# A Strategy for Optimizing and Evaluating Behavioral Interventions

Linda M. Collins

The Methodology Center and

Department of Human Development and Family Studies

The Pennsylvania State University

Susan A. Murphy

Institute for Social Research and

Department of Statistics

University of Michigan

Vijay N. Nair

Department of Statistics and

Department of Industrial and Operations Engineering

University of Michigan

Victor J. Strecher

Department of Health Behavior and Health Education

University of Michigan

## Abstract

**Background**. This article suggests a multiphase optimization strategy (MOST) for achieving the dual goals of program optimization and program evaluation in the behavioral intervention field. **Methods**. MOST consists of the following three phases: (1) *screening*, in which randomized experimentation closely guided by theory is used to asses an array of program and/or delivery components and select the components that merit further investigation; (2) *refining*, in which interactions among the identified set of components and their interrelationships with covariates are investigated in detail, again via randomized experiments, and optimal dosage levels and combinations of components are identified; and (3) *confirming*, in which the resulting optimized intervention is evaluated by means of a standard randomized intervention trial. In order to make the best use of available resources, MOST relies on design and analysis tools that help maximize efficiency, such as fractional factorials. **Results**. A slightly modified version of an actual application of MOST to develop a smoking cessation intervention is used to develop and present the ideas. **Conclusions**. MOST has the potential to husband program development resources while increasing our understanding of the individual program and delivery components that make up interventions. Considerations, challenges, open questions, and other potential benefits are discussed.

# A Strategy for Optimizing and Evaluating Behavioral Interventions

## *Introduction*

Intervention researchers, intervention targets, health care providers, and other stakeholders would agree that optimizing the potency[1] of behavioral interventions is a worthy objective. Optimized interventions will offer public health advantages by reaching more people and having a greater and more lasting impact on those they reach. Moreover, optimized interventions will benefit research by leading to larger effect sizes and therefore improved statistical power for detecting genuine treatment effects. Although currently there is no widely agreed-upon procedure for optimizing an intervention and its delivery, it is clear that an efficient and scientifically rigorous method is needed for exploring the individual and joint operation of the components of an intervention. Here the term "components" is broadly defined to include both program (i.e. aspects of the intervention program itself) and delivery (i.e. aspects of the implementation) components. Important questions include which program components are working well; which should be discarded, revised, or replaced; which dosages of program components are most appropriate; whether delivery components are enhancing, maintaining or diluting intervention efficacy; and whether individual and group characteristics interact with program or delivery components.[1] Addressing these questions is an indispensable tool in optimizing interventions. For example:

- An intervention of demonstrated potency may contain inactive program components. These components can be discarded to produce a shorter and more economical, yet equally potent, intervention, or replaced by active program components to make an enhanced intervention.
- A set of program components may work well individually, but in combination work less well, perhaps due to negative synergy or because together they impose an unduly heavy participant burden. In a case like this, intervention potency will be increased if a program component is removed.

- Two alternative delivery mechanisms applied to an efficacious intervention program may work very differently, with one undermining the intervention's effectiveness in certain minority groups, and the other maintaining effectiveness in these groups.

Of course, the question is how best to gather information to provide a basis for optimization. Currently, most of the information that forms the basis for refinement of behavioral interventions is gathered within the context of randomized confirmatory intervention trials. During such trials the opportunity is usually taken to collect a wealth of data on study participants, so that after the intervention has been completed, post-hoc, exploratory analyses can be performed in the hope of learning more about the intervention and its effects, such as what worked well, what did not work well, interactions with individual, group, or situational characteristics, and so on. The conclusions drawn from these analyses form the basis of amendments aimed at refinement of the intervention program and/or its delivery, after which the revised intervention is tested in another iteration of the confirmatory trial. At the new trial further data are collected for another round of amendments, which are then subjected to another confirmatory trial, and so on.

In our view, this strategy has some important weaknesses. In confirmatory intervention trials random assignment is aimed at providing a solid foundation for inferences about whether the intervention is effective *as a package,* not for inferences concerning the potency of the *individual components* that make up the package. In other words, even though a confirmatory trial uses random assignment, questions concerning individual program components typically must be addressed by nonexperimental comparisons. Such analyses frequently yield data that can be valuable sources of preliminary information and excellent leads for follow-up experiments, but any inferences are weaker, that is, subject to many more alternative explanations, than those based on randomized experiments.[2,3] Furthermore, this approach is time-consuming and expensive in the long run. Typically many iterations of the confirmatory trial/exploratory analysis/revision/confirmatory trial cycle are required for full optimization of a behavioral intervention, each of which may take years to complete, depending on the nature of the intervention.

