Evaluation Review
Volume 31 Number 1
February 2007 43-74
© 2007 Sage Publications
10.1177/0193841X06287188
http://erx.sagepub.com
hosted at
http://online.sagepub.com

The Devil Is in the Details

Examining the Evidence for "Proven" School-Based Drug Abuse Prevention Programs

Allison Gruner Gandhi
American Institutes for Research
Erin Murphy-Graham
University of California, Berkeley
Anthony Petrosino
WestEd
Sara Schwartz Chrismer
Carol H. Weiss
Harvard Graduate School of Education

In an effort to promote evidence-based practice, government officials, researchers, and program developers have developed lists of model programs in the prevention field. This article reviews the evidence used by seven best-practice lists to select five model prevention programs. The authors' examination of this research raises questions about the process used to identify and publicize programs as successful. They found limited evidence showing substantial impact on drug use behavior at posttest, with very few studies showing substantial impact at longer follow-ups. The authors advocate additional long-term follow-up studies and conclude by suggesting changes in the procedures for developing best-practice lists.

Keywords: *drug abuse prevention; evaluation; evidence-based policy; model programs; education*

Evidence-based practice has become a common term in many fields, including substance abuse prevention. The idea of basing practice and policy decisions on the best available evidence has obvious appeal. It is better than ignoring research and evaluation evidence, a practice that has received much scholarly attention (Weiss 1977, 1989; Birkeland, Murphy-Graham, and Weiss 2005; Husen 1994; Reimers and McGinn 1997). Using evidence to inform practice and policy is a rational way to make decisions about which programs and practices to adopt to prevent substance abuse.

The U.S. Department of Education (ED) moved in this direction in the field of school-based drug abuse prevention in 2001. The Safe and Drug-Free Schools and Communities program of ED set out its “principles of effectiveness,” a set of criteria that it expected the substance abuse programs that it funded to meet. The principles included such items as an assessment of the incidence of the problems of drug use and violence and their consequences and scientific evidence on the success of the program to be adopted in reducing drug use and violence. In 2002, the principles of effectiveness were incorporated into the No Child Left Behind education law. Schools can no longer receive federal funding for programs that are not evidence based. The intention of ED was obviously to encourage rational choice of school-based programs.

When ED put these principles of effectiveness into effect as a basis for providing funds to school districts, the districts were faced with the task of deciding what evidence is and what it showed. In the drug abuse field, there are hundreds of studies of prevention programs of varying quality and effects. To help school districts make informed choices, ED released the *List of Exemplary and Promising Prevention Programs.* The list was compiled on the basis of program evaluations with the help of a panel of eminent prevention researchers. The initial *List* contained 9 “exemplary” and 33 “promising” programs that were said to meet the principles of effectiveness. Henceforth, districts were to spend federal money on programs such of those on the *List.* Officials at ED stated that they did not require that districts choose a program from the *List*, but that is what most districts believed (Weiss, Murphy-Graham, and Birkeland 2005). In all events, it was much simpler for a district to take a program that ED had officially sanctioned than to collect and analyze its own evidence about programs.

Many other federal and academic agencies have produced “lists” that identify scientifically proven school-based drug prevention programs. Six prominent lists other than ED’s list are (a) *SAMHSA National Registry of Effective Programs*, (b) *Drug Strategies Guide to Effective School-Based Drug Prevention Curricula*, (c) *University of Colorado Blueprints for*

Authors’ Note: The Robert Wood Johnson Foundation Grant 039080 to the Study on Decisions in Education at the Harvard Graduate School of Education provided funding for this research. We thank Sarah Illingworth for research assistance, Dennis Gorman, and previous anonymous reviewers for comments. All statements and interpretations, however, are those of the authors and do not represent the official position of any institution, agency, or other individual.

Violence Prevention, (d) *University of Maryland Report to the Congress on Crime Prevention*, (e) *National Institute of Drug Abuse Guide to Effective Drug Prevention Programs*, and (f) *Youth Violence: A Report of the Surgeon General.* An earlier publication from our study summarizes these documents (Petrosino 2003). Some of these lists influence federal funding decisions at agencies such as the Substance Abuse and Mental Health Services Administration and the Office of Juvenile Justice and Delinquency Prevention.

This article considers the evidence on which these lists are based. The purpose of these lists and guides is to help decision makers, at both the federal and local levels, choose programs that are supported by the best available evidence (Petrosino 2003). Basing program selection and funding on good data is a laudable undertaking. We, like other social scientists, have long advocated such behavior. Our concern is how strong the evidence base is for providing advice to school districts and other agencies. As we have learned in many contexts, the devil is in the details. Our analysis indicates that there is a paucity of evidence that any of the "proven" drug prevention programs will have long-term effects and that the lists give a misleading aura of certainty to their recommendations.

Method

We looked at seven prominent lists to see which prevention programs were frequently named (the ED list and the six others listed above). We identified five programs that appeared on most of the lists: (a) Life Skills Training, (b) Midwestern Prevention Program, (c) Project ALERT, (d) Project Northland, and (e) CASASTART. Table 1 shows which programs were identified as either "model" or "promising" on each of the seven lists.

The seven lists in Table 1 use a variety of criteria to determine what evidence should be considered and how much of that evidence is necessary before determining whether a program is "effective" or "promising" (see Table 2). But there are two main commonalities: (a) the lists prioritize internal validity and include as valid evidence either randomized experiments or rigorous quasi-experimental designs (e.g., matching was used to form the comparison group) and (b) the lists use statistically significant positive findings as evidence of "success." For the most part, one rigorous evaluation with a positive outcome would result in a program's being classified as "promising," and two or more evaluations would result in an "effective" or "model" designation.

Each program was designated as "exemplary" or "model" on at least one of the lists, even though it might have received a "promising" designation

Table 1
Summary of Lists and Programs Selected for Review

	Drug Strategies[a]	Blueprints	Maryland Report[b]	NIDA Guide	SAMHSA NREPP	U.S. Education List	Youth Violence: A Report of the Surgeon General
Life Skills Training	A	Model	✓	Universal	Model	Exemplary	Model
Project ALERT	A	Promising[c]	✓	Universal	Model	Exemplary	
CASASTART		Promising			Model	Exemplary	Promising
Project Northland	A	Promising			Model	Exemplary	
Midwestern Prevention Program	A	Model		Universal		Promising	Model

Note: NIDA = National Institute on Drug Abuse; SAMHSA = Substance Abuse and Mental Health Services Administration; NREPP = National Registry of Evidence-Based Programs and Practices; CASASTART = Striving Together to Achieve Rewarding Tomorrows (formerly the Children at Risk [CAR] program).

a. An “A” grade means that the overall program quality is very good.

b. Although the University of Maryland report does not rank programs, it is notable in its five-point grading of evidence according to methodology. Evidence reviewed for Project ALERT received a scientific methods score of 5. Three Life Skills reports reviewed received scientific methods scores of 3, 4, and 5.

c. Project ALERT was identified as a “promising” program by Blueprints for the first time in 2004. However, no fact sheet is available identifying the studies Blueprints used to determine this classification. Therefore, in our review of Project ALERT, we consider only the studies listed by the other lists.

Table 2
Best Practice Lists: Standard for Acceptable Evidence and Amount of Evidence Needed to Designate a Program as Effective

Guide	Standard for Evidence	Evidence for Standard
Drug Strategies: Making the Grade	The main focus of the document is to assess curricular quality. Evaluation evidence is then presented in the summary. An "acceptable evaluation" is that which is extensive, is published in a peer-reviewed journal, includes pretest and posttest measures in a control group design, and measures for reductions in drug use.	All of the "acceptable evidence" is presented in summary form so that decision makers can make an informed choice. Statistically significant findings are relied on to indicate "positive" impact.
Blueprints	Randomized experiment or a matched quasi-experimental study showing a positive impact on youth violence, delinquency, or drug use at least 1 year later with no evidence effects dissipated.	A program is a Blueprint for Violence Prevention if a replication study meeting the standard for evidence is reported. A "promising program" does not include a report of replication.
Maryland Report	Studies are categorized according to methodology (largely influenced by internal validity) into five levels ranging from Level 1 studies (simple correlations) to Level 5 studies (well-implemented randomized experiments).	A program "works" if it has at least two Level 3 studies (comparison group quasi-experiments) showing positive and statistically significant results and preponderance of all other evidence in favor of a positive impact.
National Institute on Drug Abuse (NIDA) Guide	The focus of the guide is on scientific principles of prevention. One of the principles is ongoing evaluation of the program. (Because these are NIDA-funded programs, they almost always include randomized experimental or well-controlled quasi-experimental designs.)	Research-based NIDA programs are those that include an ongoing program of evaluation, reporting positive results over a "reasonable period of time."
Substance Abuse and Mental Health Services Administration	Center for Substance Abuse Prevention (CSAP) prevention expert panel reviews application packages (submitted by program	Programs are categorized as effective and placed into NREPP if they score 4.0 or higher on both "integrity" and "utility."

