probability³ of smoking marijuana ($\gamma_{300} = .146$), whereas for those age 20 to 44 the probability is lower ($\gamma_{400} = -.073$). These estimates may appear small at first, but consider these effects are for each year of age; for each decade from age 20 and older the effect is 10 times as big, or (logit = $-.730$). This represents a change in odds by a factor of $\exp(-.730) = .48$. That is, every decade the respondent ages reduces the odds of smoking marijuana by about half.

Marijuana consumption is significantly higher in FB communities than control communities at Time 1 ($\gamma_{010} = .193$). This difference is unrelated to the effectiveness of the program, because the program has started but is still in its early stage. Rather, this is the result of initial differences in spite of matching. Marijuana consumption did not change in control sites over the course of the survey ($\gamma_{100} = .009$, n.s.). The parameter for FB \times TIME ($\gamma_{010} = -.033$) is in the desired direction favoring the evaluation: Marijuana use is decreasing in the intervention communities, at the rate of $-.024 = .009 + (-.033)$ per two-year interval, while increasing in the control communities. However, this difference is not statistically significant ($SE = .046$, $t = -.717$). Table 5 displays the conversion of parameter estimates into proportions for a person who has a value of zero all of the predictors (White male, age 19, employed, high school graduate).

Variance in Program Effects

The variance of the FB average program effect is not significantly different from zero ($\tau_{\pi_{11}} = .005$, $SE = .005$). However, one can still estimate a confidence interval around the FB program effect estimate ($\gamma_{010} = -.033$): The variance for the program effect is .005; its square root (.071) is the standard deviation of the state-level effect sizes. Using the standard error of the FB \times Time term, the 95% confidence interval for the average or "typical" FB program effect is estimated as $-.123 [-.033 - (1.96 * .046)]$ to $.057 [-.033 + (1.96 * .046)]$.

Examination of program effect residuals can also be quite informative, particularly with regard to outliers. Are residuals normally distributed? Multilevel computer programs such as MLWin provide a fairly straightforward way to estimate residuals and standardized residuals for variance terms at any level. Figure 3 shows the standardized residuals for the program effect for marijuana use. The points represent the residual for the FB programs in Sites A through L. The residual for Site A is nearly two

³ Note that the coefficient does not represent the probability, rather it is the log odds which has to be transformed in order to estimate a probability.

Table 4

Parameter Estimates for Model of Marijuana Use in Last 12 Months, as Calculated by MLWin

<u>Fixed Parameters</u>				
<u>Parameter</u>		<u>Estimate</u>	<u>Standard Error</u>	
γ_{00} – Constant		-.916	.094	
<u>Program Effects</u>				
γ_{01} – FB		.193	.077	
γ_{10} – Time		.009	.030	
γ_{11} – FB * Time		-.033	.046	
<u>Individual Characteristics</u>				
γ_{20} – Female		-.600	.033	
γ_{30} – Age (16-18)		-.146	.035	
γ_{40} – Age (20-44)		-.073	.002	
γ_{50} – Black		-.392	.043	
γ_{60} – Hispanic		-.675	.053	
γ_{70} – Other Race		-.492	.072	
γ_{80} – Education		-.112	.016	
γ_{90} – Not in Labor Force		-.032	.039	
<u>Weighting Variables</u>				
γ_{100} – HH1		.151	.038	
γ_{110} – HH3		.136	.048	
γ_{120} – HH4		.194	.055	
γ_{130} – Phones		.147	.041	
<u>Random Parameter</u>				
Variance of				
<u>Random Term</u>	<u>Level</u>	<u>Random Term</u>	<u>Estimate</u>	<u>Standard Error</u>
u_{00}	3	$\tau_{\pi 00}$.054	.027
u_{11}	3	$\tau_{\pi 11}$.005	.005
r	2	τ_{β}	.023	.010
e	1	σ^2	.000	.000

Table 5

Conversion of Program Effect Parameters to Proportions

Time	FB	Parameters (In Logits)	Proportion
0	0	$-.916 + 0 + 0 + 0 = -.916$.286
1	0	$-.916 + .009 + 0 + 0 = -.907$.288
0	1	$-.916 + 0 + .193 + 0 = -.723$.327
1	1	$-.916 + .009 + .192 - .033 = -.747$.321