The purpose of this article is to suggest as an alternative an approach we call a multiphase optimization strategy (hereafter abbreviated as MOST). MOST is a framework for achieving the dual goals of program optimization and program evaluation in the behavioral intervention field. In this article we advocate the application of this approach in every aspect of behavioral intervention development, including constructing new interventions, fine-tuning existing interventions, and selecting delivery mechanisms. The following section provides a brief introduction to the MOST procedure and the philosophy behind it. Then a slightly modified version of an actual application on smoking cessation intervention is introduced and used to develop and present the ideas. The example will highlight the critical role that theory and prior evidence play in the decisions the investigator makes in MOST. We conclude with a discussion of additional considerations, challenges, open questions, and potential benefits.

## Introduction to the Multiphase Optimization Strategy (MOST)

The basic principles underlying the MOST approach are not new; they originally emerged from engineering[4,5] and have been successfully applied for years in many areas of engineering and manufacturing, as well as other fields such as pharmacology. These fields subscribe to a philosophy stressing that scientific research and discovery is an iterative process of deduction and induction where there is a constant interplay among theory, experimentation, and data analysis. As Box et al.[4] point out:

> *... the best time to design an experiment is after it is finished, the converse… is that the worst time is… the beginning, when least is known. If the entire experiment was designed at the outset, the following would have to be assumed as known: (1) which variables were the most important, (2) over what ranges the variables should be studied... The experimenter is least able to answer such questions at the outset of an investigation but gradually becomes more able to do so as a program evolves*. (p. 303)

Such observations have led to a preference for addressing important research questions by means of a chain of highly focused, strategic studies, with the results of each study informing the design of each subsequent study, rather than relying on a single large study.

MOST is an application of these ideas in the behavioral intervention field, using a strategy with the following three phases as a replacement for the cycle of confirmatory trial/exploratory analysis/revision/confirmatory trial:

- First is a *screening* phase that uses experimentation closely guided by theory to select the important components that merit further investigation.  This phase of research is devoted to randomized experiments examining a number of putatively effective components, combinations of components, candidate interventions and/or candidate implementation systems quickly and efficiently in order to identify those with a high potential for true efficaciousness (or effectiveness) and weed out the others.  In situations where the objective is to improve an existing intervention, it is possible to begin with existing program components, possibly supplemented by promising new components.  In situations where the objective is to arrive at an optimized implementation system, the starting point may be a program of optimized efficacy (perhaps the result of a previous MOST) and a set of candidate delivery components.  In some situations there may be a mix of program components and delivery components under consideration.
- Second is a *refining* phase where the interactions among the identified set of components and their interrelationships with covariates are investigated in detail, again via randomized experiments, and appropriate dosage levels and combinations of components are determined.  The refining phase is informed by the screening phase, as only components that pass the screening phase are investigated further in the second phase.  If the refining phase suggests that additional work is needed, for example, if it is discovered that an important intervention component is effective for males but completely ineffective for females, then additional small and focused screening experiments may be conducted.  The screening and refining phases may be alternated several times, particularly if development is taking place in a relatively uncharted area.  The information gained from the activities of the screening and refining phases is then used to formulate the behavioral intervention.

- Third is a *confirming* phase in which results from the screening and refining phases are used to construct an intervention. This intervention has been optimized, and furthermore, it is not a "black box," because effects of the individual components are understood. The potency of the intervention is confirmed via a standard randomized trial.

Although the MOST perspective is drawn from the field of engineering, it is congruent with existing ideas in behavioral research. For example, Onkin, Blaine, and Battjes [6] called for "a goal-oriented, systematic approach to the development and testing of behavioral therapies" (p. 479). They outlined three stages of research, consisting of therapy development, efficacy testing, and establishing transferability to real-world therapy settings. Authors in other areas of the behavioral and biomedical sciences have made similar appeals for orderly movement through stages of research.[7,8] The MOST framework presented here can be viewed as a principled and rigorous method for moving progressively through the stages of research suggested by these authors. Pilot testing, particularly as advocated by Sussman, Dent, Burton, Stacy, and Flay, [9] is similar in spirit to MOST. However, we wish to emphasize that the screening and refining phases in MOST are not pilot testing phases, at least not according to the usual definition of pilot testing as a fairly informal feasibility test or fact-finding exercise.[10] Pilot testing is important and will frequently play a role in MOST as a preliminary step before the screening phase is begun.