(continued)

Table 2 (continued)

Guide	Standard for Evidence	Evidence for Standard
National Registry of Evidence-Based Programs and Practices (NREPP)	developers or by others) against 15 criteria (see Appendix A). The criteria include characteristics of the program and the quality of the evaluations. Two criteria, “integrity” (scientific rigor) and “utility” (generalizability), are scored using a 5-point scale and are used to rank the programs.	Effective programs in which the program developer has agreed to disseminate and provide technical assistance through CSAP are “model programs.”
U.S. Department of Education (ED) List	ED expert panel reviews application packages (submitted by program developers or by others) against seven criteria in Appendix B. The criteria include characteristics of the program and quality of the evaluations. Eligibility for the list is dependent on at least one evaluation reporting positive results on drug use, violent behavior, or another problem for 1 year or longer beyond baseline.	Exemplary status is obtained by a rating of 3 on evidence of efficacy and a rating of at least 3 on three other criteria (see Appendix B). The other three criteria must receive no lower than a rating of 2. Promising programs do not meet these criteria.
Youth Violence: A Report of the Surgeon General	Researchers relied on existing best-practice lists to identify model, promising, or ineffective violence prevention programs. Evaluations had to use experimental or rigorous quasi-experimental design. Level I studies included violence or serious delinquency outcomes; if they included measures of known risk factors for violence (e.g., drug use), they were classified as Level II studies.	Model programs are those that show significant positive impact and are replicated with demonstrated and sustained effects. Promising programs demonstrate either replication or sustainability of effects but not both. Ineffective programs report null or negative effects on violence or known risk factors, with replication and a preponderance of evidence suggesting the program is ineffective or harmful.

Source: Adapted from Petrosino (2003).

or been omitted on one or more of the other lists. We reviewed the evaluations that each of the lists cited as evidence for the effectiveness of these programs. Because our objective was to review the evidence used in creating the lists, we did not take into account reports or studies published recently—after the lists were issued—or otherwise not cited in the lists. We acknowledge that subsequent to the publication of these lists, additional studies have been released that may provide a stronger base of evidence for the effectiveness of programs. However, our purpose here is not to systematically review the prevention literature. Rather, we examine the evidence used to determine the five programs' designations as "exemplary," "model," or "promising." It is these lists, not individual studies, that school districts turn to when making decisions about program implementation (Weiss, Murphy-Graham, and Birkeland 2005). One of the issues we discuss later in the article is that lists should periodically be updated to reflect new studies.

The Department of Education did not offer any evidence for the programs on its list. In March 2003, a member of our staff asked the federal coordinator of the Safe, Discipline, and Drug-Free Expert Panel for a list of the evaluations that had been used to draw up the *List* for the Safe and Drug-Free Schools program. She was told that the department was not making references to particular evaluations available and that we should check the Web sites of individual programs.

For the six other lists, we attempted to acquire copies of each evaluation study that was cited. We found in this process that some lists did not clearly differentiate between evaluation studies and references to descriptive or background material on programs. We reviewed only empirical studies, coded them, and entered the information into a central database. This database included information about the study participants, the evaluation design, the outcome measures, the evaluation findings, study limitations, and any additional notes of interest. Using this database, we reviewed the evaluation evidence for each program in an attempt to summarize (a) the quality of evaluation studies that were cited and (b) the overall findings about the program that emerged.

Review of Programs: Findings

Life Skills Training (LST)

LST is a school-based drug prevention program developed in the late 1970s by Gilbert Botvin. The program is designed to work with elementary and middle/junior high school students over the course of a 2- or 3-year period.

All seven lists that we reviewed identified LST as an exemplary or model program. The lists cited a total of 28 studies in support of these designations. Seven of these citations were conceptual or background articles.

We were unable to locate 3 empirical studies via the library or the Internet (Botvin 1994; Botvin and Cardwell 1992; Botvin, Griffin, Paul, et al. 2001), leaving 19 empirical studies (18 by Dr. Botvin, 1 by Dr. Spoth) that form the basis of this summary. Of these 19, 18 were authored by Dr. Botvin and colleagues.

Although most of these studies collect data from different samples of students, there are several that in fact report findings from the same data but focus on different time points, subgroups, or outcomes. These studies as a whole look at a variety of outcomes, including knowledge, attitudes, and behavior for cigarettes, alcohol, marijuana, and other drugs. All the studies randomly assigned schools to intervention or control conditions.

Data Set 1. One set of data looked at the results of LST after implementation and again at 1-year follow-up, for approximately 1,300 seventh-grade students from 10 suburban New York junior high schools. The schools were randomly assigned to one of three conditions: (a) LST implemented by a teacher, (b) LST implemented by an older peer, and (c) a control group. After 1 year, results showed that students in the peer-led program were less likely to engage in substance abuse and had greater drinking knowledge and better overall attitudes than those in the teacher-led program and in the control condition. The only exception was in smoking and marijuana knowledge, which were greater in both the peer-led and teacher-led program than in the control condition (Botvin, Baker, Renick, et al. 1984). After that year, researchers created two additional conditions: a peer-led series of booster sessions and a teacher-led series of booster sessions. Results showed that students in the peer booster condition were less likely to smoke cigarettes and marijuana than were students in all other conditions. Interestingly, the results also showed that for drinking, students in every condition (including the control) were less likely to drink than those in the teacher booster condition. The peer booster condition also showed the best results in terms of knowledge about substance abuse and attitudes toward smoking. The teacher booster condition, however, had the best results when it came to attitudes about drinking and marijuana (Botvin, Baker, Filazzola, et al. 1990).

Data Set 2. A second set of reports describes a long-term follow-up of nearly 6,000 seventh-grade students from 56 schools in three regions in New York State. These schools were randomly assigned to three groups: (a) LST in which training was delivered by workshop with feedback, (b) LST in which training was delivered by video with no feedback, and (c) a control group, which did not use the LST curriculum at all. One report looked at outcomes after project implementation (at the end of 9th grade), and the other looked at outcomes at the end of 12th grade, 3 years after the end of project implementation. At posttest, results for both versions of LST showed positive effects on substance abuse behavior, knowledge, and attitudes (Botvin, Baker, Dusenbury, et al. 1990). At 3-year follow-up, results showed that students in the intervention groups were less likely to smoke regularly than were those in the control group, but there were no differences between groups with respect to regular drinking and marijuana use. The subsample of students that received LST with a "high degree of fidelity," however, demonstrated lower drinking and marijuana use (Botvin, Baker, et al. 1995). Another study looked at an even smaller subsample of students at the end of the 12th grade and focused on illicit drug use. This study found that there were no statistically significant differences between groups in marijuana use, cocaine use, inhalants use, and nonmedical pill use. The study did find that students who had been in the intervention group were less likely at the end of 12th grade than those in the control group to have engaged in heroin and hallucinogen use (Botvin et al. 2000).

Data Set 3. Another article reported a follow-up using data from approximately 750 predominantly minority seventh-grade students from six New York City public schools. The schools were matched according to demographics and then randomly assigned to one of three conditions: (a) broad-spectrum LST, (b) culturally focused intervention, and (c) information-only control. One report looked at outcomes at posttest (after 4 months) and found that the culturally focused curriculum, when compared to the control condition (no LST curriculum at all), had significant effects in reducing intentions to drink beer, improving substance abuse attitudes, and reducing risk taking. It was not superior to the regular LST curriculum for any outcome (Botvin, Schinke, et al. 1994). Two years later, however, the culturally focused curriculum did have stronger effects among minority youth than the regular LST curriculum did when it came to reducing drinking frequency, amount per occasion, and drunkenness (Botvin, Schnike, Epstein, et al. 1995).

Data Set 4. A fourth set of data looked at approximately 5,200 seventh-grade students in 29 New York City public schools. These schools were blocked by high, medium, or low smoking prevalence and then randomly assigned to intervention or control conditions within blocks. At the end of the project implementation period (after eighth grade), students in the intervention condition smoked less frequently, drank less, and used marijuana and inhalants less than their counterparts in the control condition did.

The intervention students also showed greater knowledge about drugs than the control students, less intention to use drugs in the future, stronger antidrug attitudes, lower normative expectations regarding peer and adult drug use, greater efficacy at refusal, and lesser tendency to take risks (Botvin, Griffin, Diaz, and Ifill-Williams 2001a). Another report showed similar results for a subsample of girls from the same study (Botvin et al. 1999). At follow-up, a year after the program ended (at the end of ninth grade), a study showed that students in the intervention group were less likely to engage in binge drinking and had lower normative expectations regarding peer drinking than their counterparts in the control condition did. However, there was no difference in drinking knowledge or prodrinking attitudes, and no other outcomes were reported (Botvin, Griffin, Diaz, and Ifill-Williams 2001b).

