TBM

ORIGINAL RESEARCH

Evaluating individual intervention components: making decisions based on the results of a factorial screening experiment

Linda M Collins, PhD,[1,2] Jessica B Trail, MS,[1,3] Kari C Kugler, PhD,[1] Timothy B Baker, PhD,[4] Megan E Piper, PhD,[4] Robin J Mermelstein, PhD[5]

[1]The Methodology Center, Pennsylvania State University, 204 E. Calder Way, Suite 400, State College, PA 16801, USA
[2]Department of Human Development and Family Studies, Pennsylvania State University, State College, PA, USA
[3]Department of Statistics, Pennsylvania State University, State College, PA, USA
[4]Center for Tobacco Research and Intervention, University of Wisconsin, Madison, WI, USA
[5]Department of Psychology and Institute for Health Research and Policy, University of Illinois, Chicago, IL, USA

Correspondence to: L M Collins
lmcollins@psu.edu

Cite this as: *TBM* 2014;4:238–251
doi: 10.1007/s13142-013-0239-7

ABSTRACT

The multiphase optimization strategy (MOST) is a framework for not only evaluating but also optimizing behavioral interventions. A tool critical for MOST is the screening experiment, which enables efficient gathering of information for deciding which components to include in an optimized intervention. This article outlines a procedure for making decisions based on data from a factorial screening experiment. The decision making procedure is illustrated with artificial data generated to resemble empirical data. The illustration suggests that this approach is useful for selecting intervention components and settings based on the results of a factorial screening experiment. It is important to develop methods for making decisions based on factorial screening experiments. The approach demonstrated here is potentially useful, but has limited generalizability. Future research should develop additional decision making procedures for a variety of situations.

KEYWORDS

Comparative effectiveness, Multiphase optimization strategy, Factorial experiments, Behavioral interventions

Implications

Practice: Practitioners considering implementing an intervention should take into account whether it has been optimized.

Policy: Policy decisions should include consideration of whether an intervention has been optimized.

Research: Factorial screening experiments can yield outcome data on multiple behavioral intervention components; strategies developed in engineering can guide the identification of particularly promising components, a key step in intervention optimization.

For many years, there has been a heavy emphasis on evaluation of multicomponent behavioral interventions by means of the randomized clinical trial (RCT) and very little emphasis on examination of the individual components making up interventions. In most cases it is unknown whether all of the major components making up a successful intervention contribute to the overall observed effect, or whether expensive or logistically demanding components contribute enough to offset their resource requirements. When new interventions are developed, there may be little disincentive for including many components to try to ensure a significant program effect, even though a significant program effect is no guarantee that all of the components are necessary.

Today, there is a growing interest in improving the effectiveness and efficiency of behavioral health, as evidenced by the increase in research comparing the effectiveness of alternative approaches to prevention and treatment of health problems. One approach to hastening progress in this area is to optimize the performance of interventions proactively, before they are evaluated and compared to existing alternatives. The multiphase optimization strategy (MOST) [1] provides one framework for accomplishing this. MOST is a comprehensive, engineering-based approach for behavioral intervention optimization and evaluation. MOST includes the RCT as the gold standard for establishing whether an intervention as a package has a statistically significant effect in comparison to a control group, standard of care, or competing intervention. However, in MOST, additional steps are taken to optimize the intervention systematically, typically in advance of an RCT.

In MOST, the term "optimization" has a specific technical meaning: "The process of finding the best possible solution to a problem… subject to given constraints" [2]. Thus, the goal of MOST is not to build the best intervention in some absolute, and perhaps unattainable, sense; rather, the long-range goal is to build the best intervention that can be attained while working within realistic and clearly articulated limitations on resources, both for inter-

vention development and for eventual intervention implementation, dissemination, and fidelity. For example, suppose it is known that to go to scale an intervention will have to be delivered for no more than $100 per person. Then MOST can be used to engineer an intervention that maximizes effectiveness within this per-person cost limit on intervention delivery. MOST can also be used to engineer an intervention to other specifications, such as the least costly intervention that reaches a stated standard of effectiveness, an intervention that maximizes effectiveness without exceeding a certain time limit on implementation, or the most cost-effective intervention.

MOST is conducted from a resource management perspective, which holds that research should be conducted so as to use available resources to move science forward the fastest. Phased optimization approaches like MOST often take longer than the traditional approach of developing and immediately testing multicomponent intervention packages via an RCT, because they insert steps to optimize the intervention into the process. However, the optimization phase of MOST yields incremental scientific knowledge about the efficacy of specific intervention components and their combinations. This incremental progress can be built on with successive applications of MOST to form a cumulative base of knowledge about which intervention components work well and how they work together. Thus we would argue that phased optimization approaches will move science forward faster in the long run. (For a demonstration of this via statistical simulation, see Collins et al. [3]).

Intervention components

An important part of intervention optimization in MOST is examining the empirical performance of individual intervention components, with the objective of gathering the information necessary to assess the effects of the different components and make decisions about the ultimate composition of the intervention (i.e., to identify the optimal package of components). Intervention components may be drawn from the content of an intervention; features that promote engagement, adherence, or compliance; steps to improve fidelity; or aspects of implementation or dissemination. In this section, we briefly review three examples that illustrate the variety in the objectives of component screening experiments and the intervention components examined.

When the objective of a component screening experiment is to select intervention content, intervention components typically are based on translation of basic or applied scientific findings. For example, Baker et al. [4] and Collins et al. [1] described an application of MOST aimed at optimization of a primary-care-clinic-based smoking cessation intervention developed around a phase-based model of smoking cessation. Three of the six candidate components pertained to the precessation phase of the smoking cessation process, which was defined as the 3 weeks prior to the quit day. These components were nicotine patch, nicotine gum, and (precessation) in-person counseling. Two more components pertained to the cessation phase, defined as the quit day to 2 weeks postquit. These were (cessation) in-person counseling and telephone counseling. The final component pertained to the duration of nicotine replacement therapy during the maintenance phase, defined as 2 weeks postquit to 6 months postquit.

Another example is Strecher et al. [5], who examined five intervention components that were to make up an Internet-delivered smoking cessation intervention. Their components were exposure schedule (receiving the message all at once or over a 5-week period); personalization of the message source; and the depth of individual tailoring provided in three different domains: outcome expectations, efficacy expectations, and positive modeling (a story about someone who quit successfully). Tailoring refers to producing a more individualized message by altering the message in response to assessed characteristics of the smoker; tailoring depth refers to how individualized the message is.

Sometimes the objective of a component screening experiment is not selection of intervention content, but rather the translation of an efficacious intervention into an effective, scalable intervention with minimal loss of potency. For example, in the study of Caldwell et al. [6], the content of HealthWise, a school-based drug abuse and HIV prevention program developed for use in South Africa, had been established and the intervention had previously been evaluated [7]. Caldwell et al. [6] wanted to determine how to optimize the fidelity of the delivery of HealthWise before its inclusion in the curriculum of a local school district. The three components they examined were the type of training provided to the teachers, special ongoing support provided to the teachers, and school-level measures taken to create a climate supportive of the intervention.