## Statistical Tools Used in MOST

In order to make the best use of available resources, MOST relies on design and analysis tools that help maximize efficiency. The single most important tool for maximizing efficiency and reducing resource demands comes, again, from engineering and related fields. This tool is a class of experimental design techniques called fractional factorials. [4] In engineering and related fields there are typically so many possible design choices, materials, manufacturing conditions, etc., that resource limitations preclude constructing and testing all of them. Similarly, in behavioral intervention research theory suggests many possible components that are candidates for inclusion in an intervention, and testing all possible combinations and doses of program components is usually not an option. Fractional factorial designs provide a principled, theory-driven method for identifying key combinations for extensive testing, with the ultimate objective of arriving at the most effective prototype without having to invest resources in trying each and every one. These and related approaches reduce the size of the designs needed in order to conduct a randomized test of experimental hypotheses, often considerably, while still directly addressing the research questions of interest with adequate statistical power. These resource savings are predicated on the investigator's strategic choices, which in turn are based on working assumptions directly informed by theory, clinical experience, and prior research. Therefore, to a substantial degree the investigator's conceptual and substantive input determines the efficiency of MOST. The validity of the important working assumptions can be assessed in the refining phase.

Random assignment is a cornerstone of all three phases of MOST. However, in the screening and refining phases, traditional hypothesis testing is not (formal hypothesis testing usually plays an important role in the confirming phase). This suggests some additional strategies for conserving resources. In some cases the philosophy may be that in the screening and refining phases a Type II error, i.e. overlooking an active intervention component, is at least as serious as a Type I error, i.e. mistakenly concluding that an inactive component is active. An investigator who subscribes to this philosophy may choose a Type I error rate greater than the traditionally used (but arbitrary) $p$=.05. Given a fixed sample size, this increases statistical power. Another possibility is to rank components by standardized effect size and select the most important ones, rather than examining the statistical significance of each effect. Both of these approaches reduce sample size demands. Later, when the full intervention is tested in the confirming phase, the traditional Type I error rate may be used.

## An Example of the MOST Procedure

### *Overview*

In this section a slightly simplified version of a real application, currently being conducted by three of us, is presented as an example to illustrate the concepts underlying MOST as it can be applied in behavioral intervention research. This example involves development of a new intervention, but the approach is equally appropriate for optimizing an existing multi-component program, and for investigating delivery components for a set of previously optimized program components. The components being tested may be new and previously untested, or they may be drawn from existing programs.

The example involves development of a new and optimized behavioral intervention program for smoking cessation. [11] There were six intervention components of interest to the investigators. The first four, which concerned the content of the message, were program components, whereas the last two were delivery components. (1) *Outcome expectation messages* address an individual's expectations of outcomes related to quitting smoking. Such messages may be tailored to the individual according to variables such as health history and perceived health status. (2) *Efficacy expectation messages* address relevant barriers to quitting, high-risk situations, existing skills for successful quitting, and attributions for previous failed attempts at quitting. Such messages may be tailored to the individual according to variables such as smoking history, smoking behavior, and barriers to quitting in the home. (3) *Message framing* motivates the decision to quit. The framing may be in positive (e.g. quitting results in having more energy) or negative (e.g. not quitting increases risk of cancer) terms. (4) *Testimonials* about quitting are accounts of smoking-related and quitting-related experiences from former smokers. (5) *Exposure schedule* refers to whether the message is delivered in one larger message or several smaller ones. (6) *Source of message* may be either the health maintenance organization (HMO) or primary care physician (PCP).

The screening phase is used to decide which of these components to include in or drop from the intervention. The refining phase is used to work on details of dosage and pursue leads suggested by results of the screening phase. Finally, in the confirming phase the behavioral intervention program, consisting of components and dosages suggested by the screening and refining phases, is tested in a randomized trial.

### *The Screening Phase*

The question of interest in the screening phase is always simply whether or not to include a component in the intervention package. In this example several of the components could reasonably take on several different values. Outcome expectation messages and efficacy expectation messages could take on many values, corresponding to different degrees of tailoring the message to characteristics of the individual. In addition, the investigators felt that exposure schedule could range between one and six messages. This raised the question of which values to select for study. In this phase only two levels of these components were considered, corresponding to settings at or near the ends of the practical range, with the goal of first estimating the linear effect. The working assumption here was that if tailoring of outcome expectation messages, tailoring of efficacy expectation messages, and exposure schedule were important, then the difference at the low and high levels would be large enough to be meaningful. This was expected to furnish only a gross sense of whether a component was efficacious. The appropriate dose is investigated later, during the refining phase.

