Challenges to Social Experiments

A Drug Prevention Example

Phyllis L. Ellickson, Robert M. Bell

RAND

R-4173-CHF

Challenges to Social Experiments

A Drug Prevention Example

Phyllis L. Ellickson, Robert M. Bell

Supported by the
Conrad N. Hilton Foundation

RAND

The research described in this report was supported by a grant from the Conrad N. Hilton Foundation.

ISBN: 0-8330-1238-X

Published 1992 by RAND
1700 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138

CHALLENGES TO SOCIAL EXPERIMENTS: A DRUG PREVENTION EXAMPLE

PHYLLIS L. ELLICKSON
ROBERT M. BELL

Evaluations of school-based drug prevention programs have been plagued by problems that limited the validity of their findings. These limitations pose clear challenges for current prevention research. This article describes how a recent, multisite experiment conducted in 30 junior high schools met several of those challenges, specifically: evaluating the program in a variety of environments; achieving well-balanced experimental groups; implementing the program as designed; obtaining reliable outcome measures; and eliminating alternative explanations for the results. In most cases, multiple strategies were employed. Extensive analyses were conducted to assess how well the strategies worked; they indicated that each obstacle was overcome. This success implies several lessons for future experiments that are generally applicable to field studies conducted with schools and other organizations. Of particular importance are the guidelines for recruiting institutions from diverse communities and maintaining their cooperation over several years and the techniques for facilitating faithful program delivery and monitoring the implementation process. Recommended design and analysis features include using more than simple random assignment to achieve a balanced design, and employing control variables to rule out alternative explanations of the results—even under conditions of substantial pretreatment equivalence.

Over the past decade, school-based efforts aimed at preventing or reducing adolescent drug use have proliferated. However, credible evaluations of their effectiveness have been considerably less plentiful, in part because successfully carrying out a rigorous social experiment requires overcoming many—and difficult—obstacles.

This research was supported by a grant from the Conrad N. Hilton Foundation. The expertise of the Project ALERT staff and cooperation of school personnel were critical to successful implementation of the experiment. We thank the Special Issue editors, David Weisburd and Joel Garner, and the anonymous reviewers for valuable comments on a previous draft. Judy Morton provided secretarial assistance.

JOURNAL OF RESEARCH IN CRIME AND DELINQUENCY, Vol. 29 No. 1, February 1992 79-101

We recently reported the results of a multisite drug prevention trial that found a significant impact on curbing adolescent use of cigarettes and marijuana (Ellickson and Bell 1990a). As with all such studies, the credibility of the findings largely rests on the degree to which the research meets the criteria for achieving a rigorous experimental test.

Often cited weaknesses of previous school-based prevention studies include limitations in scope (too few schools and students, too little diversity), lack of random assignment, faulty implementation, unanswered questions about the accuracy of reported drug use, and inadequate statistical controls (Moskowitz 1989; Biglan and Ary 1985). Each problem poses a challenge for experimental research; each can be rephrased into a question by which the strength of specific studies can be assessed:

- Has the program been evaluated in a variety of environments?
- Is the experimental design balanced across conditions?
- Was the program implemented as designed?
- Are the outcome measures reliable?
- Have alternative explanations of the results been ruled out?

This article describes the strategies we adopted to address these questions and evaluate how well each strategy worked. It concludes with what we learned from the effort to meet and overcome key challenges to the successful completion of a field experiment.

STUDY BACKGROUND: CURRICULUM, RESEARCH DESIGN, AND RESULTS

Project ALERT, the curriculum we evaluated, was designed to motivate students against using these drugs and to teach them the skills they need to translate that motivation into effective resistance behavior (Ellickson, Bell, Thomas, Robyn, and Zellman 1988). Earlier drug prevention programs, which focused on providing students with information about drugs and/or teaching them general communication and problem-solving skills, have typically failed to reduce drug use (Polich, Ellickson, Reuter, and Kahan 1984; Goodstadt, 1986). Project ALERT differs from these programs in at least two ways: it treats information about the negative consequences of drug use as just one of several potential motivators for resisting drugs; and it teaches skills that are specifically linked to situations in which young people might feel pressured to use drugs.[1]

The program is based on the social influence model of prevention, which assumes that a decision to begin using alcohol, cigarettes, or other substances is primarily influenced by family, friends, and other societal forces (such as advertising), whose impact is reflected in the young person's own beliefs about drugs. Prevention strategies based on this model focus on: (a) puncturing common myths about drug use (such as that most people smoke, or that cigarettes are relaxing); (b) helping young people understand how drug use can affect their daily lives and social relationships; (c) enhancing awareness of the social influences at work; and (d) teaching specific techniques for resisting those influences (learning to say "no").

A unique aspect of Project ALERT is its focus on helping adolescents to identify and resist internal, as well as external, pressures to use drugs (for example, beliefs that doing drugs will help you be more "accepted," will cover over bad feelings, will alleviate boredom, etc.). Teachers help students develop these skills through a combination of "show, practice, and feedback"—watching skill demonstrations, roleplaying successful solutions to different pressure scenarios, and receiving appropriate reinforcement for successful performance. Drawn from Bandura's (1977, 1984) work on social learning and effective behavior change, these techniques foster resistance motivation by helping students believe that others have successfully resisted and that they can do it, too.

The Project ALERT curriculum was designed for seventh and eighth graders, a group that is highly vulnerable to pro-drug social influences but typically not yet regular users. It was tested with over 6,500 students from 30 schools in California and Oregon—20 treatment and 10 control schools. All students in the first group received the same curriculum, eight weekly lessons in seventh grade and three booster lessons when the students reached eighth grade. Those in 10 of the treatment schools were taught solely by an adult teacher (or health educator) hired for the experiment; in the other 10 schools, students received the training from adults who were assisted by older teens from neighboring high schools. Students in the control schools, who received no special resistance training, served as the control group.[2]

The results of the experimental test show that Project ALERT effectively prevented or reduced both cigarette and marijuana use during the junior high years (Ellickson and Bell 1990a, 1990b). It was particularly successful in delaying the onset of marijuana use and in holding down occasional and regular smoking among previous cigarette experimenters. For marijuana, the program produced generally positive effects across students exhibiting different levels of risk for future use. For cigarettes, the pattern varied by level of prior smoking experience: It was most effective with prior experimenters,

whose risk of becoming regular smokers by grade eight was four times greater than that of the baseline nonsmokers. However, among students whose commitment had gone beyond trying cigarettes once or twice, subsequent smoking was higher for students who had the program than for those who did not—a boomerang effect found in other antismoking programs (Biglan et al. 1987).