The objective of MOST

Although it may appear counterintuitive, the objective of MOST is not to identify the single best combination of intervention components out of the set of all possible combinations. Instead, the objective is to identify *one of the best* combinations of components. Definitive identification of the single best combination would require a massive RCT with intervention arms corresponding to each viable combination. Such an experiment would be prohibitively resource-intensive, particularly if there were several high-performing combinations with relatively small differences between them. However, the more modest goal of identifying of one of the best combi-

nations is attainable with a reasonable allocation of resources in most cases.

Given a set of intervention components to be examined, MOST focuses on identifying the components that demonstrate the desired effects on the outcome(s); put another way, MOST screens out any components that fail to show the desired effects. This information is then used to build an intervention out of the resulting set of components. The idea is that although the resulting set may or may not be the single best combination, it is highly likely to be a very good combination. (For verification of this by statistical simulation, see Collins et al. [3].) Once the intervention has been built, MOST calls for confirmation of intervention effectiveness via a standard RCT.

Selection of components: component screening experiments

In MOST, as in engineering and related fields, selecting the best components and screening out poorly performing components is based on the results of a *screening* experiment. Here, we refer not to screening individual subjects for eligibility to be included in a study, but to screening intervention components for inclusion in an intervention package. The purpose of a screening experiment is to provide an efficient way of gathering information that will be used in making decisions about which components should be chosen for the intervention that is to be evaluated in a subsequent RCT.

The examples of MOST reviewed above illustrate the kinds of decisions that must be made based on component screening experiments. For every component in the three studies, a choice had to be made between two settings or dosages. In some cases, the two settings were "include" and "do not include." One purpose of the Collins et al. [1] study was to determine whether or not nicotine replacement during the precessation phase led to better cessation outcomes. Caldwell et al. [6] gathered data to inform the decision about whether or not providing special ongoing support to teachers resulted in better intervention fidelity. In other cases, the decision was whether a more expensive, time-consuming, or logistically complicated version of a component should be included rather than a cheaper, briefer, or simpler alternative. In the study of Collins et al., the investigators wished to decide whether intensive in-person cessation counseling led to better cessation outcomes than minimal counseling, and whether a dosage of 16 weeks of maintenance nicotine replacement therapy led to better outcomes than the standard dosage of 8 weeks. Caldwell et al. wished to determine whether the standard teacher training of one and one-half days was sufficient, or whether a longer, more elaborate, and more costly enhanced training produced better intervention fidelity. Strecher et al. [5] were interested in whether a message was more effective at promoting smoking cessation if it was delivered in one session, which would be logistically less complicated, or over several sessions. In each of these cases, the more resource-intensive setting of the component had to demonstrate its value empirically before it could be selected; otherwise, the less resource-intensive setting would be the default.

In the MOST framework, a component screening experiment may be conducted using any experimental design, as long as the design is selected from a resource management perspective [3]. From this perspective, the best experimental design is the one that directly addresses the research questions deemed most important, while making the most efficient use of available research resources, such as subjects, money, time, equipment, and trained staff. When design alternatives are compared for use in MOST, very often factorial experiments emerge as the most economical. In fact, the resource management perspective led Collins et al. [1], Strecher et al. [5], and Caldwell et al. [6] to examine their intervention components using factorial experiments. A major source of the economy of factorial experiments is their efficient use of research subjects. For example, in the Collins et al. [1] study, one alternative to a factorial experiment would have been to conduct individual experiments, one for each intervention component. However, this would have required six times more subjects to achieve the same statistical power as a factorial experiment [8].

One tradeoff for this economy is that factorial experiments often require implementation of large numbers of experimental conditions. An experimental condition is a distinct combination of levels of the factors in an experiment (e.g., for a four-factor design with two levels in each factor, there are $2\times2\times2\times2=16$ experimental conditions. Note that in factorial experiments, unlike RCTs, the number of experimental conditions in the design bears little relation to overall sample size requirements because the entire sample is used to evaluate the effect of each factor; see Collins et al. [8].) When the overhead associated with the implementation of experimental conditions is high, this can more than negate any gains in economy that a factorial experiment provides in terms of number of subjects. However, fractional factorial designs are available that can cut the required number of experimental conditions by half or more. Collins et al. [1] used a fractional factorial design that enabled them to examine six intervention components with a 32-condition experiment, and Strecher et al. [5] used a fractional factorial design that enabled them to examine five intervention components in a 16-condition experiment. In both cases the number of experimental conditions required was cut in half, but, on the other hand, some assumptions were required that are not necessary in complete factorial designs. (For more details about the considerations associated with complete and fractional factorial

experimental designs and about selecting an appropriate fractional factorial, see Collins et al. [8] for a brief introduction or Wu and Hamada [9] for a more comprehensive treatment.)

The present article

Factorial experiments are emerging as a viable platform for component screening experiments. Although they offer advantages such as great efficiency, they also present new challenges. Data gathered through factorial experiments provide a wealth of information about the individual main effects of intervention components and interactions between components – so much that sorting through it all to decide which components to include can be daunting. In particular, because factorial experiments larger than a 2×2, like the ones in Collins et al. [1], Strecher et al. [5], and Caldwell et al. [6], are still relatively rarely used in the behavioral sciences, investigators may have little experience in considering interactions between components in decision making. Unfortunately, to the best of our knowledge there are no guidelines for how to make decisions based on the results of factorial experiments in the context of intervention science.

This article is aimed at behavioral scientists who have conducted or plan to conduct a factorial screening experiment for the purpose of intervention optimization using MOST or a similar phased experimental framework. The objective of this article is to outline an engineering-inspired procedure for making decisions about which components and component settings should make up the optimized intervention, based on estimates of main and interaction effects obtained in a classical factorial analysis of variance (ANOVA). We demonstrate the decision making approach by reviewing an artificial data example, generated to mimic the results from a study like the ones reviewed above. We conclude by discussing additional considerations that arise when evaluating components for inclusion in a behavioral intervention.

EFFECTS IN ANOVA AND DECISION MAKING

Hypothetical example

Throughout this article we will base a hypothetical example on the first five intervention components from Collins et al. [1], namely, nicotine patch, nicotine gum, precessation in-person counseling, cessation in-person counseling, and telephone counseling. In this example, the set of five components has been examined using a factorial experiment, with each independent variable, or factor, in the experiment corresponding to one of the components. We will refer to the five factors as *PATCH*, *GUM*, *PRECOUN*, *CESSCOUN*, and *PHONE*. We will reserve the term "setting" to refer to a particular dose or presentation of a component and use the term "level" to refer to a value that can be taken on by a factor.