Thus for purposes of the screening phase, there are six factors with two levels each. A full factorial experiment would require 2^6 = 64 cells (component combinations). In this case it was apparent that this large experiment made such significant demands in terms of money, time, logistics, staff training, experimental subjects, and other resources, that it was not feasible. One option would be to reduce the size of the experiment by eliminating some components, or holding them constant, and conducting a full factorial experiment involving the remaining components. But this was considered unsatisfying from a scientific point of view, because it left promising intervention components and potential synergistic interactions unstudied. Under these conditions, fractional factorial designs often present an attractive and highly economical alternative. The primary advantage of fractional factorial designs is that they allow the researcher to study a large number of factors using many fewer cells than required by a full factorial. The philosophy underlying fractional factorials is that usually it is not necessary to involve all the cells of a full factorial design in order to address the questions of greatest interest. Full factorial designs permit the estimation of every individual effect, including all interactions up to the highest order interaction, which in this example is a six-way interaction. Full factorials are necessary when it is important to study each and every effect. But in the behavioral sciences many effects, particularly higher-order interactions, are so small in size that they are negligible, and/or have such large standard errors that it would take an impractically large sample size to detect them. Frequently it is possible to articulate working assumptions that identify negligible effects based on theory, literature and/or past experience. These working assumptions form the basis for the choice of a fractional factorial design that selectively reduces the number of component combinations that must be tested in the screening phase. Then available resources are concentrated on addressing vital research questions instead of squandered on estimating effects that are trivial in size and of no scientific interest.

The reduction in the number of cells provided by fractional factorial designs results in a tradeoff in terms of "aliasing" of effects. When two effects are aliased, the design has "bundled" them together, estimating a single combined effect for both.[4] There is a variety of fractional factorial designs from which to choose, each resulting in aliasing of different effects. Ideally, a design is identified in which each aliased effect includes at most only one effect deemed important, with any other effects included in the "bundle" assumed likely to be negligible. Then if the working assumptions are reasonable, estimates of aliased effects provide useful estimates of effects of interest. Thus the articulation of working assumptions and the strategic choice of a corresponding fractional factorial design is critical. In this way, theory and prior knowledge inform design decisions in order to reduce resource demands.

In the example reviewed here, the investigators were able to make a plausible set of working assumptions, based on theory, scientific literature, and prior experience. These assumptions guided the choice of a fractional factorial design. The working assumptions included the following: First, the investigators noted a set of effects that were of particular interest and that previous research suggested would be sizeable. The set consisted of all the main effects and the following two-factor interactions: outcome expectation messages by efficacy expectation messages, and outcome expectation messages by message framing. A design was sought that concentrated resources on these effects. Second, because there was no compelling reason to believe that higher-order interactions would be sizeable, the investigators made the working assumption that all the three, four, and five-way interactions and the six-way interaction were negligible, and thus decided to put few resources into studying them. Third, previous research suggested that the exposure schedule by source of message and also testimonial by source of message interactions would be inert.

Based on the working assumptions, it is possible to identify a balanced 16-cell fractional factorial experiment – one-quarter the size of the corresponding full factorial design – that can be used to study all six components and their combinations. Table 1 shows the cells in this design, with the corresponding levels of each factor. For example, the treatment in Cell 1 consists of untailored outcome and efficacy expectation messages, messages framed in terms of loss, and no testimonials included, with exposure to the message at a single occasion, delivered by the individual's HMO. As Table 1 also shows, the term "balanced" in this context means that every factor (Outcome expectation messages, Efficacy expectation messages, and so on) occurs an equal number of times at the high and low levels. Similarly, every two-, three- and higher- factor combination also occurs an equal number of times. This balance provides many statistical advantages; for a complete discussion of these advantages, see Box et al.[4] For now, we point out some advantages enjoyed by balanced fractional factorial designs when their underlying working assumptions are appropriate.

First, in the design shown in Table 1 all main effects and the two-way interactions of interest are estimated without bias. Second, even though six different components are being studied, the sample size, and hence statistical power, for testing each main effect is the *same* as that for a single component. To see why statistical power is not lost, consider the estimation of the main effect of Message source. This factor has two levels, delivery by PCP and delivery by HMO. Because the design is balanced, ½ of the research subjects are used to estimate the mean response to a PCP delivering the message and ½ of the research subjects are used to estimate the mean response to an HMO delivering the message. This is the same number of subjects used to estimate the main effect in an experiment with just the one factor Message source at two levels with no other factors present. The same is true for each of the six main effects – ½ of the research subjects are used to estimate each mean. This is one aspect of the economy gained by using balanced designs. Not all fractional factorial designs are balanced; those that are not may not have these important characteristics.

