Prevention Science, Vol. 6, No. 3, September 2005 (© 2005)
DOI: 10.1007/s11121-005-5553-y

Standards of Evidence: Criteria for Efficacy, Effectiveness and Dissemination

Brian R. Flay,[1,10] Anthony Biglan,[2] Robert F. Boruch,[3] Felipe González Castro,[4] Denise Gottfredson,[5] Sheppard Kellam,[6] Eve K. Mościcki,[7] Steven Schinke,[8] Jeffrey C. Valentine,[9] and Peter Ji[1]

Published online: 16 May 2005

Ever increasing demands for accountability, together with the proliferation of lists of evidence-based prevention programs and policies, led the Society for Prevention Research to charge a committee with establishing standards for identifying effective prevention programs and policies. Recognizing that interventions that are effective and ready for dissemination are a subset of effective programs and policies, and that effective programs and policies are a subset of efficacious interventions, SPR's Standards Committee developed overlapping sets of standards. We designed these Standards to assist practitioners, policy makers, and administrators to determine which interventions are efficacious, which are effective, and which are ready for dissemination. Under these Standards, an efficacious intervention will have been tested in at least two rigorous trials that (1) involved defined samples from defined populations, (2) used psychometrically sound measures and data collection procedures; (3) analyzed their data with rigorous statistical approaches; (4) showed consistent positive effects (without serious iatrogenic effects); and (5) reported at least one significant long-term follow-up. An effective intervention under these Standards will not only meet all standards for efficacious interventions, but also will have (1) manuals, appropriate training, and technical support available to allow third parties to adopt and implement the intervention; (2) been evaluated under real-world conditions in studies that included sound measurement of the level of implementation and engagement of the target audience (in both the intervention and control conditions); (3) indicated the practical importance of intervention outcome effects; and (4) clearly demonstrated to whom intervention findings can be generalized. An intervention recognized as ready for broad dissemination under these Standards will not only meet all standards for efficacious and effective interventions, but will also provide (1) evidence of the ability to "go to scale"; (2) clear cost information; and (3) monitoring and evaluation tools so that adopting agencies can monitor or evaluate how well the intervention works in their settings. Finally, the Standards Committee identified possible standards desirable for current and future areas of prevention science as the field develops. If successful, these Standards will inform efforts in the field to find prevention programs and policies that are of proven efficacy, effectiveness, or readiness for adoption and will guide prevention scientists as they seek to discover, research, and bring to the field new prevention programs and policies.

KEY WORDS: standards; efficacy; effectiveness; dissemination.

[1]University of Illinois at Chicago, Chicago, Illinois.
[2]Oregon Research Institute, Eugene, Oregon.
[3]University of Pennsylvania, Philadelphia, Pennsylvania.
[4]Arizona State University, Tempe, Arizona.
[5]University of Maryland, College Park, Maryland.
[6]American Institutes for Research, Washington, District of Columbia.
[7]National Institute of Mental Health (NIMH), Bethesda, Maryland.
[8]Columbia University, New York, New York.
[9]Duke University, Durham, North Carolina.
[10]Correspondence should be directed to Brian R. Flay, D.Phil., Distinguished Professor, Institute for Health Research and Policy, University of Illinois at Chicago, 1747 W. Roosevelt Road, Suite 500, M/C 275, Chicago, Illinois 60608; e-mail bflay@uic.edu.

SPR PRESIDENT STATEMENT

The Society for Prevention Research (SPR) is committed to identifying and disseminating the most effective ways of preventing problems of human behavior. In recent years, prevention research has produced a huge body of research on practices that could prevent most of the common and costly problems. As a result, there is increasing interest in identifying and widely implementing the most effective prevention practices. Toward this end, the SPR Board of Directors appointed a task force[11] to determine the most appropriate criteria for prevention programs and policies to be judged efficacious, effective, or ready for dissemination. The Standards of Evidence that resulted from these deliberations were unanimously endorsed by the Board of Directors of SPR on April 12, 2004 and are available at the SPR website (http://www.preventionresearch.org/StandardsofEvidencebook.pdf). This paper provides a more extensive rationale and discussion of these Standards.

We, through this work, hope to provide a set of shared standards to be used by the diverse organizations seeking to identify tested and effective prevention programs and policies worthy of replication, adoption or dissemination. We also expect that these Standards will provide guidance for the research community to generate and test evidence-based prevention programs to improve the public health. We believe that the promulgation and widespread use of these criteria will lead to consistent and high standards for determining whether prevention programs and policies have been scientifically demonstrated to be efficacious, effective or ready for dissemination, thereby increasing confidence in and commitment to the use of tested and effective programs and policies to promote positive youth development and prevent health and behavior problems among young people and the general population.

J. David Hawkins, PhD
President 2003–05

INTRODUCTION

Prevention science has reached the point where our society has the potential to dramatically reduce the incidence and prevalence of the most common and costly problems of human behavior. Growing evidence suggests that a wide variety of problems can be reduced, including depression, violence and delinquency, tobacco, alcohol and other drug use, academic failure, risky sexual behavior, unemployment, injuries and accidents, and marital discord.[12] There is also increasing demand from policy makers, practitioners, and civic leaders for accountability in the expenditure of public funds on interventions[13] designed to promote health and well-being.

It is in this context that we need standards to assess how well prevention programs and policies work, whether they are ready for widespread dissemination and, if they are not ready for widespread dissemination, what further research we need to justify their widespread dissemination. It is only when effective prevention practices are widely disseminated that society will reap the potential benefits of the research conducted so far.

Government agencies and other funders have partnered with researchers to create guidelines for evaluating the validity of claims for intervention effectiveness. Examples include the Blueprints for Violence Prevention (Elliott & Mihalic, 2004), the CDC's Guide for Community Preventive Services (Benedict *et al.*, 2000), SAMHSA's National Registry of Evidence-based Programs and Practices (NREPP) program (Barkham *et al.*, 2001), the U.S. Department of Education's (1998) Safe and Drug Free Schools program and the Institute of Education Science's (Coalition for Evidence-Based Policy, 2003; U.S. Department of Education, 2003) What Works Clearinghouse.[14] However, these use

[11] Members of the SPR Standards Committee were the first nine authors, chaired by the first author and assisted by the 10th author.

[12] For each of these problems, there are one or more studies showing that a preventive intervention can reduce the likelihood of the problem. Although in most cases the evidence does not yet justify widespread dissemination of these interventions, the studies do show the potential of preventive interventions to significantly improve human wellbeing. We cite one or two such studies for each of these problems by way of example. Depression (Clarke *et al*., 1995), violence and delinquency (Flannery *et al*., 2003), tobacco, alcohol, and other drug use (Biglan *et al*., 2004; Botvin *et al*., 1995), reading failure (Gunn *et al*., 2002), risky sexual behavior (Kelly *et al*., 1997), injury and accident prevention (Edgerton *et al*., 2004; O'Malley & Wagenaar, 1991) and marital discord (Markman *et al*., 1993).

[13] This document and the Standards pertain equally to programs and policies. To avoid awkwardness, we sometimes use the term "intervention" to refer to both programs and policies.

[14] As of March 2005, each of these programs had websites as follows: http://www.colorado.edu/cspv/blueprints/, http://www.thecommunityguide.org/ tobacco/, http://www.modelprograms.

somewhat different criteria for the selection of effective programs (Greenberg, 2004) and the use of different criteria has resulted in a low degree of overlap of ratings of empirical studies when these different systems assess the same programs (Elliott & Mihalic, 2004; Mihalic, 2002–2004). It thus seems appropriate for prevention scientists to draw from these prior efforts and offer a more complete set of criteria specifically for evaluating prevention programs and policies (Hansen & Dusenbury, 2001).

Efficacy, Effectiveness, and Dissemination

Most interventions are first evaluated by developers or others under optimal conditions, such as having ample resources and well-trained and carefully supervised intervention personnel. Yet, programs worthy of dissemination must also be effective under real-world conditions. For this reason, prevention scientists distinguish between efficacy trials and effectiveness trials (Flay, 1986). *Efficacy* refers to the beneficial effects of a program or policy under optimal conditions of delivery, whereas *effectiveness* refers to effects of a program or policy under more real-world conditions (Flay, 1986; Greenberg, 2004; Holder *et al.*, 1995, 1999; Kellam & Langevin, 2003; Last, 1988; Moscicki, 1993).

Efficacy trials require a rigorous research design, a high quality of program implementation, and researcher control over confounding factors. In an efficacy trial, for example, a researcher may test a school-based program with highly trained and supervised research staff delivering the intervention under optimal conditions. By contrast, regular classroom teachers, who have many competing demands on their time and attention every day, may be expected to deliver the intervention once it is disseminated (Hansen & Dusenbury, 2001).

Effectiveness studies focus on important factors such as the quality of implementation, which will affect program outcomes when delivered under naturalistic conditions. Furthermore, issues regarding program fidelity and adaptation as programs are "taken to scale" may contribute further variation in the expected outcomes (Elliott & Mihalic, 2004; Flay, 1986). Thus, a program that produces significant effects in an efficacy trial may or may not yield similar effects under real-world conditions.

In general, prevention research has progressed from identifying efficacious programs and policies and then testing their effectiveness in increasingly real-world conditions. Accordingly, questions of whether an intervention meets efficacy or effectiveness standards are different and may involve a different set of standards. *For a program to be found effective, it must also meet all Standards for efficacy*.

As evidence of the efficacy and effectiveness of prevention programs and policies has accumulated, the question has emerged as to when programs and policies that have evidence of both efficacy and effectiveness are appropriate for dissemination (Hansen & Dusenbury, 2001; Kellam & Langevin, 2003; Lynagh *et al.*, 2002). *A program worthy of dissemination must also meet all of the Standards for effectiveness*. Not all programs of proven effectiveness are ready for widespread dissemination. For example, a program may require special materials and special training of teachers or clinicians before it can be delivered in a way that it is effective. It is important that programs be ready for dissemination so that they can be implemented effectively, that is, in a manner that achieves the expected effects.

Thus, we outline our Standards for Evidence in three sections: efficacy, effectiveness and dissemination. Our objective in writing these standards was to articulate a set of principles for identifying prevention programs and policies that are sufficiently empirically validated to merit the labels "tested and efficacious," "tested and effective," or "tested, effective, and ready for dissemination." Consistent with SPR's mission, we were interested in prevention programs and policies that have *public health importance* (e.g. Healthy People 2010 [U.S. Department of Health and Human Services, 2000], Weissberg *et al.*, 2003). These are directed to the prevention of social, physical, and mental health problems and the promotion of health, safety, and well-being.

For establishing programs that are efficacious or effective, we emphasize research designs that can establish causal effects. That is, we want to be confident that the program or policy under question, rather than some other factor, is responsible for the observed effects. Otherwise, claims of effectiveness are likely to be biased or untrue and it is necessary to consider competing explanations that could lead to the expected outcome (Bertrand *et al.*, 2002; Holland, 1986; Manski, 1995; Rubin, 1974; Shadish *et al.*, 2002).

samhsa.gov/template.cfm?page=default, http://www.ed.gov/about/ offices/list/osdfs/index.html?src=oc and http://whatworks.ed.gov/.

Our focus on research pertaining to the causal effects of programs and policies does not mean that we believe that research designs meant to uncover causal relationships are the only research tool that should be used in prevention science, or that these are the only tools that are truly "scientific" (Valentine & Cooper, 2003). To the contrary, we believe that (a) no single method can be used to address all interesting and important questions about prevention and (b) even when causal relationships are of primary interest, other types of research tools and designs are often needed to yield important information about when, why, and how prevention programs and policies work, and for whom. Because our central mission is to identify programs and policies that make a difference, our central focus is on research designs that have as their primary purpose uncovering causal relationships. However, confidence in causal relations is only one factor in evaluating prevention programs and policies; other factors such as the size of program effects, importance of outcomes obtained, whether effects last over time, etc. are also critically important.

In the following sections we list the required standards in italics, using the same numbering system as in the original Standards document (Flay *et al.*, 2004a).