In the decision making approach outlined here we assume that each factor has two levels, corresponding to two component settings (e.g., "do not include/include," "intense/minimal"). In other words, using standard notation this discussion is limited to 2^k factorial experiments, where k represents the number of factors. Although experiments involving factors with more than two levels may be useful at times, 2^k factorial experiments tend to be the most efficient in terms of use of subjects [3], and therefore are frequently used in screening experiments. (2^k factorial screening experiments may be useful even when one or more intervention components have more than two settings. We return to this point in the discussion.) We denote one level of each factor by "⊕" and the other by "⊖," where the ⊕ level always corresponds to the component setting that is hypothesized to perform better, as shown in Table 1.

In this hypothetical study, two different varieties of intervention components can be distinguished, each of which requires a somewhat different kind of decision. For the first three components, the two settings under consideration are "include" and "do not include." In contrast, for the last two components the settings are a lower-intensity and a higher-intensity version, so whichever setting is selected, a version of the component will be included in the intervention package. For simplicity, in this article we will often refer to selection and screening of components, but in every case we mean selection and screening of component settings in general.

Definition of main effects and interaction effects

Table 2 shows the design of a 2^5 factorial experiment manipulating the intervention components in Table 1. In other words, each intervention component corresponds to a factor in the design. Each row represents one of the 32 (i.e., 2×2×2×2×2) experimental conditions. For example, Condition 8 has *PATCH* and *GUM* at the "do not include" level, *PRECOUN* at the "include" level, and *CESSCOUN* and *PHONE* at the "intensive" levels. Based on the data from an experiment like this one it is possible to estimate five main effects, 10 two-factor interactions, 10 three-factor interactions, 5 four-factor interactions, and 1 five-factor interaction.

It is outside the scope of this article to provide an extensive tutorial on analysis of data from a factorial

Table 1 | Components and component settings

Component	Setting designated	
	⊖	⊕
PATCH	Do not include	Include
GUM	Do not include	Include
PRECOUN	Do not include	Include
CESSCOUN	Minimal	Intensive
PHONE	Minimal	Intensive

Table 2 | 2^5 Factorial design for hypothetical screening experiment

Experimental condition	Component				
	PATCH	*GUM*	*PRECOUN*	*CESSCOUN*	*PHONE*
1	⊖	⊖	⊖	⊖	⊖
2	⊖	⊖	⊖	⊖	⊕
3	⊖	⊖	⊖	⊕	⊖
4	⊖	⊖	⊖	⊕	⊕
5	⊖	⊖	⊕	⊖	⊖
6	⊖	⊖	⊕	⊖	⊕
7	⊖	⊖	⊕	⊕	⊖
8	⊖	⊖	⊕	⊕	⊕
9	⊖	⊕	⊖	⊖	⊖
10	⊖	⊕	⊖	⊖	⊕
11	⊖	⊕	⊖	⊕	⊖
12	⊖	⊕	⊖	⊕	⊕
13	⊖	⊕	⊕	⊖	⊖
14	⊖	⊕	⊕	⊖	⊕
15	⊖	⊕	⊕	⊕	⊖
16	⊖	⊕	⊕	⊕	⊕
17	⊕	⊖	⊖	⊖	⊖
18	⊕	⊖	⊖	⊖	⊕
19	⊕	⊖	⊖	⊕	⊖
20	⊕	⊖	⊖	⊕	⊕
21	⊕	⊖	⊕	⊖	⊖
22	⊕	⊖	⊕	⊖	⊕
23	⊕	⊖	⊕	⊕	⊖
24	⊕	⊖	⊕	⊕	⊕
25	⊕	⊕	⊖	⊖	⊖
26	⊕	⊕	⊖	⊖	⊕
27	⊕	⊕	⊖	⊕	⊖
28	⊕	⊕	⊖	⊕	⊕
29	⊕	⊕	⊕	⊖	⊖
30	⊕	⊕	⊕	⊖	⊕
31	⊕	⊕	⊕	⊕	⊖
32	⊕	⊕	⊕	⊕	⊕

experiment. Here we note that the conceptual underpinnings of a factorial experiment are quite different from those of the RCT. The design in Table 2 should not be viewed as a 32-arm RCT. The purpose of the experiment is not to compare individual experimental conditions to each other as separate study arms, as would be done in an RCT; rather, individual experimental conditions are combined in different ways to produce estimates of main effects and interactions. For more information the reader is referred to Refs. [8, 10], and the classic text penned by Kirk [11].

Consider a set of five intervention components to be examined in a factorial experiment. The five factors are labeled *A* through *E*, and each can be at either the ⊕ or ⊖ level. The main effect of factor A is defined as the difference between the ⊕ level and the ⊖ level of that factor, averaged across all the levels of factors *B* through *E*. In our example, the main effect of *PATCH* is the difference between the ⊕ ("include") and ⊖ ("do not include") levels, averaged across all the levels of the remaining four factors. (This would be computed by subtracting the mean response in conditions 1–16 in Table 2 from the mean response in conditions 17–32.)

A two-way interaction involving factors *A* and *B* occurs if the effect of *A* at the ⊕ level of *B* is different from the effect of *A* at the ⊖ level of *B*, averaged across all the levels of *C*, *D*, and *E*. For example, *PATCH* and *PRECOUN* interact if the effect of *PATCH* when *PRECOUN* is ⊖ is different from its effect when *PRECOUN* is ⊕. Similarly, a three-way interaction involving Factors *A*, *B*, and *C* occurs if the two-way interaction between *A* and *B* at the ⊕ level of *C* is different from this two-way interaction at the ⊖ level of *C*, averaged across all the levels of *D* and *E*. In our example, there would be a three-way interaction involving *PATCH*, *PRECOUN*, and *PHONE* if the two-way interaction between *PATCH* and *PRECOUN* was different depending on whether *PHONE* was at the ⊕ or ⊖ level, averaged across the levels of the remaining two factors. Interactions involving more factors are defined similarly.

Coding of effects

The above definitions of the main effect and the interaction are the classical definitions that appear in most statistics textbooks (e.g., [11]). Regression

coefficient estimates consistent with these definitions are produced by conducting a factorial ANOVA using effect coding (−1,1 for a factor with two levels) rather than dummy coding (0,1). This is the approach that is used in engineering and related fields in which decisions are based on the results of factorial screening experiments. Effect coding facilitates decision making because it produces effects that are uncorrelated when there are equal *n*s in each experimental condition and nearly uncorrelated otherwise. By contrast, when dummy coding is used, effects are often correlated even with equal *n*s, complicating interpretation. It should also be noted that when dummy coding is used in the ANOVA, the regression coefficients generally do not correspond to the classical definitions. For example, the "main effect" of a factor as estimated using dummy coding is interpreted as the effect of that factor *with all of the remaining factors set to the zero level*, whereas a main effect based on effect coding is interpreted as the effect *averaged across all remaining factors*. For these reasons we strongly recommend effect coding data from factorial screening experiments, and have used this approach in this article.