### *The Refining Phase*

The refining phase is used for activities such as obtaining a sense of the best dosages to use, assessing whether key variables moderate important effects, and resolving any lingering questions resulting from the screening phase. The nature of the experiments to be conducted in the refining phase depends on the specific components under investigation, the investigator's a priori knowledge and theory, and the results of the screening experiments. Because as of this writing the screening phase of the study to develop a smoking cessation program described above is still in progress, the remainder of this section will describe hypothetical refining phase activities based on the example.

As an example of a typical refining phase activity, suppose the screening experiments described in the example indicated that the main effects of outcome expectation messages and efficacy expectation messages as well as their interaction are important, and that there are no other significant main effects or interactions. As previously mentioned, these two components can take on any of a range of possible values, ranging from no individualized tailoring of the messages to highly individualized tailoring. Recall that for purposes of the screening experiment only two levels of these components were selected for study, leaving open the question of the appropriate joint dosage levels of the two components. This question can be addressed by conducting a *response surface experiment* [5,12] to examine how the outcome changes as the dosage levels of the two components are varied. It is common in the design literature to conserve resources by limiting the component settings (dosage levels) to three levels and approximating the response surface by using a quadratic approximation.[5,12] In this example a follow-up full factorial experiment can be run crossing three levels for each component, resulting in 3x 3 = 9 cells. The three dosage levels should be chosen appropriately using subject matter knowledge, and in this case the settings of the other components can be selected based on other considerations (least cost, current levels, etc.) because the screening experiment showed that there were no other factors that had any significant effects.

The fitted joint quadratic response surface model for outcome expectation messages and efficacy expectation messages can then be used to determine the appropriate joint dosage levels so as to maximize the outcome.

If it is desired to choose the optimal dosage level more precisely, it is of course possible to use more than three levels. For situations where there are more than two factors for which the optimal settings must be determined, there are other design possibilities, such as fractional factorial designs for more than two levels or central composite designs. [5,12] These and other implementation details on the use of fractional factorials and related designs in the behavioral intervention field will be developed in a subsequent article.

Another aspect of choosing the optimal dose, and one that is critical when an adaptive intervention is planned (where dose is dependent on individual participant characteristics [13]) is establishing whether the optimal dose varies according to characteristics of the individual or other variables. This can be investigated in the refining phase by conducting experiments using these characteristics as blocking variables. In the example, the investigators were interested in determining whether an individual’s current stage of change (e.g. precontemplator, contemplator) interacted with dose. For example, theory may suggest that a positive message framing will be more effective with precontemplators, whereas a negative message framing will be more effective with contemplators. A series of experiments to determine the extent to which characteristics of the individual, environment, delivery, etc. moderate treatment effectiveness can be highly useful even in situations where the intervention consists of a single component or for conceptual or operational reasons cannot be decomposed.

In many studies, an important activity of the refining phase is to conduct experiments aimed at verifying the working assumptions. We have been careful to use the term “working assumptions” in order to convey that these assumptions can be tested and, where indicated, revised during the course of the MOST procedure. For example, some of the aliased effect “bundles” will be composed entirely of effects that should be negligible according to the working assumptions. If one of these aliased effects unexpectedly turns out to be large, this suggests that some of the working assumptions are false. It is possible to investigate this in the refining phase by running additional cells to isolate the individual effects. Thus if a working assumption turns out to be incorrect, this usually can be dealt with, and does not invalidate the data gathered using the chosen fractional factorial design.

### *The Confirming Phase*

The confirming phase is the final phase of the MOST approach. At the point when the confirming phase is begun, the screening and refining phases have identified the important components of the intervention and their most appropriate levels or doses. Using this information, the intervention researcher can construct an optimized prototype version of the program, made up of only components determined to be active, at doses determined to be most efficacious. In the confirming phase this optimized prototype intervention is evaluated in a standard randomized intervention trial, in order to verify that it is in fact sufficiently efficacious to justify wider adoption. Generally only a single treatment group and a comparison group will be required, because only one version of the intervention will be under consideration -- the work of refining the intervention will have been done before the confirming phase, in the screening and refining phases. Thus such an evaluation can be relatively straightforward. Frequently blocking will be employed for variables that are expected to affect the outcome but are not amenable to random assignment, as was discussed in the example of the refining phase above.

## Additional Considerations, Open Questions, and Potential Benefits

### *More about Fractional Factorial Designs*

As discussed previously, MOST often involves the use of fractional factorial designs. The term “fractional factorial” may be unfamiliar to many intervention scientists, but as West and Aiken [14] have reviewed, several varieties of fractional design strategies have been used in intervention studies, particularly for behavioral preventive interventions. Many of these designs are fractional in the broadest sense of the term, that is, they are subsets of full factorials; however, not all of them are *balanced* fractional factorials. The advantage of balanced fractional factorial designs over the full factorial design is that for a given sample size, they require many fewer cells while maintaining statistical power. Thus they reduce the expense of formulating and implementing different combinations of intervention components. By contrast, unbalanced fractional factorial designs do not always maintain statistical power.