CRITERIA FOR EFFICACY

Specificity of the Efficacy Statement

1. Standard: A statement of efficacy should be of the form that "Program or policy X is efficacious for producing Y outcomes for Z population."

Our first criterion pertains to the form of the efficacy statement. Because outcome research results are specific to the program or policy actually tested, the samples (or populations and their settings from which they were drawn) and the outcomes measured, it is essential that conclusions from the research be clear regarding the program or policy, population(s) and their settings, and the outcomes for which efficacy is claimed.

The remaining standards pertain specifically to the four areas of validity described by Cook and Campbell (1979):

- Description of the program or policy and the observed outcomes;
- Clarity of causal inferences;
- Generalizability of findings;
- Precision of outcome.

Standards might change over time as methods develop and prevention science and practice advance. For this reason, we also include standards that are desirable (labeled as such) though not essential for efficacy given the current state of prevention program development and evaluation. We find some of these to be required for effectiveness or dissemination and others that may become necessary criteria in the future as our methods advance.

Intervention Description and Outcomes

2.a. Standard: The intervention must be described at a level that would allow others to implement/replicate it.

To meet this standard, authors should provide a more detailed description of the intervention (program or policy) than most research journals will publish. Authors could write such a description as a manual or as an appendix available from the authors or posted on a website. An adequate description of a program or policy includes a clear statement of the population for which it is intended; the theoretical basis or a logic model describing the expected causal mechanisms by which the intervention should work; and a detailed description of its content and organization, its duration, the amount of training required, intervention procedures, etc. The level of detail needs to be sufficient so that others would be able to replicate the program or policy. With regard to policy interventions, the description must include information on relevant variations in policy definition and related mechanisms for implementation and enforcement.

Outcomes—What is Measured

2.b.i. Standard: The stated public health or behavioral outcome(s) of the intervention must be measured.

The statement of efficacy can only be about the outcomes that are measured and reported. Some efficacy trials measure only an early form or a predictor of the intended outcomes of the prevention practice, or they measure what can be easily assessed using available records or less intrusive means of primary data collection. Such programs or policies

can be labeled efficacious only for the outcomes actually measured. Before a program or practice can be labeled efficacious for a public health or behavioral outcome claimed to be targeted (e.g., crime, substance use, violent behavior) these must be measured and demonstrated. For example, a measure of attitudes about violence cannot substitute for a measure of actual violent behavior.

Some prevention efforts aimed at children who have not yet initiated a behavior expected to emerge later might demonstrate effects on proxy measures. For example, research has demonstrated that problem behavior is a general syndrome or behavioral pattern (Gottfredson & Hirschi, 1990) in the sense that youth who engage in one form (e.g., substance use) are highly likely to engage in other forms (e.g., property crime) and that problem behavior is relatively stable across the life cycle. A demonstration that a prevention program affects an early form of either the targeted behavior or a related form of the behavior is sufficient to establish efficacy *only on the earlier forms of the outcome*. For example, an elementary school program intended to reduce adolescent delinquency might be established as efficacious for reducing first grade conduct problems only, or a policy to require broader public access to preschool facilities for single parents could be established as efficacious for increasing public access. In either case, however, additional studies or long-term follow up would be required to establish efficacy for the prevention of adolescent delinquency.

2.b.ii. Standard: For outcomes that may decay over time, there must be at least one long-term follow-up at an appropriate interval beyond the end of the intervention (e.g., at least 6 months after the intervention, but the most appropriate interval may be different for different kinds of interventions).

The positive effects of an intervention (program or policy) can vary over time after the period of intervention. They may diminish rapidly or slowly, or they may persist for long periods. Some investigators have reported intervention effects under conditions requiring adaptation to new social task demands, such as entering middle school, that may require more self-regulation of behavior. Some interventions may demonstrate effects on problems that emerge later in development, such as substance use or abuse, sexual behavior, mental disorder, criminal behavior or drunk driving (Griffin *et al.*, 2004; Olds *et al.*, 2004; Wolchik *et al.*, 2002).

Variation in the course of intervention effects requires a periodic assessment of impact rather than assessment at a single point in time. Repeated measurement provides information on the course and timing of effects, and increases confidence in inferences about the efficacy of the intervention. Before we can infer efficacy there must be evidence of significant effects for at least one long-term follow-up at an appropriate interval beyond the end of the program or the initiation of a policy. The more time points that are assessed, the greater the certainty and the details that can be inferred about the course of effects. Thus, efficacy statements should also specify the time frame within which the effects are expected to be maintained.

- *Desirable Standard: It is also desirable, though not necessary, to include measures of proximal outcomes (i.e., mediators).* The analysis of program effects on theoretical mediators is essential for establishing causal mechanisms. Although the efficacy of a program or practice for achieving its ultimate outcomes can be assessed on the basis of a study measuring only the ultimate outcomes, it is also desirable to measure the intermediate outcomes hypothesized to lead to the final outcome. For example, a substance abuse prevention program may lead to improved parenting, increased social skills or reduced externalizing behavior (as mediators). Such additional measures facilitate mediational analyses that provide valuable information about how the program works (Baron & Kenny, 1986).

  The specific outcomes that would be affected by a prevention program or policy are informed by theory and by prior empirical analyses, as described in a working model or logic framework of the putative causal processes (e.g., Conduct Problems Prevention Research Group, 1992). Such models relate antecedents or predictor variables to outcomes as they may be influenced by the preventive intervention. In such models, it is useful to measure and model the effects of more proximal intermediary variables, that is, mediator variables. The measurement (in both intervention and control conditions) and evaluation of a preventive intervention's effects on a specific mediator variable can provide information about the causal mechanism of effect, and how it is influenced by the preventive intervention. It is highly desirable that future efficacy studies include measures

of theoretically based mediator variables and tests of their hypothesized mechanisms of action.

- *Desirable Standard: It is desirable to measure implementation.* The expectation in an efficacy trial is that the implementation will be standardized and of high quality. However, it would still be desirable to measure the level of implementation to ensure that this occurred, to document any variation that does occur, and to specify the level of implementation of the program that achieved the reported effect. Given that control conditions usually involve activities relevant to the desired outcomes, researchers should also measure the level of implementation of these activities in the control condition. For example, in a study designed to test the efficacy of a classroom behavior management intervention, the control condition would also be expected to have classroom management activities, and these need to be measured.
- *Desirable Standard: It is desirable to measure potential side-effects or iatrogenic effects.* Most past efficacy trials of behavioral programs and policies have not measured potential negative effects. Although such effects may not be obvious, emerging evidence in prevention science suggests that negative effects are not uncommon. Iatrogenic effects may be the reverse of the intended outcome for whole groups (e.g., Dishion *et al.*, 1999; Goodstadt, 1978) or subgroups (Kellam *et al.*, 1994). They may also be negative effects unrelated to the intended outcome (e.g., side-effects of vaccines) or unanticipated consequences of systems change (e.g., substitutions between problem behaviors in response to implementation and enforcement, Holder, 1998). *To ensure the safety of prevention programs or policies*, it is highly desirable that measures of potential side-effects and iatrogenic effects be included in future efficacy trials.

Outcomes—Measurement Properties

2.c. Standard: Measures must be psychometrically sound.

The measures used must either be of established quality, or the study must demonstrate their quality. Quality of measurement consists of construct validity and reliability.

2.c.i. Standard: Construct validity—Valid measures of the targeted behavior must be used, following standard definitions within the appropriate related literature.

2.c.ii. Standard: Reliability—Internal consistency (alpha), test–retest reliability, and/or reliability across raters must be reported.

Researchers can obtain evidence of validity (e.g., that the measurement assesses what it is intended to assess) and reliability (e.g., the ability of a measurement process to obtain similar responses by retest or with different raters) from test manuals or prior studies that use the same instruments if these sources report properties for samples similar to that used in the efficacy trial of interest. Alternatively, investigators can conduct pilot tests prior to an actual efficacy trial to assess the quality of measurement or they can provide evidence produced in the efficacy trial itself.

Measurement processes need to be equally valid and reliable across conditions. For some kinds of interventions, the measure may come to have a different meaning for people in treatment and control conditions. For example, intensive supervision can lead to a closer observation, or more accurate measurement, of technical infractions. This can lead to the program appearing to have a negative effect—where it is merely a consequence of the increased quality of measurement in the intervention group.

- *Desirable Standard: It is desirable to use multiple measures and/or sources.* Multiple measures of the same construct, particularly from multiple sources (e.g., student self-reports, parent reported and teacher ratings of student behavior, ratings by independent observers), can increase confidence in both the validity and reliability of measures and in the robustness of findings.

2.c.iii. Standard: Where "demand characteristics" are plausible, there must be at least one form of data (measure) that is collected by people different from the people who are applying or delivering the intervention. This is desirable even for standardized achievement tests.

Demand characteristics refer to possible reactivity to the measurement or its context (Rosnow, 2002). This is likely to occur when people known to subjects collect information from them about sensitive behaviors, or in which implementers have a stake in the outcome. Under such conditions, measuring the impacts of a preventive intervention requires methods

and data collectors independent of the intervenors. For example, in evaluations of school-based programs, the staff delivering the program should not collect outcome data from their students. To the extent possible, the independent observers should be *blinded or masked* to the intervention condition (Meinert, 1986).

Clarity of Causal Inference

3. Standard: The design must allow for the strongest possible causal statements.

The research design must be the strongest possible given the nature of the program or policy, research question, and institutional framework within which the intervention/research occurs. The design must also be well executed, and any remaining threats to causal inference, or alternative explanations for observed effects, should be addressed.

Control[15] *Condition*

3.a. Standard: The design must have at least one control condition that does not receive the tested intervention.

The control condition can be no-treatment, attention-placebo or wait-listed. Or, it can be some alternative intervention or usual care (e.g., what the participants would have received had the new programs and policies not been introduced), in which case the research question would be, "Is the new program or policy better than a current program or policy?" In time-series studies, the control condition may be the same group that does not get the intervention for a while.

Assignment

3.b. Standard: Assignment to conditions needs to minimize the statistical bias in the estimate of the relative effects of the intervention and allow for a legitimate statistical statement of confidence in the results.

Researchers should assign units to conditions in such a way as to minimize systematic selection, for example, self-selection or unexplained selection. Such assignment reduces the plausibility of alternative explanations for the causes of observed outcomes. This then increases the plausibility of causal inference about the intervention. The design and the assumptions embedded in the design must take into account exactly how people or groups were selected into intervention and control conditions and how influences on the treatment and control conditions might differ apart from the intervention.

3.b.i. Standard: For generating statistically unbiased estimates of the effects of most kinds of preventive interventions, random assignment is essential.

Within the context of ethical research, it is necessary to use randomization whenever possible to ensure the strongest causal statements and produce the strongest possible benefits to society (Fisher *et al.*, 2002). Many objections to randomization may be unfounded (Cook & Payne, 2002). Randomization is possible in most contexts and situations. For example, the Cochrane registry (www.cochrane.org) contains over 350,000 entries on randomized trials; the Campbell Collaboration (Petrosino *et al.*, 2000) registry (www.campbellcollaboration.org) contains over 13,000 entries of randomized trials in the social sciences, many of which involve randomization of larger social institutions such as schools and communities (Boruch, 2005a).

Randomization requires support from many different stakeholders. Different stakeholders should be involved in the planning of the research and sanction the randomization process (Berends & Garet, 2002; Kellam, 2000; Madison *et al.*, 2000; Petosa & Goodman, 1991; Schinke *et al.*, 1983; Towne & Hilton, 2004).

The level of randomization should be driven by the nature of the program or policy and the research question. Randomization can be of individuals or of intact groups like classrooms, schools, worksites, neighborhoods or clinics (Boruch, 2005a). Publications should specify exactly how the randomization was done. It is not sufficient to simply state that participants/units were randomly assigned to conditions.

