Efficacy and Effectiveness Trials (and Other Phases of Research) in the Development of Health Promotion Programs

BRIAN R. FLAY, D. PHIL.

Health Behavior Research Institute, University of Southern California, 35 North Lake Avenue, Pasadena, California 91101

The concepts of efficacy and effectiveness are examined from the viewpoints of the traditions and philosophies of health-care research and social program evaluation. Consideration of the status of the program being assessed, its availability to, and its acceptance by the target audience leads to the derivation of four levels of health promotion program testing: (a) efficacy trials, under optimum conditions of program implementation and recipient participation; (b) treatment effectiveness trials, with expected variation in target audience acceptance; (c) implementation effectiveness trials, under varying conditions of implementation; and (d) program evaluation of previously untested programs. These four levels of testing, together with experience in one area of health promotion research (smoking prevention), suggest eight phases of research for the development of health promotion programs: (I) basic research, (II) hypothesis development, (III) pilot applied research, (IV) prototype evaluation studies, (V) efficacy trials, (VI) treatment effectiveness trials, (VII) implementation effectiveness trials, and (VIII) demonstration evaluations. Issues of design, the use of random assignment, the use of blinding procedures, and of the role of process evaluation in these different research levels, particularly efficacy and effectiveness trials, are considered in light of the terminologies and methods of health-care and social program evaluation research. Suggestions are made for improved health promotion research.

INTRODUCTION

Assessing the value or worth of a health promotion or education program requires many levels and types of research (36, 37, 52–55). Two of these are efficacy trials and effectiveness trials. 2 Efficacy trials 3 provide tests of whether a technology, treatment, procedure, or program does more good than harm when delivered under optimum conditions (79, 92, 102). Effectiveness trials provide tests of whether a technology, treatment, procedure, intervention, or program does more good than harm when delivered under real-world conditions. Efficacy is necessary to, but not sufficient for, effectiveness. The major purposes of this article are to examine and analyze these definitions and to clarify our understanding of the steps involved in research on, and development of, health promotion programs. (See Appendix for glossary of terms.)

Researchers interested in assessing the efficacy or effectiveness of health pro-

1 Supported in part by National Cancer Institute Grants CA34622 and CA38268 and National Institute on Drug Abuse Grant DA03468.
2 Cochrane (19) appears to have been the first to make this distinction, although he used the terms effectiveness and efficiency where we use efficacy and effectiveness. Thanks to John McKinlay for bringing this to my attention.
3 This use of the term efficacy is not to be confused with self-efficacy or outcome efficacy, terms Bandura (3) has used to refer to specific types of cognitions.
motion programs have usually been trained in one of two broad disciplinary areas—the health sciences (medicine and public health) or the social sciences (particularly psychology, sociology, and education). For the purposes of this article, I refer to the first area as health-care research and the second as social program evaluation. Each area approaches the task of health promotion research with different traditions and philosophies of research training, methods, and terminology. A secondary purpose of this article is to clarify the terminologies and methods used by these different disciplines in the assessment of health promotion programs. The terminologies and methods of each area will be considered with a view to improving future assessments of health promotion program efficacy and effectiveness.

EFFICACY TRIALS

An efficacy trial is designed to evaluate what an intervention achieves under optimum conditions. For example, an efficacy trial of a new drug "would make sure of the chemical purity of the drug samples used in the study, would closely screen patients so as to examine only those with the relevant medical problems, would closely monitor the frequency with which the experimental drugs were taken, would take special pains in examining the self-help activities of control group patients, and would carefully measure the physical health of patients at different points in time" (24). Thus, an efficacy trial of a health promotion program would provide a test of whether the program does more good than harm, i.e., leads to positive, or reduces negative, changes in knowledge, attitudes, behavioral skills, behavior, morbidity, or mortality when delivered under optimum conditions. An efficacy trial provides a test of (a) a well-specified and standardized treatment/program that (b) is made available in a uniform fashion, within standardized contexts/settings, to a specified target audience which (c) completely accepts, participates in, complies with, or adheres to the treatment/program as delivered. Tests of efficacy also usually involve randomized comparison (or control) groups and utilize, where possible, blinding procedures and placebos.

Perhaps the most well-known approach to evaluative research in medicine is the randomized controlled trial (17, 73, 74), also called the randomized clinical trial (RCT) (2, 14, 25, 88, 89, 100). An RCT is a randomized, well-controlled experimental test of the efficacy of a treatment procedure or program that may take place in either a laboratory or field setting. Complete random assignment to experimental conditions is often achieved and maintained by the use of double-blind procedures.

In a double-blind trial or experiment, neither the recipient (patient, client, subject) nor the treatment procedure provider is aware of into which experimental condition the recipient is placed. Recipients in the control condition of a double-blind clinical trial are treated in exactly the same way as recipients in the treatment or experimental condition, except for the experimental manipulation. This is often achieved by the use of a placebo, i.e., something perceived by the recipient to be a real treatment but one the researcher considers to be inert with regard to intended experimental effects.

 Often the comparison condition may consist of the best known available treatment, rather than a placebo. The test of efficacy is then concerned with whether
or not the new (or experimental) treatment does more good and/or less harm than the alternative treatment. Such tests often cannot be double-blind as the provider must be aware of what treatment is being delivered; however, recipients in such an experiment could still be randomly assigned and blind to experimental manipulation. Furthermore, currently in many countries, and until recently in the United States recipients may be kept blind as to which condition—experimental or comparison—they are in; they may not even be aware that they are participating in a trial.