Some intervention scientists may be uncomfortable specifying which effects are expected to be negligible because, as West and Aiken [14] pointed out, “If assumptions about which effects are negligible are not reasonable, the estimates of interest will be biased because they will be confounded with higher order interactions (p. 179).” Although the possibility of bias is a risk, in the MOST procedure the risk is less than it may appear on first consideration if, as described briefly above, the refining phase is used as an opportunity to investigate the validity of some of the assumptions. In cases where it appears that an assumption may be invalid, additional cells can be run and added to the study. Then the data from these cells can be combined with the previously collected data to arrive at unbiased effect estimates.

We wish to comment that every a priori prediction made in science is accompanied by some risk that the prediction is incorrect. The question is whether the benefits, in this case greatly increased economy of time and other resources, are worth the risk, or, put another way, which strategy ultimately will move science forward faster.

### *Open Questions and Challenges*

One challenge associated with the MOST approach is the extent of the net resource requirements needed as compared to the standard approach. Certainly the use of multistage experimentation is expected to delay the launching of an experimental confirmatory trial to evaluate a new intervention, because of the time spent in the screening and refining phases. Furthermore, the screening and refining phases may require recruiting research subjects, training staff to deliver multiple versions of an intervention, and so on. The tradeoff is that ultimately fewer full-blown confirmatory trials must be undertaken, and these will be undertaken with greater confidence that the intervention will be efficacious. Moreover, as discussed above, these confirmatory trials are likely to be more straightforward than today’s complex program evaluations. We are convinced that in the long run, MOST will be less resource-intensive than if the intervention were devised, implemented, subject to post-hoc analysis, redesigned, and tested again in the traditional way, with the cycle repeated until the desired level of efficaciousness is reached. We argue that for these reasons, MOST is likely to be cost-effective; however, this has not yet been verified empirically.

A related issue is that of timing. When there is a public health crisis, it may be necessary to go in immediately with the best available intervention program, even if it is not optimal, rather than wait until a MOST procedure has been completed. Although MOST may seem like an unaffordable luxury under such circumstances, we would argue that these are precisely the circumstances under which an intervention of the highest possible potency is most badly needed. MOST can always be undertaken to optimize a program that is currently in the field. The optimized, more potent intervention can eventually replace the “rough and ready” version in order to reap increased public health benefits.

Another matter of timing concerns federal funding cycles. Currently, there appears to be a widespread perception among scientists that funding sources such as the National Institutes of Health expect a full-blown confirmatory intervention trial to be conducted within the typical five-year funding period. In most cases, this will be insufficient for carrying out all three phases of MOST. One way that the National Institutes of Health could support the idea of experimental development of intervention programs prior to a full-blown confirmatory intervention trial would be to allow the possibility of funding for one or, depending on the area, possibly more five-year periods of screening and refining. Later a subsequent application could be made for funding to test the refined intervention, provided that an argument could be made that the screening and refining phases were productive and clearly point the way to a credible intervention. Funding agencies may be increasingly open to these ideas. As mentioned above, the example of a MOST implementation presented above is based on a project to develop a behavioral intervention to prevention cancer that is currently funded by the National Cancer Institute.

When MOST is used, a series of experiments may, in some cases, take place over a period of years. This requires a steady stream of research subjects; but, we emphasize, if fractional factorial designs are used judiciously the sample size requirements will remain modest. In some areas of behavioral intervention, cohort effects may emerge to complicate interpretation of results. When there are cohort effects, it is crucial to investigate their source. They may be due to differences in the population from year to year, perhaps attributable to historical trends. For example, a component of a program to prevent teen pregnancy may not work as well if recent cohorts are more sophisticated about reproductive health than previous cohorts. Even if the population has not changed, nonrepresentative sampling can create the illusion of cohort effects. Or, cohort effects may be due to implementation differences from year to year, for example because staff have become more comfortable with the program. Reasons like the former call for continually revising/updating the program to accommodate the changing population, whereas reasons like the latter may call for increased vigilance about implementation.