Other LST studies. The remainder of the LST reports looked at short-term outcomes for the program with various samples of students (ranging from 200 to 3,500, mostly from public schools in the New York/New Jersey area), all of which used random assignment in their evaluation design. Several looked at outcomes for samples of minority students, finding positive results in terms of reducing smoking (Botvin, Batson, et al. 1989; Botvin, Dunesbury, et al. 1989; Botvin et al. 1992) and other drug use (Botvin et al. 1997). Some looked at smoking alone and found positive results (Botvin and Eng 1982; Botvin, Eng, and Williams 1980), and some looked at drinking alone and found positive results (Botvin, Baker, Botvin, et al. 1984). One report looked at the effect of scheduling format and found that a mini course format was more effective at reducing smoking than one in which the curriculum was integrated into regular course content (Botvin, Renick, and Baker 1983). Finally, the one study mentioned earlier that looked at LST alone and in combination with Strengthening Families showed that both versions were successful at delaying the initiation of substance use (Spoth et al. 2002).

Midwestern Prevention Project (MPP)

The MPP, also known as Project STAR, is a school and community-based drug prevention program developed by a group of researchers at the University of Southern California. The program is designed to work with sixth- and seventh-grade students over the course of 5 years.

Five lists identified the MPP as an effective program, citing 11 evaluations across them. Of these, 3 are conceptual/background articles, leaving 8 empirical studies. These 8 evaluations look at one large data set of students in 50 Kansas City schools. This data set consists of one sample of students in 42 junior high or middle schools and one panel sample of all students in 8 junior high or middle schools. These schools were each randomly assigned to experimental and control conditions. Three reports look at data from 1,600 sixth- and seventh-grade students from 8 Kansas City schools, reporting outcomes 1 and 3 years after the program ended. One study looked at results 1 year after the program ended and found that the relative odds of using cigarettes were less for students who had been in the treatment group than for those who had been in the control group. This same study found no evidence of a program effect on alcohol use and mixed evidence of a program effect on marijuana use (Dwyer et al. 1989). Another study focused only on cigarette use 1 year after the program ended, finding lower rates for students in the treatment group than controls (Pentz, MacKinnon, Flay, et al. 1989). The 3-year report found that students in the intervention group were less likely than those in the control group to have used cigarettes or marijuana during the past 30 days (Johnson et al. 1990).

The remaining five reports used data from approximately 5,400 sixth- and seventh-grade students in schools that were randomly assigned to the intervention and to a delayed-intervention control. One report looks solely at substance abuse outcomes and reports that students in the intervention group were less likely to report smoking cigarettes, using alcohol, or smoking marijuana during the past month and during the past week (Pentz, Dwyer, et al. 1989). Another report shows that 1 year after the program ended, students in the intervention group had increased their use of cigarettes, alcohol, and marijuana at about half the rate of their control group peers (Pentz, Johnson, et al. 1989). A third report showed that the percentage of students in the intervention group who reported smoking in the past month, week, and day was lower than for students in the control group 1 year and 2 years after the program ended (Pentz, MacKinnon, Dwyer, et al. 1989).

Another article reports that students who received the intervention had more positive attitudes and greater knowledge about substance abuse (MacKinnon et al. 1991). Yet another study compared a high implementation group, low implementation group, and a control group (Pentz et al. 1990). The results show that the high implementation group was less likely than the control group to have smoked cigarettes, used alcohol, or smoked marijuana during the past month and during the past week. There were no significant results for the low implementation group (Pentz et al. 1990).

Project ALERT

Project ALERT is a school-based drug prevention program that was developed in the mid-1980s at RAND Corporation, principally by Phyllis Ellickson. The program works with middle-school students, grades 6 to 8, and is taught by trained teachers.

Four lists identified Project ALERT as an exemplary or model program, citing five evaluations across them. These five evaluations use the same data—6,500 seventh-grade students from 30 schools drawn from eight urban, rural, and suburban communities in California and Oregon—but report outcomes at different follow-up periods. Schools were randomly assigned to one of three conditions: (a) ALERT taught by teacher alone, (b) ALERT delivered by teachers assisted by teen leaders, and (c) a control group. Results were reported for each substance (cigarettes, alcohol, and marijuana) by the level of students' prior use (nonusers, experimenters, and users). Three reports provide results at the end of project implementation, that is, at the end of 8th grade (Ellickson and Bell 1990a, 1990b; Ellickson, Bell, and Harrison 1993), one report presents results 1 year after the end of the project in 9th grade (Bell, Ellickson, and Harrison 1993), and the other gives results 4 years after the end of the project in 12th grade (Ellickson, Bell, and McGuigan 1993).

Of the three reports that present results at the end of project implementation in eighth grade, two report the same findings. At the end of project implementation, students classified as "experimenters" in both Project ALERT programs were less likely than students in the control group to have smoked in the past month or week and were more likely to have quit using cigarettes altogether. The opposite effect, however, was shown for "users." Users, in the teen-assisted Project ALERT program, were more likely than those in the control group to have used cigarettes in the past month or week. No significant findings were reported after 15 months related to alcohol use. Nonusers of marijuana and cigarettes in both Project ALERT programs were less likely than those in the control group to report ever having used the drug at the end of project implementation (Ellickson and Bell 1990a, 1990b).[1]

Another report on outcomes after eighth grade looks at measures of knowledge and attitudes. The report divides students into the same three risk levels (users, experimenters, and nonusers) and also reports results for the full sample. In general, students in the Project ALERT conditions had greater knowledge about the risks associated with substance abuse as well as more positive attitudes toward prevention than those in the control condition did (Ellickson, Bell, and Harrison 1993).

One report on the effects of Project ALERT 1 year after the end of project implementation showed that students in the teen-assisted Project ALERT had more positive attitudes than those in the control group. However, there were several instances in which students in the adult-led Project ALERT had less positive attitudes than those in the control group did. For example, nonusers in the adult-led Project ALERT were more likely than nonusers in the control group to believe that there was no risk from occasional alcohol use, and experimenters in the adult-led group were more likely to report that they were likely to use alcohol during the next 6 months. More experimenters and users in the adult-led group reported that their friends tolerated alcohol use. More users in the adult-led group reported that it was difficult to resist alcohol at a party and marijuana on a date. Also, 1 year after the program ended, nonusers in the adult-led group were more likely to report monthly use of marijuana (Bell, Ellickson, and Harrison 1993). Thus, Project ALERT seems to have some negative effects on attitudes when led by an adult. As for behavior, the teen-assisted program showed some reduction in cigarette use for nonusers and experimenters, but users in the teen-assisted program were more likely than those in the control group to report weekly and daily cigarette use.

Finally, one report based on a survey of the students when they were in the 12th grade found virtually no statistically significant findings related to substance abuse behavior; in other words, students in the adult-led, teen-assisted, and control groups seemed to perform similarly. The only exceptions were that nonusers in the adult-led program were less likely than those in the control group to report weekly alcohol use, and users in the adult-led program were more likely than those in the control group to report weekly cigarette use (Ellickson, Bell, and McGuigan 1993).

Project Northland

Project Northland is a school- and community-based alcohol use prevention program developed in 1990 by Cheryl Perry and Carolyn Williams. The program works with middle/junior high school students between sixth and eighth grade.

Four lists identified Project Northland as a promising or exemplary program, citing six evaluations across them. Two of these are background/conceptual pieces, and one is not an evaluation of Project Northland but an unnamed alcohol education intervention in four countries sponsored by the World Health Organization (Perry et al. 1989; Komro et al. 1994; Perry et al. 1993). This left three empirical studies for our review.

Of the remaining three evaluations, two described results from a 3-year implementation of Project Northland, using the same set of data from 2,300 students who were in the sixth grade at the beginning of the implementation period. Ten districts in northeast Minnesota were randomly assigned to the intervention condition and 10 to control. Some control districts used other substance abuse programs, such as D.A.R.E., and some eventually used Project Northland, but not until after the implementation period. One article reported results after 1 year of implementation (Williams et al. 1995) and the other after 3 years of implementation (Perry et al. 1996). We found no reports of long-term follow-up after implementation was complete.

After 1 year of implementation, there were no significant effects on substance use. The report showed that students in the intervention group had more knowledge and positive attitudes on statements including "Beer, alcohol and wine advertisements try to get people my age to think it's cool to drink" and "My parents talk with me about problems alcohol can cause young people." Students, however, were also more likely to agree with the statement "Drinking alcohol gives people energy" and less likely to agree that "People my age who drink alcohol are likely to get into trouble with the police" (Williams et al. 1995).

