Standards of Evidence

CRITERIA FOR EFFICACY, EFFECTIVENESS AND DISSEMINATION

Preparation of this document was supported by funding from the National Institutes of Health and the Robert Wood Johnson Foundation. NIH funding was coordinated through the National Science Foundation.

TABLE OF CONTENTS

Preface i

Criteria for Efficacy 1

Criteria for Effectiveness 7

Criteria for Dissemination 10

References 11

Standards of Evidence Committee 12

PREFACE

The Society for Prevention Research is committed to the advancement of science-based prevention programs and policies through empirical research. Increasingly, decision-makers and prevention service providers seek tested and efficacious or effective programs and policies for possible implementation. However, until now, somewhat different standards have been used by different organizations seeking to identify and list programs and policies that have been tested and shown to be efficacious or effective. As part of SPR's strategic plan, in 2003, the SPR Board of Directors appointed a committee of prevention scientists, chaired by Brian Flay, to determine the requisite criteria that must be met for preventive interventions to be judged tested and efficacious or tested and effective. The Standards of Evidence developed by this committee have been unanimously adopted by the Board of Directors of SPR on April 12, 2004 as the standards which SPR asserts should be met if a program or policy is to be called tested and efficacious or tested and effective.

We hope, through this work, to provide a set of shared standards to be used by diverse organizations seeking to identify tested and effective prevention programs worthy of replication, adoption or dissemination. We believe that the promulgation and widespread use of these criteria will lead to consistent and high standards for determining whether programs have been scientifically shown to be efficacious, effective or ready for dissemination. This process should increase confidence in and commitment to the use of tested and effective policies, programs and actions to promote positive youth development and prevent health and behavior problems among young people.

We are pleased to make these Standards of Evidence available to all interested parties and we welcome your comments, including suggestions for additional steps SPR might take to foster the use of these standards by policy makers, stakeholders and service providers.

We are grateful to SPR's committee for developing these standards and to the National Institutes of Health, the Robert Wood Johnson Foundation, and the National Science Foundation for support for the preparation of this document.

J. David Hawkins, Ph.D.
President

CRITERIA FOR EFFICACY

Our objective in writing these standards is to articulate a set of principles for identifying prevention programs and policies that are sufficiently empirically validated to merit being called "tested and efficacious." Consistent with SPR's mission, we are interested in prevention programs or policies (interventions) of public health importance. These are directed to the prevention of social, physical and mental health problems and the promotion of health, safety and well-being. We place a special focus on issues that have high prevalence in the population and/or high costs to society, and the population(s) with these problems.

Our focus on research pertaining to the causal effects of interventions does not mean that we believe that research designs meant to uncover causal relations are the only research tool that should be used in prevention science, or that these are the only tools that are truly "scientific" (Valentine and 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 relations are of primary interest, other types of research tools and designs are often needed to yield important information about when, why, and how interventions work, and for whom. Because our central mission is to uncover causal statements about prevention programs and policies, our central focus is on research designs that have as their primary purpose uncovering causal relations. Other types of research are appropriate before and after efficacy trials.

1. **Specificity of Efficacy Statement**

   Our first criterion pertains to the form of the efficacy statement. Efficacy is the extent to which an intervention (technology, treatment, procedure, service, or program) does more good than harm when delivered under optimal conditions (Flay, 1986; Last, 1988). Efficacy is distinguished from effectiveness, which refers to program effects when delivered under more real-world conditions. Because outcome research results are specific to the samples (or populations from which they were drawn) and the outcomes measured, it is essential that conclusions from the research be clear as to the population(s) and outcomes for which efficacy is claimed. Therefore, a statement of efficacy should be of the form that "Program X is efficacious for producing Y outcomes for Z population." The remaining standards pertain specifically to the four areas of validity described by Cook and Campbell (1979): the description of the intervention and outcomes, the clarity of causal inferences, the generalizability of findings, and the precision of outcome measurement. Standards might change over time as methods develop and prevention science and practice advance. *Italics indicate items that are desirable though not essential for efficacy currently*. Some of these may become necessary criteria in the future.

2. Intervention Description and Outcomes

a. Program or policy description

The intervention must be described at a level that would allow others to implement/replicate it (this may require a more detailed description than what is presented in most research journals). An adequate description of an intervention includes a clear statement of the population for which it is intended and a description of its content and organization, its duration, the amount of training required, etc.

b. Outcomes – What is measured?

i. The stated public health or behavioral outcome(s) of the intervention must be measured. For example, a measure of attitudes about violence cannot substitute for a "measure" of actual violent behavior.

ii. 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).