For some kinds of policy or community-wide interventions, where randomization is impossible, other approaches may be acceptable, but only when used with caution and methodological expertise, and when careful attention is given to ruling out

[15]Use of the terms "control" and "comparison" varies within and across the social, behavioral, educational and medical sciences. Some use "control" only for randomized trials and others use "control" only for no-treatment control conditions. Authors sometimes use the term "comparison" to refer only to a nonrandomized comparison condition and sometimes as an umbrella term to refer to any type of comparison condition, including a no-treatment control group. We use "control" throughout this document to refer to any control or comparison condition or group.

plausible alternative explanations (see next three Standards).

3.b.ii. Standard: For some kinds of large-scale interventions (e.g., policy interventions, whole-state interventions) where randomization is not practical or possible, repeated time-series designs without randomization can be convincing—given large effects and long baselines (Biglan et al., 2000). *Even with these designs, randomization to multiple conditions or times is still preferable, especially if long baselines are not available*.

The logic of the so-called "interrupted time-series designs" is that the effect of a program or policy can be judged by whether it affects the intercept or slope of an outcome that is repeatedly measured (Greene, 1993; Nerlove & Diebold, 1990; Shadish *et al.*, 2002). For example, Wagenaar and Webster (1986) evaluated the effects of Michigan's mandatory automobile safety seat law for children under 4 by comparing the rate of injuries to passengers 0–3-years old for the 4 years prior to enactment of the law and a year-and-three quarters after its enactment.

Confidence that a program or policy affected an outcome is enhanced if there are additional comparisons of intervention and nonintervention phases. For example, policies may be implemented and reversed in a given geopolitical entity. These "reversals" constitute further tests of the effects of the policies (Wagenaar, 1983, 1993). If the dependent variable changes with each withdrawal and reinstatement, then we can be more confident that the policy, in fact, accounted for the effect. A shortcoming of reversal designs is the inability to estimate long-term effects.

The effects of intervening in a given community or state also can be compared with time-series data for control group entities that do not receive the program or policy, or receive it at a different time. Such designs allow one to examine whether the time-series data in control communities or states remains unchanged at the point at which the program or policy is introduced in the entity receiving the intervention.

Time-series analyses and archival data can provide means to test the efficacy of population-based policy interventions over time and, ultimately, through comparisons of intervention and nonintervention sites. One example of such an efficacy trial was the initial evaluation of a higher minimum legal drinking/alcohol purchase age in Michigan (Wagenaar, 1983). Moreover, confidence in the efficacy of the policy can be high with multiple replications of the intervention effect when the policy is introduced into other entities later.

3.b.iii. Standard: Well-conducted regression-discontinuity designs also can be convincing because, as in randomized studies, the selection model is completely known.

This design involves determining who receives an intervention based on a cutoff score on a pre-intervention measure. The cutoff score might be based on merit or need, or on some other consideration negotiated with the other research partners. For example, students with reading scores below the 25th percentile might be assigned to a tutoring program while the remaining students serve as a control. Treatment effects are inferred by observing differences in the slopes and/or intercepts of the regression lines for the different groups. Angrist & Lavy (1999) and Riecken and Brouch (1974) provide examples of applications. Regression discontinuity designs have important assumptions that require a degree of statistical expertise to assess (e.g., that the functional form of the relationship between the assignment and outcome variable be properly specified; Shadish *et al.*, 2002, Trochim, 1984, 2000), and should, therefore, be undertaken by researchers with specific training in the design and analysis of these types of studies.

3.b.iv. Standard: Matched control designs with demonstrated pretest equivalence using adequately powered tests on multiple baselines or pretests of multiple outcomes and important covariates can be credible—as long as assignment was not by a process that results in a correlation between unmeasured variables and condition.

Estimates of effects from studies using any type of equating (e.g., matching, analysis of covariance, propensity scoring, selection modeling) are often wrong. At a minimum, the results of these studies vary more widely than the results of randomized experiments. In other words, a few randomized trials will probably provide a more precise answer about whether or not an intervention works than many nonrandomized experiments using equating (e.g., Glazerman *et al.*, 2003). At worst, the estimates from nonrandomized experiments are more likely to be wrong, sometimes with serious consequences. A prominent recent example is the research on hormone replacement therapy for women where prior nonrandom trials suggested positive effects but a randomized trial found negative effects (Shumaker *et al.*, 2003). For these reasons (and several others), randomized experiments should be considered to be

the method of choice for answering questions about whether or not an intervention is efficacious.

Matched control designs are credible only when there is a pretest demonstration of group equivalence. For example, if researchers match students on socioeconomic status and prior achievement, the researcher should conduct a statistical test comparing the intervention and control groups on those variables, and the statistical test should have sufficient power to detect a relatively small difference between them. Another common strategy is for researchers to use statistical techniques (such as analysis of covariance) to remove the variability in the outcomes associated with group differences at baseline. For matching to be credible, researchers must thoughtfully select variables on which to match participants. As an example, if researchers match students on age but fail to match students on the severity of the problem at baseline, it is unlikely that reviewers will view their procedures as credible.

Empirical research comparing estimates from randomized trials against those from nonrandomized trials is currently underway (Boruch, 2005b; Glazerman *et al.*, 2003; Jacob & Ludwig, 2005). This work helps us to understand the magnitude and direction of the biases in nonrandomized trials. Readers need to be alert to advances in this arena.

Generalizability of Findings

Sample is Defined

4. Standard: The report must specify the sample and how it was obtained.

It needs to be clear how well the study sample does or does not represent the intended population. This is an essential component of the efficacy statement. An exemplary standard format can be found in the CONSORT statement adopted by the journals of the American Medical Association (Moher *et al.*, 2001). An intervention shown to be efficacious can claim to be so only for groups similar to the sample (including the geographic and temporal context) on which it was tested.

To establish the generalizability of findings, researchers must describe in appropriate detail their sample source and how they recruited their sample. For example, drawing intervention and control groups from clinical or other self-selected populations limits generalizability to nonclinical populations (Berkson, 1946; 1958; Meinert, 1986).

Being able to characterize the study sample is essential not only for identifying the population for which the program or policy is intended, but also for the identification of subgroups. Risk group and subgroup characteristics need to be defined with as much care as the outcome (Kellam *et al.*, 1999; Moscicki, 1993). Researchers should describe their sample in terms of the age distribution, developmental stage, sex, race/ethnicity, socioeconomic characteristics (which may include social class, educational attainment or a proxy, and/or income or poverty status), marital status, and any other known risk characteristics relevant to the program or policy being tested. Such characteristics may include, for example, contact with the juvenile justice or child welfare systems, residence in a high crime area, presence of diagnosed major depressive disorder where intervention is to prevent comorbidity, etc. As with outcome measurement, risk characteristics and sample selection criteria should be assessed using rigorous tools with high reliability and validity.

- *Desirable Standard: It is desirable that subgroup analyses demonstrate efficacy for subgroups within the sample—gender, ethnicity/race, risk levels.* A small main effect may involve a large effect for a particular (e.g., high-risk) subgroup and small or no effects for other subgroups. If an investigator anticipates implementing an intervention in specific population subgroups defined by sociodemographic or risk characteristics, the sample in which efficacy is tested needs to include participants from those subgroups. The subgroup sample size should be large enough to allow for sufficient statistical power to conduct meaningful analyses by subgroup. In addition, the subgroup samples need to be described in appropriate detail to determine generalizability of the findings. For example, in the Elmira Study that tested the efficacy of nurse home visitation in preventing a wide range of adverse maternal and child outcomes, the intervention conveyed the strongest benefits in a subgroup of mothers with the highest levels of risk (Olds *et al.*, 1998). Other programs have found similar subgroup effects (Dawson-McClure *et al.*, 2004; Dolan *et al.*, 1993; Eddy *et al.*, 2000; Kellam *et al.*, 1998; Segawa *et al.*, 2005). It is also possible that strong positive effects for one subgroup are accompanied by negative effects for another subgroup.

Precision of Outcome

Statistical Analysis

5.a. Standard: Statistical analysis must be based on the design and should aim to produce a statistically unbiased estimate of the relative effects of the intervention and a legitimate statistical statement of confidence in the results.

5.a.i. Standard: In testing main effects, the analysis must take into account the level of the randomization and include all cases assigned to treatment and control conditions (except for attrition—see below).

In many contexts in which prevention researchers carry out their work, the participants belong to naturally occurring groups, which often must be taken into account when conducting statistical tests. For example, if a researcher is testing a drug prevention curriculum in six 3rd-grade classrooms, the fact that the students belong to the classrooms means those student responses may not be independent of other students in the same classroom, and this has an important impact on the validity of the statistical tests. Often, researchers will randomize at a higher level (e.g., the clinic) but analyze the data at a lower level (e.g., individuals). Doing so almost always results in a violation of the assumption of the statistical independence of observations. Even small violations of this assumption can have very large impacts on the standard error of the effect size estimate (Kenny & Judd, 1986; Murray, 1998), which in turn can greatly inflate the Type I error rate (e.g., Scariano & Davenport, 1987). In these situations, analysts must at least conduct analyses at the level of randomization. Further, multilevel models can improve the analysis at the level of randomization by taking into account observations on units that might be clustered at a different level (Brown, 1993; Bryk & Raudenbush, 1992; Hedeker *et al.*, 1994; Zeger *et al.*, 1988). For example, if an intervention is delivered at the clinic level (e.g., some clinics deliver a new intervention, others do not), then clinics should be randomly assigned to conditions, and the statistical analyses should take into account that patients are nested within clinics.

Statistical analyses should also be conducted using all of the cases assigned to the treatment and control conditions. This is commonly known as an "intent-to-treat" analysis. Strictly speaking, the statistical statements about biases and probability are applicable only if one does an intent-to-treat analysis. That is, analysts should analyze data for the groups as randomized to condition, regardless of what treatment they did or did not receive.

Often, investigators cannot follow cases assigned to a condition to the end of the study (due to death, moving, inability to locate, etc.). Ignoring this attrition can result in biased statistical tests. Authors have identified a variety of methods for analyzing data in these situations (Hollis & Campbell, 1999; Schafer & Graham, 2002). [See Standard 5.a.iv.]

Sometimes researchers present results only for "high fidelity" subsamples of the data, that is, for subsamples who received high levels of the intervention. This standard requires that all cases be included in main outcomes analyses. After a complete-case analysis, it may be useful to conduct further analyses investigating dosage effects. However, these analyses are of nonrandomized samples, and authors should label results from "high fidelity" subsamples as such in reports of outcomes.

5.a.ii. Standard: Test for pretest differences and adjust for them if necessary.

Random assignment, when properly carried out, yields groups that are similar on (that is, have the same expected value for) all observed and unobserved characteristics, within the limits of sampling error. Observed differences between groups are, thus, a function of sampling error. Because sampling error is a factor, "unhappy" random assignment (or a "bad draw") may in fact lead to groups that differ in important ways on pretests. If these are identified, it is essential to adjust for these differences statistically (e.g., covariance analysis). Even if there are no pretest differences, adjusting on a set of covariates will control for chance variations and improve the precision of the impact estimates.

5.a.iii. Standard: When multiple outcomes are analyzed, there must be adjustment for multiple comparisons, i.e., correction of the experiment-wise (Type I) error rate.

5.a.iv. Standard: Analyses to minimize the possibility that observed effects are significantly biased by differential measurement attrition are essential.

Measurement attrition refers to the fact that people or other units in the intervention and control conditions may differ because of attrition (Hansen *et al.*, 1985). Differences in the nature and magnitude of attrition can bias estimates of intervention effects if they are not taken into account. Note that differential measurement attrition can occur even

when the rates of attrition are comparable across groups.

- *Desirable Standard: It is desirable that the extent and patterns of missing data from sources other than attrition be reported and handled appropriately.* Examples of this include participants who are missing at particular waves of data collection, failure to complete particular items or individual measures, and equipment failure. Schafer and Graham (2002) discuss methods of analyzing these kinds of data.

Statistically Significant Effects

5.b.i. Standard: Results must be reported for every measured outcome, regardless of whether they are positive, nonsignificant or negative.

All measures should be described and results reported, whether they reveal significant results or not, not merely those showing positive effects. We recognize the difficulty of meeting this standard given contemporary journal practices. However, this full disclosure policy seems both desirable and feasible given the development of electronic forms of publication. Reporting only statistically significant results is misleading.