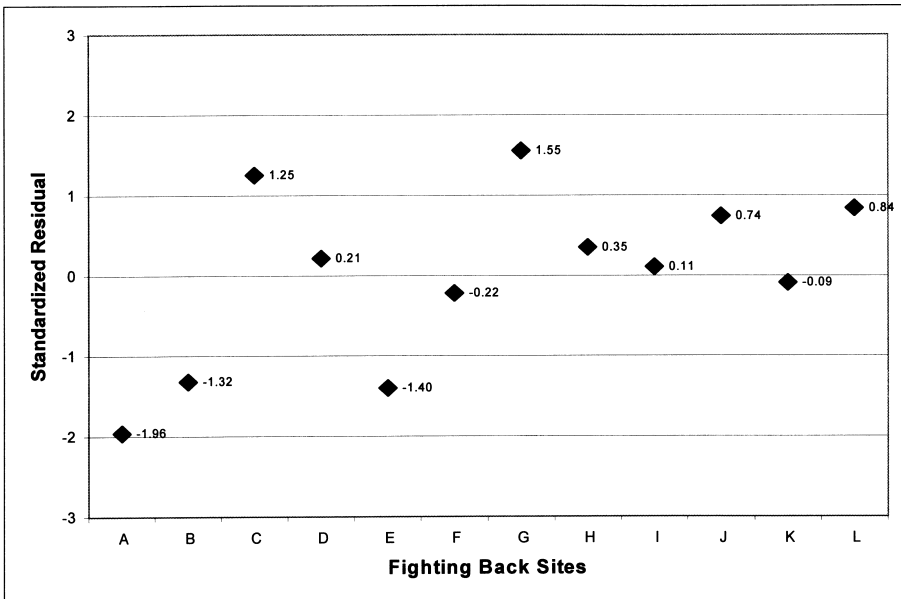


Figure 3
Standardized Residuals of Program Effects for Marijuana Use

SDs below the average program effect, indicating that the marijuana use trend in that treatment site is notably lower than in control sites – a desirable program effect. Statistical explanations for such effects should be explored first. For example, the intercept may be less meaningful in Site A: There may be relatively few White males and, as a result, marijuana use may be estimated poorly. This presents a real issue in making coding decisions as to the characteristics of the typical person.

Specification of the statistical model is an issue of more general concern. It may be the case that differences between sites are due to individual

characteristics that were not adequately controlled for by the covariates in the model. This may mislead evaluators into examining city-wide characteristics or aspects of the treatment program when differences in fact are due to inadequate specification at Level 1.

Implausible Variances and Covariances

Because many of the evaluation outcomes are relatively low incidence events (e.g., cocaine use), we have sometimes encountered implausible estimates of the correlation between the residuals, such as -1.33 , which may indicate one of two problems. First, one of the variances may be so near zero that there is instability in the estimates of variances and covariances. Setting the offending variance and covariances to zero will solve the problem. Another potential offender is an outlier; one must estimate the residuals and examine plots of them to determine whether this is a problem. One possible strategy is to include a dummy variable for the outlying site (at Level 2) or state (at Level 3).

Presentation of Site-Level Estimates

In addition to presentation of program effects, it was useful to present simple graphs showing estimated rates of evaluation outcomes (e.g., alcohol and drug use rates) across the treatment and control sites in the survey. Figure 4 displays the estimates of 12 month marijuana use for a “typical” respondent. For simplicity of presentation, we only showed rates based on the most recent wave of the survey. The FB and control communities are arrayed across the horizontal axis in random order with each letter representing a different community. The logits have been translated into percentages. Two estimates appear for each community. The triangles represent empirical Bayes (EB) estimates, which are produced by substituting the appropriate values into the combined model (18) and deriving the Level 2 residual for each community and the Level 3 residuals for state and for the FB effect. In contrast, the horizontal bars show estimates based upon a single-level logistic regression with dummy coding for each community. Examination of communities C and L demonstrate the “shrinkage” associated with Bayesean estimation (Efron & Morris, 1977): EB estimates are closer to the grand mean than those produced by a single-level logistic regression. This graph provides a straightforward way of displaying differences among states and among cities within those states, after taking into account the demographic composition (as modeled at Level 1) of each community.