The program also produced modest inroads against drinking during grade seven. But those gains eroded within 12 months, a result that reflects alcohol's comparatively greater social acceptability in the nation at large, as well as among the West Coast students who participated in the study. Variations in the methods of program delivery (whether it was taught by adults only or by adults with the assistance of nonusing teen leaders) did not yield significantly different program effects on actual use.

TESTING THE PROGRAM IN DIVERSE ENVIRONMENTS

In designing the experiment, our first challenge was to ensure that the program would be tested across diverse environments. To avoid the limitation of previous studies that had focused primarily on White, middle-class suburban communities with few minority students (Botvin and Wills 1985; Flay 1985), we recruited school districts representing a variety of community types, socioeconomic levels, and ethnicities.

However, the integrity of the experiment also required that the districts abide by certain conditions; namely, to submit to random assignment, to forego any existing drug prevention curriculum if assigned to a treatment condition, to provide mixed-gender classes with a maximum of 40 students, and to accommodate our data collection requirements (including the collection of physiological samples) and a weekly lesson schedule. Inevitably, therefore, some districts that we identified as likely candidates for inclusion in the experiment were bound to find one or more of these conditions unacceptable. Indeed, 11 of the districts we approached chose not to participate in the experiment because they could not meet these requirements.[3]

To maximize diversity, we identified potential districts for the sample on the basis of demographic characteristics compiled for all California and Oregon school districts. We sought variety on the following dimensions: geography, community type (urban/suburban/rural), racial/ethnic composition, socioeconomic level, school size, and grade span for the middle or junior high schools. Promising districts were contacted by phone and then visited to determine whether the district was willing to meet the experimental requirements. As some districts joined the experiment and others dropped

TABLE 1: Characteristics of Districts Forming the Sample

Characteristic	*Number of Districts*
Location	
Oregon	3
Northern California	3
Southern California	2
Locale type	
Large city (over 100,000)	2
Medium city (50,000-100,000)	1
Suburb	3
Small city (under 50,000)	1
Rural	1
Grade span	
6-8	2
7-8	4[a]
7-9	2

a. One of these districts includes two schools with grades seven to 12.

out of contention, we contacted new districts that would complement those already included.

In all, eight school districts ranging in size from three to nine middle schools signed up.[4] As Tables 1 and 2 show, the resulting group of participating districts and schools encompassed a variety of community and school environments. The eight districts included urban, suburban, and rural communities from five regions in California and Oregon, and the participating schools included a wide range of socioeconomic levels and ethnic and racial groups. Of the 19 California schools, 11 drew from neighborhoods with median family income below the 1979 median for the state ($21,537); 7 of the 11 schools from Oregon served neighborhoods with median family income below the Oregon median ($20,027).[5] The range extended from $11,000 to $28,000. In addition, 9 of the 30 schools had a minority population of 50% or more. Among the participating seventh graders, 41% came from disrupted families (did not live with both natural parents), and one-third had a minority background.

The selection process also yielded substantial variation on other school characteristics for which no information was available at the time of district selection. For example, the percentage of students who had used cigarettes in the month prior to the baseline survey ranged from less than 10% in five schools to more than 30% in three other schools. Neither the sample of schools nor the resulting sample of students can be considered a random or representative sample from California and Oregon. However, our goal was

TABLE 2: Characteristics of Schools Forming the Sample

Characteristic	*Number of Schools*
Seventh-grade enrollment	
Under 200	5
200-299	16
300-399	6
400-499	3
Non-White[a]	
0-9%	6
10-19%	9
20-49%	6
50-82%	9
Some college[b]	
37-49%	5
50-59%	11
60-69%	7
70-80%	7

a. Combined percentage of Blacks, Hispanics, American Indians, and multiethnic students at baseline.
b. Percentage of students who reported on the baseline survey that at least one parent had attended some college.

to evaluate Project ALERT's effectiveness across different school and district environments rather than to estimate use in a specific population. Thus the purposive nature of the sample enhanced the study's generalizability while not affecting its internal validity.

ACHIEVING WELL-BALANCED EXPERIMENTAL GROUPS

Random assignment is a cornerstone of treatment evaluation experiments. Without it, selection bias poses an unanswerable threat to a study's findings. For social influence programs, schools form the natural unit for treatment and thus for experimental assignment. However, when the number of experimental units is limited, simple random assignment is not sufficient to ensure adequate balance across the experimental groups on baseline characteristics related to later substance use. This problem has severely handicapped earlier prevention evaluations with as few as 6 to 10 schools.

To reduce differences among experimental groups in both measured and unmeasured school attributes, we augmented simple randomized assignments in two ways: (a) using district as a blocking factor and (b) restricting

randomization to those designs with relatively little imbalance on several key factors.

Blocking by District

School districts form blocks in Project ALERT because schools within a district often have similar underlying substance use rates.[6] Local socioeconomic status and community norms may influence substance use in ways that other variables could not explain. Also, district policies—for example how severely school officials penalize substance use on campus—may influence the rate of substance use.

To take advantage of within-district homogeneity, we spread each district's schools as evenly as possible among the three experimental conditions. When we used three schools from a district, we considered only those designs that assigned exactly one school to each cell. When we used four schools, we required that two cells receive exactly one school while the third cell received two schools. For districts with five schools, one cell received one school and the other two received two schools each. To maintain the randomization property, we selected at random the one or two cells that received two schools. In some districts with four or five schools, we could divide the district into two relatively homogeneous subblocks, one of which contained exactly three schools. In those cases, we required that each member of the three-school subblock be assigned to a different cell.