Non- or quasi-experimental (15, 23) clinical trials are also carried out, though they result in weaker causal inferences. The most common of these is the historical control trial (HCT), where the comparison condition is one or more conditions from a previous trial, and recipients are obviously not assigned randomly to conditions. While HCTs certainly have their advocates [e.g., Refs. (45, 46)], the problems with such designs are well known in both program evaluation [e.g., Refs. (15, 23)], and medical/health research [e.g., Refs. (16, 61)]. However, because of recent legal constraints involving ethics and informed consent, and the increasing use of second-opinion referrals by patients, the practical problems of conducting RCTs are making HCTs more common in medical research. Similar problems with regard to social and educational programs provide a major reason for the common use of other quasi-experimental designs (15, 23) for program evaluations.

Tests of efficacy are regarded in health-care research as necessary steps in the development of a new treatment or technology. At least two divisions of the U.S. National Institutes of Health (NIH) [the National Heart, Lung, and Blood Institute (NHLBI) and the National Cancer Institute (NCI)] suggest a sequence of research phases for the development of health education/promotion programs (see Tables 1 and 2). At the NCI, controlled intervention trials are designed to "test the efficacy of an intervention on a group of individuals. . . . Case-control

<table>
<thead>
<tr>
<th>Phase</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td><strong>Basic research.</strong> Research seeking new knowledge about normal and abnormal functions of the heart, lungs, and blood and the etiology and pathogenesis of their diseases.</td>
</tr>
<tr>
<td>II</td>
<td><strong>Applied research and development.</strong> Research seeking to develop new ways of using basic research results to achieve specific practical goals.</td>
</tr>
<tr>
<td>III</td>
<td><strong>Clinical trials.</strong> Trials to determine the efficacy and safety of clinical interventions in samples of patients drawn from larger population groups.</td>
</tr>
<tr>
<td>IV</td>
<td><strong>Prototype studies.</strong> Small-scale tests of refined programs using components suggested by Phase II research to be efficacious (and further development of methods for future research).</td>
</tr>
<tr>
<td>V</td>
<td><strong>Demonstration and education research.</strong> Tests of the effectiveness of interventions designed to promote health or prevent disease in defined populations. The interventions selected for such testing should be those that have already been found to be efficacious in other studies and include, but are not limited to, education strategies and modifications in health care and health-related practices.</td>
</tr>
</tbody>
</table>

From Ref. (78) p. 2.
# TABLE 2
THE NATIONAL CANCER INSTITUTE CANCER CONTROL RESEARCH PHASES

<table>
<thead>
<tr>
<th>Phase</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td><em>Hypothesis development.</em> Identifies and synthesizes available scientific evidence about a specific cancer. A testable hypothesis is then formulated about the effectiveness of applying an intervention.</td>
</tr>
<tr>
<td>II</td>
<td><em>Methods development.</em> Characterizes the variables that must be controlled or monitored in subsequent intervention studies, and ensures that accurate and valid procedures are available before the actual study is begun.</td>
</tr>
<tr>
<td>III</td>
<td><em>Controlled intervention trials.</em> Test hypotheses developed in Phase I, using methodology validated in Phase II. Phase III studies test the efficacy of an intervention. Case-control methodology is generally used.</td>
</tr>
<tr>
<td>IV</td>
<td><em>Defined population studies.</em> Quantitatively measure the impact of an intervention in a sizable, distinct, and well-characterized population.</td>
</tr>
<tr>
<td>V</td>
<td><em>Demonstration and implementation.</em> Application of the intervention in a community at large and to measure the public health impact.</td>
</tr>
</tbody>
</table>


Methodology is generally used. ... [R]andomized controlled intervention trials are likely to give the most convincing results” (55). At the NHLBI, *clinical trials* are designed to “determine the efficacy and safety of interventions” (78).

In program evaluation, efficacy tests are often regarded as “basic research” preceding the implementation of a program and its evaluation. It is assumed, either explicitly or implicitly, that the treatment procedure being utilized in a program to be evaluated is one that has already been proven to do more good than harm when delivered under optimum conditions. Unfortunately, however, many health promotion programs are based on combinations of principles, approaches, or procedures that have not been subjected to such tests. If a program evaluation produces negative or null results, is it because the program as implemented was ineffective or because the kind of program being implemented was not efficacious? Clearly, if treatment efficacy is not established, this question cannot be answered easily. If treatment efficacy has not been established, a program evaluator might be well advised to perform such a test rather than be concerned with program effectiveness. The aim would be to determine whether or not the treatment procedure does more good than harm when delivered under optimum conditions. It can be argued that randomized experiments (clinical trials) are necessary for tests of efficacy, and that only when a treatment procedure has been shown to cause desirable changes under optimum conditions is it worthwhile to evaluate its effectiveness when delivered in real-world settings.

In clinical drug trials, all of the conditions of a pure efficacy trial are feasible. In health promotion program evaluation, some of these conditions are difficult, though usually not impossible, to meet. Reasons for these difficulties are numerous [for an extensive discussion see Ref. (36)] and include the following: First, health promotion programs are often complex and multifaceted in nature, particularly when compared with a drug treatment or even a surgical procedure.
Second, variations in delivery/implementation or the context of delivery are often unavoidable. Third, the theoretical causal mechanisms underlying health promotion programs are long and complex, and not totally understood (5, 32, 52, 72). Fourth, the target population is often widely dispersed, making it difficult to reach, and for some population members, the issue may be of low salience. Despite these difficulties, a pure efficacy trial must attempt to control all these factors—the less control obtained, the closer to an effectiveness trial the evaluation becomes (60).

EFFECTIVENESS TRIALS

Whereas efficacy trials are concerned with testing whether a treatment or procedure does more good than harm when delivered under optimum conditions, effectiveness trials are concerned with testing whether a treatment does more good than harm when delivered via a real-world program. In contrast to an efficacy trial, an effectiveness trial of a new drug, for example, would place more emphasis on testing the drug under more typical conditions of production, delivery, or availability; compliance or acceptance; and measurement.