After 3 years of implementation, students in the intervention group were less likely than those in the control group to have used alcohol at least once during the past month or week. They were also less likely to report polydrug use or being heavily influenced by their peers (Perry et al. 1996).

Another report on Project Northland compared outcomes for 1,200 seventh-grade students in 20 Minnesota schools who were "planners" (students involved in peer planning for the project), "attenders" (who attended peer-planned events although not involved in the planning), and "nonparticipants" (students who chose not to participate in any of the peer program activities). Planners were less likely than attenders—and in some cases nonparticipants—to report that they had used alcohol in the past month or that they intended to use alcohol in the future. Caution is urged in interpreting these results, as the students were not randomly assigned but self-selected into these three exposure groups (Komro et al. 1996).

CASASTART

The CASASTART program (formerly known as CAR or Children at Risk) was developed at the National Center on Addiction and Substance

Abuse at Columbia University in 1992. As a school- and community-based program, it focuses on the prevention of drug abuse and violence among youth aged 8 to 13.

Four lists identified CASASTART as an exemplary or promising program and cited seven evaluations across them. Five of these seven are background/conceptual pieces, leaving two empirical studies for our review. Of these two evaluations, one was published in the *Georgia Academy Journal* (Murray 1999), which we were unable to locate through the Harvard library system, the University of California library system, the interlibrary loan network, or via the Internet. The *Comprehensive Final Report*, written and funded by the program's developers, is the only evaluation study cited that we could locate. This study compared outcomes for 656 youths aged 11 to 13 years from a treatment group, a randomly assigned control group, and a quasi-experimental comparison group, at the end of the program and at 1-year follow-up. When compared to the randomly assigned control group, CASASTART participants showed less past year and past month drug use (including cigarettes, alcohol, and marijuana). When compared to the quasi-experimental comparison group, CASASTART participants showed less lifetime drug use. The report also looked at peer influence variables and generally showed that youths in the treatment condition were less likely to be negatively influenced by their peers than were those in either the control or comparison groups (Harrell, Cavanagh, and Sridharan 1998).

Summary of Findings

Table 3 provides a summary of the evidence that was cited by the developers of the lists for each program. In the first row, we indicate the number of lists in which the program was designated as effective or promising. In the second row, we identify the number of empirical studies we reviewed. In the third row, we identify the number of distinct data sets used across the evaluations (because multiple reports were issued using the same data set), and in the fourth row, we identify the number of evaluations that were conducted by outside researchers rather than the program-developers.

In the last four rows, we list findings with "substantial impact." We chose to define substantial impact as a statistically significant reduction in substance use behavior (for either cigarettes, alcohol, marijuana, or other drugs) for students in a full-sample treatment group compared to a control group. As noted earlier, many of the reports find positive effects of the programs on outcomes such as knowledge, attitudes, or personal and psychological variables. Given that the lists and programs primarily

Table 3
Summary of Lists-Based Evidence for Five Drug Prevention Programs

	Program				
	Life Skills Training	Midwestern Prevention Project	Project ALERT	Project Northland	CASASTART
Number of lists on which program appears	7	7	3	4	4
Number of evaluative reports	19[a]	8	5	3	1[a]
Number of distinct data sets	13	2	1	3	1
Number of outside evaluations	1	0	0	0	1
Substantial impact at posttest	7 C, 1 A, 3 CAM	1 C, 1 CAM	—[b]	1 A[b]	—
Substantial impact at 1-year follow-up	2 C, 2 A, 3 CAM	3 C, 1 CAM	—	—	1 CAM
Substantial impact at 2-year follow-up	1 A	2 C	—	—	—
Substantial impact after 2 years	1 CA	1 CM	—	—	—

Note: C = cigarettes; A = alcohol; M = marijuana. Dashes indicate either that no substantial impact was found or that the articles did not report outcomes of interest at this time point.

a. For Life Skills Training, there were three articles that we could not locate, and for CASASTART, there was one article we could not locate. Because these numbers reflect the evaluations that we reviewed, we could be underestimating the number of evaluations that exist.

b. These numbers reflect the fact that all the Project ALERT reports and some of the Project Northland reports showed findings for subsamples only (e.g., nonusers, experimenters, and users; drinkers, and nondrinkers; planners, attendees, and nonparticipants). In this table, we counted only studies that reported findings for the full sample.

focus on the prevention of substance abuse behavior, we give priority to behavioral outcomes. Finally, Table 3 notes the number of times a substantial impact was found at posttest, at 1-year follow-up, at 2-year follow-up, and later than 2 years. Furthermore, it separates these out into type of drug ("C" for cigarettes, "A" for alcohol, "M" for marijuana, and any combination of the three, whereas "other" is used for any other drugs).

In Table 3, we did not pay attention to effect size in our designations of substantial impact. First, none of the lists rely on effect size in their designations. Second, standards regarding what are considered "big," "medium," and "small" effect sizes vary across fields and contexts. Third, studies often did not include enough data to calculate effect size. Therefore, we used statistical significance alone. We note that the use of statistical significance as the touchstone weights our table in favor of the programs, as some studies report findings that are listed as a "substantial impact" although they may not be practically important. For example, one report on LST found that the control group on average scored a 2.0 on a 9-point scale of drinking frequency, compared to an average score of 1.73 for the treatment group (Botvin et al. 1997). The result, because of the size of the sample, was statistically significant. A study on the MPP reported that 12.8% of students in the control group used alcohol twice or more in the past month, compared to 10% of students in the treatment group, again a statistically significant result (MacKinnon et al. 1991). The likelihood of obtaining statistical significance even when the effect is in actuality quite small is common when dealing with school-based studies involving thousands of students. Some of these findings have little practical significance. Relying only on statistically significant findings (which is all some of the lists require) to designate a program as effective might mislead consumers about the extent to which using the intervention could contribute to a meaningful reduction in substance abuse (for a discussion of statistical significance and practical significance, see Gorman 1995b; Lipsey 1990).

Discussion

Although several evaluations we reviewed provided solid evidence pointing to the effectiveness of a particular program, others made us question the process used to consider evidence and select programs for inclusion. As we mentioned earlier in this article, the two most common criteria for selection on the lists that we examined were (a) use of either randomized experiments or rigorous quasi-experimental design and (b) positive statistically significant effects. All the evaluations that we describe in this article did meet these criteria. However, our review of the evaluation evidence led us to wonder if these two criteria are sufficient for deeming a program "model" or "exemplary."

We raise questions about three features of the lists. First, we question the limited criteria for selection to the lists. Second, we question the extent to which some of the lists actually adhered to these criteria fully in selecting programs. And third, we question the validity of some of the scientific evidence itself. Although in some ways, these three issues can be considered distinct, in our analysis, they become entangled into one general concern, which is that much of the evidence on which the lists are based is insufficient for deeming a program "model" or "exemplary." In what follows, we discuss five patterns that we discovered in the evaluation evidence cited across these lists. Each pattern in some way touches on the three questions described above.

Limited Evidence of Program Effectiveness

We found that for every program except LST, there were few empirical evaluations cited on the lists from which to draw conclusions about a program's effectiveness. The lists generally require only one or two evaluations to designate a program as "effective." Our examination of the evaluations raises some questions about such a standard. Our review suggests that there is limited evidence of the effectiveness of all of these programs on reducing substance use in the long term, particularly CASASTART, Project Northland, MPP, and Project ALERT. The standard of one or two evaluations for designating a program as effective is not high enough, particularly if the reports are based on the same data set and/or are conducted by the program developers.

Absence of Independent Evaluators

Program developers often served as evaluators of their own programs, which is common in the field of prevention science. Program developers conduct evaluations to find out how well their program is working, to try out different versions of the program during the development process, and to demonstrate their success to others. But, in many cases, there were no outside evaluations of the programs. When the developers' evaluations were eliminated, there remained few independent evaluations to judge the merit of the program. If independent evaluations had been a criterion for program inclusion on a best-practice list, we wonder if any of these programs would have been selected.[2] There are several implications of this, including conflict of interest and a potential for bias in reporting. As a recent meta-analysis of evaluations of comprehensive school reform found, “studies performed by the developer yielded considerably stronger effects than studies performed by others” (Borman et al. 2003). It may very well be that the program developers achieve “high fidelity” conditions necessary for treatment effectiveness (e.g., Lipsey 1995), but again, at the very least, it raises questions about what the results would be if the program were evaluated by individuals who did not develop it.

Moreover, program developers are likely to study well-implemented versions of their programs rather than run-of-the-mill implementations. Inasmuch as their interest is in exploring the outcomes of a faithfully run version of the program, they are likely to gravitate to well-run sites.