Restricted Assignments

Despite the tendency for schools within a district to resemble each other on various characteristics, substantial differences do exist among schools within districts. Because of this fact and the limited number of units per cell (10), blocking and randomization alone would not ensure a well-balanced design.[7] To improve cell similarity, we balanced school characteristics across districts as well—for example, by linking "unlike" schools from two districts into pairs and randomly assigning the pairs to the experimental conditions. We restricted the assignments to a subset of designs with good balance across cells and selected randomly from among those assignments. Thus each school in the sample had a one-third probability of being assigned to any particular cell.

Data used for assignment included a mixture of publicly available data and some gathered specifically for this experiment. The most important data

source was an 11-question survey administered to eighth-grade students at potential Project ALERT schools late in the academic year preceding baseline. This survey provided information about the mobility of the students between schools, the educational level of both parents, whether English was spoken at home, and lifetime and recent student use of cigarettes and marijuana.

Census files provided demographic and socioeconomic status (SES) data for the census tracts supplying students to each school, which we matched with each school's cachement area to produce estimates of racial/ethnic composition, median family income, educational levels of adults, and the proportion of families that included both parents. Sixth-grade reading, writing, and math achievement test scores were collected for the schools that supplied students to our California sample. The state of California also supplied data on the percentage of seventh graders who possessed limited English proficiency. Each school provided data on current and past enrollment in the seventh grade.

Because districts and schools joined the project sequentially, we assigned schools to treatments in four phases. In each phase, schools from two of the eight districts were assigned.

In Phase I, we assigned six schools, three from each of two districts. With district as a blocking factor, each cell included exactly one school from each district. Temporarily ignoring assignments to experimental condition, we find that each school from the first district (District A) must be paired with a school from the second district (District B). This yields the six potential configurations shown below:

Configuration					
1	2	3	4	5	6
A1-B1	A1-B1	A1-B2	A1-B2	A1-B3	A1-B3
A2-B2	A2-B3	A2-B1	A2-B3	A2-B1	A2-B2
A3-B3	A3-B2	A3-B3	A3-B1	A3-B2	A3-B1

Some potential configurations would produce better balance across cells than would others. For example, suppose that schools A1 and B1 had the lowest SES in their respective districts and that schools A3 and B3 each had the highest SES. In that case, configuration 1 would produce a poor design because the mean SES for the cell with schools A1 and B1 would fall much below the mean SES for the cell with schools A3 and B3. In contrast,

TABLE 3: Baseline Survey Data by Experimental Cell

Item	Cell		
	Adult Only Curriculum	Teen Leader Curriculum	Control
Mean school size	234	234	203
Parents who refused consent (%)	7	8	10
Percentage Black	11	9	11
Percentage Hispanic	12	7	12
Percentage Asian	8	9	7
One parent who attended college	63	66	60
Used marijuana in lifetime	20	19	22
Used cigarettes in lifetime	52	52	52
Used alcohol in lifetime	74	74	73
Used marijuana in past month	7	6	7
Used cigarettes in past month	16	16	14
Used alcohol in past month	21	24	23

configuration 6, which pairs opposites, produces SES means that agree closely across the three cells.

We selected the configuration that produced the least imbalance on the assignment variables as measured by a weighted sum of the variances of the cell means. Once we had determined the best configuration, we randomly assigned each of the three pairs to a different experimental condition. The appendix of Ellickson et al. (1988) describes the later assignment phases.

Results of the Assignment Process

Table 3 compares the three experimental cells on several baseline sample characteristics. It shows that the assignment procedure achieved substantial comparability on a variety of measures that were unavailable when the assignments were made. For each of the substance use variables, the between-cell differences are small compared with differences that might constitute "meaningful" treatment effects.

Use of blocking and restricted randomization substantially improved the balance across cells. For most variables, the amount of variation among cells is several times smaller than would have occurred, on average, with the use of simple random assignment.[8] Greater imbalance than expected occurred for only one variable: the percentage of parents who refused consent. Even for this variable, however, simple random assignment would have created greater imbalance about 20% of the time.

ENSURING IMPLEMENTATION FIDELITY

Experience tells us that prevention programs can fail for one or both of two reasons: (a) the underlying model reflects a faulty diagnosis of the problem and its solution; or (b) the program is poorly implemented. Unfortunately, many evaluations do not provide information that allows us to distinguish between the two (Schaps, DiBartolo, Moskowitz, Palley, and Churgin 1981; Polich et al. 1984). To eliminate this ambiguity, we devoted substantial resources to ensuring that the program was faithfully implemented.

Training for both the adult teachers and the teen leaders emphasized the importance of adhering to the curriculum's content and delivery style. At the same time, the training illustrated that there was room to inject one's own personality and style into the teaching process.[9] Overall, the training was designed to provide the health educators with a firm grasp of the curriculum by explaining each lesson's rationale, modeling key activities, and providing plenty of teaching practice. Videotapes of actual classes offered a realistic view of student reactions to the curriculum, as well as examples of different teachers' ways of making the curriculum "their own."

To assess whether the curriculum was delivered as designed, we monitored 41% of the 2,300 classroom sessions scheduled during Grades 7 and 8. The monitors used standardized observation forms that had been tested and refined during the program's pilot test. Each form broke the lesson into its major activities (for example, introduction, video and discussions, role playing, and wrap-up); for each activity, the monitor assessed whether its key components had been covered and the extent to which it generated student interest and participation.