A DECISION MAKING APPROACH ROOTED IN ENGINEERING

According to the classical (i.e., effect-coded) definition, when two or more factors do not interact, their combined effect is purely additive. This means that their combined effect is equal to the sum of their respective main effects. In a purely additive model, selection of intervention components is straightforward: the decision can be based simply on the main effects. If a factor's main effect is sufficiently large and in the desired direction, the ⊕ setting of the corresponding component is selected; otherwise, the ⊖ setting is selected. The presence of interactions complicates this decision. When two or more factors interact, their combined effect is either greater than or less than the sum of their respective main effects; in other words, the performance of certain components may be different depending on which other components or component settings occur with them. When sufficiently large interactions are present, it is necessary to take these interactions into account in decision making. (We define "sufficiently large" below.)

The decision making logic we suggest is summarized as follows (see Table 3; hypothetical example is given below).

First, examine main effects to determine whether there is evidence that a factor has an effect overall, averaged across the other factors. If a factor has a sufficiently large main effect in the desired direction, tentatively select the corresponding component's ⊕ setting for inclusion in the intervention; if the factor has no main effect or an effect in the wrong direction, tentatively select the ⊖ setting.

Second, systematically reconsider these tentative decisions in the light of interactions. An interaction may suggest that when combined with the ⊕ setting of another component, the effect of a particular component may be diminished or enhanced. If the effect is diminished, consideration may be given to changing from the ⊕ setting to the ⊖ setting of the component. If the effect is enhanced, consideration may be given to changing from the ⊖ setting to the ⊕ setting of a component.

This decision making logic is adapted from general principles developed in the field of engineering, which has used factorial screening experiments for decades. These principles are reviewed in [9]. The *hierarchical ordering principle* specifies that decision making is based primarily on simpler effects, with more complex effects brought in as needed. In factorial experiments, this means that decisions are based primarily on main effects, with interactions brought in as needed, starting with the lowest-order interactions. The *heredity principle* states that interactions are of interest for decision making purposes only if all factors involved in the interaction have sufficiently large main effects.

As will be demonstrated below, the primary difference between our approach and the approach used in engineering is a modification of the heredity principle, so that an interaction has a role in the decision making process if *at least one* of the factors involved has a sufficiently large main effect. Both the engineering-based approach and our modification are intended to apply to situations in which theory does not provide all the necessary guidance, and could be superseded by a priori theory-based considerations.

Establishing what effects are sufficiently large

We recommend determining a priori thresholds that operationally define which main effects and interactions are considered sufficiently large. In this article, we will simply consider any main effect or interaction that is statistically significant at $p \leq 0.05$ to be sufficiently large. However, hypothesis testing may not be strictly necessary, particularly if the results are to be used to build an intervention that will subsequently be evaluated by means of an RCT. A threshold effect size could be used instead. A meaningful raw difference on the key outcome variable could even be used; although standardization may be helpful heuristically, it is not strictly necessary because the same standard deviation, derived from the ANOVA mean squared error, would always be in the denominator. Approaches that involve selecting components based on relative magnitudes of effects may be used if an intervention must be identified for clinical purposes whether or not components exceed a criterion (e.g., statistical significance). For instance, it may be decided a priori that the three components with the largest effect sizes will be selected. Naturally, any approach taken for arriving at an operational definition of "sufficiently large" must be justifiable on both scientific and practical grounds.

Table 3 | Template for deciding whether to select ⊕ or ⊖ setting of component B when main effect of component A is sufficiently large

	Type of $A \times B$ interaction		
	None	Synergistic	Antagonistic
Component B main effect exceeds threshold	Select $B\oplus$	Select $B\oplus$	Select based on examination of incremental effect of B at $A\oplus$
Component B main effect does not exceed threshold	Select $B\ominus$	Select based on examination of incremental effect of B at $A\oplus$	Select $B\ominus$

ARTIFICIAL DATA FOR HYPOTHETICAL EXAMPLE

How the artificial data were generated

We generated artificial data for the 2^5 factorial experiment manipulating *PATCH*, *GUM*, *PRECOUN*, *CESSCOUN*, and *PHONE* shown in Table 2. The hypothetical outcome variable is a scale measuring the subjects' beliefs about their self-efficacy for quitting smoking, so that a higher score indicates more self-efficacy. For simplicity, we will evaluate all five components based on a single outcome variable, but it is possible to evaluate different components based on different outcomes. (For example, it would be possible to evaluate the three nicotine replacement components using a measure of cigarette craving rather than self-efficacy.) We generated the data so that the raw self-efficacy score of participant j in the ith experimental condition was modeled as $\mu_i + \epsilon_{ij}$, where the μ_i are given by

$$\begin{aligned}\mu = {} & 5.00 + 1.25 * PATCH + 1.00 * \mathrm{GUM} + 0.90 \\ & * CESSCOUN - 0.40 * PATCH \times GUM + 0.50 \\ & * CESSCOUN \times PHONE + 0.75 * PATCH \\ & \times CESSCOUN - 0.75 * PATCH \times CESSCOUN \\ & \times PRECOUN,\end{aligned}$$

and the errors are independent and $N(0,4^2)$. Using this model, we generated data for a total of 512 subjects, with 16 subjects randomly assigned to each of the 32 experimental conditions.

Main effects

A standard ANOVA was performed on the artificial data, with "do not include" or "minimal" coded −1 and "include" or "intensive" coded 1 (see Table 1). The results are shown in Table 4. There are significant main effects of *PATCH*, *GUM*, and *CESSCOUN*. Thus, the tentative decisions are include the nicotine patch and nicotine gum, include intensive cessation counseling, do not include precessation counseling, and include phone counseling at the minimal setting.

Two-way interactions

Next, these decisions are systematically reconsidered in the light of any sufficiently large two-way interactions. There are three significant two-way interactions: *CESSCOUN* × *PATCH*, *CESSCOUN* × *PHONE*, and *PATCH* × *GUM*. Let us consider them in turn, and see whether any of them suggest reconsidering the tentative decisions made on the basis of the main effects. We recommend examining plots of all sufficiently large interactions.

CESSCOUN × *PATCH*–The sign of the *b*-weight for the *CESSCOUN* × *PATCH* interaction is positive, indicating that this is a *synergistic* interaction. In synergistic interactions, the effect of two or more factors combined is greater than the sum of the main effects; in other words, the combined effect of the two factors is greater than would be expected based on the main effects alone. This is evident in Fig. 1, which shows a plot of this interaction. Because the main effects of both *CESSCOUN* and *PATCH* exceed the a priori threshold, and the interaction suggests that when combined their effect is even stronger, the tentative decisions to select the intensive setting of cessation counseling and to include the nicotine patch are upheld.