Multiple Outcome Measures Reported

Another problem was the large number of different outcome measures reported. Not only do different measures make comparisons difficult, but many of the reports run the risk of “capitalizing on chance” by conducting scores of analyses. This questions the use of relying on statistically significant findings as an index of “success,” which might have been the result of only chance probability (i.e., 1 of 20 findings might be statistically significant at the .05 level by chance).

In these cases, we recognize that the stated criteria of having a positive statistically significant program effect was most likely adhered to. However, many studies included multiple measures of drug use behavior, such as daily, weekly, or monthly use; amount of drug use; frequency of drug use; or frequency of intoxication. In some studies, only a few of these measures are statistically significant. For example, an evaluation of Project ALERT conducted significance tests for 100 comparisons between a program and control condition (including six outcome measures for three different substances, three risk levels, and two types of programs). Out of these 100 tests, only 2 were statistically significant: one showing a decrease in weekly alcohol use for nonusers in the adult-led program and the other showing an increase in weekly cigarette use for users in the adult-led program (Ellickson, Bell, and McGuigan 1993). This means that only 1% of all the comparisons made showed a positive program effect, and given the number of comparisons made, this positive finding might be the result of chance rather than a program effect. Granted, this study looked at effects for a long-term follow-up, and the authors acknowledge that most effects disappear. Nevertheless, the fact that a best-practice list chose to include this particular study as evidence of the program's effectiveness leads us to question the process used for reviewing and selecting evidence.

Programs Effective Only Under Specific Conditions

Studies sometimes compared different treatment groups to determine what "dosage" of the program was most effective. If the basic program is effective, comparison of different variations may actually yield results that are less positive because all students receive some version of the program. However, if the treatment is not helpful or even harmful, the results may downplay these effects. We found that studies varied in the conditions or context under which interventions were determined to be effective, and in some cases, students who received certain variations of the program were worse off than those in a no-treatment control group.

Some studies showed that programs were harmful for particular groups of students. For example, two Project ALERT evaluations reported findings for each of the three risk groups (users, experimenters, and nonusers). As described earlier, although Project ALERT has been shown to be effective in reducing substance use among "experimenters," the opposite effect has been shown for students classified as "users" (Ellickson and Bell 1990a, 1990b). Consumers need to be aware of such important details when implementing a prevention program. The lists do not stipulate the specific conditions under which a program is effective but rather give the impression that programs are effective for all students.

Some programs were also found to be harmful when led by teachers. For example, evaluations of LST and Project ALERT looked at the effectiveness of their program when implemented by teachers as opposed to implemented by older students or peers. In some cases, the peer-led program was found to be more effective than the teacher-led program, and the teacher-led program was found to be as effective or less effective than the control (Botvin, Baker, Filazzola, et al. 1990; Bell, Ellickson, and Harrison 1993). This apparent ineffectiveness of the program when implemented by teachers should be made clear to consumers of the lists.

Other studies were concerned with how effective the program was when implemented with greater and lesser degrees of fidelity. For example, LST and MPP studies reported outcomes for "high-fidelity" or "high-implementation" subsamples, in other words, best-case scenarios of program implementation (Botvin, Baker, et al. 1995; Botvin, Baker, Dusenbury, et al. 1990; Botvin, Baker, Filazzola, et al. 1990; Botvin, Batson, et al. 1989; Pentz et al. 1990). By presenting the findings in this way, these reports are attempting to find out how "stronger dosages" of their program affect outcomes. But such analyses raise difficult issues of interpretation, not the least being the selection bias created by focusing on subgroups that were not randomly assigned to conditions (see Gorman 2002; Sherman 2003). From these studies, we learn that that the program may be effective only when implemented under a very specific set of conditions.

Finally, in some studies, it simply was not clear of what the control group consisted: whether the control group was receiving no drug prevention program at all or some other form of drug prevention education (perhaps as part of the school's general health curriculum). It is even more difficult to interpret results and judge a program's effectiveness when we do not know to what the intervention or treatment group is being compared.

Potential Implementation Problems For Randomized Designs

As stated earlier in this article, the lists that we reviewed prioritize randomized experiments or rigorous quasi-experimental designs when reviewing the evidence of a program's effectiveness. Indeed, nearly all of the studies we reviewed for this article used random assignment. However, it is widely noted in the literature that randomized experiments are not immune to issues that can arise during the course of a study and potentially threaten its internal validity. For example, potential problems include group randomization, lack of consent to participate, attrition from the study, and influential interactions among participants within a study (Cook 2002).

For a large majority of the studies we reviewed, the unit of randomization was the school or district as opposed to the individual student. Because the programs involve classroom curricula that are designed to be administered to groups of students, it makes sense that study participants would be randomized in groups. However, there are potential problems with group randomization, one being that students within the same class, school, or district will naturally share characteristics other than the fact that they were all assigned to the same experimental group. Such within-group correlation can make it difficult to assess how much of the effect is a result of the treatment versus some other shared characteristic of the students in the group. This issue is particularly problematic when there are few groups from which to randomize.

It is not clear to us that the lists we reviewed considered the potential problems described above. Although the criteria for deeming a program effective include rigorous study design, we did not see any evidence that the lists considered the implementation of the study design (see Table 2). Indeed, one could argue that a study employing a randomized design is only as good as the care with which it is carried out. In a worst-case scenario, problems such as those described above can fundamentally destroy the internal validity of a randomized experiment. Best-practice lists should take such potential problems into account when reviewing the evidence for program effectiveness.

Few Long-Term Follow-Up Studies

Few studies looked at outcomes more than 2 years after the end of project implementation. At immediate posttest, few reports showed substantial impact, and even fewer studies showed substantial impact at longer follow-ups. The study with the most reports showing substantial impact was LST, although this still represented only a portion of the total studies on LST that were cited across the lists. The few evaluations of all the programs that showed substantial impact generally reported that any positive early results dissipated after a few years.

Reports on two of the programs show a substantial impact at long-term follow-up. In the case of LST, the long-term follow-up study using the first data set found no statistically significant effect of the program on use of marijuana, cocaine, inhalants, and nonmedical pills. The long-term follow-up study using the third data set reported a significant effect of LST on tobacco use but not on alcohol or marijuana.[3] The one report of long-term follow-up for MPP found positive effects on use of marijuana and cigarette use (use in the past 30 days) but not on alcohol use. An earlier study using this same data reported positive findings only on tobacco use. We find the results of the follow-up study surprising because no positive shorter-term effects on marijuana were reported. Most of the programs are more effective in changing attitudes and increasing knowledge than they are in changing drug use behavior.

In summary, much of the evidence cited across the lists demonstrates positive effects for these five programs. However, as the follow-up periods get longer, many of those positive effects disappear. Furthermore, many programs seem to have more positive effects on outcomes such as knowledge and attitudes than they do on actual substance abuse behavior. Given this pattern, it is not clear what exactly the lists mean when designating a program as “effective.”

Conclusion and Recommendations

Our examination of the evidence cited across seven prominent lists in the prevention field raises questions about the criteria used by these lists to designate programs as effective. We wonder whether the standard of one or two studies is sufficient before categorizing programs as effective or promising. Clearly, when contemplating one or two studies, and one or a few of the outcomes, most of these programs will meet the criteria of the lists. Although we fully support basing programming on good evaluation data, when we look at all of the evaluations cited across the lists, we are disturbed by the frailty of evidence for some of the "proven" programs.

We are delighted that program developers, government agencies, and educational institutions have undertaken rigorous evaluations of the impacts of prevention programs. We encourage the continued use of evaluation. But when we look more closely, beyond a study or two, or beyond a few selected outcomes, the totality of evidence becomes more uncertain and our confidence in the designations of programs as "effective" or "promising" less sure. Yet there is pressure on school districts to use a "proven" program from the lists, leading them to switch from whatever they were previously running in their district to a program that (supposedly) is more effective. Our analysis suggests that we still have a lot to learn about the effectiveness of school-based drug prevention programs. We advocate an increase in the number of rigorous long-term evaluations of these programs. Perhaps then we can make better distinctions between good, fair, and poor practice.

We claim no special expertise for prescribing programmatic action. Our recommendations concern development of the lists. We believe it is important to note that the lists we used in writing this article give a misleading aura of certainty to their recommendations. School districts that follow the recommendations are probably at best only slightly better off than they were before. But now that their practice accords with government guidelines and professional prevention advice, districts have no incentive to evaluate outcomes again.