- *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 mechanism.*

- *It is desirable to measure implementation.*

- *It is desirable to measure potential side-effects or iatrogenic effects.*

c. Outcomes - Measurement properties

Measures must be psychometrically sound. The measures used must either be of established quality, or the study must demonstrate their quality.

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

ii. Reliability
Internal consistency (alpha), test-retest reliability, and/or reliability across raters must be reported.

- *It is desirable to use multiple measures and/or sources.*

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

3. **Clarity of Causal Inference**

The design must allow for unambiguous causal statements. It must be the strongest design possible given the nature of the intervention, research question, and institutional framework within which the intervention/research occurs. The design must also be well implemented.

a. **Comparison Condition**

The design must have at least one comparison condition that does not receive the tested intervention. (Usually this is a comparison group. In time-series studies it may be the same group that does not get the intervention for a while.) The comparison group can be no-treatment, usual care, attention-placebo, or wait-listed. Or, it can be the best available or some alternative intervention. In this case, then the research question is, "Is the new intervention better than a current intervention?"

b. **Assignment**

The assignment to conditions needs to be done in such a way as to maximize confidence that the intervention, rather than some other alternative explanation, causes the reported outcomes. It also needs to minimize self-selection or unexplained selection. This requires a clear description of how people or groups are selected into intervention and comparison conditions.

i. For most kinds of interventions, random assignment (of sufficient sample size without significant pretest differences) is essential. Level of randomization should be driven by the nature of the intervention. Randomization can be of individuals or of intact groups like schools. It is clearly established that randomization is possible in many, or perhaps even most, contexts and situations. For some kinds of interventions where randomization is impossible, other approaches may be acceptable, when used with caution and methodological expertise, and when careful attention is given to ruling out plausible alternative explanations (see below).

ii. 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, Ary, & Wagenaar, 2000). *Even with these designs, randomization to multiple conditions or times is still preferable, especially if long baselines are not available.*

iii. Well-conducted regression-discontinuity designs also can be convincing because as in randomized studies, the selection model is completely known. These designs have important assumptions that require a fair degree of statistical expertise to assess (e.g., that the functional form of the relations between the assignment and outcome variable be properly specified) (Shadish et. al., 2002 and Trochim, 1984).

iv. Matched control designs are credible only with demonstrated pretest equivalence using adequately powered tests on multiple baselines or pretests of multiple outcomes and important covariates, and as long as assignment was not by self-selection, but instead is by some other factor (e.g., geography, every 2nd case, or all sites applying for a service every alternate month).

4. **Generalizability of Findings**

a. **Sample is defined**

The report must specify what/who the sample is and how it was obtained. It needs to be clear how well the sample does or does not represent the intended population. This is an essential component of the efficacy statement (see #1). An intervention shown to be efficacious can claim to be so only for groups similar to the sample on which it was tested.

- *It is desirable that subgroup analyses demonstrate efficacy for subgroups within the sample (e.g., 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.*

5. **Precision of Outcome**

a. **Statistical analysis must allow us to unambiguously establish the causal relations between the intervention and the outcomes (main effects).**

i. In testing main effects, the analysis must be at the same level as the randomization and include all cases assigned to treatment and control conditions (except for attrition, see below).

ii. Test for pretest differences. Random assignment, when properly carried out, yields groups that are similar on all observed and unobserved characteristics, within the limits of sampling error. Because sampling error is a factor, random assignment may in fact lead to groups that differ in important ways on pretest. If these are identified, it is essential to adjust for these differences statistically (e.g., covariance analysis) before conducting other analyses.

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

same sample size, this adjustment will result in lower power, so researchers are advised to consider these factors (e.g., sample size, number of outcomes measured) when planning their studies.

iv. Analyses to minimize the possibility that observed effects are significantly biased by differential measurement attrition, which occurs when the characteristics of participants that are not available for the post-test are not equally distributed across study groups, are essential. Note that differential measurement attrition can occur even when the rates of attrition are comparable across groups.

- *It is desirable that the extent and patterns of missing data from sources other than attrition be reported and handled appropriately.*

b. **Statistically significant effects**

i. Results must be reported for every measured outcome, regardless of whether they are positive, non-significant or negative.

ii. Efficacy can be claimed only for constructs with a consistent pattern of statistically significant positive effects. That is, when multiple indicators are used, most or all must be in the positive direction and at least one must be statistically significant.

iii. For an efficacy claim, there must be no negative (iatrogenic) effects on important outcomes.

c. **Practical value**

It is necessary to demonstrate practical significance in terms of public health impact.