"Thus, drug purity would be as mass produced, practicing physicians would prescribe the drug with variable degrees of appropriateness of the patient's true symptoms, patients would be free to take the drug in whatever dosages and whenever they wanted, control groups would be free to initiate their own self-help activities that might include gaining access to the same or similar drugs, and measures of physical health might be taken from routine records that are less reliable and less complete than most records collected by researchers for research purposes." (24)

Similarly, a health promotion program effectiveness trial would provide a test of whether the program does more good than harm when delivered under real-world conditions.

As noted earlier, efficacy is necessary to, but not sufficient for, effectiveness. A program will be effective only if an efficacious treatment/program is delivered/implemented in such a way as to be made available to an appropriate target audience in a manner acceptable to them (i.e., that they will be receptive to, participate in, comply with, or adhere to). Thus, the observed effects, or lack thereof, of a program may be due to one or more of the following: (a) the efficacy level of the evaluated treatment; (b) the availability of the treatment to the target audience (which may be affected by the mode and extent of treatment delivery/implementation); or (c) the level of acceptance of (participation in, compliance with, or adherence to) the treatment by the target audience.

Effectiveness trials are termed health-care trials by some health-care researchers (92, 100) and are similar to what program evaluators call outcome/impact evaluations. However, some key differences between the health-care and evaluation approaches have important implications. When health-care researchers conduct an effectiveness evaluation, they usually have already determined the efficacy level of the treatment or program being tested. The effectiveness evaluation is then concerned with whether or not more good than harm is done by the efficacious treatment when delivered via a program in the real-world setting. An efficacious treatment may still have differential effects because of
differences in its availability or acceptance, or because of the effects of other uncontrolled environmental or social influences.

The NIH suggests levels or phases of research on health promotion programs that are similar to effectiveness trials (see Tables 1 and 2). In the NHLBI research spectrum, demonstration and education research is defined as testing "the effectiveness of interventions designed to promote health or prevent disease in defined populations . . . that have already been found to be efficacious in other studies" (78). Greenwald and Cullen (55) write that in the NCI phases, defined population studies "are designed to quantitatively measure the impact of an intervention in a sizeable, distinct, and well-characterized population. . . . the results can be generalized to the entire target population. . . . [Defined population] studies provide further validation of . . . the efficacy determined in [controlled intervention trials], and they resolve new issues that arise when applying interventions to population groups larger than those required for [controlled intervention trials]" (55). These authors do not define the new issues that arise when implementing new interventions in real-world settings, but the analysis of effectiveness evaluations provided here suggests that these issues concern questions such as (a) for whom is the treatment effective, and (b) under what conditions or within what contexts of delivery.

In the program evaluation field, on the other hand, outcome/impact evaluation often concerns an assessment of the effects of a program that is already operational. That is, program evaluators do not always predetermine a treatment's efficacy level, so that any one of low treatment efficacy, availability, or acceptance, can be possible explanations of observed patterns of program effects. Thus, outcome/impact evaluations, as often conducted by program evaluators, can be higher risk propositions than health-care trials. Negative results are more likely from outcome/impact evaluations of programs of unknown efficacy, and such results are often less interpretable than those from health-care trials, because they have several possible explanations. Of course, if positive results are obtained, it may generally be assumed that an efficacious treatment was made available to the target audience and that they accepted it without the interference of other environmental or social influences.

The above discussion suggests that good effectiveness trials require assessments of program implementation, availability, and acceptance as well as program effects. Health-care researchers, however, have suggested that availability need be assessed only after it has been determined that a program is effective (92). The question for health-care researchers, then, concerns whether or not a good program is being made available to the target audience. We argue that availability should be assessed in both types of situations: (a) within effectiveness trials to increase the interpretability of results, as well as (b) after program effectiveness has been determined to ensure adequate delivery. Implementation evaluation, as practiced in social program evaluations, usually includes assessments of program availability and has been seriously advocated by program evaluators (80). Without adequate implementation evaluation, it is impossible to determine whether a lack of program effects is due to inadequate program delivery or an inefficacious treatment (18, 44, 60, 80, 93). In any comprehensive effectiveness
trial, acceptance of a program by recipients also needs to be assessed. Without indicators of program acceptance, a lack of program effects could be due either to treatment inefficacy or lack of acceptance by recipients.

Both program evaluators and health-care researchers usually attempt to implement randomized trials of program effectiveness. Often, however, quasi-experimental designs (15, 23) are used. Double or single-blind procedures and placebos are occasionally used in effectiveness trials, but this is unusual because such procedures are so much more difficult to implement in the less-controlled, real-world settings of effectiveness trials than they are in the more controlled atmosphere of efficacy trials.

APPLICATION TO HEALTH PROMOTION PROGRAMS

The above discussion suggests four possible levels of experimental or quasi-experimental tests of health promotion programs. These four are defined by which of three major factors—the program as implemented, its availability, and its acceptance—remain unstandardized (Table 3).

**Efficacy Trials**

An efficacy trial has been defined to include standardization and optimization of all three of the above factors. This set of conditions is possible in the assessment of health promotion programs only when a clearly defined program is delivered, in a standardized way, to a sample of the target audience which is either highly motivated to participate or is "captive" (e.g., schoolchildren in classrooms), yet is representative of the target audience. These types of conditions have been rare in past research on health promotion programs, but they need to be achieved more often. The health promotional sciences will advance more quickly, and successful health promotion programs will be developed more quickly, if some approaches can be proven to be more efficacious than others. We can have more confidence in those conclusions about efficacy that are derived from true efficacy trials than we can in those inferred from lower-order effectiveness trials.