5.b.ii. Standard: Efficacy can be claimed only for constructs with a consistent pattern of statistically significant positive effects.

When multiple indicators are used, most or all effects must be in the expected direction and at least one must be statistically significant.

5.b.iii. Standard: For an efficacy claim, there must be no serious negative (iatrogenic) effects on important outcomes.

Some programs have unintended negative effects. If those effects are large, or are on important outcomes, then efficacy cannot be claimed. For example, if a program decreased alcohol use but increased marijuana use, it might not be appropriate to claim it as an efficacious substance abuse prevention program; a risk-benefit analysis is needed to determine this. It must be clear that the benefit reasonably outweighs the negative side effects. Negative side effects, while very common and accepted in medicine, need to be outweighed by the benefits.

Practical Value

5.c. Standard: It is necessary to demonstrate practical significance in terms of public health impact.

It is not sufficient for program effects to be statistically significant because statistical significance conveys no information about the practical significance of the effects. Instead, researchers should strive to present their results in terms that a wide range of consumers could understand, such as reports of standardized effect sizes, odds ratios, confidence intervals or cost-effectiveness.

Researchers should also keep in mind that effect sizes that may appear small might actually be quite important. For example, in the Physician's Health Study, the correlation between taking a daily aspirin and having a second heart attack was only about $r = .03$, but the odds of experiencing a second heart attack were about 1.8 times greater among participants who took the placebo (Steering Committee of the Physicians' Health Study, 1988). Alternatively, very large effects may be unimportant if their associated costs are such that few people are likely to implement the intervention.

- *Desirable Standard: It is desirable to have/report cost and cost-effectiveness information.* Researchers do not usually estimate cost-effectiveness in efficacy trials, because the focus is on testing causal mechanisms under highly controlled conditions. Even at the efficacy level, however, researchers should give some consideration to estimating the potential costs that would be involved if the intervention were eventually taken to scale.

Duration of Effect

5.d. Standard: In general, for outcomes that may decay over time, there must be a report of significant effects for at least one long-term follow-up at an appropriate interval beyond the end of the intervention (e.g., at least 6 months). [Also see *Standard2.b.ii*]

For a program to claim efficacy, it must show effects at a meaningful long-term follow-up. The issue is how the program affects developmental course—whether it affects a meaningful social outcome at the time that that outcome should be expected to occur developmentally. For example, a drug abuse prevention program that reduces attitudes and early use but is never demonstrated to reduce harm or abuse might be of questionable value. Researchers should measure the outcome at the time it is developmentally expected. As an extension to this standard, it would be desirable to include multiple follow-ups to examine the nature of the time-course of the program effects.

Replication

5.e.i. Standard: Consistent findings are required from at least two different high-quality studies/replicates that meet all of the above criteria and each of which has adequate statistical power.

Replication to confirm findings is an important scientific principle, and any finding in science must be replicated to rule out chance findings before it can be widely accepted with confidence (Hunter, 2001; Larsen, 2004). Indeed, one could say that the unit of advancement in any field is the systematic review, not an individual study. It is important to encourage more replication studies of programs and practices whose evidence of efficacy is based on a single study. In its current state, prevention research has produced fewer replication studies than needed to reach the eventual goal of offering a wide variety of evidence-based programs and practices to the field. Any claims for program effects on population subgroups should also be replicated.

The most relevant level of replication for efficacy trials is known as exact, full or "statistical" (Hunter, 2001) replication. Exact replication refers to a replication of the same intervention on a new sample of the same population, delivered in the same way by the same kinds of people, with the same training, as in the original study. Exact replication is rarely possible. "Scientific" replication (Hunter, 2001) differs only in that the study samples come from similar populations rather than the exact same population. Thus, multi-site evaluations of a program, if the planned statistical power is adequate to analyze data from each site separately, could be considered as scientific replications. "Conceptual" replications allow for differences in procedures or measures and, if the differences do not lead to differences in results, they subsequently may be considered as scientific replications. Many replications of prevention programs, including those by independent investigators, are of this type. "Systematic" replications allow for systematic variation in the intervention, intervenors, procedures or measures, and are designed to assess the generalizability of a finding. Systematic replications with different populations also contribute to (and broaden) the efficacy statement for a program. Systematic variation in other dimensions is appropriate for effectiveness trials (See Effectiveness section later).

For some policy interventions, replication might be the only means of securing the scientific integrity of an observation of policy effectiveness. A dramatic demonstration of this has been the series of time series studies conducted on impacts of privatization legislation on alcohol use and problems. When a single observation for one state was called into question, multiple replications across states led to the convincing demonstration of the harms that arise from this legislation (Wagenaar & Holder, 1996).

Although recognizing the importance of the replication standard, we note that flexibility may be required in the application of this standard for some kinds of interventions until enough time passes to allow the prevention research enterprise to meet this high standard.

For some kinds of programs, the replication can occur in an effectiveness trial that meets all of the conditions of an efficacy trial.

- *Desirable Standard: More studies are desirable.* It is also desirable that at least one replication be conducted by independent investigators, and that organizations which choose to adopt a prevention program based on a single study seriously consider undertaking a replication study as part of the adoption effort so as to add to the body of knowledge. Ultimately, developers and investigators need to create a body of evidence to maintain a claim of efficacy.

5.e.ii. Standard: When more than two efficacy and/or effectiveness studies are available, the preponderance of evidence must be consistent with that from the two studies of highest quality.

CRITERIA FOR EFFECTIVENESS

Effectiveness trials test whether programs or policies are effective under "real-world" conditions or in "natural" settings (Flay, 1986). Effectiveness trials may also establish for whom, and under what conditions of delivery, the program or policy is effective (Flay, 1986).

Program or policy developers may or may not be involved in effectiveness studies. For broad dissemination, it is desirable eventually to have some effectiveness trials that do not involve the developer—to establish whether programs are sustained and still effective when the developer is not involved.

Every effort should be made to apply the same standards to effectiveness trials that are applied to efficacy trials, although we recognize that the challenges of doing so may be greater in real-world

settings. Effectiveness trials are heavily dependent on the relationship between the host environment and the research team, such that the intervention and measurement must be harmonious with the mission and vision of the host institution.

Efficacy Criteria

1. Standard: *To claim effectiveness, studies must meet all of the conditions of efficacy trials plus the following standards*.

Efficacy trials are not necessary before conducting effectiveness trials as long as the effectiveness trial meets all of the standards of efficacy trials. In addition, a rigorous effectiveness trial that produces results similar to those found in a single efficacy trial can provide needed replication.

Program Description and Outcomes

Program Definition

2.a. Standard: Manuals and, as appropriate, training and technical support must be readily available.

The information available must be sufficient such that *practitioners* in the field, not just other researchers, could implement the program or policy.

Intervention delivery

2.b. Standard: The intervention should be delivered under the same types of conditions as one would expect in the real world (*e.g., by teachers rather than research staff*).

Effectiveness trials should not be implemented using staff, materials or other resources that are unlikely to be available to implementers in natural settings.

Theory

2.c.i. Standard: A clear theory of causal mechanisms should be stated.

The specific outcomes that would be affected by a prevention program or policy are informed by theory and by prior empirical analyses, as described in a working model or logic frame of the presumed causal process. Such models relate antecedents or predictor variables to outcomes.

2.c.ii. Standard: A clear statement of "for whom" and "under what conditions" the intervention is expected to be effective should be stated.

For example, some interventions work for boys but not for girls (Farrell & Meyer, 1997; Flay *et al.*, 2004b; Kellam *et al.*, 1998; Perry *et al.*, 2003). An effectiveness statement about such interventions must specify that it is effective only for boys, albeit under conditions of program delivery to the mixed group. [Also see *Standard 4.b.i*]

Measures

2.d. Standard: Level of exposure should be measured, where appropriate, in both treatment and control conditions.

Two factors determine level of exposure: (1) level and integrity of implementation and (2) engagement (acceptance, compliance, adherence or involvement) of the target audience in the intervention. Effectiveness trials generally have much more variation in exposure than efficacy trials. Therefore, evaluators should document the level of variation, as it may affect ultimate program impact.

2.d.i. Standard: It is essential to measure the integrity and level of implementation/delivery of the intervention.

2.d.ii. Standard: It is essential to measure acceptance, compliance, adherence and/or involvement of the target audience and subgroups of interest in the intervention activities.

Careful documentation of the quality and quantity of the implementation of the program or policy *in both treatment and control conditions* is critical. Implementation and engagement are especially important to measure in an effectiveness trial because they provide information about the degree of difference between the treatment and control groups that produced the observed outcomes. Such information may also help explain null effects when found. Variation in program dose or quality also may provide useful information about (a) the key elements of the intervention and (b) the dosage amount and quality required to achieve observed outcomes.

- *Desirable Standard: It is desirable to measure appropriate mediators (if suggested by the theory of cause).* [See Efficacy 2.a.ii]

- *Desirable Standard: It is desirable to measure appropriate moderators (if suggested by the theory of cause)*. Moderators are variables whose values condition the size of the effect of the program or policy. That is, there is a statistical interaction between levels of the moderating variable and experimental condition. For example, a parenting skills program may have more beneficial outcomes for parents who perceive a behavioral problem in their child than for families whose parents believe that their children's behavior is under control. A tax increase to reduce harmful alcohol use may be effective in states where other controls are relatively lax, but ineffective in states in which alcohol control is strict (Gruenewald & Treno, 2000).

It is increasingly common for researchers to report stronger effects (or the only significant effects) for high-risk groups in efficacy trials. In such cases, follow-up effectiveness trials should replicate that effect, either by reporting analyses for the same high-risk group or by conducting the effectiveness trial only on the high-risk group.

We consider measurement of mediator and moderator variables to be desirable rather than essential in establishing effectiveness because not all programs have hypothesized mediating or moderated effects. When researchers or developers hypothesize mediated or moderated effects, measurement and tests of them are strongly desired.

Clarity of Causal Inference

3. Standard: The same standards as stated for efficacy apply, though the challenges are greater.

In the study of the effectiveness of prevention programs or policies, randomization is still the best approach, but the other alternatives suggested for efficacy (regression-discontinuity, time-series, or high quality matched controlled designs) may be used as well. The bottom line is that researchers should use maximally powerful designs and acknowledge the limits of the chosen design for causal inference.

Generalizability of Findings

Representative Sample

4.a. Standard: *The real-world target population and the method for sampling should be explained in order to make it as clear as possible how closely the sample represents the specified real-world target population*.

The study sample should come from the population to which the outcomes of the effectiveness trial will be generalized, and should reflect the composition of that population. This is best achieved by probability sampling (Kalton, 1983; Last, 1988) from a defined population.

Generalizability of Findings

4.b.i. Standard: The degree to which findings are generalizable should be evaluated.

One of the objectives of effectiveness studies is to establish for whom the program or policy is effective (Flay, 1986). Effectiveness trials present an important opportunity to address the specific impact of an efficacious program or policy by testing it in a variety of settings and populations. Assessing generalizability involves (1) conducting the study on a representative sample of the population to which the evaluator wants to generalize (see above standard), and/or (2) identifying subgroups to which the evaluator wants to generalize, assuring their presence in the sample, and analyzing effects by subgroup.

Absent a probability sample and appropriate a priori subgroup analyses, there are no statistical methods for assessing the generalizability of a program's effects. Therefore, we ultimately need replication in different populations. Without replication, understanding whether an effect found in one population generalizes to another population is a matter of speculation. The speculation may be informed by theory, subgroup analyses, experience, expert judgment, fragmentary data, etc. The problem of generalizability remains an important area for scientific definition and investigation. Therefore, in addition to the criteria for generalizability delineated above, investigators should conduct one or more of the following to establish the conditions and populations for which a program or policy is effective: postfacto subgroup analyses, dosage studies, or replication.