In the short term, we know from our own research on school district choices of drug abuse prevention programs that decision makers at the district level are using these lists to help them identify a "scientifically proven" program (Weiss, Murphy-Graham, and Birkeland 2005). In several cases, these decision makers switched from the D.A.R.E. program to one identified by these lists.[4] D.A.R.E. has been evaluated much more often than any other school-based drug prevention program (which is not surprising given its diffusion and popularity). Many of the studies of D.A.R.E. find early positive effects on some measures of attitudes, knowledge, or behavior (Aniskiewicz and Wysong 1990; Clayton et al. 1991; Ringwalt et al. 1994, Rosenbaum and Hanson 1998). However, long-term follow-ups suggest that D.A.R.E. is ineffective in sustaining the effects over time (Clayton et al. 1991; Clayton, Cattarello, and Johnstone 1996; Ennett et al. 1994; Ringwalt et al. 1994, Rosenbaum and Hanson 1998).

The findings from our review cause us to wonder if the programs cited across the best-practice lists are any more effective than D.A.R.E. We find it plausible that some of these programs may seem more effective than D.A.R.E. simply because they have been studied less, over a shorter length of time, and by the developers of the programs themselves. Furthermore, using the list criteria, cherry-picking positive findings from some of the D.A.R.E. evaluations would also lead to a designation as effective (e.g., see Dukes, Stein, and Ulhman 1997). Furthermore, as Gorman (2003) pointed out, if multiple subgroup analysis and post hoc sample refinement were used in the D.A.R.E. evaluations, would not these also yield positive results? We are not advocates for D.A.R.E. Rather, we wish to point out that the evidence used to place the other programs on these lists does not convince us that these programs are anything more than marginally better. We use this comparison with D.A.R.E. to further illustrate our point that the recommendations of the lists give a misleading aura of certainty.

In principle, the lists are useful in pointing school district decision makers toward effective programs. But the lists should be fair. Those who release lists must ensure that they are developed through credible procedures. Above all, the procedures need to be consistent, transparent, and rigorous. Probably the best way to develop lists of effective programs is through use of expert panels whose members have no stake in any candidate programs. The panels must be consistent in selecting valid evidence and interpreting it. The panel should publish the list of evaluations they considered and the process used to make their determinations.[5] Finally, panels should consider the most rigorous scientific research and look at the composite evidence rather than discrete findings. Furthermore, study implementation is as important as study design. Future expert panels must consider implementation issues, such as participant attrition, which can affect the validity of study results.

Panels must also make a serious effort to determine which program effects are large enough to matter. In addition, the balance of evidence determined through meta-analysis or syntheses is more compelling than single studies. When a program has been evaluated only a few times, especially with"special" samples, panels should withhold classification or label it "promising" rather than "effective" or "model." Lists should also be regularly updated to take into account new evidence. These lists get "stale" very quickly.

Finally, our findings also raise the question of whether any school-based prevention program will substantially reduce the number of teens who experiment with drugs. It seems at least possible that no school-based prevention program has a chance of fending off drug use 5 years down the road.[6] Certainly, we must lower our expectations for such programs. Reports by the Columbia University Center for Addictions and Substance Abuse, RAND Corporation, and the Maryland Report all indicated that the effects from even the best of the school-based drug prevention curricula are likely to be modest (e.g., National Center on Addiction and Substance Abuse at Columbia University 2001). Refresher courses in later grades evidently help to sustain whatever gains there are, but even with refreshers, there may be a limit to what school programs can do. We cannot expect small programs to have large effects. It may take broader scale, community-wide efforts to marshal the energies of multiple agencies to combat drug abuse.

If this is the case, if no single short-term program has much chance of significantly changing adolescent drug behavior, this conclusion is vital for schools—and the rest of us—to know. It is possible that some percentage of students will try alcohol or marijuana in their teens. If the country is seeking to prevent youth drug use, we are not sure that the best expenditure of public money is for schools to offer programs that are only marginally better, if better at all, than current offerings. Simply substituting one low-gain program for another does not seem worth it.

Addendum

We reviewed the lists for this article in the summer of 2003. Since that time, most of the lists we reviewed have remained intact with no updated information that conflicts with the information contained in the article. There is one exception, however. The National Registry of Evidence-Based Programs and Practices (NREPP) sponsored by the Substance Abuse and Mental Health Services Administration (SAMHSA) is currently undergoing a major overhaul, with a new submission and review process, and new Web site set to be launched in early 2007. Rather than categorize programs as “Promising,” “Effective,” or “Model,” programs will now be reviewed and assigned quantitative ratings on a 0 to 4 ordinal scale for a number of criteria related to the program’s strength of evaluation evidence and its readiness for dissemination. Programs that previously were reviewed by NREPP will be invited to undergo the new review procedure and receive a revised rating. At the time of publication of this manuscript, we do not know the extent to which this new NREPP rating system will affect the status of the programs that received NREPP “Model” ratings previously.

Appendix A
Criteria Used by the Center for Substance Abuse Prevention (Csap) Expert Panel to Classify Programs as Effective for the National Registry of Evidence-Based Programs and Practices (Csap, 2001)

Theory
Intervention fidelity
Process evaluation
Sampling strategy and implementation
Attrition
Outcome measures
Missing data
Data collection
Analysis
Other plausible threats to validity
Integrity
Utility
Replication
Dissemination capability
Cultural and age appropriateness

Appendix B
Safe, Disciplined, and Drug-Free Schools Expert Panel Criteria (Expert Panel on Safe, Disciplined, and Drug-Free Schools, 1999)

1. *Evidence of efficacy*. The program reports relevant evidence on efficacy or effectiveness based on a methodologically strong evaluation. This evaluation must be rated 3 on a 3-point scale, indicating overall that there is strong evidence that conditions are being met. These four conditions were the following:
   a. The program evaluation indicates a measurable difference in outcomes that is based on statistical significance testing or a credible indicator of magnitude of effect.
   b. The program evaluation used a design and analysis that adequately controls for threats to internal validity, including attrition.
   c. The program evaluation used reliable and valid outcome measures.
   d. The program evaluation used analyses appropriate to the data.

Appendix B (continued)

2. *Quality of program goals*. The program's goals with respect to changing behavior (or risk and protective factors) are clear and appropriate for the intended population and setting.
   a. The program's goals are explicit and clearly stated.
   b. The program's goals are appropriate to the intended population and setting.
3. *Quality of program rationale*. The rationale underlying the program is clearly stated, and the program's content and processes are aligned with its goals.
   a. The rationale underlying the program is clearly stated and includes appropriate documentation.
   b. The program's content and processes are aligned with its goals.
4. *Quality of program content and appropriateness*. The program's content takes into consideration the characteristics of the intended population and setting and the needs implied by these characteristics.
5. *Quality of program implementation methods*. The program's implementation process effectively engages the intended population.
   a. The program provides a relevant rationale to participants for its implementation.
   b. The program actively engages the intended population.
   c. The program attends to participants' prior knowledge, attitudes, and commonly held conceptions.
   d. The program implementation methods promote participants' collaboration, discourse, and reflection.
6. *Educational significance*. The application describes how the program is integrated into schools' educational mission.
7. *Usefulness/replicability*. The program provides necessary information and guidance for replication in other appropriate settings.

Notes

1. See Gorman (1994, 1995a, 1998) and Gerstein and Green (1993) for reanalysis and critique of these data.

2. Furthermore, in some cases, program developers served as members of the "expert panels" that drew up these lists, highlighting a further conflict of interest. Although they left the room when their own program was being discussed for inclusion on the list, their membership on the panel could well have had an effect.

3. With the exception of the "high-fidelity" group.

4. The D.A.R.E. (Drug Abuse Resistance Education) program is widely implemented but does not appear on any of these lists.

5. The U.S. Department of Education's What Works Clearinghouse is currently in the process of selecting effective programs and has been very transparent in its selection of evaluation evidence (see www.whatworks.ed.gov).

6. Denise Gottfredson (1997) has argued this to be the case. In her chapter on the University of Maryland report on preventing crime, she concludes that “no instructional program is likely to have a dramatic effect on substance use. . . . Rather, such programs should be embedded within more comprehensive programs” (pp. 5-36). See also Brown and Horowitz (1998) and Gorman (1997, 1998).

References

Aniskiewicz, R., and E. Wysong. 1990. Evaluating D.A.R.E: Drug education and the multiple meanings of success. *Policy Studies Review* 9 (4): 727-47.

Bell, R. M., P. L. Ellickson, and E. R. Harrison. 1993. Do drug prevention effects persist into high school? How Project ALERT did with ninth graders. *Preventive Medicine* 22:463-83.

Birkeland, S., E. Murphy-Graham, and C. Weiss. 2005. Good reasons for ignoring good evaluation: Evidence from a study of the Drug Abuse Resistance Education (D.A.R.E.) program. *Evaluation and Program Planning* 28:3.

Borman, G. D., G. M. Hewes, L. T. Overman, and S. Brown. 2003. Comprehensive school reform: A meta-analysis. *Review of Educational Research* 73 (2): 125-230.