**Treatment Effectiveness**

The easiest type of effectiveness trial to interpret is one in which only acceptance can vary—that is, one in which a technology of proven efficacy is known to

<table>
<thead>
<tr>
<th>Level of experimental assessment</th>
<th>Program implementation</th>
<th>Availability</th>
<th>Acceptance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Efficacy</td>
<td>Standardized</td>
<td>Optimized</td>
<td>Optimized</td>
</tr>
<tr>
<td>Treatment effectiveness</td>
<td>Efficacious</td>
<td>Optimized</td>
<td>Variable</td>
</tr>
<tr>
<td>Implementation effectiveness</td>
<td>Efficacious</td>
<td>Variable</td>
<td>Variable</td>
</tr>
<tr>
<td>Program evaluation</td>
<td>Unproven efficacy or nonstandardized</td>
<td>Variable</td>
<td>Variable</td>
</tr>
</tbody>
</table>
have been made available to the target audience. Such trials provide assessments of how acceptable a program is to the target audience, for whom the program is acceptable (and effective), or whether the program does more good than harm for those to whom it is offered. Thus, outcome variables of primary interest must concern receptivity of, compliance with, adherence to, or participation in program activities by program recipients; program effectiveness is, in some senses, of secondary interest. These types of effectiveness trials may be called “treatment effectiveness trials” (92). They provide the strongest possible test of program effectiveness in real-world settings, because only one factor can vary. In health promotion terms, when a tested technology/approach of recognized efficacy is known to have been made available to the target audience, then any variation in observed effectiveness can unambiguously be attributed to lack of acceptance/participation in program activities by the program recipients.

**Implementation Effectiveness**

When program delivery/implementation is not standardized (or is deliberately manipulated), the availability of the program to the target audience is also likely to vary. Variation in availability will tend to lead also to variation in acceptance or participation, and hence program effectiveness. The primary outcome variables of such a trial must concern program availability; acceptance and effectiveness must be of secondary interest. These types of effectiveness trials might be called “implementation effectiveness trials,” because any observed variation in program effectiveness will most likely be due to variations in how the program was implemented/delivered. It is worth reemphasizing that implementation effectiveness trials can be uncontrolled in the sense that variations in program delivery are not controlled, or they can be controlled experiments in which different approaches to the delivery or dissemination of an efficacious technology are tested.

**Program Evaluation**

A fourth type of experimental (or quasi-experimental) assessment of health promotion programs is one wherein the program being tested is not standardized or is based on a technology/approach of unknown efficacy. Technically speaking, this is not an effectiveness trial, because effectiveness trials are tests of programs of proven efficacy. As many program evaluations are of this nature, however, I shall use this term to refer to this type of assessment of any health promotion program. As discussed above, program evaluations are the least interpretable of the four types of studies considered here, in that a lack of program effects could be due to any one of four factors: program inefficacy, poor implementation, low availability, or low acceptance.

**Sequencing of Studies**

The NCI and NHLBI have both suggested that the development of any new treatment or program should ideally result from a sequence of studies, ranging from basic research to demonstration studies. The analysis provided in this article suggests a refined sequence of studies for developing health promotion programs (Table 4).
<table>
<thead>
<tr>
<th>Phase</th>
<th>Description</th>
<th>Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Basic research</td>
<td>Disciplinary-based research on basic mechanisms (e.g., etiology, epidemiology, social psychology, education). Defined by discipline.</td>
</tr>
<tr>
<td>II</td>
<td>Hypothesis development</td>
<td>Development of hypotheses about new approaches to health promotion for a specific health problem. Review; synthesis of basic research; exploratory research.</td>
</tr>
<tr>
<td>III</td>
<td>Pilot applied research</td>
<td>Preliminary tests of new approaches toward using basic research results to achieve specific immediate effects related to specific health promotion goals (and methods development for future research). Pilot test (pre- or quasi-experimental) of innovative manipulations; very small scale (few individuals or aggregated units per condition).</td>
</tr>
<tr>
<td>IV</td>
<td>Prototype studies</td>
<td>Small-scale tests of refined programs using components suggested by Phase III research to be efficacious (and further development of methods for future research). Experimental or quasi-experimental tests of complete &quot;programs&quot;; small number of aggregated units (e.g., schools) per condition; measurement to include behavioral outcomes.</td>
</tr>
<tr>
<td>V</td>
<td>Efficacy trials</td>
<td>Trials to determine the efficacy of programs or approaches suggested to be effective by earlier phases. Pure experimental trials with random assignment of aggregated units to conditions in sufficient number for practical, significant behavioral effects to be detected.</td>
</tr>
<tr>
<td>VI</td>
<td>Treatment effectiveness trials</td>
<td>Trials to determine the effectiveness and acceptability of efficacious programs on a broader population. Large-scale experimental or quasi-experimental trials in real-world settings; delivery/implementation optimized/standardized as much as possible (and carefully assessed); morbidity/mortality outcomes may be assessed.</td>
</tr>
<tr>
<td>VII</td>
<td>Implementation effectiveness trials</td>
<td>Trials to determine the effectiveness of an efficacious and acceptable program under real-world conditions of delivery/implementation. Large-scale experimental or quasi-experimental trials in real-world settings; delivery/implementation can vary naturally or involve planned comparisons (deliberate variations); careful assessment of delivery/implementation; morbidity/mortality may be assessed.</td>
</tr>
<tr>
<td>VIII</td>
<td>Demonstration studies</td>
<td>Studies to determine the effects of an efficacious program on public health when implemented in whole systems (schools, cities, states, nations). &quot;Naturalistic&quot; quasi-experimental program evaluation; morbidity/mortality definitely assessed; natural variation in delivery/implementation may be studied; diffusion patterns may be studied.</td>
</tr>
</tbody>
</table>
As the NHLBI spectrum suggests, the first step in health promotion program development research is basic research. This research will usually be conducted by basic research scientists in basic disciplines. For social psychological smoking prevention programs, for example, the basic research consisted of the social psychological research that supported the theories applied to smoking prevention by Evans (28) and the researchers who followed him (33, 34). Examples include social learning theory (3), social inoculation theory (71), and persuasion and attitude change theories (72). Hypothesis development (Phase II) would consist of the steps of synthesizing available basic research to apply to the solution of a practical goal and the first small-scale pilot test(s) of the derived solution. Measurement might be confined to those cognitive/attitudinal, and possibly skill development, variables hypothesized to be directly affected by the manipulation. The article by Evans (28) and the study by Evans et al. (29) are examples of this phase in smoking prevention research. While Phase II studies would be confined to pilot tests of components of potential programs, Phase III studies would consist of pilot tests of early versions of the program and lead to further development of the program. The “second generation” tests of social psychological approaches to smoking prevention are examples (33, 34) (see Table 5). Phase IV would consist of prototype studies, designed to test improved versions of the program and to re-