- *Desirable Standard: Subgroup analyses: If the study sample is heterogeneous with respect to important variables (such as age, gender, ethnicity/race, risk levels), it is desirable to report subgroup analyses of these groups.* Such analyses can be used to support claims that the program/policy is effective for these subgroups

(which might support statements in Standard 2.c.ii). The heterogeneity of community populations on parameters such as race/ethnicity and risk levels is an opportunity to address a central issue of prevention trials by specifying for whom the intervention works and under what conditions. All subgroup analyses should be identified either as a priori or post hoc. Subgroup analyses not only contribute to the prevention science knowledge base, but also can provide useful information to policy makers faced with making choices among prevention or treatment programs. To be able to report subgroup analyses, effectiveness trials should be designed so that sufficient numbers of individuals can be recruited from each subgroup to provide adequate statistical power for meaningful analyses and reporting of findings.

- *Desirable Standard: Experimental dosage analyses: It is desirable to conduct experimental dosage analyses.* Often, researchers will examine the relationship between levels of exposure to the intervention and outcomes. However, natural variation in exposure to the intervention may be confounded with unmeasured characteristics of the participants and/or the host environment. Therefore, variations in outcomes related to different levels of program exposure should not necessarily be interpreted as dosage effects. Statistical techniques such as propensity score analyses (Huppler-Hullsiek & Louis, 2002; McCaffrey *et al.*, 2004; Rosenbaum, 2002; Rosenbaum & Rubin, 1985), that use information about variation in dose from both treatment and control groups, are more informative than bivariate correlational analyses within the treatment group alone. Ideally, investigation of dosage effects should be done by randomly assigning individuals to different levels of the intervention (e.g., Metropolitan Area Research Group, 2002).
- *Desirable Standard: Replication with different populations: It is desirable to have one or more investigations (replications) of the manner in which findings are or are not replicated in qualitatively different population(s).* For example, Botvin and his colleagues have replicated Life Skills Training with African American and Hispanic groups in New York State and found it to be effective with both groups (Botvin *et al.*, 1994, 1995).
- *Desirable Standard: Replication with different intervention delivery agents or modes: For some types of interventions, it may be desirable to have one or more investigations (replications) of the manner in which findings are or are not replicated when delivered by different types of people or under different conditions.* One example of testing preventive intervention effectiveness using different delivery agents is the randomized controlled trial of the nurse home visitation program in Colorado (Korfmacher *et al.*, 1999; Olds, 2002; Olds *et al.*, 2004). The trial addressed a public health issue with important policy implications: whether the intervention for first-time mothers delivered by trained lay home visitors was as effective as when delivered by nurses. Findings indicated that the intervention delivered by nurse home visitors conveyed substantially more benefits to the participants than the one delivered by lay home visitors.

Given the diversity of populations, providers, and settings in which any given intervention is implemented, it is unlikely that even the best-supported intervention will be effective in every implementation. Moreover, it is possible that the effectiveness of interventions will deteriorate over time in cases where providers abandon critical procedures or otherwise make changes that are not guided by well-conceived adaptation strategies. Therefore, *ongoing multiple replications of effectiveness studies are desirable.*

Precision of Outcome

Practical Value

5.a. Standard: To be considered effective, the effects of an intervention must be practically important. Evaluation reports should report some evidence of practical importance.

Researchers should go beyond reporting the results (as required for efficacy studies), and also provide an analysis of the practical importance or public health impact of the findings (McCartney & Rosenthal, 2000). These may take the form of standardized effect sizes, confidence intervals, odds ratios, percent relative change, cost-effectiveness or other practical measures of the magnitude of the practical value.

- *Desirable Standard: It is desirable to have reports of costs and cost-effectiveness analyses.* Cost information is essential for policy makers if a promising intervention is to go to scale. This is despite the fact that an appropriate economic analysis is difficult to do well when testing preventive interventions (Aos *et al.*, 2004; Caulkins *et al.*, 2004; Foster *et al.*, 2003; Greenwood, 2005; Zarkin & Hubbard, 1998). Increasingly, efforts to develop lists of "model" or "promising" preventive interventions include information on economic analyses (Aos *et al.*, 2004; Benedict *et al.*, 2000). As one example, the Nurse-Family Partnership program has been shown to be a good investment (Aos *et al.*, 2004). In the Colorado effectiveness test of the nurse home visitation program mentioned above (Olds *et al.*, 2004), the costs of program delivery by lay home visitors were initially lower than those by nurse home visitors. Because lay visitors required more training and monitoring, and had greater turnover, however, the long-term costs associated with lay visitors were comparable to those associated with nurse home visitors.

Replication

5.b. Standard: Consistent findings are required from at least two different high-quality trials that meet all of the above criteria and each of which has adequate statistical power.

Replication reduces the likelihood that chance alone can explain the findings, and increases our confidence in the effectiveness statement. Effectiveness can be claimed only for those outcomes for which there are similar effect sizes in the preponderance of evidence from effectiveness trials, within the constraints of sampling error.

- *Desirable Standard: It is desirable to have more than two replications.*

CRITERIA FOR BROAD DISSEMINATION

For the purposes of these Standards, we define dissemination broadly as incorporating scaling up, adoption, implementation and sustainability. To be ready for broad dissemination, a program or policy must not only be of proven effectiveness, but it must also meet other criteria that ensure that agencies will adopt it and providers (teachers, counselors, social workers, service agencies, etc) can effectively use it. Broad dissemination requires prevention programs or policies that are effective and that lend themselves to predictably effective use in the field. Successful dissemination is more likely when measures of program delivery, fidelity, and proximal goals are designed into implementation efforts. Other characteristics that increase the probability of effective prevention program dissemination include an understanding of system elements that can foster adoption and sustainable program delivery, and a smoothly functioning relationship between program developers and those responsible for dissemination in the field.

Despite wide agreement that dissemination is the ultimate purpose of efforts to develop effective programs or policies, little empirical work has focused on the process through which interventions are adopted, implemented, and sustained. Greater research investment must be made in research on how organizations and individuals adopt effective interventions. Of equal value will be studies that examine the process through which programs or policies are consistently implemented with high quality—so that they have their desired effect. In addition, because programs or policies can only achieve lasting effects if they are sustained, we need new work on strategies to ensure the continuation of effective delivery over time, in different contexts, and in the hands of new providers in the field. In its strategic plan, SPR articulates its commitment to the promotion of research on prevention program and policy adoption, implementation, and sustainability.

Effectiveness

1. Standard: To be ready for broad dissemination, a program or policy must meet all of the criteria for effectiveness and must be supported by relevant provider materials and by evidence that the program or policy can be implemented with fidelity.

Effectiveness is essential for programs or policies to be disseminated. But these interventions must also include complete and user-friendly materials necessary for their replication. Effective programs or policies should not be widely disseminated until there is evidence that replications can be done with the fidelity that originally attended the intervention's effectiveness testing. Though developers cannot assure

that an effective program or policy will achieve positive results whenever it is implemented, they must equip providers with the wherewithal to replicate the program or policy in the manner it was originally tested.

Going to Scale

2. Standard: The program or policy must have the ability to go to scale, including providing all materials and necessary services including, as appropriate, a manual, training and technical support.

Each of the following may be necessary for successful adoption, implementation and sustainability of effective programs or policies (Elliott & Mihalic, 2004; Fagan & Mihalic, 2003; Greenberg *et al.*, 2001):

i. A standard training and technical assistance process.
ii. Training manuals that provide information beyond what is included in the implementation manual about how to prepare for and implement the activity.
iii. A statement of the presumed causal mechanisms or logic model relating the intervention to the outcomes.
iv. An infrastructure for adequately managing training, technical assistance, and materials.
v. Listing of the conditions and resources needed to support adoption, implementation and sustainability.

Given the current state of research on dissemination, these elements may not be sufficient for successful program or policy adoption and implementation. For example, substantial, high-quality partnerships between providers and adopters may be necessary such that the program or policy becomes "owned" by the adopting agency and is seen to meet their mission.

Cost Information

3. Standard: Clear cost information must be readily available.

Information about all costs to implement a program or policy in the field is essential. Knowing what financial and staff resources an intervention consumes will allow practitioners and policy makers to make informed decisions about adopting an effective prevention program or policy. Further, cost information allows comparisons of interventions based on their relative impact and benefits.

Intervention cost information, for example, might encompass nonresearch investments in delivery staff training, on-site time, any necessary facility, equipment, or resource rental and maintenance, and reproduction of materials (Plotnick *et al.*, 1998). Programs must estimate the value of volunteer labor and donated space and equipment to assess opportunity costs (Chatterji *et al.*, 2001; Lillie-Blanton *et al.*, 1998; Foster *et al.*, 2003). Additionally, programs should compile attendant delivery costs for consultants, clerical staff, and physical plants. School-based programs, for example, would also need to similarly assess delivery costs to encompass teacher time, training workshops and materials, classroom use, and school building overhead. Among policy interventions, costs often go far beyond the policies themselves, like enforcement costs unanticipated by the original policy action.

Full disclosure of the costs of programs or policies, therefore, should encompass not simply the value of intervention materials, but all burdens placed on the potential delivering organization and staff. In this way, those charged with selecting and implementing effective interventions can make informed decisions based on financial considerations.

Ongoing Monitoring and Evaluation

4. Standard: Monitoring and evaluation tools must be available to providers.

The dissemination of effective preventive interventions is incomplete without proper tools to monitor and evaluate replications. Unless prevention programs and policies are implemented with fidelity, their impact can be lessened (Elliott & Mihalic, 2004; Greenberg, 2004). Fidelity is best assessed with implementation monitoring tools. Practitioners and policy makers are facing increased pressure to demonstrate that public and private dollars entrusted to them are well spent. When those scarce dollars are devoted to effective preventive interventions, the surest method of determining whether they are spent wisely is through evaluations of program or policy implementation and outcomes.

Monitoring tools should include checklists, process data-gathering forms, and related procedures to assess program delivery. Equipped with these tools, those involved in implementations can determine (1) whether the program was delivered as originally

designed, (2) what changes in delivery are indicated when sufficient time remains to effect changes, (3) how the delivery agents, end-use consumers, and other interested parties responded to the program or policy, and (4) the outcomes obtained. Such data can be fed back to implementers to maintain or improve their implementation on a continual basis. Effective programs and policies must provide detailed guidelines and actual data collection measures to aid in these monitoring tasks.

Evaluation tools should include illustrative designs, measurement schedules and protocols, analytic strategies, and detailed guidelines. Notwithstanding their own prior research on effective programs or policies, developers cannot satisfy the need for evaluation tools by merely offering up their own measures and methodologies. Rather, they must make available to implementers expressly designed tools for use in the applied field settings. These tools should consider the real-world differences between academic research and applied evaluation. What is more, any evaluation tool should have been previously tested and include templates and illustrative or sample end products so implementers can bring the evaluation to its full conclusion by properly reporting their findings.

Replication Studies

- *Desirable Standard: Organizations choosing to adopt programs or policies that do not necessarily meet all criteria should consider undertaking a replication study.* In particular, organizations adopting programs, possibly in partnership with the original developer or others, should consider designing and executing a study of the effects of the program or policy in their own site. Replications are particularly helpful when interventions are implemented with new populations, in novel settings, or by providers previously not engaged in intervention delivery.

Programs or policies that require program changes (including translation into another language), or otherwise undergo significant adaptations to their protocol or format, should also be targets of replication studies. When programs or policies are greatly changed to respond to emerging issues, it might be appropriate to cycle back to the efficacy or effectiveness stage because the intervention is no longer the same as tested in prior efficacy and/or effectiveness trials. By examining the processes and outcomes of program replications, such studies will add to the body of knowledge on an intervention and strengthen its empirical foundation.

Sustainability

- *Desirable Standard: It is desirable to have a clear statement of the factors that are expected to assure the sustainability of the program or policy once it is implemented.* We expect this issue to become the focus of future dissemination research efforts.