- *It is desirable to have/report cost and cost-effectiveness information.*

d. **Duration of effect**

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).

e. **Replication**

i. 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. Any finding in science must be replicated to rule out chance findings before it is widely accepted with confidence. Replication to confirm findings is an important scientific principle. SPR believes that it is

important to set a high standard to encourage more replication studies of programs and practices whose evidence of efficacy is based on a single study. SPR also recognizes that in its current state, prevention research has produced fewer replication studies than is needed to reach the eventual goal of offering a wide variety of efficacious programs and practices to the field. Recognizing the importance of the replication standards, 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 research enterprise to meet this high 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.*

ii. When more than two efficacy and 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 interventions are effective under "real-world" conditions or in "natural" settings. Effectiveness trials may also establish for whom, and under what conditions of delivery, the intervention is effective.

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

Every effort should be made to apply the same standards as applied in efficacy trials, although it is recognized 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.

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

2. **Program Description and Outcomes**

   a. **Program definition**

   Manuals and, as appropriate, training and technical support must be readily available.

   b. **Intervention delivery**

   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).

   c. **Theory**

      i. A clear theory of causal mechanisms should be stated.

      ii. A clear statement of "for whom?" and "under what conditions?" the intervention is expected to be effective should be made.

   d. **Measures**

   Level of exposure should be measured, where appropriate, in both treatment and control conditions. Effectiveness trials generally have much more variation in these elements than efficacy trials; therefore the level of variation should be documented, as it affects ultimate program impact. Level of exposure is determined by two factors, both of which it is essential to measure:

i. Integrity and level of implementation/delivery of intervention.

ii. Acceptance/compliance/adherence/involvement of target audience and subgroups of interest in the intervention activities.

- *It is desirable to measure appropriate mediators (if suggested by the theory of cause).*

- *It is desirable to measure appropriate moderators (if suggested by the theory of cause).*

3. **Clarity of Causal Inference**

The same standards as stated for efficacy apply, though the challenges are greater. Randomization is still the best approach, but the other alternatives suggested (regression-discontinuity, time-series, high quality matched controlled designs) may be used.

4. **Generalizability of Findings**

a. **Representative sample**

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

b. **Generalizability of findings**

The degree to which findings are generalizable should be evaluated. One of the objectives of effectiveness studies is to establish for whom the intervention is effective.

- *Subgroup analyses. If the study sample is heterogeneous with respect to important variables (e.g., 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 2.c.ii).*

- *Experimental dosage analyses. It is desirable to conduct experimental dosage analyses. These analyze effects of differential program delivery. Propensity score analyses of variation in dose (that use information from both treatment and control groups) are better than correlational analyses within the treatment group alone, but not as good as a separate randomized study.*

- *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).*

- *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.*

5. **Precision of Outcome**

a. **Practical value**

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

- *It is desirable to have reports of costs and cost-effectiveness analyses.*

b. **Replication**

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. Effectiveness can be claimed only for those outcomes for which there are similar effect sizes in all trials (which may be lower than the effect sizes found in efficacy studies). This reduces chance findings.

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

CRITERIA FOR BROAD DISSEMINATION

To be ready for broad dissemination, a program must not only be of proven effectiveness, but it must also meet other criteria that ensure that it can be appropriately used by providers (teachers, counselors, social workers, service agencies, etc.).

1. To claim readiness for broad dissemination, a program must meet all of the criteria for effectiveness plus the following.

2. The program must have the ability to go to scale, including providing all program materials and necessary services (e.g., manual, training and technical support).

3. Clear cost information must be readily available.

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

    - *It is desirable that organizations which choose to adopt a prevention program based on barely or not quite meeting all criteria seriously consider undertaking a replication study as part of the adoption effort so as to add to the body of knowledge.*

    - *It is desirable to have a clear statement of the factors that are expected to assure the sustainability of the program once it is implemented.*

REFERENCES

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.

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

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.

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

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

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

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).

STANDARDS OF EVIDENCE COMMITTEE

Brian R. Flay, D. Phil. University of Illinois at Chicago, Chair

Anthony Biglan, Ph.D. Oregon Research Institute.
Robert F. Boruch, Ph.D., University of Pennsylvania.
Felipe González Castro, Ph.D., M.P.H., Arizona State University.
Denise Gottfredson, Ph.D., University of Maryland.
Sheppard Kellam, M.D., American Institutes for Research.
Eve K. Moscicki, Sc.D., M.P.H., National Institute of Mental Health, NIH.
Steven Schinke, Ph.D., Columbia University.
Jeff Valentine, Ph.D., Duke University.

With assistance from Peter Ji, Ph.D., University of Illinois at Chicago.

This publication can be accessed electronically at
http://www.preventionresearch.org.

For additional copies of this document, please contact:

Society for Prevention Research
7531 Leesburg Pike, Ste 300
Falls Church, VA 22043
703-228-0801
info@preventionresearch.org