At the end of the lesson, the monitor completed three series of overall session assessments: (a) completeness and fidelity of program content (3 items); (b) fidelity of the teaching process—did the delivery process adhere to Project ALERT's teaching style? (3 items); and (c) an overall session evaluation (1 item). All evaluative judgments were rated on 5-point scales anchored by behavioral indicators that specified how to judge the end points. For example, indicators that the teacher's knowledge of the curriculum was inadequate (a score of 1) included omitting key session points and activities, constantly referring to the manual, reading the lesson verbatim, demonstrating an inability to answer questions for which the curriculum provided answers, and making incorrect statements about lesson goals. Superior knowledge (a score of five) was indexed by the opposite criteria.

Monitors were taught how to fill out the form for each session and how to resolve rating discrepancies. In addition, experienced staff accompanied

TABLE 4: Delivery of Project ALERT

Variable	Grade	
	Seventh	Eighth
Number of classes taught	1,666	639
Number of classes missed	0	0
Number of classes monitored	657	289
Percent of classes monitored where:		
All activities were presented	84	83
All activities except wrap-up were presented	93	92
No activities were rushed	60	59
No activities went on too long	81	80

new monitors on their initial assignments to further discuss and resolve rating discrepancies in actual classroom situations. To assess intermonitor reliability, we compared the variance between sessions with the variance within sessions using the analysis-of-variance (ANOVA) -based intraclass correlation (Winer 1971). The intraclass correlation estimates the reliability of a single monitor's ratings on the basis of the agreement observed among monitors rating the same session. For the overall session effectiveness rating (averaged across three components of the sessions), we obtained an estimated reliability for a single monitor's rating of 0.67. The estimated reliability of overall session effectiveness ratings for two monitors was 0.81.

Completeness of Program Delivery

These assessments indicated that Project ALERT was indeed delivered to participating students. Classroom logs and field coordinators' reports showed that every scheduled class was presented, although several had to be rescheduled to accommodate special school events or unanticipated occurrences such as blizzards. Over the 2-year period, the health educators and teen leaders delivered the seventh- and eighth-grade Project ALERT lessons in 2,305 classes, reaching over 5,000 students each year (see Table 4).

On the basis of monitoring assessments done in about 950 classes—41% of the lessons taught—we can also conclude that all curriculum activities were presented in the vast majority of classes. As Table 4 shows, the monitors recorded missed activities in 16% of the observed classes, but in half the cases, the omitted activity was the lesson wrap-up, which reviews the session's important points but introduces no new material. Hence, substantive material was omitted in less than 8% of the observed classes.

TABLE 5: Fidelity of Program Implementation

Variable	Percentage of Classes Rated Satisfactory or Above[a]	
	Seventh	Eighth
Established intended classroom environment	97	90
Conveyed intended material	96	95
How well it went overall	93	92

a. Rated ≥ 3.0 on a 5-point scale.

The majority of health educators also succeeded in teaching the curriculum lessons within the times suggested for each segment. In 80% of the classes, the health educators avoided stretching any activity too long (becoming repetitive and diminishing the time available for the rest of the lesson). In about 40%, however, the monitors judged that one or more activities had been rushed (foregoing or cutting off student comments in the interest of time).

Fidelity of Program Delivery

The fact that health educators met every scheduled class and generally covered all the lesson activities provides evidence that the entire curriculum was implemented. However, if they had substantially modified specific activity components or failed to follow the guidelines for facilitating student involvement, the underlying curriculum model would still have received a less than ideal test. Consequently, we also examined whether the health educator conveyed the intended substance and established a facilitative climate in the monitored classrooms. We then assessed the lesson's overall "gestalt"—the session blend of intended substance and delivery style. Table 5 shows the results.

In 90% or more of the monitored classes, the 17 observers felt that the health educator and teen leaders had established the intended classroom environment, had conveyed the intended substance, and had effectively combined substance and process. By this we mean that the classroom session received a rating of 3 or higher on each of three 5-point scales: how warm and open the classroom environment was, how well the session's purposes were conveyed, and how the session went overall. The midpoint on each scale marked a satisfactory performance. The mean ratings for each scale were as follows: 4.2 for classroom climate, 4.2 for substance, and 3.9 for overall gestalt.[10]

Although these data provide evidence that Project ALERT was implemented with fidelity, it is possible that the ratings were reliable but consistently positive. Other evidence indicates, however, that the monitors observed the lessons critically, noting both strengths and weaknesses in program delivery. For example, they faithfully recorded when curriculum activities were omitted or rushed, reporting that the latter occurred in 40% of the observed classes. In addition, they noted clear differences in overall lesson quality, giving lessons 1 and 6 significantly lower effectiveness ratings than the other six sessions. Both lessons had problems that explain the results: lesson 1 had more instances in which activities were rushed than any other, while feedback from the health educators indicated that lesson 6 lacked a consistent flow and contained an activity that failed to sufficiently engage students. Hence the monitors accurately noted problems in lesson content that impinged on the quality of curriculum delivery.

OBTAINING RELIABLE MEASURES OF DRUG USE

Our primary outcome measures came from student self-reports of whether, when, and how often they had used the target substances—alcohol, cigarettes, and marijuana. Because adolescent use or purchase of each of these substances is illegal, thoughtful observers frequently raise concerns about the truthfulness and accuracy of self-reported use. In fact, researchers have found student reports of drug use to be generally accurate (Single, Kandel, and Johnson 1975; Williams, Eng, Botvin, Hill, and Wynder 1979). Nevertheless, such fears underscore the need to reduce incentives for under- or overreporting drug use behavior, as well as the need to assess the extent to which either occurs. Additional threats to data accuracy and reliability can arise if students consciously try to "help" (or hinder) the experiment or if they have difficulty recalling past behavior or reading and understanding the questions.

Steps to Enhance Data Quality

Following well-tested survey research procedures, we took several steps to address these problems. To minimize intentional concealment (or bragging), we collected saliva samples from the students, explaining that marijuana and tobacco use could be detected in saliva. Such procedures have been shown to improve the accuracy of self-reported tobacco use (Bauman and Dent 1982; Murray, O'Connell, Schmid, and Perry 1987).