<table>
<thead>
<tr>
<th>Generation</th>
<th>Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Hypothesis development and the first tests.</td>
</tr>
<tr>
<td>II</td>
<td>First pilot tests of improved programs in small-scale studies (one school or classroom per experimental condition). Further program improvements. Methods development.</td>
</tr>
<tr>
<td>III</td>
<td>Prototype studies of the improved programs developed and pilot tested in the second generation. Two or three schools or classrooms per experimental condition. Further methods development.</td>
</tr>
<tr>
<td>IV</td>
<td>Efficacy trials. Five or more schools or classrooms per experimental condition, randomly assigned. Simple program vs. control designs. Less methodological problems than previous studies.</td>
</tr>
</tbody>
</table>

*a Adapted from Flay (33).
fine research methodologies to attain high internal validity for the efficacy trials to follow in Phase V (see the third generation of smoking prevention studies in Table 5 as examples). Assessment of the desired behavioral outcomes should be added at this phase (if not already introduced in Phase III). This step is necessary in the development of health promotion programs, just as it is in the development of many clinical interventions, because methods should be well developed before embarking on efficacy trials, which cannot help but be large and expensive.

The widespread practice of developing and implementing health promotion programs without prior efficacy trials means that subsequent effectiveness trials or program evaluations are often uninterpretable. In addition, an inefficacious program may be implemented at considerable cost, and then become difficult to remove or change because it has acquired its own supporters for other political reasons (73). Health promotion program developers and evaluators would be advised to follow the example of medical and health-care researchers and conduct efficacy trials (Phase V) more often. Social program evaluators might argue that a particular efficacy trial is not desirable because it will often have low external validity—that is, its results cannot easily be generalized to many real-world settings. While that may be true of any particular case, generalizability is improved when multiple replications of efficacy trials are considered (reviews, which synthesize findings from many important studies, are the major unit of progress in any applied science, not individual studies), and when later effectiveness trials are completed.

The NHLBI suggests only one level beyond efficacy trials, and the NCI suggests two. Three levels are suggested by the analysis provided here because of the importance of separating treatment assessments from implementation effectiveness. After the efficacy of a new health promotion program has been established, the first effectiveness trials would best be confined to testing treatment effectiveness (Phase VI). That is, the efficacious program should be implemented/delivered in an optimum fashion within the constraints of a real-world setting, such that only recipient acceptance or participation can vary. An easily interpretable test of the new program’s effectiveness for those to whom it is delivered is then possible. In a treatment effectiveness trial of a school-based smoking prevention program, for example, extensive training would be provided to the teachers delivering the curriculum. Their implementation of the program would be monitored, so that the only major variations would be in the levels of participation by students and their families in out-of-class activities (such as homework or viewing of television segments).

After successful treatment effectiveness trials, Phase VII studies, implementation effectiveness trials, would then allow, or deliberately manipulate, variation in level or mode of implementation (i.e., how an intervention is delivered, but not what is delivered). The best of such studies would provide experimental tests of program effectiveness by varying the personnel, training, and setting, and adding supplementary activities. Such studies, which might also be called dissemination trials, would provide rich data on the generalizability of the program—that is, under what conditions and for whom the program is effective and the best method of dissemination.
The final phase of research proposed here consists of demonstration and implementation studies (Phase VIII). (These are most like NCI Phase V studies; NHLBI Phase V studies are most like our treatment and implementation effectiveness trials.) A program of proven efficacy and effectiveness is implemented in whole communities, states, or nations, and its effects on public health are monitored; i.e., morbidity and/or mortality measures are included.

The suggestion that a particular sequence of different research study types be engaged in to ensure the development of efficacious and effective health promotion programs has important implications for progress in the health promotional sciences. The obvious implication is that the current state of the art be considered when deciding what type or level of study to conduct next and when reviewing research proposals or reports. It would be unwise, for example, to propose a large-scale efficacy trial of an innovative program that has not yet been tested. To emphasize a point already made, it would also be unwise to plan an effectiveness trial of a program of unproven efficacy. Similarly, it would be unfair for reviewers to insist on a large-scale randomized trial of an innovative approach or to expect a test of any type of intervention during Phase II research. That is, the level of methodological rigor expected needs to be consistent with the phase of research being attempted which, in turn, needs to be consistent with the current state of knowledge.

IMPLICATIONS FOR ASSESSMENT OF HEALTH PROMOTION PROGRAMS

This brief review and comparison of terminology and methods from health-care research and social science program evaluation, and the refinement of definitions of efficacy and effectiveness, has three major implications for research on health promotion programs. They concern reliance on causal inference, use of blinding procedures, and use of process evaluation.