DISCUSSION

Prevention science has advanced greatly in a short time. Not long ago, practitioners and policy makers alike wondered whether preventive interventions could deliver positive, lasting outcomes. More recently, the field has been concerned about whether programs and policies developed and tested in research settings could be implemented with predictable success in schools, social agencies, and communities. Today, we have ample evidence that not only do many preventive interventions work, but also that they lend themselves to wide ranging application in everyday settings. The field is now challenged by questions about which interventions work, how well those programs and policies work with different populations and in various delivery settings, and the extent to which the quality of their research support warrants their wide dissemination.

The search to answer these questions scientifically motivated the Society for Prevention Research to develop the present Standards of Evidence for prevention programs and policies. We intend the resulting standards presented in this report to guide decisions about causal effects of interventions delivered under ideal and naturalistic conditions. This paper is an expanded discussion of the Standards accepted by the SPR Board in April 2004 (as such, the standards are stated here in the same form and with the same numbering as in the document approved by the Board). Although this paper was written by and for researchers, we hope that policy makers and funders will also find the report and its findings useful.

The Committee on Standards determined that different but overlapping sets of standards were

needed to establish that a preventive intervention has been (1) tested and proven efficacious under optimal conditions, (2) tested and proven effective under real-world conditions, or (3) tested and proven to be ready for dissemination. Given the state-of-the-field in prevention program or policy development and testing, we also found that it would be unrealistic to require some standards at this time. Consequently, we also developed some desirable standards. Because the setting of standards is a dynamic process, standards that are desirable today may become required at some time in the future. Indeed, some users of the Standards may determine that the area of prevention research they wish to summarize is already at a stage where some standards we list here as desirable should be required. Alternatively, some users of the Standards may decide that some standards identified by SPR as required cannot be met in their particular area of prevention research. Ultimately, users must decide which of the standards to apply to their particular area. However, the Standards Committee cautions against using only a subset of the "required" standards or otherwise weakening the standards. Standards listed in this document should provide the base upon which future standards should build.

Under these Standards, a preventive intervention can be considered efficacious when it has been tested in at least two rigorous trials that involved: (1) defined samples from defined populations; (2) psychometrically sound measures and data collection procedures; (3) rigorous statistical approaches appropriate to the research and sampling design; (4) consistent positive effects (without serious iatrogenic effects); and (5) significant findings maintained through at least one long-term follow-up. Moreover, an effective intervention will not only meet all standards for efficacious interventions, but will also: (1) offer manuals, appropriate training, and technical support available to allow third parties to adopt and implement the interventions; (2) be evaluated under real-world conditions in studies that included sound measurement of the level of implementation and engagement of the target audience in both the intervention and control conditions; (3) demonstrate the practical importance of intervention outcome effects; and (4) specify the population to whom intervention findings can be generalized. Finally, preventive interventions recognized as ready for broad dissemination will not only meet all standards for efficacious and effective interventions, but will also provide: (1) evidence of the ability to "go to scale;" (2) clear cost information; and (3) monitoring and evaluation tools so that adopting agencies can monitor or evaluate how well the intervention works within their settings. The standards are listed in Table 1 and the numbers of standards in broad categories are summarized in Table 2.

To summarize, SPR's Committee on Standards developed

- a set of 21 required standards for determining that an intervention has been tested and proven efficacious, with an additional 9 standards (for a total of 30) that are desirable;
- an additional 10 required standards (for a total of 31 required standards) for determining that an intervention is tested and effective, with an additional 12 standards that are desirable (for a total of 43); and
- an additional three required standards (for a total of 34 required standards) for determining that an intervention is of proven effectiveness and ready for broad dissemination, with an additional 13 desirable standards (for a possible total of 47).

As these Standards meet the Society's expectations, they will advance the science of prevention and will increase the use of effective preventive interventions to improve the public health. For example, by emphasizing replications, these guidelines will ideally shape a new agenda of a too-long neglected area of prevention research. We also need more studies of effectiveness and dissemination that evaluate the outcomes of differential implementation. Similarly, the development and validation of measures of program implementation and adaptation are also critically important and much needed within our field.

Administrators, communities and policy makers can use these Standards to select prevention programs and policies to improve the public health. The Standards facilitate the identification of effective programs and policies that can meet local community needs. In so doing, the Standards set the stage to move programs and policies from demonstrations of efficacy and effectiveness to widespread public use.

Notwithstanding their potential benefits for the field, the present Standards could be misused to stultify prevention science and practice (Rosenstock & Lee, 2002). For example, the Standards could be employed to dismiss programs as yet untested, to remove funding or resources for developing new

Table 1. Forty-Seven SPR Standards for Efficacy, Effectiveness and Dissemination

		Standards	Efficacy	Effectiveness	Dissemination
Specificity of the efficacy statement					
1	EY	1. Statement of efficacy is of the form: Program X is efficacious for producing Y outcomes for Z population	X	X	X
	EV	1. To claim effectiveness, studies must meet all conditions of efficacy trials plus others		X	X
	DI	1. To claim readiness for dissemination, must meet all criteria for effectiveness			X
Program description and measures					
2	EY	2.a. Described at a level that would allow others to implement/replicate it	X	X	X
3	EV	2.a. Manuals and appropriate training and technical support readily available		X	X
4	EV	2.b. Intervention delivered under conditions expected in the real world		X	X
5	EV	2.c.i. Stated theory of causal mechanisms		X	X
6	EV	2.c.ii. Statement of "for whom?" and "under what conditions?" intervention is effective		X	X
7	BDI	2. Evidence of ability to go to scale			X
8	EY	2.b.i. The stated public health or behavioral outcome(s) of the intervention must be measured	X	X	X
9	EY	2.b.ii. There must be at least one long-term follow-up measure	X	X	X
	EY	2.c. Psychometrically sound measures			
10	EY	2.c.i. Valid measures of the targeted behavior	X	X	X
11	EY	2.c.ii. Internal consistency (alpha), test-retest reliability, and/or reliability across raters	X	X	X
12	EY	2.c.iii. At least one form of data collected by people independent of the intervention	X	X	X
	EV	2.d. Level of exposure measured in both treatment and control conditions			
13	EV	2.d.i. Integrity and level of implementation/deliver of the intervention	*x*	X	X
14	EV	2.d.ii. Engagement of the target audience and subgroups of interest	*x*	X	X
15	ODC	*Measures of mediating variables (or immediate program effects)*	*x*	*x*	*x*
16	ODC	*Measures of potential side-effects*	*x*	*x*	*x*
17	ODC	*Multiple measures of constructs*	*x*	*x*	*x*
18	ODC	*Measures of moderating variables*		*x*	*x*
Clarity of causal inference					
	EY	3. Research design allows for unambiguous causal statements			
19	EY	3.a. Design has at least one comparison condition that does not receive the tested intervention	X	X	X
20	EY	3.b. Assignment to conditions maximizes confidence in causal statements	X	X	X
	EY	3.b.i. For most kinds of interventions, random assignment (of sufficient N without sig pretest differences)			
	EY	3.b.ii. For some kinds of large-scale interventions, repeated time-series designs without randomization			
	EY	3.b.iii. Well-conducted regression-discontinuity designs (selection model is completely known)			
	EY	3.b.iv. Matched control designs with pretest equivalence and when assignment not by self-selection			
Generalizability of findings					
21	EY	4.a. Report specifies what/who the sample is and how it was obtained	X	X	X
22	EV	4.a. Real-world target population and the method for sampling it is explained		X	X
23	EV	4.b.i. Degree to which findings are generalizable is evaluated		X	X
24	ODC	*Reports of subgroup analyses*	*x*	*x*	*x*
25	ODC	*Experimental dosage studies/analyses*		*x*	*x*
26	ODC	*Replication with different populations*		*x*	*x*
27	ODC	*Replication with different program providers*		*x*	*x*
Precision of outcomes					
	EY	5.a. Statistical analysis allows unambiguous causal statements			
28	EY	5.a.i. Main effects analysis at the same level as the randomization and includes all cases	X	X	X
29	EY	5.a.ii. Tests of pretest differences and adjustments for them if necessary	X	X	X
30	EY	5.a.iii. When multiple outcomes are analyzed, adjustments for multiple comparisons	X	X	X
31	EY	5.a.iv. Analyses minimize possibility that effects are due to differential measurement attrition	X	X	X

Table 1. Continued

		Standards	Efficacy	Effectiveness	Dissemination
32	ODC	*Report the extent and patterns of missing data*	*x*	*x*	*x*
Statistically significant effects					
33	EY	5.b.i. Results reported for every measured outcome	X	X	X
34	EY	5.b.ii. A consistent pattern of statistically significant positive effects	X	X	X
35	EY	5.b.iii. No serious negative (iatrogenic) effects on important outcomes	X	X	X
36	EY	5.c. Demonstrated practical public health impact	X	X	X
37	EV	5.a. Evaluation reports some evidence of practical importance		X	X
38	BDI	3. Clear cost information readily available	*x*	*x*	X
39		*Report costs and cost-effectiveness analyses*		*X*	*X*
40	EY	5.d. Significant effects for at least one long-term follow-up	X	X	X
41	EY	5.e.i. At least 2 high-quality studies/replicates that meet all of the above criteria for efficacy	X	X[a]	X[a]
42	EY	5.e.ii. Preponderance of evidence consistent with that from the 2 highest quality studies	X	X	X
43	EV	5.b. Consistent findings from at least 2 different high-quality effectiveness trials		X	X
44	ODC	*The more replications the better*	*x*	*x*	*X*
45	ODC	*Independent replications by organizations adopting programs*			*x*
46	BDI	4. Monitoring and evaluation tools available to providers		X	
47	ODC	*Statement of factors expected to assure program sustainability*		*x*	

Note. Desirable standards are shown in italics and small italicized x's. Numbering in first column is simply a running count. Numbers in columns 2–4 are numbers used in the text. EY: Efiicacy; EV: Effectiveness; BDI: Broad dissemination; ODC: Other desirable criteria.
[a]A program may have results from effectiveness trials without separate efficacy trials.

interventions, or to halt programs on the developmental course from efficacy to broad implementation. The field of prevention science has made vast progress in developing and testing preventive interventions in the past decade. These efforts have already yielded a number of clearly efficacious programs. Using these Standards, the field will be able to rapidly increase the number of clearly efficacious studies and add several more that are also effective and ready for broad dissemination.

Table 2. Number[a] of Criteria by Major Category and Level

	Efficacy	Effectiveness	Dissemination
Efficacy statement	1	1	1
Program description	1	5	6
Measures	5 (5)	7 (4)	7 (4)
Design	2	2	2
Generalizability	1 (1)	3 (4)	3 (4)
Unambiguous causality	4 (1)	4 (1)	4 (1)
Significance/practicality	4	5	5 (1)
Cost/cost-effectiveness	0 (1)	0 (2)	1 (1)
Long-term effects	1	1	1
Replication	2 (1)	3 (1)	3 (2)
Evaluation tools	0	0	1
Sustainability	0	0	0 (1)
Total required	21	31	34
Total desirable	9	12	13
Possible total	30	43	47

[a]First number indicates the number of required standards, parenthetical are the additional number of desirable standards.

The relative emphasis in this paper on efficacy trials, effectiveness studies, and dissemination studies reflects the state of the field. When funding agencies and researchers give greater emphasis to conducting more effectiveness and dissemination trials, these Standards will be strengthened in those areas. In the meantime, we have offered some forward-looking standards for effectiveness and dissemination, and hope to revisit and refine them as the field matures.

Ultimately, these Standards should serve to increase the quality of prevention research and, by extension, to improve prevention program/policy implementation and outcomes and contribute to the reduction of health disparities. This process can occur, for example, as intervention adopters and researchers form partnerships in the scientific enterprise and move toward better understanding of program adoption, implementation and sustainability. Of greatest and overarching importance, the present Standards should lead to the adoption and delivery of effective programs and policies that will reach their intended end users. These are the children, youth, and families who have much to gain from evidence-based prevention efforts aimed at increasing their health, psychological functioning, and social wellbeing. To achieve these goals is the reason that the Society for Prevention Research was founded and to which it is dedicated.