Before administering the surveys in the classroom, the data collectors described our measures for protecting data privacy and stressed the importance of telling the truth. We assured students that none of their teachers, principals, or parents would see their responses; we used numbers rather than names as identifiers; and we distributed the questionnaires in a group setting rather than in a face-to-face interview. These techniques have been shown to enhance students' confidence that their answers will remain anonymous (Johnston and O'Malley 1985; Mensch and Kandel 1988).

As a final measure to ensure the data's privacy, we obtained a certificate of confidentiality from the Department of Health and Human Services. Students also had the option of refusing to participate at any time.[11] To keep nonparticipating students from distracting others during data collection, we escorted nonparticipants to some other location designated by school officials.

The survey instruments were carefully designed to accommodate the reading levels and experience of seventh and eighth graders. To identify possible problems with wording or instructions, we pilot-tested various versions of the baseline questionnaire. These pretests allowed us to compare alternative versions for the frequency of problem indicators—such as missing data, internal inconsistencies, and student questions indicating confusion—and to select the most successful items for the final instrument. To improve response accuracy, we used objective, explicit anchors on the response scales. Thus responses indicated the number of times the students used drugs in specific time periods, rather than "sometimes" or "often." Inconsistent interpretations of subjective choices would have reduced the comparability of responses across both students and waves.

Finally, we used separate personnel for data collection and program delivery to reduce the possibility that students in treatment schools might try to provide the answers that their Project ALERT teachers would like (or dislike). Two-person teams collected the saliva samples and administered the written survey during regular classroom periods. These teams followed a scripted protocol that standardized procedures across classrooms and schools. To further ensure uniformity, the data collectors attended a 3-day training session before each wave of data collection.

Assessment of Data Quality

These procedures appear to have worked. Independent verification that the vast majority of students accurately reported recent cigarette use comes from the laboratory assessments of saliva cotinine. Using a cut-off point of

TABLE 6: Percentage of Students Who Retracted Previously Admitted Use

Data Collection Wave	Substance		
	Cigarettes	Alcohol	Marijuana
2	3.8	6.0	2.1
3	5.2	7.1	2.5
4	7.5	8.3	4.2

10 nanograms per milliliter (chosen to avoid identifying as smokers nonusing adolescents who were exposed to second-hand smoke at home), we identified only 17 students out of more than 6,500 as probable liars—that is, they denied recent use of tobacco, but their saliva tests argued otherwise. Further, of students identified as recent tobacco users based on the laboratory tests ($N = 257$), 95% also admitted to recent cigarette smoking or use of other tobacco products.

However, the saliva cotinine test may not detect students who have used a relatively small dosage, and offers no evidence about the accuracy of self-reported alcohol and marijuana use. For these reasons, we also assessed the consistency of student reports for all three substances—both within a questionnaire and over time.

The results of the inconsistency checks also support the argument that the great majority of students provided honest and accurate self-reports. Less than 5% of the students provided incomplete or inconsistent responses within questionnaires, rates that reflect those found in other studies of self-reported drug use by adolescents (Barnea, Rahav, and Teichman 1987; Single et al 1975). Moreover, the rates declined after we eliminated skip patterns from drug use item batteries, suggesting that the problems largely reflected confusion and carelessness, rather than deliberate misrepresentation.

Severe inconsistencies between questionnaires were also rare.[12] Although more than 40% of the students committed at least one longitudinal inconsistency over four data collection periods, over 95% of these discrepancies were minor: They involved inconsistent reporting of experimental use (most of which occurred long ago) or errors in placing previously reported use within the appropriate 12 month period.

Table 6 shows the frequency and type of retractions by wave (data collection period) and substance. At two of the three waves, they averaged less than 5% across substances; at the fourth, they averaged less than 7%. Throughout this period, there were comparatively fewer retractions for marijuana, largely because fewer students had used marijuana at each data

collection period. Only 5% of the total number of discrepancies reflected reversals of previously admitted *frequent* use (at least 11 times in the past year or 6 times in the past month). Overall, these reversals involved about 1% of the students. Thus the findings support the conclusions of earlier studies, that the majority of inconsistencies are committed by students who have used infrequently (Collins, Graham, Hansen, and Johnson 1985; Mensch and Kandel 1988; O'Malley, Bachman, and Johnston 1983; Single et al. 1975).

RULING OUT ALTERNATIVE EXPLANATIONS OF THE RESULTS

Although randomized assignment eliminates systematic differences among experimental groups, we argue that analysts should use regression methods to provide more accurate estimates of program effects and to help rule out alternative explanations of the results. Unadjusted estimates are unbiased unconditionally; that is, across many realizations of the experiment, true treatment effects will be estimated without bias. However, any one experiment assigns treatments only once. After the actual experimental groups are set, it is natural, and appropriate, to ask, "How much would we expect these groups to differ in outcome if the true treatment effect is nil?" If baseline differences forecast substantial outcome differences, then unadjusted estimates lose much of their credibility.

Even combined with other methods, random assignment cannot eliminate all pretreatment differences between experimental conditions. Some valuable predictors of outcomes do not become available until after the assignment process. And, no matter how well experimental groups are balanced when assigned, later attrition can produce differences among conditions. The potential imbalance is greatest when the number of experimental units is limited. Thus, we felt that it was essential to check for preintervention equivalence and to control for multiple characteristics frequently related to drug use. For similar reasons, analysis of covariance and other adjustment methods are commonly advocated for and used in clinical trials and other experiments (Cochran 1957; Armitage 1979; Lavori, Louis, Bailar, and Polansky 1983).

In Project ALERT's case, attrition rates and types of students lost did not differ significantly across experimental conditions, but attrition did yield small differences in the propensity to use the target substances (Ellickson and Bell 1990a). We adjusted for these differences when conducting both school-level and individual-level analyses, which generally produced similar results.