Reliance on Causal Inference

In tests of efficacy and effectiveness, causal inference is highly desirable. We must be able to conclude, without much doubt, that a certain treatment or program causes a desired effect (that leads to more good than harm). Medical researchers seem to accept the use of random assignment to experimental conditions more readily than many social or educational program evaluators. Thus, clinical trials are most often randomized experiments, often with blinding procedures and placebo conditions. It must be noted, however, that while medical researchers accept the need for randomized trials, practicing physicians and other health-care workers are not always so accepting. On the other hand, medical providers are slow to accept results from nonrandomized studies (88), so medical researchers have a remarkable record of convincing providers of the need for randomization. Unfortunately, health promoters and health educators do not seem readily to accept the need for randomized trials and may be too quick to accept results from nonrandomized studies.

It is generally accepted that, overall, it would be unethical never to know for certain whether a treatment does more harm than good (8). The acceptance of subexperimental evidence could lead to decisions that result in more harm than
good being done (17, 73). Gilbert et al. (47) and Peto (88) provide analyses of results of randomized vs nonrandomized studies of innovative therapies that demonstrate this. Fortunately, most rapidly adopted health promotion programs do not cause demonstrable harm to individuals,\(^4\) though they might to the credibility of the profession (and to the economy). Health promotion researchers are advised to insist on randomized studies more often than they do. They could educate instructors and administrators about the need for randomization rather than accept objections to such trials as justification for not performing them. In a very real sense, as stated by Chalmers, “poor research is unethical practice” (17, p. 325).

It is recognized, however, even by medical and health researchers, that randomized trials are not always possible. The major alternative to the randomized trial used by health-care researchers is the historical case-control design. This is a design that might more readily be used in health promotion research. The major threat to the validity of such designs is history—that is, something else occurring at the same time as the program can cause the observed effects. However, when historical controls exist, enough other data also usually exist to enable trends to be plotted for a long period. Such data can allow for either (a) a reduction in the plausibility of history as an alternative explanation of observed effects, or (b) estimation of the portion of any observed effect that is likely to be attributable to historical trends. Smoking prevention researchers might soon need to make greater use of such data as smoking decreases in both adolescent and adult populations (22). When randomized trials are not possible, there does not seem to be any compelling reason for health-care researchers not to make greater use of quasi-experimental designs developed by social and educational researchers (15, 23). As long as care is taken to rule out plausible alternative explanations, other quasi-experimental designs can be just as powerful as the historical control design. It must be emphasized, however, that randomized experiments are preferable whenever possible and that quasi-experimentation should be considered only as an alternative when pure randomization is impossible.

**Use of Blinding Procedures**

One major methodological difference between health-care research and social program evaluation concerns the use of blinding procedures. These are associated almost exclusively with clinical trials of medical treatments. Double-blind procedures obviously are not possible whenever extensive social interaction between provider and recipient is an inherent part of the treatment. It is just not possible to have both parties unaware of whether one type of educational package or another is being provided. It is possible, however, for recipients, and sometimes for providers, to be kept unaware as to which treatment condition, i.e., experimental or comparison, they are assigned (35, 97, 98). This is preferable to (a) not providing an alternative program to a comparison group, which would make treatment recipients differ from comparison subjects in receiving attention

\(^4\) A notable exception might be drug education, where many programs have probably increased drug use by adolescents rather than decreased it as intended (50, 51).
as well as receiving the treatment, or (b) providing an alternative treatment and
having subjects aware that they are receiving the "new" one or the "old" one. In
the former situation, any differences between treatment and comparison condi-
tions could be due to a Hawthorne effect—the major remaining alternative inter-
pretation of the positive results reported by most studies of social psychological
approaches to smoking prevention (33, 34). In the latter situation, subject aware-
ness poses threats to internal validity because comparison subjects can easily act
differently than they otherwise would, just because they know that they are not
receiving the desired program5 (23). These particular threats to validity are major
ones that are not removed by random assignment to conditions. They, therefore,
deserve much more serious consideration than they have received to date—sub-
jects (and providers, where possible) should be kept blind to their experimental
condition.

The concept of blindness also applies to measurement. Data collectors and in-
terviewers should be kept unaware of the subject's experimental condition. This
is rarely done in health promotion research but is necessary for the field to ad-
vance as a science. Interviewers or data collectors are too often aware of the
subject's condition, sometimes even reminding their subjects, leading to expect-
tancies or demands well known to social psychological researchers (38, 91, 108).
Only in assessments of the particulars of program implementation and availability
need data collectors be aware of the subject/respondent condition. Such ques-
tions could routinely be left to the end of interview schedules or questionnaires,
or the information could be collected by a different interviewer. The use of ar-
chival data, such as medical records,6 or the use of other unobtrusive measures
(107) offers other viable approaches to maintaining blindness.

Process Evaluation

The term process evaluation has been used in a confusing number of ways in
both the program evaluation and health-care research literatures (99, 103). Some
of the usages are discussed here.7

An assessment of how a program is implemented, what treatment is provided,
under what conditions, by whom, to what audience, and with what level of effort
is one form of process evaluation that is often also called implementation evalua-

5 Some types of threats to internal validity are diffusion or imitation of treatments, compensatory
equalization of treatments, compensatory rivalry by respondents receiving less desirable treatments,
and resentful demoralization of respondents receiving less desirable treatment (23).

6 Records are subject to contamination (20, 21); however, even record keepers might need to be
kept blind as to the potential use of those records in determining the relative efficacy or effectiveness
of treatments/programs.