ACKNOWLEDGMENTS

All points of view are those of the Society for Prevention Research and the authors, and do not necessarily reflect those of their employers or their funders. Preparation of these Standards and this paper were sponsored by the Society for Prevention Research with support from the National Institutes of Health and the Robert Wood Johnson Foundation, coordinated through the National Science Foundation. We thank Hendricks Brown, Bob Granger, Joel Grube, Paul Gruenewald, Harold Holder, Cheryl Perry, Rick Price, John Reid, Bob Saltz, and Irwin Sandler for helpful comments.

REFERENCES

Angrist, J. D., & Lavy, V. (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly Journal of Economics*, *114*(2), 533–575.

Aos, S., Lieb, R., Mayfield, J., Miller, M., & Pennucci, A. (2004). Benefits and costs of prevention and Early intervention programs for youth. *Washington State Institute for Public Policy*. Available at http://www.wsipp.wa.gov/rptfiles/04-07-3901.pdf.

Barkham, M., Margison, F., Leach, C., Lucock, M., Mellor-Clark, J., Evans, C., Benson, L., Connell, J., Audin, K., & McGrath, G. (2001). Service profiling and outcomes benchmarking using the CORE-OM: Toward practice-based evidence in the psychological therapies. *Journal of Consulting and Clinical Psychology*, *69,* 184–196.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology*, *51,* 1173–1182.

Truman, B. I., Smith-Akin, C. K., Hinman, A. R., *et al.* (2000). Developing the guide to community preventive services—Overview and rationale. *American Journal of Preventive Medicine*, *18*(1S), 18–26.

Berends, M., & Garet, M. S. (2002). In (re)search of evidence-based school practices: Possibilities for integrating nationally representative surveys and randomized field trials to inform educational policy. *Peabody Journal of Education*, *77*(4), 28–58.

Berkson, J. (1946). Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*, *2,* 47–53.

Berkson, J. (1958). Smoking and lung cancer: some observations on two recent reports. *Journal of the American Statistical Association*, *53,* 28–38.

Bertrand, M., Duflo, E., & Mullainathan, S. (2002). *How much should we trust differences in differences estimates?* National Bureau of Economic Research (NBER) Working Paper 8841. Washington, DC: NBER.

Biglan, A., Ary, D., & Wagenaar, A. C. (2000). The value of interrupted time-series experiments for community intervention research. *Prevention Science*, *1*(1), 31–49.

Biglan, A., Brennan, P. A., Foster, S. L., Holder, H. D., Miller, T. L., Cunningham, P. B., *et al.* (2004). *Helping adolescents at risk: Prevention of multiple problems of youth*. New York: Guilford.

Boruch, R. F. (Ed). (2005a). Place randomized trials: special issue. *Annals of the American Academy of Political and Social Sciences*, *599*(1), whole issue.

Boruch, R. F. (2005b). Comments on Jacob and Ludwig. In D. Ravitch (Ed.), *Brookings papers on educational policy* (pp. 67–73). Washington, DC: Brookings Institution Press.

Botvin, G. J., Schinke, S. P., Epstein, J. A., Diaz, T., & Botvin, E. M. (1994). Effectiveness of a culturally focused and generic skills training approaches to alcohol and drug abuse prevention among minority youths. *Psychology of Addictive Behaviors*, *8,* 116–127.

Botvin, G. J., Schinke, S. P., Epstein, J. A., Diaz, T., & Botvin, E. M. (1995). Effectiveness of cultural 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,* 183–194.

Brown, C. H. (1993). Statistical methods for prevention trials in mental health. *Statistics in Medicine*, *12,* 289–300.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods.* Newbury Park, CA: Sage.

Caulkins, J. P., Pacula, R. L., Paddock, S., & Chiesa, J. (2004). What we can–and cannot–expect from school-based drug prevention. *Drug and Alcohol Review*, *23*(1), 79–87.

Chatterji, P., Caffray, C. M., Jones, A. S., Lillie-Blanton, M., & Werhamer, L. (2001). Applying cost analysis methods to school-based prevention programs. *Prevention Science*, *2*(1), 45–56.

Clarke, G. N., Hawkins, W., Murphy, M., Sheeber, L. B., Lewinsohn, P. M., & Seeley, J. R. (1995). Targeted prevention of unipolar depressive disorder in an at-risk sample of high school adolescents: A randomized trial of a group cognitive intervention. *Journal of American Academy of Child and Adolescent Psychiatry*, *34,* 312–321.

Coalition for Evidence-Based Policy. (2003). *Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide*. Washington, DC: The Council for Excellence in Government.

Conduct Problems Prevention Research Group. (1992). A developmental and clinical model for the prevention of conduct disorders: The Fast Track Program. *Development and Psychopathology*, *4,* 509–527.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Skokie, IL: Rand McNally.

Cook, T. D., & Payne, M. R. (2002). Objecting to the objections to using random assignment in educational studies. In F. Mosteller & R. Boruch (Eds.), *Evidence matters: Randomized trials in education research* (pp. 150–178). Brookings Institution Press: Washington, DC.

Dawson-McClure, S. R., Sandler, I. N., Wolchik, S. A., & Millsap, R. E. (2004). Prediction and reduction of risk for children of divorce: A six-year longitudinal study. *Journal of Abnormal Child Psychology*, *32,* 175–190.

Dishion, T. J., McCord, J., & Poulin, F. (1999). When interventions harm: Peer groups and problem behavior. *American Psychologist*, *54*(9), 755–764.

Dolan, L. J., Kellam, S. G., Brown, C., Werthamer-Larsson, L., *et al.* (1993). The short-term impact of two classroom-based preventive interventions on aggressive and shy behaviors and poor achievement. *Journal of Applied Developmental Psychology*, *14*(3), 317–345.

Elliott, D. S., & Mihalic, S. (2004). Issues in Disseminating and Replicating Effective Prevention Programs. *Prevention Science, 5*(1), 47–52.

Eddy, J. M., Reid, J. B., & Fetrow, R. A. (2000). An elementary school-based prevention program targeting modifiable antecedents of youth delinquency and violence: Linking the Interests of Families and Teachers (LIFT). *Journal of Emotional and Behavioral Disorder*, *8*(3), 165–176.

Edgerton, E. A., Orzechowski, K. M., & Eichelberger, M. R. (2004). Not all child safety seats are created equal: The

potential dangers of child booster seats. *Pediatrics*, *113,* e153–e158.

Fagan, A., & Mihalic, S. (2003). Strategies for enhancing the adoption of school-based prevention programs: Lessons learned from the Blueprints for Violence Prevention replications of the Life Skills Training Program. *Journal of Community Psychology*, *31,* 235–253.

Farrell, A., & Meyer, A. (1997). The effectiveness of a school-based curriculum for reducing violence among urban sixth-grade students. *American Journal of Public Health*, *87*(6), 979–984.

Fisher, C. B., Hoagwood, K., Boyce, C., Duster, T., Frank, D. A., Grisso, T., Levine, R. J., Macklin, R., Spencer, M. B., Takanishi, R., Trimble, J. E., & Zayas, L. H. (2002). Research ethics for mental health science involving ethnic minority children and youths. *American Psychologist*, *57*(12), 1024–1040.

Flannery, D. J., Vazsonyi, A. T., Liau, A. K., Guo, S., Powell, K. E., Atha, H., Vesterdal, W., & Embry, D. (2003). Initial behavior outcomes for the PeaceBuilders universal school-based violence prevention program. *Developmental Psychology*, *39,* 292–308.

Flay, B. R. (1986). Efficacy and effectiveness trials (and other phases of research) in the development of health promotion programs. *Preventive Medicine*, *15,* 451–474.

Flay, B. R., Biglan, A., Boruch, R. F., Castro, F. G., Gottfredson, D., Kellam, S. G., Moscicki, E. K., Schinke, S., Valentine, J. C., & Ji, P. (2004a). Standards of evidence: Criteria for efficacy, effectiveness and dissemination. Falls Church, VA: Society for Prevention Research. Available at http://www.preventionresearch.org/StandardsofEvidencebook.pdf.

Flay, B. R., Graumlich, S., Segawa, S., Burns, J. L., Holliday, M. Y., & Aban Aya Investigators. (2004b). Effects of two prevention programs on high-risk behaviors among African-American youth: A randomized trial. *Archives of Pediatric and Adolescent Medicine*, *158*(4), 377–384.

Foster, E. M., Dodge, K. A., & Jones, D. (2003). Issues in the economic evaluation of prevention programs. *Applied Developmental Science*, *7*(2), 76–86.

Glazerman, S., Levy, D. M., & Myers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. *The annals of the American Academy of Political and Social Science*, *589,* 63–93.

Greenwood, P. (2005). *Changing lives: Delinquency prevention as crime control policy*, Chicago, IL: University of Chicago Press.

Goodstadt, M. (1978). Alcohol and drug education: Models and outcomes. *Health Education Monographs*, *6,* 263–279.

Gottfredson, M. R., & Hirschi, T. (1990). *A general theory of crime*. Stanford: Stanford University Press.

Griffin, K. W., Botvin, G. J., & Nichols, T. R. (2004). Long-term follow-up effects of a school-based prevention program on adolescent risky driving. *Prevention Science*, *5,* 207–212.

Greene, W. H. (1993). *Econometric analysis*. New York: MacMillan.

Greenberg, M. T. (2004). Current and future challenges in school-based prevention: The researcher perspective. *Prevention Science*, *5*(1), 5–13.

Greenberg, M., Domitrovich, C., Graczyk, P., & Zins, J. (2001). *The study of implementation in school-based preventive interventions: Theory, research and practice*. Washington, DC: Center for Mental Health Services, Substance Abuse and Mental Health Administration, U.S. Department of Health and Human Services.

Gruenewald, P. J., & Treno, A. J. (2000). Local and global alcohol supply: Economic and geographic models of community systems. *Addiction*, *95,* S537–S549.

Gunn, B., Smolkowski, K., Biglan, A., & Black, C. (2002). Supplemental instruction in decoding skills for Hispanic and non-Hispanic students in early elementary school: A follow-up. *The Journal of Special Education*, *36*(2), 69–79.

Hansen, W. B., Collins, L. M., Malotte, K. C., Johnson, C. A., & Fielding, J. E. (1985). Attrition in prevention research. *Journal of Behavioral Medicine*, *8,* 261–275.

Hansen, W. B., & Dusenbury, L. (2001). Building capacity for prevention's next generation. *Prevention Science*, *2*(4), 207–208.

Hedeker, D., Gibbons, R. D., & Flay, B. R. (1994). Random-effects regression models for clustered data: With an example from smoking prevention research. *Journal of Consulting and Clinical Psychology*, *624,* 57–765.

Holder, H. D. (1998). *Alcohol and the community: A systems approach to prevention*. Cambridge, UK: Cambridge University Press.

Holder, H., Boyd, G., Howard, J., Flay, B., Voas, R., & Grossman, M. (1995). Alcohol-problem prevention policy: The need for a phases research model. *Journal of Public Health Policy*, *16*(3), 324–346.

Holder, H. D., Flay, B., Howard, J., Boyd, G., Voas, R. B., & Grossman, M. (1999). Phases of alcohol problem prevention research. *Alcoholism: Clinical and Experimental Research*, *23*(1), 183–194.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81,* 945–960.

Hollis, S., & Campbell, F. (1999). What is meant by intention to treat analysis? Survey of published randomised controlled trials. *British Medical Journal*, *319,* 670–674.

Hunter, J. E. (2001). The desperate need for replications. *Journal of Consumer Research*, *28*(1), 149–158.

Huppler-Hullsiek, K., & Louis, T. A. (2002). Propensity score modeling strategies for the causal analysis of observational data. *Biostatistics*, *2,* 1–15.

Jacob, B., & Ludwig, J. (2005). Can the federal government improve government research? In D. Ravitch (Ed.), *Brookings papers on educational policy* (pp. 47–66). Washington, DC: Brookings Institution Press.

Kalton, G. (1983). *Introduction to survey sampling*. Beverly Hills, CA: Sage.

Kellam (2000). *Community and institutional partnerships for school violence prevention.* In Preventing School Violence: Plenary Papers of the 1999 Conference on Criminal Justice. Research and Evaluation–Enhancing Policy and Practice Through Research, Volume 2 NCJ 180972 (pp. 1–21). Washington, DC.