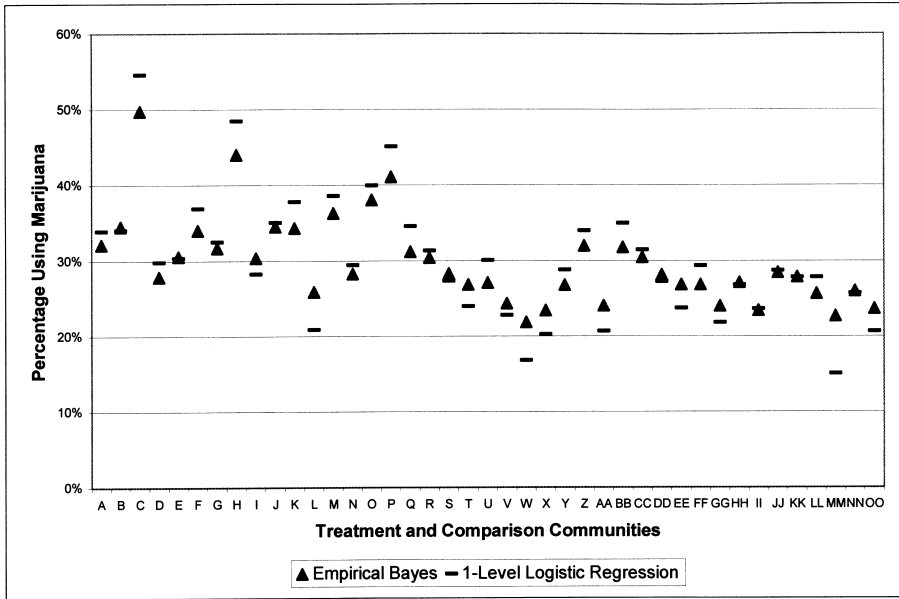


Figure 4
Empirical Bayes Estimates

General Remarks

The application of multilevel models is becoming more popular, and justifiably so. By now, most sophisticated analysts would probably apply these methods to community studies such as the FB evaluation. But the analytic task is more complicated than merely getting a computer program to run and presenting the output. A variety of steps must be taken to ensure a useful design, and to translate the results from the output into useful tables and figures.

A critical issue is that the design of a study is crucial. Without randomization, causal inference is seriously weakened. With only one matched control city per site, some effects cannot be estimated, or can be estimated only by making untestable assumptions. As was so well-stated by Light, Singer, and Willett (1990), *“You can’t fix by analysis what you bungled by design.”* (italics in original; p. viii).

Although the most important evaluation question may be “Did it work here?,” many evaluators of community-based interventions are interested in generalizing across communities and would like to know “Would it work

somewhere else like this?” The ability to answer the latter question will depend, in part, on whether communities are treated in statistical models as fixed or random effects (Murray, 1998). A potential point of controversy in our evaluation is the application of multilevel methods in a study where, in fact, only respondents were randomly selected. Although we treated cities and states as random factors, in fact there was nothing random about their selection. Treatment communities were selected by RWJF through an application process; sites were not randomly selected and then assigned to a treatment or control condition. Further, comparison sites were drawn from the mid-sized communities within the same state which demographically resembled the intervention site. These communities may represent a random selection from some population, but we do not know what it is. At best we could describe the cities and states, and allow others to make inferences about the generality of results. This is a problem of interpretation: As with most studies, whether these effects could be considered random is open to debate.

A statistical problem would arise from our treatment of cities as random effects if the selection of cities or states produced odd-shaped distributions that violated the assumptions of the analysis. Examination of residuals is an important step in determining whether this occurs. An alternative that would be possible here, though not practical for much larger studies, is to treat cities and states as fixed effects. This results, however, in a proliferation of parameters that can become unwieldy in terms of interpretation and presentation.