CESSCOUN × *PHONE*–The sign of the *b*-weight for the *CESSCOUN* × *PHONE* interaction is positive, again indicating a synergistic interaction. This situation differs from the one above, because the main effect of *CESSCOUN* exceeds the a priori threshold but that of *PHONE* does not. Thus, as mentioned above, the tentative decisions are to include intensive cessation counseling and minimal phone counseling. However, as Fig. 2 shows, it appears that when *CESSCOUN* is at the intensive level there is a positive difference between the minimal and intensive levels of *PHONE*. Does this mean that the decision about *PHONE* should be reconsidered and the intensive setting should be selected?

To make this decision, it is helpful to examine the incremental effect of *PHONE* when *CESSCOUN* is at the intensive level. This incremental effect is indicated by a bracket in Fig. 2. Examination of means shows that this represents a raw difference of approximately 1.22 on the self-efficacy scale. To put this in perspective, the main effect of *GUM*, the smallest of the significant main effects, represents a difference of approximately 1.81 raw units. It is a good idea to set an a priori threshold for incremental effects, in either raw or standardized units. Let us assume that this difference of 1.22 is sufficiently large to justify changing the previous tentative

Table 4 | ANOVA effect estimates and hypothesis tests

Effect	*b*-weight	*t*	*p*	Cohen's *d*
Intercept	4.82	26.69	<.01	
PATCH	**1.06**[a]	**5.88**	**<.01**	**0.52**
GUM	**0.90**	**4.98**	**<.01**	**0.44**
PRECOUN	0.06	0.34	0.73	0.03
CESSCOUN	**1.15**	**6.35**	**<.01**	**0.56**
PHONE	0.01	0.03	0.98	0.00
PATCH × GUM	**−0.36**	**−2.00**	**0.05**	**−0.18**
PATCH × PRECOUN	−0.04	−0.19	0.85	−0.02
PATCH × CESSCOUN	**0.79**	**4.39**	**<.01**	**0.39**
PATCH × PHONE	0.12	0.68	0.50	0.06
GUM × PRECOUN	−0.24	−1.34	0.18	−0.12
GUM × CESSCOUN	0.09	0.48	0.63	0.04
GUM × PHONE	0.12	0.64	0.52	0.06
PRECOUN × CESSCOUN	−0.02	−0.09	0.92	−0.01
PRECOUN × PHONE	0.08	0.46	0.65	0.04
CESSCOUN × PHONE	**0.61**	**3.35**	**<0.01**	**0.30**
PATCH × GUM × PRECOUN	−0.08	−0.45	0.65	−0.04
PATCH × GUM × CESSCOUN	−0.07	−0.37	0.71	−0.03
PATCH × GUM × PHONE	−0.13	−0.74	0.46	−0.07
PATCH × PRECOUN × CESSCOUN	**−0.55**	**−3.06**	**<.01**	**−0.27**
PATCH × PRECOUN × PHONE	−0.09	−0.52	0.61	−0.05
PATCH × CESSCOUN × PHONE	0.24	1.31	0.19	0.12
GUM × PRECOUN × CESSCOUN	0.23	1.29	0.20	0.11
GUM × PRECOUN × PHONE	−0.05	−0.30	0.77	−0.03
GUM × CESSCOUN × PHONE	0.06	0.36	0.72	0.03
PRECOUN × CESSCOUN × PHONE	−0.12	−0.59	0.55	−0.05
PATCH × GUM × PRECOUN × CESSCOUN	0.04	0.22	0.83	0.02
PATCH × GUM × PRECOUN × PHONE	−0.16	−0.86	0.39	−0.08
PATCH × GUM × CESSCOUN × PHONE	0.08	0.44	0.66	0.04
PATCH × PRECOUN × CESSCOUN × PHONE	0.19	1.08	0.28	0.10
GUM × PRECOUN × CESSCOUN × PHONE	0.19	1.05	0.29	0.09
PATCH × GUM × PRECOUN × CESSCOUN × PHONE	−0.07	−0.40	0.69	−0.04

[a] Bolded entries correspond to effects significant at $p \leq 0.05$

decision, and revise the decision so that phone counseling will be included at the intensive setting.

PATCH × GUM–Both *PATCH* and *GUM* have positive main effects, and so both have been tentatively selected for inclusion in the intervention. The sign of the *b*-weight for the *PATCH × GUM* interaction is negative, indicating an *antagonistic* interaction. In antagonistic interactions, the effect of two or more factors combined is less than the sum of the main effects of the factors; in other words, the combined effect of the two factors is smaller than would be expected based on the main effects alone. Does this mean we should remove either the nicotine patch or nicotine gum from the intervention package?

The presence of an antagonistic interaction is not automatically a reason to reject components when the corresponding factors have demonstrated sufficiently large main effects, because even when there is an antagonistic interaction the combined effect of the factors may nevertheless exceed the effect of either factor alone. Examination of Fig. 3 shows that when *GUM* is at the "include" level, *PATCH* has an incremental positive effect; for *PATCH* = "include" mean self-efficacy = 6.9, as opposed to 5.2 for *PATCH* = "do not include," a difference of 1.7. The interaction is antagonistic because the incremental effect of *PATCH* is larger when *GUM* is at the "do not include" level (for *PATCH* = "include" mean self-efficacy=5.7, as opposed to 2.4 for *PATCH* = "do not include" – a difference of 3.3). In this case, the antagonistic interaction does not suggest that the initial decision should be revised, because *PATCH* demonstrates an incremental positive effect when *GUM* = "include," and the mean self-efficacy score of 6.9 for the conditions in which both *PATCH* and *GUM* are included is the largest of the four possibilities.

Three-way and higher-order interactions

It is not always straightforward to categorize interactions that involve more than two factors as synergistic or antagonistic, so it is particularly important to examine plots. In the artificial data example there is one significant three-way interaction, *PATCH × PRECOUN × CESSCOUN*.

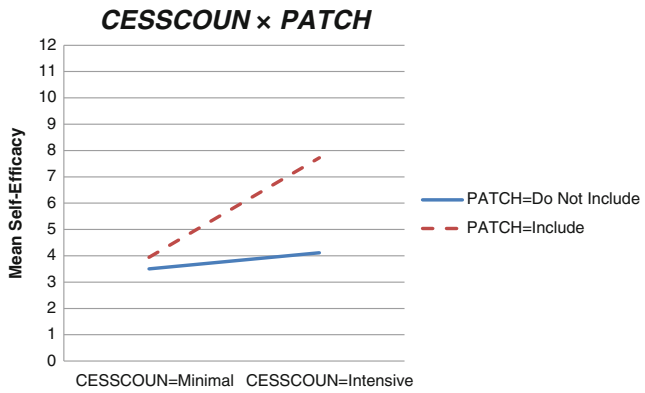

Fig 1 | Plot of the synergistic interaction between *CESSCOUN* and *PATCH*

At this point in the decision making process, we plan to include the nicotine patch in the intervention, not to include precessation counseling, and to include intensive cessation counseling. The purpose of examining this three-way interaction is to see whether it suggests that any of these decisions should be reconsidered. Figure 4 shows that when *PATCH* is at the "include" level and *CESSCOUN* is at the "intensive" level, the mean outcome is highest when *PRECOUN*="do not include." Thus, the three-way interaction suggests that the previous decisions should stand.