For the school-level analysis, the control variables were limited by the available degrees of freedom. They included district, percentage of minorities in the school, and the school mean of a composite variable summarizing 64 items measuring exposure to drug use, attitudes about the target substances, and various background items. Because it allowed more extensive controls and ultimately yielded more stable standard errors than the school-level analysis (see below), we reported results from the individual level analysis.[13] Each of these equations included demographic information, measures of baseline substance use, and the composite variable described above. For each specific substance, we also included intentions to use, offers, and a substance-specific scale of other items. The control variables were chosen to reflect existing knowledge about the antecedents of drug use.

The impact of these controls was to substantially lower our estimates of some program effects. For students who had not previously tried marijuana or cigarettes, for example, those in the treatment schools who tried marijuana within the next 12 months amounted to 6.6% and 6.9% in the two treatment groups before the adjustments for individual differences. After the adjustments, each proportion rose to 8.3%. These changes, while small in absolute terms, take on greater significance when translated into percent reductions from the control school rate of 12%: they shift that reduction from 43% or 45% down to 31%. Other pre- and postcontrol differences typically produced shifts in the same direction.[14] We conclude that using theoretically relevant control variables to adjust program effect estimates is essential when there are few experimental units—even if random assignment has been used.

A related issue involves accounting for the unit of assignment. Correlation of outcomes within schools violates the independence assumption of analysis of covariance, yielding a downward bias in standard-error estimates for school-level variables like treatment (Cornfield 1978; Biglan and Ary 1985). Thus analyses that fail to account for school effects risk sharply overstating the significance of estimated treatment effects. This criticism holds for many drug prevention studies. A common solution is to use school as the unit of analysis.

For the reasons noted above, we preferred estimates based on an analysis of individual data. To account for the unit of assignment, we estimated the size of the within-school correlations and computed factors for multiplying standard errors and dividing t statistics associated with treatment effects (Kish 1965).[15] To improve precision in estimating those factors, we combined information across samples and outcomes (Kish, Groves, and Krotki 1976). Our analysis indicated that the outcomes shared a common within-school correlation of only 0.0032 after controls for baseline use, district, and other

covariates. As a result, this procedure produced more precise standard-error estimates than did the school level analysis, allowing the use of *z* tests rather than *t* tests with few degrees of freedom.

The resulting adjustment factors were small, ranging from 1.04 to 1.11. Nevertheless, those adjustments substantially lowered the number of significant program effects. Twenty-four percent of the differences that were significant at the .05 level *before* the adjustment were dropped after recalculating the *t* statistics. Thus, when school is the unit of assignment, analysts must either use school as the unit of analysis or account appropriately for school effects.

LESSONS FROM PROJECT ALERT

Learning from the successes, as well as the failures of social experiments, helps the scientific community build on past experience and improve future work. What lessons emerge from our experience?

First, successfully putting together a broad test of this program entailed considerable effort and expense. This involved identifying potential candidates, contacting them by phone, and, when interest was indicated, visiting the community at least once to present the proposed study and curriculum in detail to administrators, principals, and curriculum coordinators. Our contact-to-success recruitment ratio was slightly more than two to one. That ratio reflected our dual requirements: (a) to obtain a diverse collection of communities and (b) to recruit communities and schools that would adhere to strict experimental requirements. Clearly, recruiting diverse schools within one large political entity is a more cost-effective approach. The tradeoff, of course, is reduced generalizability across geographic locations and community types. That calculation needs to be explicitly addressed in research proposals.

Second, communities and schools will abide by experimental conditions if they understand the importance of doing so and feel that their burden has been minimized within that constraint. Such cooperation is critical because the loss of a single school after assignments are complete threatens the credibility of the entire experiment. Once they had agreed to the requirements, the organizations that participated in Project ALERT fully cooperated with the study over several years. They helped us schedule (and reschedule) classes for program delivery and data collection, arrange for parent meetings, and supervise nonparticipating students. We, in turn, fulfilled our promises to them—to avoid publicity,[16] to make the curriculum available to successive

cohorts, to minimize data collection burdens on school personnel, and to act as troubleshooters when community and/or parental concerns arose. Setting forth clear agreements at the beginning of the study and following through on each obligation was central to developing and maintaining a cooperative relationship.

Third, for field trials in which the units of assignment are large and limited in number, producing a balanced experimental design requires more than simple random assignment. Although the best mix of blocking, balancing, and randomization will vary from one study to another, some effort to limit the preintervention differences pays a generous reward in terms of credibility. Even when an adjustment procedure is used in the analysis stage, a well-balanced assignment reduces the sensitivity of results to whether and how adjustments are made.

Fourth, it is possible to achieve faithful implementation while also generating teacher "ownership" of the curriculum. The literature on effective innovation highlights ownership and the ability to adapt an innovation to its particular organizational environment as key to successful implementation (Berman and McLaughlin 1978; Ellickson and Petersilia 1983; Sarason 1982). Thus we worried that our requirements for implementing the curriculum as designed might lead to teacher resistance and inadequate, non-facilitative program delivery. During training, therefore, we repeatedly encouraged teachers to inject their own styles into the delivery process while also following the written lessons. The judgments of seventeen different monitors indicate that the great majority of teachers both conveyed the curriculum's substance and created a facilitative classroom environment in which student participation was encouraged and rewarded.

Fifth, as shown by several other studies, careful efforts to encourage truthful and accurate reporting can yield highly reliable drug use self-reports. We used strict privacy guarantees, collection of physiological samples before survey administration, and pilot testing of questionnaire items. The overall results proved to be very successful, but we cannot judge which of these activities was the most or least effective in improving accuracy.

Sixth, the results demonstrate the importance of controlling for theoretically or empirically relevant variables when estimating the effects of a social experiment with small numbers of assignment units. By theoretically or empirically relevant, we mean that the variable has been theoretically postulated as a key determinant of the experimental outcome or empirically linked with that outcome. Despite the achievement of pretreatment equivalence in measured variables, attrition can produce substantial differences across conditions over time. By controlling for that occurrence in the variables most

strongly related to the behavior, the analyst produces more credible estimates of program effects. Our estimates of percent reductions in drug use dropped considerably after such controls were implemented. They might just as easily have risen. But the fact that they changed by substantial amounts provides strong evidence for the procedure we advocate.