Parameter estimation is critical when employing multilevel models for experimental and evaluation designs. It is issue both for design and analysis. In our case, using multiple control sites for each FB site ensured that we could estimate how much variability existed in the program effects.

As with any multilevel model, choice of coding for each variable is important, because intercepts and slopes at each level are explained at the next highest level; if intercepts or slopes have no useful interpretation, then explaining them at a higher level is worthless. One choice was to use dummy variables at the individual level; this made the intercept a predicted value for a “real” person, rather than some sort of “average” person had we used effects coding. Although others may prefer to make different choices, the analyst should be aware of the statistical and interpretive implications inherent in each coding scheme.

References

- Aguirre-Molina, M. & Gorman, D. M. (1996). Community-based approaches for the prevention of alcohol, tobacco, and other drug use. *Annual Review of Public Health, 17*, 357-358.
- Beveridge, A., Bucuvalas, M., Kadushin, C., Trippel, K., & Livert, D. (1997, May). *Targeting respondents in small defined geographic areas: Experiences from a large scale telephone survey*. Paper presented at the Annual Meeting of the American Association for Public Opinion Research, Norfolk, VA.
- Bryk, A. S. & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage Publications.
- Center for Substance Abuse Prevention. (1996). *Fourth annual report of the National Evaluation of the Community Partnership Demonstration Grant Program*. Rockville, MD: Center for Substance Abuse Prevention, U.S. Department of Health and Human Services.
- Connell, J. P., Kubisch, A. C., Schorr, L. B., & Weiss, C. H. (Eds.) (1995). *New approaches to evaluating community initiatives*. Washington, DC: The Aspen Institute.
- Cook, T. D. & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand-McNally College Publishing Company.
- Efron, B. & Morris, C. (1977, May). Stein's Paradox in statistics. *Scientific American, 119*-128.
- Gfroerer, J. C. & Hughes, A. L. (1992). Collecting data on illicit drug use by phone. In C. F. Turner, J. T., Lessler, & J. C. Gfroerer, *Survey measurement of drug use: Methodological studies* (pp. 277-295). (DHHS Publication No. ADM 92-1929). Washington, DC: Government Printing Office.
- Hallfors, D., Reber, E., Saxe, L., & Watson, K. (1998). *Making the mark: Analysis of strategies in a comprehensive community substance abuse prevention program*. Report prepared for the Robert Wood Johnson Foundation. Waltham, MA: Brandeis University, Fighting Back National Evaluation.
- Jellinek, P. S. & Hearn, R. P. (1991). Fighting drug abuse at the local level: Can communities consolidate their resources into a single system of prevention, treatment, and aftercare? *Issues in Science and Technology, 7*(4), 78-84.
- Kadushin, C., Tighe, E., Saxe, L., Livert, D., Trudeau, K., Ford, J., Barreras, R., & Buteau, E. (1999). *General Population Survey: 1999 codebook*. Unpublished.
- Kaftarian, S. J. & Hansen, W. B. (1994). Improving methodologies for the evaluation of community-based substance abuse prevention programs. *Journal of Community Psychology, CSAP Special Issue, 3*-5.
- Kenny, D. A. (1996). The design and analysis of social-interaction research. *Annual Review of Psychology, 47*, 59-86.
- Kish, L. (1965). *Survey sampling*. New York: John Wiley & Sons.
- Lee, E. S., Forthofer, R. N., & Lorimor, R. J. (1989). *Analyzing complex survey data*. Newbury Park, CA: Sage.
- Light, R. J., Singer, J. D., & Willett, J. B. (1990). *By design: Planning research on higher education*. Cambridge, MA: Harvard University Press.
- Massey, J. T. & Botman, S. L. (1989). Weighting adjustments for random digit dialed surveys. In R. M. Groves, P. P. Biemer, L. E. Lyberg, J. T. Massey, W. L. Nicholls, & J. Waksberg (Eds.) *Survey errors and survey costs* (pp.143-160). New York: Wiley.