The final selection

Based on the above decision making process, the final selection is include the nicotine patch and nicotine gum, do not include precessation counseling, include intensive cessation counseling, and include intensive phone counseling.

The goal of the decision making process is to identify one of the best combinations of components and component settings. Because this is artificial data, we know the true mean on the outcome variable associated with each combination, and can determine whether the decision making process was successful at identifying one of the combinations with the highest mean. In this case, the decision making process was successful; in fact, it identified the best combination (means available upon request). One decision that could be considered a judgment call was whether to include phone counseling at the intensive setting. If we had decided to select the minimal setting of phone counseling, perhaps based on concerns related to the resources needed to deliver that component of the intervention, the resulting combination would have produced the second highest mean self-efficacy score.

DISCUSSION

The factorial screening experiment is an important tool for optimizing behavioral interventions within the MOST framework. Once the data from a screening experiment have been analyzed via ANOVA, the results can form the basis for making decisions about which components and component settings will form the optimized behavioral intervention. This article has outlined and illustrated a procedure for selecting components and component settings based on the results of a factorial ANOVA.

Using main effects and interactions in decision making

In this article we recommend beginning decision making based primarily on main effects, and then revisiting the decisions in the light of observed

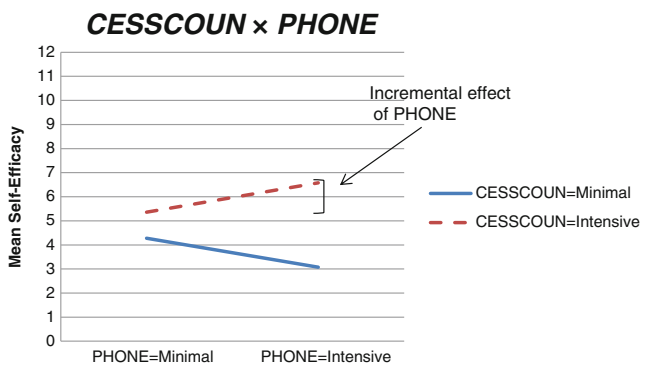

Fig 2 | Plot of the synergistic interaction between *CESSCOUN* and *PHONE*

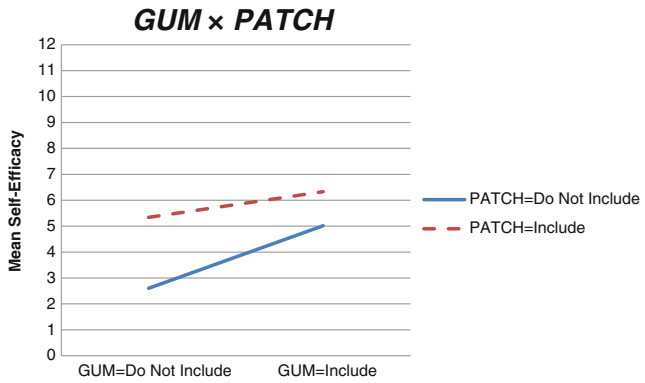

Fig 3 | Plot of the antagonistic interaction between *GUM* and *PATCH*

interactions between factors. Many behavioral scientists have been taught to begin with interactions, and to examine main effects only if there are no large interactions. These perspectives appear to be polar opposites. One reason for this difference in perspectives may be related to the use of dummy coding as opposed to effect coding.

As mentioned above, in a factorial ANOVA including interaction terms, dummy coding produces estimates of different effects than are produced by effect coding. The only exception is the hypothesis test of the highest-order interaction, which is the same when dummy coding and effect coding are used. Kugler et al. [12] discussed how to express the estimates produced by dummy coding in terms of the classical effects produced by effect coding. They showed that the "main effect" (in quotes because it is not a main effect by the classical definition) estimate for a component produced by dummy coding is a weighted combination of that component's classical main effect and all of the classical interactions involving that component. As the number of factors increases, the dummy coded "main effect" includes an increasing number of terms corresponding to interactions. Similarly, each dummy coded "interaction" is a weighted combination of the corresponding classical interaction and certain higher-order interactions.

As a result, dummy coded effects are usually correlated, sometimes substantially, even when based on data collected in a balanced factorial experiment. With this entangling of effects it would seem advisable not to interpret the dummy coded "main effects" without first reviewing all higher-order effects. By contrast, when effect coding is used this entangling does not occur: main effects and interactions are uncorrelated with a balanced factorial experiment and nearly uncorrelated even if the *n*s vary somewhat across experimental conditions. Of course it is always wise to consider large interactions when interpreting main effects; the point here is that because the main effects and interactions are more separate when effect coding is used than when dummy coding is used, it is reasonable to consider them separately in decision making. In other words, the hierarchical ordering and heredity principles discussed above, which provide the foundation for the decision making approach suggested here, apply primarily when effect coding is used.

Because different effects are estimated when dummy coding and effect coding are used, they are interpreted differently. The "main effect" of a factor as estimated using dummy coding, which as mentioned above is interpreted as the effect of that factor *with all of the remaining factors set to the zero level*, is

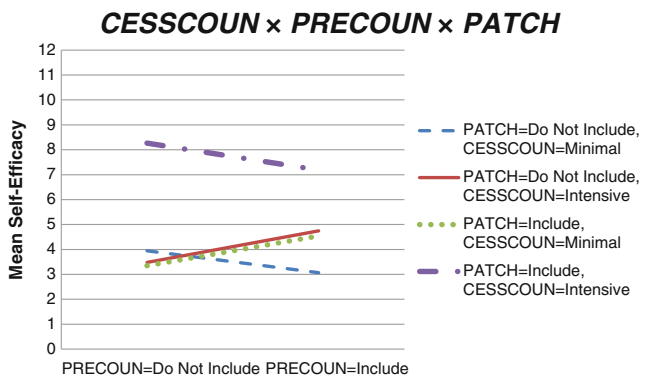

Fig 4 | Graph of three-way interaction involving *CESSOUN*, *PRECOUN*, and *PATCH*

essentially a simple effect [11]. Unless only one component is to be selected, this quantity is of limited utility for component selection until interactions are carefully considered. By contrast, as discussed above, the effect coded main effect is *averaged across all remaining factors*. Thus, as Fisher [13] noted, in factorial experiments a large main effect suggests that a component is in a sense robust, because its effect persists on average across many other combinations of components.