Another challenge is that in some interventions, such as those aimed at drug abuse prevention or cancer prevention, the outcome of interest is often months or years away, whereas the MOST approach relies on being able to do a series of experiments in rapid succession, with the outcome of each experiment informing the design of the next in the series. In some cases, it may be possible to rely on proximal rather than distal outcomes, such as six-month smoking cessation rates rather than one-year rates. Where there is strong a priori theory about mediation effects, it may be possible to look for short-term effects on mediators, reasoning that if the mediator is affected in the hypothesized direction, the outcome will eventually be affected in turn. This approach is consistent with the MOST philosophy of making design and measurement decisions that are closely informed by theory. It should be noted that we are not advocating abandonment of long-range outcomes in favor of more convenient proximal outcomes. The question of whether the intervention has an impact on the outcome of ultimate interest is addressed in the confirming phase.

The nested structure of many samples in intervention research is another consideration. For example, children may be nested within classrooms and/or schools, and patients may be nested within clinics. This nested structure raises many practical and conceptual issues, concerning at which level effects are defined, at which level random assignment is to be done, and how dependence among observations sharing the same nesting unit (e.g. classroom) is to be handled. For the behavioral intervention scientist considering the MOST approach, it is unclear whether these issues require the involvement of many groups in the screening and refining phases. This is an important open question.

***Potential Benefits of MOST***

We strongly believe that any investment in retooling required by MOST is justified by the promise this approach holds for increasing the potency of interventions. This belief rests on three signature characteristics of the MOST approach.

First, *resources are targeted strategically via the use of fractional factorial and related design and analysis approaches,* thereby keeping resource demands at the lowest level needed to get the job done. Besides the obvious benefits of economy, a means of testing program-related hypotheses that does not require the investment of a full confirmatory intervention trial opens up new possibilities for behavioral scientists. One possibility is testing more creative approaches that may be long shots, but also may have the possibility of a larger payoff. In the screening and refining phases it is feasible to incorporate new ideas, even new intervention components, and then test them in a cost-effective and relatively rapid way. Innovations that show potential to enhance a program can be incorporated into the refined intervention that will be tested in the confirming phase; innovations that do not work out can simply be discarded. Another possibility is ongoing experimentation aimed at continually updating a successful intervention program in order to allow it to adapt to rapidly evolving societal trends. This can maintain the freshness of approach that particularly appeals to younger target populations, such as adolescent participants in substance use prevention programs.

Second, *a full confirmatory randomized intervention evaluation trial is mounted only when an optimized intervention has been reached,* and only then when there is sufficient potential for efficacy (or effectiveness), based on the information gathered in the screening and refining phases. In other words, program evaluation resources are reserved for interventions that have been demonstrated to have a high probability of success. This reduces the probability of spending resources to evaluate a program that turns out to have little or no effect. It also reduces the probability of needing major revisions to an intervention to salvage it while the randomized evaluation trial is in full swing.

Finally, in every phase of the process MOST *relies solely on a series of carefully randomized experiments* rather than observational data or post-hoc analyses. Any important naturally occurring variables not amenable to random assignment (e.g. gender) are included as blocking factors. This approach provides a much stronger basis for inference than post hoc, exploratory analyses of non-experimental data. Moreover, MOST offers a means of “packing/unpacking” interventions by helping researchers to understand which components are operating. As the field moves increasingly toward adaptive, or tailored, interventions, [11,13,15] this information becomes particularly critical. Adaptive interventions are designed to tailor the intervention to the audience and/or the context in such a way that the dose delivered is sufficient without being so large as to be wasteful or create iatrogenic effects. These interventions hold much promise; however, their success relies heavily on selection of appropriate tailoring variables and accurate articulation of expected dose response. The use of randomized experiments with blocking helps clarify for whom a particular component is effective, which dose is most effective depending on participant characteristics, and the circumstances under which a component is most effective, findings which directly inform the design of adaptive interventions.

A thorough understanding of the individual program and delivery components that make up an intervention is in and of itself of scientific interest. As Flay [1] has pointed out, most of today’s successful interventions have never been subjected to testing of individual components. A theory-driven program of randomized, controlled experiments on intervention components holds up the possibility of accumulating a coherent, replicable, research-based body of knowledge that will inform the design of behavioral interventions. This will move intervention science further toward fulfilling its potential for improving the human condition.