Kellam, S. G., & Langevin, D. J. (2003). A framework for understanding "evidence" in prevention research and programs. *Prevention Science*, *4*(3), 137–153.

Kellam, S. G., Ling, X., Merisca, R., Brown, C., & Ialongo, N. (1998). The effect of the level of aggression in the first grade classroom on the course and malleability of aggressive behavior into middle school. *Development and Psychopathology*, *10*(2), 165–185.

Kellam, S. G., Koretz, D., & Moscicki, E. K. (1999). Core elements of developmental epidemiologically based prevention research. *American Journal of Community Psychology*, *27*(4), 463–482.

Kellam, S. G., Rebok, G. W., Mayer, L. S., & Ialongo, N. (1994). Depressive symptoms over first grade and their response to a developmental epidemiologically based preventive trial aimed at improving achievement. *Development and Psychopathology*, *6*(3), 463–481.

Kelly, J. A., Murphy, D. A., Sikkema, K. J., McAuliffe, T. L., Roffman, R. A., Solomon, L. J., Winett, R. A., & Kalichman, S. C. (1997). Randomised, controlled, community-level HIV-prevention intervention for sexual-risk behaviour among homosexual men in US cities. Community HIV Prevention Research Collaborative, *Lancet*, *350,* 1500–1505.

Kenny, D. A., & Judd, C. M. (1986). Consequences of violating the independence assumption in analysis of variance. *Psychological Bulletin*, *99*(3), 422–431.

Korfmacher, J., O'Brien, R., Hiatt, S., & Olds, D. (1999). Differences in program implementation between nurses and paraprofessionals in prenatal and infancy home visitation: A randomized trial. *American Journal of Public Health*, *89*(12), 1847–1851.

Larsen, K. S. (2004). What is replication? The development of valid and reliable social theories. *PsycCRITIQUES*, [np].

Last, J. L. (1988). *A dictionary of epidemiology*. New York: Oxford University Press.

Lillie-Blanton, M. L., Werthamer, L., Chatterji, P., Fienson, C., & Caffray, C. (1998). Issues and methods in evaluating costs, benefits, and cost-effectiveness of drug abuse prevention programs for high-risk youth. In W. J. Bukoski & R. I. Evans (Eds.), *Cost-benefit/cost effectiveness research of drug abuse prevention*: *Implications for programming and policy* [NIDA Research Monograph Series no. 176] (pp. 184–213). Washington, DC: U.S. Government Printing Office.

Lynagh, M., Perkins, J., & Schofield, M. (2002). An evidence-based approach to health promoting schools. *Journal of School Health*, *72,* 300–302.

Madison, S. M., McKay, M. M., Paikoff, R., & Bell, C. (2000). Basic research and community collaboration: necessary ingredients for the development of a family-based HIV prevention program. *AIDS Education and Prevention*, *12,* 281–298.

Mihalic, S. (2002–2004). Matrix of programs as identified by various federal and private agencies. Boulder, CO: Center for the Study and Prevention of Violence, University of Colorado. Available at http://www.colorado.edu/cspv/blueprints/matrix/overview.html

Manski, C. (1995). *Identification problems in the social sciences*. Cambridge, MA: Harvard University Press.

Markman, H. J., Renick, M. J., Floyd, F. J., Stanley, S. M., & Clements, M. (1993). Preventing marital distress through communication and conflict management training: A 4- and 5-year follow-up. *Journal of Consulting and Clinical Psychology*, *113,* 153–158.

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, *9*(4), 403–425.

McCartney, K., & Rosenthal, R. (2000). Effect size, practical importance, and social policy for children. *Child Development*, *71*(1), 173–180.

Meinert, C. L. (1986). *Clinical trials: Design*, *conduct, analysis*. New York: Oxford University Press.

Metropolitan Area Research Group [Eron, L. D., Huesmann, L. R., Spindler, A., Guerra, N. G., Henry, D., Tolan, P., & VanAcker, R.] (2002). A cognitive-ecological approach to preventing aggression in urban settings: Initial outcomes for high-risk children. *Journal of Consulting and Clinical Psychology*, *70*(1), 179–194.

Moher, D., Schultz, K. F., & Altman, D. (2001). The CONSRT statement: Revised recommendations for improving the quality of reports of parallel-group randomized trials. *Journal of the American Medical Association*, *285*(15), 1987–1991.

Mościcki, E. K. (1993). Fundamental methodological considerations in controlled clinical trials. *Journal of Fluency Disorders*, *18,* 183–196.

Murray, D. M. (1998). *Design and analysis of group-randomized trials*. New York: Oxford University Press.

Nerlove, M., & Diebold, F. (1990). Unit roots in economic time series: A selective survey. In T. Bewley (Ed.), *Advances in econometrics* (Vol. 8). New York: JAI.

Olds, D., Henderson, C. Jr., Kitzman, H., Eckenrode, J., Cole, R., & Tatelbaum, R. (1998). The promise of home visitation: Results of two randomized trials. *Journal of Community Psychology*, *26*(1), 5–21.

Olds, D. L. (2002). Prenatal and infancy home visiting by nurses: From randomized trials to community replication. *Prevention Science*, *3*(3), 153–172

Olds, D. L., Robinson, J., Pettitt, L., Luckey, D. W., Holmberg, J., Ng, R. K., Isacks, K., Sheff, K., & Henderson, C. R. (2004). Effects of home visits by paraprofessionals and by nurses: age 4 follow-up results of a randomized trial, *Pediatrics*, *114*(6), 1560–1568

O'Malley, P. M., & Wagenaar, A. C. (1991). Effects of minimum drinking age laws on alcohol use, related behaviors and traffic crash involvement among American youth: 1976–1987. *Journal of Studies of Alcohol*, *52,* 478–491.

Perry, C., Komro, K., Veblen-Mortenson, S., Bosma, L., Farbakhsh, K., Munson, K., Stigler, M., & Lytle, L. (2003, February). A randomized controlled trial of the middle and junior high school D.A.R.E. and D.A.R.E. plus programs. *Archives of Pediatric and Adolescent Medicine*, *157,* 178–184.

Petosa, R., & Goodman, R. M. (1991). Recruitment and retention of schools participating in school health research. *Journal of School Health*, *61,* 426–429.

Petrosino, A., Boruch, R., Rounding, C., McDonald, S., & Chalmers, I. (2000). The Campbell collaboration social, psychological, educational and criminological trials register (C2-SPECTR) to facilitate the preparation and maintenance of systematic reviews of social and educational interventions. *Evaluation and Research in Education*, *14*(3), 206–219.

Plotnick, R. D., Young, D. S., Catalano, R. F., & Haggerty, K. P. (1998). Benefits and costs of a family-focused methadone treatment and drug abuse prevention program: Preliminary findings. In W. J. Bukoski & R. I. Evans (Eds.), *Cost-benefit/cost effectiveness research of drug abuse prevention: Implications for programming and policy* [NIDA Research Monograph Series no. 176] (pp. 161–183). Washington, DC: U.S. Government Printing Office.

Riecken, H. W., & Brouch, R. F. (Eds.). (1974). *Social experimentation*. New York: Academic Press.

Rosnow, R. L. (2002). The nature and role of demand characteristics in scientific inquiry. *Prevention and Treatment, 502.* Retrieved January 7, 2005 from http://gateway.ut.ovid.com/gw2/ovidweb.cgi

Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). New York: Springer-Verlag.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, *39,* 33–38.

Rosenstock, L., & Lee, L. J. (2002). Attacks on science: The risk to evidence-based policy. *American Journal of Public Health*, *92*(1), 14–18.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66,* 688–701.

Scariano, S. M., & Davenport, J. M. (1987). The effects of violations of the independence assumption in the one-way ANOVA. *The American Statistician*, *41,* 123–128.

Segawa, E., Ngwe, J. E., Li, Y., Flay, B. R., & Aban Aya Investigatgors. (2005) Evaluation of the effects of the Aban Aya Youth Project in reducing violence among African American adolescent males using Latent Class Growth Mixture Modeling Techniques. *Evaluation Review*, *29,* 128–148.

Schinke, S. P., Gilchrist, L. D., Lodish, D., & Bobo, J. K. (1983). Strategies for prevention research in service environments. *Evaluation Review*, *7,* 126–136.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7,* 147–177

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin.

Shumaker, S. A., Legault, C., Rapp, S. R., Thal, L., Wallace, R. B., Ockene, J. K., Hendrix, S. L., Jones, B. N., Assaf, A. R., Jackson, R. D., Kotchen, J. M., Wassertheil-Smoller, S., & Wactawski-Wende, J. (2003). Estrogen plus progestin and the incidence of dementia and mild cognitive impairment in postmenopausal women: The women's health initiative memory study: A randomized controlled trial. *Journal of the American Medical Association*, *289*(20), 2651–2662.

Steering Committee of the Physicians' Health Study Research Group. (1988). Preliminary report: Findings from the aspirin component of the ongoing physicians' health study. *New England Journal of Medicine*, *318,* 262–264.

Towne, L., & Hilton, M. (Eds.). (2004). *Implementing randomized field trials in EDUCATION: Report of a workshop committee on research in education*. Washington, DC: National Research Council.

Trochim, W. M. K. (1984). *Research design for program evaluation: The regression-discontinuity approach.* Newbury Park, CA: Sage.

Trochim, W. (2000). *The research methods knowledge base* (2nd ed.). Cincinnati, OH: Atomic Dog Publishing. Also available at http://trochim.human.cornell.edu/kb/index.htm (version current as of August 2004).

US Department of Education. (1998). Safe and drug-free schools program: Notice of final principles of effectiveness. *Federal Register 63*(104), 29901–29906.

US Department of Education. (2003). Identifying and implementing Educational practices supported by rigorous evidence: A user friendly guide. Washington, DC: US. Department of Education.

US Department of Health and Human Services. (2000). *Healthy People 2010. With Understanding and Improving Health and Objectives for Improving Health* (2nd ed*.,* 2 Vols). Washington, DC: U.S. Government Printing Office.

Valentine, J. C., & Cooper, H. (2003). *What works Clearinghouse study design and implementation assessment device* (Version 1.0). Washington, DC: U.S. Department of Education. Available at http://www.w-w-c.org/standards.html (retrieved 01/06/04).

Wagenaar, A. (1983). *Alcohol, Young Drivers, and Traffic Accidents: Effects of Minimum Age Laws*. Lexington, MA: D.C. Heath.

Wagenaar, A. C. (1993). Research affects public policy: the case of the legal drinking age in the United States. *Addiction*, *88*(Suppl.), 75S–81S.

Wagenaar, A. C., & Webster, D. W. (1986). Preventing injuries to children through compulsory automobile safety seat use. [erratum appears in Pediatrics 1987 Jun;79(6):863]. *Pediatrics*, *78*(4), 662–672.

Wagenaar, A., & Holder, H. (1996). The scientific process works: Seven replications now show significant wine sales increases after privatization. *Journal of Studies on Alcohol*, *57*(5), 575–576.

Weissberg, R. P., Kumpfer, K. L., & Seligman, M. E. P. (2003). Prevention that works for children and youth: An introduction. *American Psychologist*, *58*(6–7), 425–432.

Wolchik, S. A., Sandler, I. N., Millsap, R. E., Plummer, B. A., Greene, S. M., Anderson, E. R., *et al.* (2002). Six-year follow-up of a randomized, controlled trial of preventive interventions for children of divorce. *Journal of the American Medical Association*, *288,* 1–8.

Zarkin, G. A., & Hubbard, R. L. (1998). Analytic issues for estimating the benefits and costs of substance abuse prevention. In W. J. Bukoski & R. I. Evans (Eds.), *Cost-benefit/cost effectiveness research of drug abuse prevention: Implications for programming and policy* [NIDA Research Monograph Series no. 176] (pp. 141–160). Washington, DC: U. S. Government Printing Office.

Zeger, S. L., Liang, K. Y., & Albert, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, *44,* 1049–1060.