Interactions when there is no main effect

The decision making procedure proposed here involves examining interactions only if at least one of the factors in the interaction demonstrates a sufficiently large main effect. Suppose two factors do not demonstrate sufficiently large main effects but do have a synergistic interaction, such that when combined they demonstrate a large positive effect. This would suggest that neither of the corresponding components is effective without the other. Using the decision making framework suggested here, these components would not be eligible for selection based on the interaction alone (each could, however, be eligible based on interactions with components that do have sufficiently large main effects, as was demonstrated above). Our reasoning is that the purpose of the screening experiment is to build an intervention made up primarily of robust intervention components that can stand alone and are likely to maintain effectiveness even if implemented with a different set of other components in the future. In our view, this is particularly important when either component is costly to implement. However, if it is necessary to identify a "best available" intervention based on the results of a screening experiment, for instance if no effective intervention exists and there is a pressing clinical need for one, then it might be advisable to relax the guidelines offered above.

In the current example, we ultimately decided to include intensive phone counseling even though *PHONE* did not have a sufficiently large main effect, because the experimental evidence suggested that intensive phone counseling was effective when paired with intensive in-person cessation counseling. In a case like this, we would likely consider combining these into one component in the future, to ensure that intensive phone counseling would never be implemented without intensive cessation counseling. A similar strategy could be considered for two or more components that show sufficiently large effects only when combined.

Interactions between intervention components and measured variables

The discussion in this article has emphasized interactions involving two or more factors that were manipulated in the screening experiment. Investigators may also be interested in interactions between factors and other variables that have been measured rather than manipulated. Examples of such variables are gender, depression diagnosis, and previous quit attempts. An interaction between, for example, depression and *CESSCOUN* would mean that the effect of *CESSCOUN* is different for those with a diagnosis of depression than those without a diagnosis of depression.

In a factorial screening experiment, interactions between factors and measured variables can be coded and included in the ANOVA. However, the role of these interactions in decision making should be considered carefully before any analyses are undertaken. A screening experiment is carefully planned to provide enough information so that decisions can be made reasonably confidently about a limited number of intervention components. We strongly recommend that if an interaction with a measured variable is considered critically important in decision making, this should be treated as an a priori hypothesis, so that resources can be devoted to ensuring that the interaction can be examined with sufficient statistical power.

It may be desirable to conduct exploratory analyses to investigate a host of measured variables for possible interactions with experimental factors. These analyses may be valuable for informing future research, for example, development of a future adaptive intervention [14]. However, post-hoc exploratory analyses are in general not a firm basis for decision making about selection of intervention components and settings.

Why observed mean outcomes for each experimental condition cannot play a major role in decision making

At first glance, it may appear that there is little need for a decision making framework like the one suggested here. A factorial experiment produces an observed mean on key outcome variables for each implemented experimental condition. Can these means simply be rank ordered, and the condition with the highest mean selected as the best combination of components?

We see at least three difficulties with this suggestion. First, factorial experiments, unlike RCTs, are not powered for direct comparison of the means of individual experimental conditions (i.e., comparison of the observed means associated with any of the combinations of factor levels shown in Table 2). In fact, even well-powered factorial experiments with multiple factors may have relatively few subjects in each experimental condition. Thus, the estimates of main effects and interactions, which are based on all of the subjects in the experiment rather than a small fraction, are much more stable than estimates of the means of individual experimental conditions, and therefore provide a better basis for decision making. For example, consider the artificial data that have formed the basis for the discussion in this article. These data were generated so that the outcome variable, self-efficacy, had a standard deviation of

σ=4. Let us compare the standard errors associated with the mean for *PATCH*="include" and the mean of Experimental Condition 32 in Table 2, in which each factor is set to the ⊕ level. As was discussed above, the mean for a single level of a factor is computed using data from all of the subjects who are assigned that level. Because each level of a factor is assigned to half of the subjects, the mean for *PATCH*="include" is estimated based on n=256. Thus the standard error of the sample mean self-efficacy for *PATCH*="include" is given by $\sigma/\sqrt{n} = 4/16 = 0.25$. By contrast, the mean of each experimental condition is based on n=16, so the standard error of the sample mean self-efficacy for Experimental Condition 32 is 4/4=1, which is four times larger than the standard error of the sample mean for *PATCH*="include." This reasoning applies to each level of every factor and to all of the experimental conditions. Moreover, a sample size of n=16 per experimental condition may not be large enough to justify the assumption of normality of the sample mean under the central limit theorem, making statistical inference pertaining to individual experimental conditions difficult. However, this per-condition sample size is large enough to justify the assumption of normality for inferences concerning the mean of a level of a factor, because these inferences are based on aggregate combinations of individual experimental conditions.

A second difficulty with the suggestion of making decisions based on the means of the individual experimental conditions is that if a fractional factorial experiment is used, at least half of the possible combinations of components will not be implemented and therefore will not have an observed mean. A third difficulty is that in many decision making situations it is not only the outcome on a key variable that is a consideration. If another important consideration must be factored in, such as the differential cost of components, then a simple rank ordering on a single dimension does not provide sufficient information for decision making.

When components have more than two settings

In our example, "minimal" and "intensive" represent the low and high ends of cessation counseling and phone counseling. Between these two ends there are many possible settings representing intermediate dosages of counseling. However, we did not examine any of these possible intermediate settings in the hypothetical screening experiment, because this would have required considerably more resources. Screening experiments that include factors with more than two levels generally require at least half again as many subjects as a comparable 2^k factorial experiment [9], as well as many additional experimental conditions. For this reason, we encourage investigators to consider alternatives carefully before conducting a screening experiment with more than two levels in any factor. An approach that conserves resources is to begin by using a 2^k screening experiment to establish whether there is a difference between the low and high ends of the range of settings. If there is a difference, follow-up experimentation can be conducted to compare some intermediate settings. If there is no difference between the low and high settings, it is probably not worthwhile to expend resources on follow-up experimentation.

Decision making and hypothesis testing

In this article we used statistical significance at the $p \leq 0.05$ level as the criterion for establishing that an effect was sufficiently large. We chose this approach primarily to make the exposition more straightforward. In practice, our view is that the conventional $p \leq 0.05$ level of significance pertains mainly to purely scientific inquiry and is less relevant to selecting intervention components and settings. We propose that the information gathered in a screening experiment can be used in any rational, principled manner.

If the investigators feel most comfortable within a hypothesis testing framework, they might consider the relative cost of mistakenly selecting an ineffective component (Type I error) as compared to the cost of overlooking an effective component (Type II error). For example, if in a given situation the cost of a Type II error is greater than a Type I error, it may be practical to operate using a Type II error rate of, say, β=0.10, even if to accomplish this without exceeding available resources the Type I error rate must be raised to, say, α=0.15. The decision of what Type I and Type II error rates to use must be made a priori, and definitely should not be made after any results have been examined, because of the danger of capitalizing on chance findings.