## References

1. Flay BR. Efficacy and effectiveness trials (and other phases of research) in the development of health promotion programs. Preventive Medicine 1986; 15: 451-474

2. Cook TD, Campbell DT. Quasi-experimentation: Design and analysis issues for field settings. Chicago: Rand McNally, 1979

3. Shadish WR, Cook TD, Campbell DT. Experimental and quasi-experimental designs for generalized causal inference. New York: Houghton Mifflin, 2002

4. Box GEP, Hunter WG, Hunter JS. Statistics for experimenters: An introduction to design, data analysis, and model building. New York: Wiley, 1978

5. Box GEP, Draper NR. Empirical model-building and response surfaces. New York: Wiley, 1987

6. Onken LS, Blaine JD, Battjes, RJ. Behavioral therapy research: A conceptualization of a process. In: Henggeler SW, Santos AB, editors. Innovative approaches for difficult-to-treat populations. Washington, DC: American Psychiatric Press, 1997

7. Greenwald P, Cullen JW. The scientific approach to cancer control. CA: A Cancer Journal for Clinicians 1984; 34: 328-332

8. UK Medical Research Council. A framework for development and evaluation of RCTs for complex interventions to improve health. Available from URL: http://www.mrc.ac.uk/mrc_cpr.pdf London: Medical Research Council, 2000

9. Sussman S, Dent CW, Burton D, Stacy, AW, & Flay, BR. Developing school-based tobacco use prevention and cessation programs. Thousand Oaks, CA: SAGE, 1995

10. Vogt WP. Dictionary of statistics and methodology: A nontechnical guide for the social sciences. Newbury Park: Sage, 1993

11. Strecher VJ. Computer-tailored smoking cessation materials: A review and discussion. Patient Education and Counseling 1999; 36: 107-117

12. Myers RH, Montgomery DC. Response surface methodology. New York: Wiley, 1995

13. Collins LM, Murphy SA, Bierman KA. A conceptual framework for adaptive preventive interventions. Prevention Science 2004; 5: 181-192

14. West SG, Aiken LS. Toward understanding individual effects in multicomponent prevention programs: Design and analysis strategies. In Bryant KJ, Windle M, West SG, editors. The science of prevention: Methodological advances from alcohol and substance use research. Washington, DC: American Psychological Association, 1997

15. Kreuter M, Strecher V, Glassman B. One size does not fit all: The case for tailoring print materials. Annals of Behavioral Medicine 1999; 21: 276-283

## Author Notes

Correspondence concerning this article may be sent to Linda M. Collins, The Methodology Center, College of Health and Human Development, Penn State, University Park, PA 16802, LMCollins@PSU.EDU. This article has benefited from discussion at the 2003 Snowbird Conference and the 2003 Society for Prevention Research Conference, particularly the comments of Gilbert Botvin, Richard Clayton, Delbert Eliot, Brian Flay, David Hawkins, Elizabeth Robertson, and Richard Spoth. It has also benefited from the comments of John Graham and many other colleagues at Penn State's Methodology Center and Prevention Center. This work has been supported by National Institute on Drug Abuse grants P50 DA10075 (Dr. Collins and Dr. Murphy) and K02 DA15674 (Dr. Murphy), National Cancer Institute grant P50 CA 101451 (Dr. Strecher, Dr. Murphy, and Dr. Nair), and National Science Foundation grant DMS 0204247 (Dr. Nair).

# Footnote

[1] For the sake of simplicity, we use the term “potency” to refer to either efficacy or effectiveness, when the point made applies to either.

Table 1

Sixteen Cell Resolution IV Fractional Factorial Design with Six Two-level Independent Variables

| | Program Components | | | | Delivery Components | |
|---|---|---|---|---|---|---|
| Cell | Outcome expectation messages | Efficacy expectation messages | Message framing | Testimonials | Exposure schedule | Source of message |
| 1 | Untailored to individual | Untailored to individual | Negative | Absent | One occasion | Health Maintenance Organization (HMO) |
| 2 | Untailored | Untailored | Negative | Present | Six occasions | Primary Care Physician (PCP) |
| 3 | Untailored | Untailored | Positive | Absent | One | PCP |
| 4 | Untailored | Untailored | Positive | Present | Six | HMO |
| 5 | Untailored | Tailored | Negative | Present | One | PCP |
| 6 | Untailored | Tailored | Negative | Absent | Six | HMO |
| 7 | Untailored | Tailored | Positive | Present | One | HMO |
| 8 | Untailored | Tailored | Positive | Absent | Six | PCP |
| 9 | Tailored | Untailored | Negative | Present | One | HMO |
| 10 | Tailored | Untailored | Negative | Absent | Six | PCP |
| 11 | Tailored | Untailored | Positive | Present | One | PCP |
| 12 | Tailored | Untailored | Positive | Absent | Six | HMO |
| 13 | Tailored | Tailored | Negative | Absent | One | PCP |
| 14 | Tailored | Tailored | Negative | Present | Six | HMO |
| 15 | Tailored | Tailored | Positive | Absent | One | HMO |
| 16 | Tailored | Tailored | Positive | Present | Six | PCP |