- Murray, D. M. (1998). *Design and analysis of group-randomized trials*. New York: Oxford Books.
- Murray, D. M. & Hannan, P. J. (1990). Planning for the appropriate analysis in school-based drug-use prevention studies. *Journal of Consulting and Clinical Psychology*, 58(4), 458-468.
- O'Connor, A. (1995). Evaluating comprehensive community initiatives: A view from history. In J. P. Connell, A. C. Kubisch, L. B. Schorr, & C. H. Weiss (Eds.) *New approaches to evaluating community initiatives: Concepts, methods, and contexts* (pp. 23-64). New York: The Aspen Institute.
- Office of National Drug Control Policy. (1999). *The National Drug Control Strategy, 1999*. Washington, DC: Executive Office of the President of the United States.
- Osgood, D. W. & Smith, G. L. (1995). Applying hierarchical linear modeling to extended longitudinal evaluations: The Boys Town follow-up study. *Evaluation Review*, 19(1), 3-38.
- Pentz, M. A., Dwyer, J. H., MacKinnon, D. P., Flay, B. R., Hansen, W. B., Wang, E. Y. I., & Johnson, C. A. (1989). Multicommunity trial for primary prevention of adolescent drug abuse: Effects of drug use prevalence. *Journal of the American Medical Association*, 261, 3259-3266.
- Perry, C. L., Williams, C. L., Veblen-Mortenson, S., Toomey, T. L., Komro, K. A., Anstine, P. S., McGovern, P. G., & et al. (1996). Project Northland: Outcomes of a community-wide alcohol use prevention program during early adolescence. *American Journal of Public Health*, 86, 956-965.
- Rasbash, J., Healy, M., Browne, & Cameron, B. (1998). *MLWin* (Version 1.02.0003) [Computer Software]. London, England: Institute of Education, University of London.
- Rindskopf, D. & Saxe, L. (1998). Zero effects in substance abuse programs: Avoiding false positives and false negatives in the evaluation of community-based programs. *Evaluation Review*, 22, 78-94.
- Rossi, P. H., Freeman, H. W., & Lipsey, M. W. (1999). *Evaluation: A systematic approach* (6th Ed.). Thousand Oaks, CA: Sage Publications.
- Saxe, L., Reber, E., Hallfors, D., Kadushin, C., Jones, D., Rindskopf, D., & Beveridge, A. (1997). Think globally, act locally: Assessing the impact of community based substance abuse prevention. *Evaluation and Program Planning*, 20, 357-366.
- Saxe, L. & Winick, C. (in press). Drug Abuse. In E. Borgatta (Ed.) *Encyclopedia of Sociology* (2nd ed.).
- Seltzer, M. H. (1994). Studying variation in program success: A multilevel modeling approach. *Evaluation Review*, 18, 342-261.
- Shah, B. V., Barnwell, B. G., & Bieler, G. S. (1997). *SUDAAN Users Manual. Release 7.5*. Triangle Park, NC: Research Triangle Institute.
- Siddiqui, O., Hedeker, D., Flay, B. R., & Hu, F. B. (1996). Intraclass correlation estimates in a school-based smoking prevention study: Outcome and mediating variables by gender and ethnicity. *American Journal of Epidemiology*, 14, 425-433.
- Spickard, W. A., Dixon, G. L., & Sarver, F. W. (1994). Fighting back against America's public health enemy number one. *Bulletin of the New York Academy of Medicine: A Journal of Urban Health*, 71(1), 111-135.
- Substance Abuse and Mental Health Services Administration. (1995). *National Household Survey on drug abuse: Main findings (1993)*. Rockville, MD: Office of Applied Statistics SAMHSA, Department of Health and Human Services, Public Health Service.

D. Livert, D. Rindskopf, L. Saxe, and M. Stirratt

Thornberry, O. T. & Massey, J. T. (1988). Trends in United States telephone coverage across time and subgroups. In R. M. Groves, P. P. Biemer, L. E. Lyberg, J. T. Massey, W. L. Nicholls, & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 25-39). New York: John Wiley & Sons.

Winick, C. & Larson, M. J. (1997). 75 community action programs. In J. H. Lowinson, P. Ruiz, R. B. Millman, & J. G. Langrod (Eds.), *Substance abuse: A comprehensive textbook* (3rd ed., pp. 755-764). Philadelphia, PA: Williams and Wilkins.