STATISTICS IN MEDICINE
Statist. Med. 2009; **28**:2687–2708
Published online 3 July 2009 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3643

Developing multicomponent interventions using fractional factorial designs

Bibhas Chakraborty[1,*,†], Linda M. Collins[2], Victor J. Strecher[3] and Susan A. Murphy[1]

[1]*Department of Statistics and The Institute for Social Research, University of Michigan, Ann Arbor, MI, U.S.A.*
[2]*The Methodology Center and Department of Human Development and Family Studies, Pennsylvania State University, University Park, U.S.A.*
[3]*Center for Health Communications Research, University of Michigan, Ann Arbor, MI, U.S.A.*

SUMMARY

Multicomponent interventions composed of behavioral, delivery, or implementation factors in addition to medications are becoming increasingly common in health sciences. A natural experimental approach to developing and refining such multicomponent interventions is to start with a large number of potential components and screen out the least active ones. Factorial designs can be used efficiently in this endeavor. We address common criticisms and misconceptions regarding the use of factorial designs in these screening studies. We also provide an operationalization of screening studies. As an example, we consider the use of a screening study in the development of a multicomponent smoking cessation intervention. Simulation results are provided to support the discussions. Copyright © 2009 John Wiley & Sons, Ltd.

KEY WORDS: multicomponent intervention; experimental design; fractional factorial design; screening; follow-up studies

1. INTRODUCTION

Multicomponent or complex interventions are increasingly being used in many health domains, e.g. AIDS [1], cardiovascular diseases [2], depression [3], diabetes [4], drug abuse [5], gerontology [6], obesity [7], and smoking cessation [8]. While some components may involve a medication, many components are behavioral, implementation, or delivery factors [2, 3, 8]. As has been recognized in the literature [6, 9–12], development and evaluation of these multicomponent interventions pose additional design challenges over those of single-component interventions, and these challenges

*Correspondence to: Bibhas Chakraborty, Department of Statistics, 439 West Hall, 1085 S. University Avenue, Ann Arbor, MI 48109-1107, U.S.A.
†E-mail: bibhas@umich.edu

Contract/grant sponsor: National Institutes of Health; contract/grant numbers: RO1 MH080015, P50 DA10075, P50 CA101451

Received 7 January 2009
Accepted 20 May 2009

tend to be addressed poorly by the standard two-group randomized controlled trial. In particular one important challenge is to whittle down a large list of potential components, by screening out the least active components. Factorial designs are ideally suited to this endeavor [13].

The primary goal of this paper is to consider the use of full and fractional factorial designs (FFDs) in screening out inactive components so as to aid in the development of high-quality multicomponent interventions. We discuss how many of the criticisms prevalent in the literature concerning the use of full and FFDs no longer hold or are of lesser importance in *screening*[‡] trials. A secondary goal is to provide an operationalization of screening trials using full and FFDs.

The present work is motivated by our participation in the design of a web-based smoking cessation study called *Project Quit* [8] that utilized FFDs. For illustrative purposes, we present a slightly modified version of *Project Quit*, following [14]. The investigators decided to study six components: *Depth of outcome expectations*, *Depth of efficacy expectations*, *Depth of success stories*, *Personalization of message source*, *Mode of message framing*, and *Exposure schedule* (depth refers to the degree to which the communication was tailored to the background information on each individual). Since varying all six components across all possible levels in a single study was logistically prohibitive, the investigators decided to move forward in phases, where results of the research conducted in the first phase would inform the second, and so on [14, 15]. The goal of the first phase was to identify the active components and screen out inactive components. Each component was varied at two levels as is common in screening studies. In addition all individuals were provided a 10-week free supply of nicotine patches. The investigators decided to use a 16-cell FFD (see Section 3 for details). The primary outcome was self-reported 7-day point-prevalence abstinence at the 6-month follow-up from the date of randomization. More information on this study can be found in [8, 14].

The remainder of the paper is organized as follows. Section 2 addresses common criticisms against the use of full and FFDs for developing multicomponent interventions. We provide an operationalization of the screening trials in Section 3. Examples of possible follow-up studies are given in Section 4. We conclude with an overall discussion in Section 5. Technical review material on FFDs appears in the appendix.

2. FACTORIAL DESIGNS FOR SCREENING STUDIES

Factorial designs were originally developed in the context of agricultural experiments [16, 17] and are now used in other areas including engineering [13, 18] and marketing research [19]. Their use in the medical and behavioral fields has been limited; however, there have been a number of papers discussing the usefulness of these designs in medication and intervention trials [12, 20–24].

Prior to discussing common criticisms and concerns, we provide a brief review of both the design and analysis of screening studies. In screening, two-level factorial designs, where all components are studied at two levels (these levels can be either present vs absent, or two ethically acceptable doses of the component), are usually used since the goal is to identify important components rather than identify the optimal level of each component. If a two-level factorial design involves k components, then the total number of treatment combinations studied is 2^k. Each of the 2^k cells in the design corresponds to a group of subjects assigned to a particular treatment combination.

[‡]Here the term *screening* refers to screening of intervention components, not screening of study participants.