*Botvin, G. J. 1994. *Smoking prevention among New York Hispanic youth: Results of a four-year evaluation study*. Unpublished manuscript.

Botvin, G. J., E. Baker, E. M. Botvin, A. D. Filazzola, and R. B. Millman. 1984. Prevention of alcohol misuse through the development of personal and social competence: A pilot study. *Journal of Studies on Alcohol* 45:550-52.

Botvin, G. J., E. Baker, L. Dusenbury, E. M. Botvin, and T. Diaz. 1995. Long-term follow-up results of a randomized drug abuse prevention trial in a white middle-class population. *Journal of the American Medical Association* 273:1106-12.

Botvin, G. J., E. Baker, L. Dusenbury, S. Tortu, and E. M. Botvin. 1990. Preventing adolescent drug abuse through a multimodal cognitive-behavioral approach: Results of a 3-year study. *Journal of Consulting and Clinical Psychology* 58:437-446.

Botvin, G. J., E. Baker, A. D. Filazzola, and E. Botvin. 1990. A cognitive-behavioral approach to substance abuse prevention: One-year follow-up. *Addictive Behaviors* 15:47-63.

Botvin, G. J., E. Baker, N. L. Renick, A. D. Filazzola, and E. M. Botvin. 1984. A cognitive-behavioral approach to substance abuse prevention. *Addictive Behaviors* 9:137-47.

Botvin, G. J., H. W. Batson, S. Witts-Vitale, V. Bess, E. Baker, and L. Dusenbury. 1989. A psychosocial approach to smoking prevention for urban Black youth. *Public Health Reports* 104:573-82.

*Botvin, G. J., and J. Cardwell. 1992. *Primary prevention (smoking) of cancer in Black populations*. Grant contract number N01-CN-6508. Final report to the National Cancer Institute. Ithaca, NY: Cornell University Medical College.

Botvin, G. J., L. Dusenbury, E. Baker, S. James-Ortiz, E. M. Botvin, and J. Kerner. 1992. Smoking prevention among urban minority youth: Assessing effects on outcome and mediating variables. *Health Psychology* 11:290-99.

Botvin, G. J., L. Dusenbury, E. Baker, S. James-Ortiz, and J. Kerner. 1989. A skills training approach to smoking prevention among Hispanic youth. *Journal of Behavioral Medicine* 12:279-96.

Botvin, G. J., and A. Eng. 1982. The efficacy of a multicomponent approach to the prevention of cigarette smoking. *Preventive Medicine* 11:199-211.

Botvin, G. J., A. Eng, and C. L. Williams. 1980. Preventing the onset of cigarette smoking through life skills training. *Preventive Medicine* 9:135-43.

Botvin, G. J., J. A. Epstein, E. Baker, T. Diaz, and M. Ifill-Williams. 1997. School-based drug abuse prevention with inner-city minority youth. *Journal of Child and Adolescent Substance Abuse* 6:5-20.

Botvin, G. J., K. W. Griffin, T. Diaz, and M. Ifill-Williams. 2001a. Drug abuse prevention among minority adolescents: Posttest and one-year follow-up of a school-based preventive intervention. *Prevention Science* 2:1-13.

———. 2001b. Preventing binge drinking during early adolescence: One- and two-year follow-up of a school-based preventive intervention. *Psychology of Addictive Behaviors* 15:360-65.

Botvin, G. J., K. W. Griffin, T. Diaz, N. Miller, and M. Ifill-Williams. 1999. Smoking initiation and escalation in early adolescent girls: One-year follow-up of a school-based prevention intervention for minority youth. *Journal of the American Medical Women's Association* 54:139-43.

Botvin, G. J., K. W. Griffin, T. Diaz, L. M. Scheier, C. Williams, and J. A. Epstein. 2000. Preventing illicit drug use in adolescents: Long-term follow-up data from a randomized control trial of a school population. *Addictive Behaviors* 25:769-74.

*Botvin, G. J., K. W. Griffin, E. Paul, and A. P. Macaulay. 2001. Preventing tobacco and alcohol use among elementary school students through Life Skills Training. *Journal of Child & Adolescent Substance Abuse.*

Botvin, G. J., N. L. Renick, and E. Baker. 1983. The effects of scheduling format and booster sessions on a broad-spectrum psychosocial approach to smoking prevention. *Journal of Behavioral Medicine* 6:359-79.

Botvin, G. J., S. P. Schinke, J. A. Epstein, and T. Diaz. 1994. Effectiveness of culturally focused and generic skills training approaches to alcohol and drug abuse prevention among minority youths. *Psychology of Addictive Behaviors* 8:116-27.

Botvin, G. J., S. P. Schinke, J. A. Epstein, T. Diaz, and E. M. Botvin. 1995. Effectiveness of culturally focused and generic skills training approaches to alcohol and drug abuse prevention among minority adolescents: Two-year follow-up results. *Psychology of Addictive Behaviors* 9:183-94.

Brown, J. H., and J. E. Horowitz. 1998. Deviance and deviants: Why adolescent substance use prevention programs do not work. *Evaluation Review* 22 (1): 3-14.

Center for Substance Abuse Prevention. 2001. *Toward the 21st century: A primer on effective programs*. Rockville, MD: Center for Substance Abuse Prevention.

Clayton, R. R., A. M. Catarello, L. E. Day, and K. P. Walden. 1991. Persuasive communication and drug prevention: An evaluation of the D.A.R.E. program. In *Persuasive communication and drug abuse prevention*, ed. Lewis Donohew, Howard E. Sypher, and Willima J. Bukoski, 295-312. Hillsdale, NJ: Lawrence Erlbaum.

Clayton, R., A. M. Cattarello, and B. M. Johnstone. 1996. The effectiveness of drug abuse resistance education (Project D.A.R.E.): 5-year follow up results. *Preventive Medicine* 25:307-18.

Cook, T. D. 2002. Randomized experiments in educational policy research: A critical examination of the reasons the educational evaluation community has offered for not doing them. *Educational Evaluation and Policy Analysis* 24 (3): 175-99.

Dukes, R. L., J. A. Stein, and J. B. Ulhman. 1997. Long-term impact of Drug Abuse Resistance Education (DARE). *Evaluation Review* 21:483-500.

Dwyer, J. H., D. P. MacKinnon, M. A. Pentz, B. R. Flay, W. B. Hansen, E. Y. I. Wang, and C. A. Johnson. 1989. Estimating intervention effects in longitudinal studies. *American Journal of Epidemiology* 130 (4): 781-95.

Ellickson, P. L., and R. M. Bell. 1990a. Drug prevention in junior high: A multi-site longitudinal test. *Science* 247:1299-305.

———. 1990b. *Prospects for preventing drug use among young adolescents*. Santa Monica, CA: RAND Corporation.

Ellickson, P. L., R. M. Bell, and E. R. Harrison. 1993. Changing adolescent propensities to use drugs: Results from Project ALERT. *Health Education Quarterly* 20:227-42.

Ellickson, P. L., R. M. Bell, and K. McGuigan. 1993. Preventing adolescent drug use: Long-term results of a junior high program. *American Journal of Public Health* 83:856-61.

Ennett, S., D. Rosenbaum, R. Flewelling, G. Bieler, S. Ringwalt, and S. Bailey. 1994. Long term evaluation of Drug Abuse Resistance Education. *Addictive Behaviors* 19:113-25.

Expert Panel on Safe, Disciplined, and Drug-Free Schools. 1999. *Guidelines for submitting safe, disciplined, and drug-free schools programs for designation as promising or exemplary*. Washington, DC: U.S. Department of Education.

Gerstein, D. R., and L. W. Green. 1993. *Preventing drug abuse: What do we know?* Washington, DC: National Academy Press.

Gorman, D. M. 1994. Preventing adolescent drug use: The effectiveness of Project ALERT. *American Journal of Public Health* 84:500.

———. 1995a. Are school-based resistance skills training programs effective in preventing alcohol misuse? *Journal of Alcohol and Drug Education* 41 (1): 74-98.

———. 1995b. On the difference between statistical and practical significance in school-based drug abuse prevention. *Drugs: Education, Prevention, and Policy* 2:275-83.

———. 1997. The failure of drug education. *Public Interest* 129:50-60.

———. 1998. The irrelevance of evidence in the development of school-based drug prevention policy, 1986-1996. *Evaluation Review* 22:118-46.

———. 2002. The “science” of drug and alcohol prevention: The case of the randomized trial of the Life Skills Training program. *International Journal of Drug Policy* 13:21-26.

———. 2003. Alcohol and drug abuse: The best of practices, the worst of practices—The making of science-based primary prevention programs. *Psychiatric Services* 54:1087-98.