Finally, with school as the unit of assignment, many of our objectives could not have been achieved in a moderate scale study of, say, 10 schools in two or three sites. Having 30 schools provided the flexibility needed to produce balanced experimental groups and to account reliably for within-school correlations. They also allowed for sample diversity and the resulting generalizability. In addition, comparing results in high versus low minority schools called for multiple districts (at least three per group) and multiple schools within each district.

This conclusion suggests why conducting experiments with schools and other institutions is often such a difficult undertaking. Funding and implementing large-scale experiments requires great time, expense and effort. Nonetheless, well-designed and well-implemented studies provide the best information available to decision makers. Thus they can be as valuable to the development of public policy as clinical trials are to medical science.

NOTES

1. The motivational elements of the curriculum reflect the theoretical perspectives of the Health Belief Model (Janz and Becker 1984) and the Self-efficacy Theory of Behavior Change (Bandura 1977). From the former, we drew the curriculum's focus on helping adolescents to identify the benefits of not using (as well as the personal costs of drug use) and to reduce perceived barriers to effective resistance. From the latter, we drew our emphasis on building resistance self-efficacy (one's belief that he or she can successfully resist pressures to use drugs).

2. The control schools were allowed to continue their regular drug curriculum, if they had one, as long as it did not focus on resistance skills and motivation. Four of the 10 control schools implemented drug prevention activities during grades 7 or 8; all four focused on providing information about drugs. To be judged effective, therefore, the curriculum had to exceed the performance of traditional programs delivered in control schools.

3. Five declined because they already had a drug prevention program they did not want to drop; six others felt they could not comply with other experimental requirements—random assignment of schools (one), collection of physiological samples (one), or weekly scheduling of all curriculum sessions during one semester (four).

4. Because resource constraints limited the total sample size to 30 schools, we did not include all the potential schools in some districts.

5. See U.S. Bureau of the Census (1983). The schools were selected before the 1990 census data became available.

6. This assertion is less likely to hold in large districts with schools that draw from highly diverse neighborhoods; hence the assignment procedure also took into account variation among

schools within districts (see the discussion that follows on the restriction of assignments to well-matched cells).

7. For example, simple random assignment within the blocking constraint might easily have produced 8 of 10 control schools that ranked below average in percent minority for their districts.

8. Although blocking by district is responsible for much of the improvement over simple randomized assignment, balancing across districts contributed as well. That is, the imbalance for most variables is less than would occur on average with blocking and randomization alone.

9. The interactive nature of curriculum activities, which allowed teachers to assess, and then build on, student knowledge and capabilities, enhanced the options for adapting the teaching process to the individual teacher's style and to different classroom environments.

10. Lessons taught by the teen leaders generally received higher ratings than those taught only by adults. However, the differences were small enough to be explained by variation among the adult teachers, most of whom taught in only one type of treatment school (adults only or adult plus teen leaders). When we controlled for that fact, none of the differences achieves statistical significance.

11. About 1% of students refused to fill out the surveys at each wave. At baseline, 8.5% of the students did not participate because they lacked parental consent. Analysis of total baseline nonresponse showed that it altered sample characteristics only slightly and had no effect on the balance across treatment conditions. Students who refused to participate or were absent had higher rates of substance use than respondents, but those whose parents refused for them tended to be at similar or less risk for future drug use (Bell, Gareleck, and Ellickson 1990).

12. Longitudinal inconsistencies occurred when: (a) students retracted previously admitted lifetime use; (b) they admitted lifetime use but the new admission occurred within a time period covered by previous denials; or (c) they denied use in the past year after having admitted to current use three to nine months earlier. The third type was most frequent, reflecting problems in placing use within the precise time period.

13. The individual-level analysis produced slightly more conservative estimates of program effects than the school-level analysis.

14. This tendency was weaker in the school-level analysis, which tended to favor the program a bit more strongly.

15. Another example of correcting individual-level analyses for within-cluster correlation is provided by the Health Insurance Experiment, which assigned families to insurance plans. That study used a nonparametric approach that yields corrections similar to those from a random effects model (Brook et al. 1984; Manning et al. 1987).

16. Several districts feared that publicity about their participation in the experiment would tarnish them with a drug abuse image. Hence we guaranteed district, as well as student anonymity, and did not seek or contribute to media coverage in the participating communities.

REFERENCES

Armitage, P. 1979. "The Analysis of Data From Clinical Trials." *The Statistician* 28:171-83.

Bandura, A. 1977. "Self-Efficacy: Toward a Unifying Theory of Behavioral Change." *Psychological Review* 84:191-215.

Bandura, A. 1984. "Social Learning Theory." In *Conference Report: Unhealthful Risk-Taking Behaviors in Adolescence*, edited by N. Maccoby, C. Albright, J. Farquhar, G. Elliot, C. Taylor, and A. Mortimer. Stanford University.

Barnea, Z., G. Rahav, and M. Teichman. 1987. "The Reliability and Consistency of Self-Reports on Substance Use in a Longitudinal Study." *British Journal of Addictions* 82:891-98.

Bauman, K. and C. Dent. 1982. "Influence of an Objective Measure on Self-Reports of Behavior." *Journal of Applied Psychology* 67:623-28.

Bell, R. M., C. Gareleck, and P. L. Ellickson. 1990. *Baseline Nonresponse in Project ALERT: Does It Matter?*, N-3291-CHF. Santa Monica, CA: RAND.

Berman, P. and M. W. McLaughlin. 1978. *Federal Programs Supporting Educational Change: Implementing and Sustaining Innovations*, R-1589/8-HEW. Santa Monica, CA: RAND.

Biglan, A. and D. Ary. 1985. "Methodological Issues in Research on Smoking Prevention." In *Prevention Research: Deterring Drug Abuse Among Children and Adolescents*, edited by C. Bell and R. Battjes. Rockville, MD: NIDA.