7 Process evaluation is often assumed to be synonymous with formative evaluation. The terms
formative and summative refer to roles that evaluative results can play. When evaluative results are
used to form or improve a program, they are used in a formative way. When evaluative results are
used to make a final judgment about the worth of a program, which is presumably in its final form,
they are used in a summative way. The tendency to equate formative with process probably has arisen
because information from process evaluation is the most useful in providing clues on how to improve a
program. However, it is certainly not the only useful information, and information from any type of
evaluation may be used to help improve programs.
tion. The standard of acceptability is appropriate conduct of practice, and the strategies of evaluation include peer review, program monitoring, management information systems, and audit. An assessment of availability may include all the elements of an implementation evaluation or might include only an assessment of treatment actually received by the target audience. Both health-care researchers and program evaluators recognize that whether or not an efficacious treatment is actually delivered to the target audience ought to be assessed (though different conceptions of its role were noted earlier). Collecting such data would make negative or marginal results more interpretable, in that it would be known whether or not a program had been made available to the appropriate target audience, under appropriate conditions, and whether or not they had accepted it. Even in the most highly controlled trial, it cannot be guaranteed that all subjects receive exactly the same treatment under the same conditions, especially in effectiveness trials. The School Health Education Evaluation (43, 90) provides a recent demonstration of these points.

Another component of process evaluation concerns the relationships between different levels of outcomes and impact. Any health promotion program should lead to changes at many different levels. For example, changes might be expected in knowledge, beliefs, attitudes, behavioral intentions, performance, physiological or biochemical indicators, morbidity, and longevity. Outcomes might also be reflected at the community level, including changes in community organization, consumption patterns, physician and hospital visits, and mortality. Effects at these higher levels of aggregation, such as the family, the community, or society as a whole, are sometimes called the "impact" of the program, though social program evaluators and health promotion researchers seem to use the terms "outcome" and "impact" in exactly opposite ways.8 The investigation of the relationships among levels of outcome/impact, perhaps leading to the testing of causal chains, might be called process analysis (62). Health promotion researchers have not yet been too concerned with this type of process analysis or process evaluation (6, 33, 34, 70). They probably should be, because most health promotion programs are based on very complex chains of effects which need to be tested. Process evaluation is also potentially a very important way of supplementing, perhaps validating, the rather weak epidemiological data available so far on many of the links between behavior and health status.

A third component of process evaluation, not necessarily independent of the previous two, is concerned with the testing of theory. Most health promotion programs are based, either explicitly or implicitly (perhaps too often the latter), on a model or theory of how they are supposed to work. Sometimes those models have been nothing more than hypothesized chains of effects as described in the previous paragraph; at other times, more complex theories are invoked. Good evaluations provide good tests of such theories. Unfortunately, not many evaluations or health-care trials have provided tests of anything more than a simple

8 This distinction between outcomes and impact is fairly well established in the program evaluation literature, but some writers in the health promotion area have used these terms in exactly the opposite way [e.g., Refs. (53, 54, 103)].
"program causes outcome" hypothesis. Very few have taken the additional step of attempting to test theories of how or why a program caused a particular outcome or of the mechanisms mediating such causal processes. Such tests are necessary if health promotion research is to add to the knowledge base necessary for significant advances in a science of health promotion.

Collection of process variables in order to explain how or why an observed change occurred is another component of process evaluation (64), and again, health promotion researchers have not been as concerned with it as they probably should be (70). In fact, one explanation for the paucity of tests of theory is that they often require the collection of additional data on variables that indicate how and why a target behavior is changed. For example, what did clients do to cut down on drinking—stop going to parties, learn to say no, join AA? These are certainly not equivalent processes, and there may be important lessons to be learned, by both the client and the evaluator or researcher, by looking closely at the ways in which change is accomplished (35).

To summarize this section, health promotion researchers are encouraged to do more process evaluation. Very few programs work with 100% efficacy. There should, therefore, be a concern for the conditions under which the program works best, for whom it works, and so on. Most health-care and social programs are based on very complex chains of effects—these deserve much more explanation and testing. Finally, our knowledge of why certain treatments and programs produce the effects they do is often imperfect. Process evaluation provides a relatively inexpensive approach to furthering our knowledge. Of course, much information from process evaluation might be at the hypothesis generation level, and researchers might cycle back to earlier phases of research to confirm such hypotheses. This does not reduce the value of process evaluation.

CONCLUSIONS

Four general conclusions can be drawn from the analysis and discussion of terminology and methods presented above. First, efficacy trials, which test the effectiveness of a program delivered under optimum conditions, need to be attempted more often by health promotion researchers. Too many health promotion and health education programs either have been implemented on a wide scale without such tests or have been tested under far less than optimum conditions, leading to multiple explanations for observed effects. The full spectrum of research and development discussed here requires balanced support. However, if any one level of research is to receive greater support and attention it should be efficacy trials—they provide a central focus toward which early exploratory research should aim and from which subsequent effectiveness trials should emanate.

Second, four phases of research must precede efficacy and effectiveness trials in any well-planned program of research that has the development of effective health promotion or education programs as its goal. These are (I) basic research and (II) hypothesis development, to assure a strong theoretical and empirical basis for innovative programs; and (III) pilot applied research and (IV) prototype studies, to provide preliminary tests of the likely effectiveness of the innovative...
programs. Much untapped basic research in the social, educational, medical, and basic sciences provides a wealth of data on which to base new hypothesis development and pilot applied research. The NIH might be advised to support more such research. This would be more desirable than the current practice of leaving the support of developmental research (Phases II and III) to the few local health agencies who support educational, psychosocial, or behavioral research, while the bulk of federal funds is concentrated in Phases IV and VI–VIII studies. A few well-placed dollars in support of initial tests of new, creative ideas might provide a substantial payoff in the long run. Without such support, younger scientists are not provided the opportunities they deserve to test their ideas, and the future development of the health promotional sciences runs the danger of being thwarted, forever frozen in a cycle of tests of minor variations of current approaches. The last decade of research on social psychological approaches to smoking prevention provides an example of the application of these four phases (plus the fifth). Few other areas of health promotion program development research have been this systematic, and no area yet provides an example of orderly progress through Phases V–VIII.