Gottfredson, D. 1997. School-based crime prevention. In *Preventing crime: What works, what doesn’t, what’s promising—A report to the United States Congress* (NCJ 171676), ed. L. W. Sherman, D. Gottfredson, D. MacKenzie, J. Eck, P. Reuter, and S. Bushway, 5-1–5-74. Washington, DC: U.S. Department of Justice, Office of Justice Programs.

Harrell, A.V., S. Cavanagh, and S. Sridharan. 1998. *Impact of the children at-risk program: Comprehensive final report II*. Washington, DC: The Urban Institute.

Husen, T. 1994. Educational research and policy making. In *International encyclopedia of education,* ed. T. Husen and N. Postlethwaite, 1857-64. Oxford, UK: Pergamon.

Johnson, C. A., M. A. Pentz, M. D. Weber, J. H. Dwyer, N. Baer, D. P. MacKinnon, W. Hansen, and B. R. Flay. 1990. Relative effectiveness of comprehensive community programming for drug abuse prevention with high-risk and low-risk adolescents. *Journal of Consulting and Clinical Psychology* 58:447-56.

Komro, K. A., C. L. Perry D. M. Murray, S. Veblen-Mortenson, C. L. Wiliams, and P. S. Anstine. 1996. Peer-planned social activities for preventing alcohol use among young adolescents. *Journal of School Health* 66:328-34.

Komro, K. A., C. L. Perry, S. Veblen-Mortenson, and C. L. Williams. 1994. Peer participation in Project Northland: A community-wide alcohol use prevention project. *Journal of School Health* 64:318-22.

Lipsey, M. 1990. *Design sensitivity: Statistical power for experimental research*. Newbury Park, CA: Sage.

———. 1995. What do we learn from 400 research studies on the effectiveness of treatment with juvenile delinquents? In *What works? Reducing reoffending*, ed. J. McGuire, 63-78. New York: John Wiley.

MacKinnon, D. P., C. A. Johnson, M. A. Pentz, J. H. Dwyer, W. B. Hansen, B. R. Flay, and E. Y. Wang. 1991. Mediating mechanisms in a school-based drug prevention program: First-year effects of the Midwestern Prevention Project. *Health Psychology* 10:164-72.

*Murray, L. 1999. Preventing substance abuse using a community-based collaborative alternative. *Georgia Academy Journal* 6:8-11.

National Center on Addiction and Substance Abuse at Columbia University. 2001. *Malignant neglect: Substance abuse and America's schools*. New York: National Center on Addiction and Substance Abuse at Columbia University.

Pentz, M. A., J. H. Dwyer, D. P. MacKinnon, B. R. Flay, W. B. Hansen, E. Y. Wang, and C. A. Johnson. 1989. A multicommunity trial for primary prevention of adolescent drug abuse. *Journal of the American Medical Association* 261:3259-66.

Pentz, M. A., C. A. Johnson, J. H. Dwyer, D. M. MacKinnon, W. B. Hansen, and B. R. Flay. 1989. A comprehensive community approach to adolescent drug abuse prevention: Effects on cardiovascular disease risk behaviors. *Annals of Medicine* 12:219-22.

Pentz, M. A., D. P. MacKinnon, J. H. Dwyer, E. Y. I. Wang, W. B. Hansen, B. R. Flay, and C. A. Johnson. 1989. Longitudinal effects of the Midwestern Prevention Project on regular and experimental smoking in adolescents. *Preventive Medicine* 18:304-21.

Pentz, M. A., D. P. MacKinnon, B. R. Flay, W. B. Hansen, C. A. Johnson, and J. H. Dwyer. 1989. Primary prevention of chronic diseases in adolescence: Effects of the Midwestern Prevention Project on tobacco use. *American Journal of Epidemiology* 130:713-24.

Pentz, M. A., E. A. Trebow, W. B. Hansen, D. P. MacKinnon, J. H. Dwyer, C. A. Johnson, B. R. Flay, S. Daniels, and C. Cormack. 1990. Effects of program implementation on adolescent drug use behavior. *Evaluation Review* 14:264-89.

Perry, C. L., M. Grant, G. Ernberg, R. U. Florenzano, M. C. Langdon, A. D. Myeni, R. Waahlberg, S. Berg, K. Andersson, K. J. Fisher, D. Blaze-Temple, D. Cross, B. Saunders, D. R. Jacobs Jr., and T. Schmidt. 1989. WHO Collaborative Study on Alcohol Education and Young People: Outcomes of a four-country pilot study. *International Journal of the Addictions* 24:1145-71.

Perry, C. L., C. L. Williams, J. L. Forster, M. Wolfson, A. C. Wagenaar, J. R. Finnegan, P. G. McGovern, S. Veblen-Mortenson, K. A. Komro, and P. S. Anstine. 1993. Background, conceptualization and design of a community-wide research program on adolescent alcohol use: Project Northland. *Health Education Research* 8:125-26.

Perry, C. L., C. L. Williams, S. Veblen-Mortenson, T. L. Toomey, K. A. Komro, P. S. Anstine, P. G. McGovern, J. R. Finnegan, J. L. Forster, A. C. Wagenaar, and M. Wolfson. 1996. Project Northland: Outcomes of a community-wide alcohol use prevention program during early adolescence. *American Journal of Public Health* 86:956-65.

Petrosino, A. 2003. Standards for evidence and evidence for standards: The case of school-based drug prevention. *Annals of the American Academy of Political and Social Science* 587:180-207.

Reimers, F., and N. McGinn. 1997. *Informed dialogue: Using research to shape education policy around the world*. Westport, CT: Praeger.

Ringwalt, R. L., J. M. Greene, S. T. Ennett, and R. Iachan. 1994. *Past and future directions of the D.A.R.E. program: An evaluation review*. Prepared for the U.S. Department of Justice, National Institute of Justice, RTI Report 5192.

Rosenbaum, D. P., and G. S. Hanson. 1998. Assessing the effects of school-based drug education: A six-year multilevel analysis of Project D.A.R.E. *Journal of Research in Crime and Delinquency* 35:381-412.

Sherman, L. W. 2003. Misleading evidence and evidence-led policy: Making social science more experimental. *Annals of the American Academy of Political and Social Science* 589:6-19.

Spoth, R. L., C. Redmond, L. Trudeau, and C. Shin. 2002. Longitudinal substance initiation outcomes for a universal preventive intervention combining family and school programs. *Psychology of Addictive Behaviors* 16:129-34.

Williams, C. L., C. L. Perry, B. Dudovitz, S. Veblen-Mortenson, P. S. Anstine, K. A. Komro, and T. L. Toomey. 1995. A home-based prevention program for sixth-grade alcohol use: Results from Project Northland. *Journal of Primary Prevention* 16:125-47.

Weiss, C. H., ed. 1977. *Using social research in public policy making*. Lexington, MA: Lexington Books.

———. 1989. Congressional committees as users of analysis. *Journal of Policy Analysis and Management* 8:411-31.

Weiss, C. H., E. Murphy-Graham, and S. Birkeland. 2005. An alternative route to policy influence: How evaluations affect D.A.R.E. *American Journal of Evaluation* 26 (1): 12-30.

*An asterisk denotes studies that we were unable to locate but were nonetheless included on one or more lists.

Allison Gruner Gandhi, EdD, is a senior research analyst at the American Institutes for Research. She has been a major contributor to policy reports and projects for the U.S. Office of Special Education Programs, National Center for Education Statistics, and U.S. Substance Abuse and Mental Health Services Administration. She is a coeditor of Special Education for a New Century, published in 2005.

Erin Murphy-Graham, EdD, is an adjunct assistant professor at the Graduate School of Education, University of California, Berkeley. She has coauthored articles on evaluation and policy influence, gender and education, and secondary education in Latin America.

Anthony Petrosino, PhD, is a senior research associate for Learning Innovations at WestEd. He has worked as a research associate at the Study on Decisions in Education at Harvard, a research fellow at the Center for Evaluation at the American Academy of Arts & Sciences, and was founding coordinator and steering committee member for the Campbell Collaboration Crime and Justice Group. He has published a wide range of articles on evaluation research, crime and justice topics, and research synthesis.

Sara Schwartz Chrismer, EdM, is an advanced doctoral student at the Harvard University Graduate School of Education where she is studying K-12 policy and evaluation. She is co-chair and has served as an editor of Harvard Educational Review. She is the author of case studies on charter schools and No Child Left Behind.

Carol H. Weiss, PhD, is the Beatrice B. Whiting Professor of Educational Policy at the Harvard University Graduate School of Education. She is the author of Evaluation: Methods for Studying Programs and Policies (1998) and 11 other books. She has published more than 140 articles, chapters, and reviews on evaluation, the influence of social science research on policy, reporting of social science research in the media, and the decision-making process.