Biglan, A., R. Glasglow, D. Ary, R. Thompson, H. Severson, E. Lichtenstein, W. Weissman, C. Faller, and C. Gallison. 1987. "How Generalizable Are the Effects of Smoking Prevention Programs? Refusal Skills Training and Parent Messages in a Teacher-Administered Program." *Journal of Behavioral Medicine* 10:613-28.

Botvin, G. and T. Wills. 1985. "Personal and Social Skills Training: Cognitive-Behavioral Approaches to Substance Abuse Prevention." In *Prevention Research: Deterring Drug Abuse Among Children and Adolescents*, edited by C. Bell and R. Battjes. Rockville, MD: NIDA.

Brook, R. H., J. E. Ware, Jr., W. H. Rogers, E. B. Keeler, A. R. Davies, C. D. Sherbourne, G. A. Goldberg, K. N. Lohr, P. Camp, and J. P. Newhouse. 1984. *The Effect of Coinsurance on the Health of Adults*, R-3055-HHS. Santa Monica, CA: RAND.

Cochran, W. G. 1957. "Analysis of Covariance: It's Nature and Uses." *Biometrics* 13:261-81.

Collins, L. M., J. W. Graham, W. B. Hansen, and C. A. Johnson. 1985. "Agreement Between Retrospective Accounts of Substance Use and Earlier Reported Substance Use." *Applied Psychological Measurement* 9:301-9.

Cornfield, J. 1978. "Randomization by Group: A Formal Analysis." *American Journal of Epidemiology* 108:100-102.

Ellickson, P. L. and R. M. Bell. 1990a. "Drug Prevention in Junior High: A Multi-Site Longitudinal Test." *Science* 247:1299-1305.

———. 1990b. *Prospects for Preventing Drug Use Among Young Adolescents*, R-3896-CHF. Santa Monica, CA: RAND.

Ellickson, P. L., R. M. Bell, M. A. Thomas, A. E. Robyn, and G. L. Zellman. 1988. *Designing and Implementing Project ALERT: A Smoking and Drug Prevention Experiment*, R-3754-CHF. Santa Monica, CA: RAND.

Ellickson, P. and J. Petersilia. 1983. *Implementing New Ideas in Criminal Justice*, R-2929-NIJ. Santa Monica, CA: RAND.

Flay, B. 1985. "What We Know About the Social Influences Approach to Smoking Prevention: Review and Recommendations." In *Prevention Research: Deterring Drug Abuse Among Children and Adolescents*, edited by C. Bell and R. Battjes. Rockville, MD: NIDA.

Goodstadt, M. 1986. "School-Based Education in North America: What Is Wrong? What Can Be Done?" *Journal of School Health* 56:278-80.

Janz, N. K. and M. H. Becker. 1984. "The Health Belief Model: A Decade Later." *Health Education Quarterly* 11:1-47.

Johnston, L. D. and P. M. O'Malley. 1985. "Issues of Validity and Population Coverage in Student Surveys of Drug Use." In *Self-Report Methods of Estimating Drug Use: Meeting Current Challenges to Validity*, NIDA Research Monograph 57. Rockville, MD: NIDA.

Kish, L. 1965. *Survey Sampling*. New York: Wiley.

Kish, L., R. M. Groves, and K. P. Krotki. 1976. *Sampling Errors for Fertility Surveys*. Ann Arbor: University of Michigan.

Lavori, P. W., T. A. Louis, J. C. Bailar III, and M. Polansky. 1983. "Designs for Experiments—Parallel Comparisons of Treatment." *New England Journal of Medicine* 309:1291-97.

Manning, W. G., J. P. Newhouse, N. Duan, E. B. Keeler, A. Leibowitz, and M. S. Marquis. 1987. "Health Insurance and the Demand for Medical Care: Evidence From a Randomized Experiment." *American Economic Review* 77:251-77.

Mensch, B. S. and D. B. Kandel. 1988. "Underreporting of Substance Use in a National Longitudinal Youth Cohort: Individual and Interviewer Effects." *Public Opinion Quarterly* 52:100-124.

Moskowitz, J. 1989. "The Primary Prevention of Alcohol Problems: A Critical Review of the Research Literature." *Journal of Studies on Alcohol* 50:54-88.

Murray, D. M., C. M. O'Connell, L. A. Schmid, and C. L. Perry. 1987. "The Validity of Smoking Self-Reports by Adolescents: A Reexamination of the Bogus Pipeline Procedure." *Addictive Behaviors 12:7-15*.

O'Malley, P. M., J. G. Bachman, and L. D. Johnston. 1983. "Reliability and Consistency in Self-Reports of Drug Use," *International Journal of the Addictions* 18:805-24.

Polich, J. M., P. L. Ellickson, P. Reuter, and J. P. Kahan. 1984. *Strategies for Controlling Adolescent Drug Use*, R-3076-CHF. Santa Monica, CA: RAND.

Sarason, S. B. 1982. *The Culture of the School and the Problem of Change*, 2nd ed. Boston, MA: Allyn & Bacon.

Schaps, E., R. DiBartolo, J. Moskowitz, C. Palley, and S. Churgin. 1981. "A Review of 127 Drug Abuse Prevention Program Evaluations." *Journal of Drug Issues* 11:17-43.

Single, E., D. Kandel, and B. D. Johnson. 1975. "The Reliability and Validity of Drug Use Responses in a Large Scale Longitudinal Survey." *Journal of Drug Issues* 5:426-43.

U.S. Bureau of the Census. 1983. *1980 Census of Population*, Vol. 1, Chap. C (PC-80-1-C). Washington, DC: Author.

Williams, C. L., A. Eng, G. J. Botvin, P. Hill, and E. L. Wynder. 1979. "Validation of Students' Self-Reported Cigarette Smoking Status With Plasma Cotinine Levels." *American Journal of Public Health* 69:1272-74.

Winer, B. 1971. *Statistical Principles in Experimental Design*. New York: McGraw-Hill.

RAND/R-4173-CHF