Third, there is also a great need for health promotion programs to be subjected to more effectiveness trials, which test the value of programs under more real-world conditions. Too many health promotion or education programs have been widely adopted without the benefit of such tests. Even efficacious programs will not always prove effective in real-world settings because of difficulties with implementation or recipient acceptance. Some efficacious programs or approaches will prove to be effective only for a certain subset of the target audience or under certain conditions of implementation. Effectiveness trials can be conducted at two different levels, depending on the level of control over program availability and acceptance. Treatment effectiveness trials control for, or standardize, implementation and, therefore, availability, to ensure that any variation in program effectiveness is not due to unknown variation in either. When neither implementation nor acceptance is standardized, and the program being tested is likewise not standardized or is of unknown or unproven efficacy, any program evaluation results have multiple interpretations. The most informative implementation effectiveness trials are those that test the relative effectiveness of different modes or mechanisms of implementation/delivery of an efficacious program or approach. Such research is needed to determine those variables most related to ultimate dissemination and to ensure proper dissemination of efficacious approaches.

Fourth, researchers and reviewers need to give greater attention to matching methodological requirements with phases of research, and phases of research with the current state of the art. Future efficacy trials, effectiveness trials, and program evaluations of health promotion programs can be improved by including comprehensive implementation assessment and process evaluation. Knowledge of exactly what was delivered, how it was implemented, who it reached (availability), and how the recipients responded to it (acceptance), as well as a range of outcome/impact variables, makes program effectiveness results—or lack of them—much more interpretable and provides leads as to the most efficient means of program dissemination. Blinding procedures and placebo programs also need to
be utilized in future health promotion research. These procedures provide means of ruling out the alternative explanations of the Hawthorne effect, expectancy effects of instructors, demand characteristics of testing, and treatment by testing interactions, all of which are plausible explanations for results of most health promotion programs currently accepted as being effective. Health promotion researchers trained from the health-care research perspective might consider greater use of the quasi-experimental designs developed by social scientists. This use should be limited as much as possible, however, to Phases III and IV, pilot applied research and prototype studies, and then always with careful attention to removal of plausible alternative explanations of observed effects. Most quasi-experimental designs should not be relied upon for efficacy or effectiveness trials, particularly the former. In a similar vein, health promotion researchers trained in the social sciences might consider use of the HCT methodology often used in medical research. Again, however, this should be limited to Phase III and IV research, and great care needs to be taken to remove or reduce plausible alternative explanations of observed effects.

Current fiscal conditions make it timely for the adoption of a more systematic approach to the development of health promotion/education programs or approaches. More, and improved, efficacy and effectiveness trials of promising approaches to health promotion are needed to guide government, health, and educational policy decisions. It is also time that health promotion research became a more scientific enterprise. This will require greater emphasis on randomized experimentation, with a focus on efficacy trials, toward which earlier phases of research can aim and from which subsequent effectiveness trials can emanate.

APPENDIX: GLOSSARY OF TERMS

The following terms have been defined and used in this article as follows. All definitions are standard for the literature from which they have been drawn [see Ref. (56) for other definitions].

Acceptance. Attention to, compliance with, or adherence to a treatment or program by the exposed audience. Sometimes assessed by asking subjects to rate their interest in, liking of, enjoyment of, and judgments as to likelihood of success of the program.

Availability. Level at which a program is implemented or delivered to the target population, or ease with which the target population can gain access to the program.

Effectiveness. Level of good over harm (or benefits over costs) that a program achieves when received under typical real-world conditions of availability and acceptance.

Effectiveness trial. Test of the effectiveness of a program under real-world conditions of availability and acceptance. See distinction between treatment and implementation effectiveness below.

Efficacy. Level of desired effects (good over harm, benefits over costs) of a program when delivered and received under optimum conditions (i.e., when availability and acceptance are maximized).

Efficacy trial. Test of efficacy. That is, test of level of effects achieved by a
program when availability and acceptance are maximized and do not vary significantly.

*Historical case–control trial.* Trial where cases from a previous trial, or archival data, are used as controls.

*Implementation effectiveness trial.* Test of the effectiveness of an efficacious program when implementation can vary, or is deliberately varied, so that both availability and acceptance can vary.

*Implementation evaluation.* Assessment of how, and at what level, a program is implemented, and what and how much were received by the target population; i.e., a type of process evaluation.

*Program evaluation.* Test of the effects of a program as implemented/delivered. Limited in this article to refer to tests of the effects of a program of unknown efficacy, availability, and acceptance.

*Treatment effectiveness trial.* Test of the effectiveness of an efficacious program under standardized conditions of implementation and, therefore, availability, but where acceptance can vary.

**ACKNOWLEDGMENTS**

Thanks to Allan Best, Erv Bettinghaus, Tony Biglan, Linda Collins, Don Iverson, John McKinlay, and Steven Sussman for comments on an earlier draft of this article.

**REFERENCES**


106. Vartiainen, E., Pallonen, V., McAlister, A., Koskela, K., and Puska, P. Effect of two years of
education intervention in adolescent smoking (The North Karelia Youth Project). Bull. WHO
109. Wills, T. A. Stress, coping, and tobacco and alcohol use in early adolescence, in "Coping and
110. Windsor, R. A., Baranowski, T., Clark, N., and Cutter, G. "Evaluation of Health Promotion and