Some investigators may feel most comfortable focusing on effect size estimates rather than formal hypothesis testing, particularly if it is clear what effect size corresponds to clinical significance. Confidence intervals about the effect size estimate can be computed directly [15] or obtained using statistical software. The cut-off for clinical significance must be made a priori.

Formal hypothesis testing and examination of confidence intervals about effect sizes both involve quantification of the uncertainty associated with decision making. This is particularly helpful with relatively small sample sizes. However, investigators may opt to take a less formal approach. For example, it was mentioned above that in some circumstances the investigators may decide in advance that they will select, say, the three components that show the largest effect sizes, irrespective of uncertainty in the effect size estimates. It should be noted that even without hypothesis testing, the selection of components and component settings based on the results of a carefully conducted randomized screening experiment is still more empirically rigorous than the typical selection

procedure used today. If the MOST framework is used, the intervention that is eventually developed will be evaluated using an RCT in a later phase of the process. This will formally test the effectiveness of the multicomponent intervention package, and formal hypothesis testing with a Type I error rate of $\alpha=0.05$ can be adhered to strictly in this phase.

A practical suggestion

Investigators considering or already conducting a factorial screening experiment may benefit from practicing decision making before they have to make decisions based on their own data. We have found that practice sessions on artificial data sets can be tremendously helpful. On The Methodology Center's web site (http://methodology.psu.edu), we have placed several artificial data sets generated using models that are plausible in intervention science. Readers are invited to download these data sets, treat them as if they were their own empirical results, and apply the decision making framework reviewed here or a different one. The true data generation model and cell means are provided on the web site for each data set, but we suggest refraining from looking at them until after selection of intervention components and settings, to avoid biasing the decision making process.

Limitations and future directions

The decision making approach described here has several important limitations. First, although one of the purposes of MOST is to enable intervention scientists to consider resource constraints when optimizing an intervention, we have not discussed the specifics of how to incorporate cost into decision making. We are currently developing models for incorporating cost into such decision making. However, it should be noted that researchers can use less formal approaches, such as specifying a maximum cost for an intervention and identifying the most effective set of components that costs less than this maximum. Second, the decision making approach described here does not extend to situations in which a component is to be evaluated based on more than one outcome variable. For example, Collins et al. [1] listed several outcome variables of interest for evaluating components of their smoking cessation intervention, including ability to establish initial cessation; number of days abstinent in the 2-week post-quit period; post-quit self-efficacy; withdrawal/craving; and latency to (a) first cigarette and (b) seven consecutive days of smoking after the target quit day. Results may be inconsistent across these outcome variables, complicating the decision making process and, in some cases, calling for difficult trade-offs. Third, our suggested decision making framework does not consider multiple decision makers. A group of decision makers may have difficulty reaching consensus on what constitutes a sufficiently large effect, how to manage multiple outcome variables, and so forth. Fourth, the proposed framework to date has been considered within the normal model only. It is less clear how to make decisions based on other models, such as hazard or Poisson. Fifth, it is not clear how to incorporate outcomes from multiple levels of analysis, for example, outcomes pertaining to patients (e.g., number of cigarettes smoked) and outcomes pertaining to the clinics in which they are seen (e.g., staff ratings of ease of implementation of an intervention component). Multiple outcome variables, multiple decision makers, non-normal models and multi-level data are all common in intervention science. Research is critically needed in all of these areas to enable intervention scientists to make increasingly more well-informed decisions about selection of intervention components based on factorial experiments. Finally, this general approach to decision making may not suit every situation in which investigators plan to use the results of a screening experiment to optimize a behavioral intervention.

In our view, MOST and similar phased approaches for optimization of behavioral interventions have much potential to facilitate translational behavioral medicine and to increase the efficiency and public health impact of interventions. Decision making based on the results of factorial screening experiments plays an important role in MOST. We hope that although this article falls far short of offering guidance for every scenario that might occur, it provides support for behavioral scientists who wish to use factorial experiments to screen intervention components.

Acknowledgments: This project was supported by Award Number P50CA143188-3 from the National Cancer Institute and by Award Number P50DA010075-15 from the National Institute on Drug Abuse. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute, the National Institute on Drug Abuse, or the National Institutes of Health. This work has benefitted from discussions with John Dziak and other colleagues at The Methodology Center. The authors thank Amanda Applegate for editorial assistance.

1. Collins LM, Baker TB, Mermelstein RJ, et al. The multiphase optimization strategy for engineering effective tobacco use interventions. *Ann Behav Med*. 2011; 41(2): 208-226.
2. Clapham C, Nicholson J. *The Concise Oxford Dictionary of Mathematics*. 4th ed. New York: Oxford University Press; 2009.
3. Collins LM, Chakraborty B, Murphy SA, Strecher V. Comparison of a phased experimental approach and a single randomized clinical trial for developing multicomponent behavioral interventions. *Clin Trials*. 2009; 6(1): 5-15.
4. Baker TB, Mermelstein R, Collins LM, et al. New methods for tobacco dependence treatment research. *Ann Behav Med*. 2011; 41(2): 192-207.
5. Strecher VJ, McClure JB, Alexander GL, et al. Web-based smoking-cessation programs: results of a randomized trial. *Am J Prev Med*. 2008; 34(5): 373-381.
6. Caldwell LL, Smith EA, Collins LM, et al. Translational research in South Africa: evaluating implementation quality using a factorial design. *Child Youth Care For*. 2012; 41(2): 119-136.
7. Smith EA, Palen L, Caldwell LL, et al. Substance use and sexual risk prevention in Cape Town, South Africa: an evaluation of the HealthWise program. *Prev Sci*. 2008; 9: 311-321.

8. Collins LM, Dziak JJ, Li R. Design of experiments with multiple independent variables: a resource management perspective on complete and reduced factorial designs. *Psychol Methods*. 2009; 14(3): 202-224.
9. Wu CFJ, Hamada M. *Experiments: Planning, Analysis, and Parameter Design Optimization*. New York: Wiley; 2011.
10. Chakraborty B, Collins LM, Strecher VJ, Murphy SA. Developing multicomponent interventions using fractional factorial designs. *Stat Med*. 2009; 28(21): 2687-2708.
11. Kirk RE. *Experimental Design: Procedures for the Behavioral Sciences*. 4th ed. Los Angeles: Sage; 2013.
12. Kugler KC, Trail JB, Dziak JJ, Collins LM. *Effect Coding Versus Dummy Coding in Analysis of Data from Factorial Experiments*. [Technical Report No. 12–120]. University Park: The Methodology Center, Penn State: The Methodology Center, Penn State; 2012.
13. Fisher JO. *The Design of Experiments*. New York: Hafner; 1971.
14. Collins LM, Murphy SA, Bierman KL. A conceptual framework for adaptive preventive interventions. *Prev Sci*. 2004; 5(3): 185-196.
15. Kirk RE. Effect magnitude: a different focus. *J Stat Plan Infer*. 2007; 137: 1634-1646.