DOI: 10.1002/sim

In screening experiments, k is often large, rendering a full factorial design with 2^k cells infeasible. In such cases, FFDs [25] offer a nice alternative since they use fewer cells (see below for more discussion). For example, in the *Project Quit* study, a full factorial design with six components would need $2^6=64$ cells. But by using an FFD, it was possible to restrict the study to only 16 cells, and still be able to estimate all the main effects and some two-way interactions under reasonable assumptions.

In case of a continuous outcome, the analysis of a 2^k full factorial design (or a 2^{k-p} FFD) with total sample size n can be done using a linear regression model. One can use a model of the form $\mathbf{Y}=\mathbf{X}\beta+\varepsilon$, where $\mathbf{Y}$ represents the $n\times 1$ vector of observations on the outcome, $\mathbf{X}$ is the $n\times m$ design matrix, β is a $m\times 1$ vector of unknown parameters ($m=2^k$ for full factorial and $m=2^{k-p}$ for FFD, with more parameters if baseline variables are included in the analysis model), and ε is the $n\times 1$ vector of errors. It is assumed that $E(\varepsilon)=0$ and $\mathrm{Cov}(\varepsilon)=\sigma^2\mathbf{I}$. The design matrix consists of an intercept column, plus columns corresponding to each component and their interactions of different order coded in $-1/1$ (i.e. different factorial effects). The least-squares estimator for β along with its covariance matrix is given by

$$\hat{\beta}=(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y},\quad \mathrm{Cov}(\hat{\beta})=\sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$$

Note that the estimates of usual ANOVA quantities of interest like the main effect of a component or the interaction between two or more components are directly related to the least-squares regression estimates $\hat{\beta}$, provided the design matrix is coded in $-1/1$. As discussed in [26], the main effect of a component A_1 is estimated by $2\hat{\beta}_{A_1}$, A_1A_2 interaction is estimated by $4\hat{\beta}_{A_1A_2}$, and so on. In general, a p-component ($1\leqslant p\leqslant k$) interaction, say $A_{i_1}\ldots A_{i_p}$ (with $1\leqslant i_1,\ldots,i_p\leqslant k$), is estimated by $2^p\hat{\beta}_{A_{i_1}\ldots A_{i_p}}$. If variance heterogeneity across different cells is anticipated in a study, one can use a robust estimator, e.g. *sandwich estimator* [27] of the covariance matrix given by

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}(\mathrm{diag}(\mathbf{Y}-\mathbf{X}\hat{\beta}))^2\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$$

in the linear regression model. But sample sizes should not be too small for this estimator to work well. Wu and Hamada [18] provide alternative methods to deal with variance heterogeneity. As discussed by Montgomery *et al.* [28], the regression approach can be used for *unbalanced*[§] data, and can estimate the factorial effects controlling for baseline or stratification variables. In case of binary (more generally, categorical) outcomes, the regression approach can be generalized via a generalized linear model. For example, if the outcome is binary, a logistic regression model can be used to analyze the data from a factorial design [18, Chapter 13]. See [8, 14] for examples of such analyses.

2.1. Criticisms against factorial designs

Within the biostatistics literature, factorial designs were assessed primarily in the context of medication trials; the objective was to *evaluate* the usefulness of a combined medication over a single medication. In contrast, our objective is to screen out inactive components of a multicomponent intervention and thus full and FFDs play a different role from that of evaluation. In this context,

[§] *Balance* means that each level of each component appears in same number of cells and is assigned to the same number of subjects.

DOI: 10.1002/sim

many of the common concerns re cost, feasibility, ethics, toxicity of combination drugs, interpretation of main effects in presence of active interactions, and concern about power for detecting interactions become moot or of lesser importance. Indeed many different complaints against factorial designs stem from a few fundamental issues and hence can be categorized as follows:

1. There are attractive alternatives to FFD.
2. It is not feasible to simultaneously implement multiple multicomponent interventions.
3. Some components cannot be crossed due to toxicity or ethical considerations.
4. The interpretation of main effects when interactions exist is complicated.
5. Power is low or alternately the required sample size is high.

In the following, we address these broad classes of criticisms against factorial designs in the present context of screening studies for developing multicomponent interventions.

2.2. Attractive alternatives to FFD

The traditional approach of empirically developing multicomponent interventions, sometimes called the *treatment package strategy* [12, 23, 29], is to formulate a 'likely best' intervention based on existing literature, theory, and clinical experience. Additionally investigators may use information from limited experimentation with some of the components either in stand-alone trials or in trials in which one component is varied at a time while the remaining components are set at fixed levels. Implicitly, one often assumes that more treatment is always better so the 'likely best' intervention includes many components. An additional implicit assumption is that any ill effects due to including inactive components are minor. The developed multicomponent intervention is then evaluated in a standard two-arm randomized trial. These two-arm trials are confirmatory in that they are designed to provide high-quality information on whether the multicomponent intervention performs better than the standard; they are not designed to provide direct information on which components are active, whether they have been set at optimal levels, and whether there is any interaction between the components [11]. To address the latter questions, investigators may use observational analyses, such as a dose-response with the level of subject adherence to the treatment as the dose [5, 6, 30, 31], or theory-based mediational analysis [12, 32]. The intervention is often refined based on the findings of these analyses, and then the refined version is tested in another two-arm randomized trial. Sometimes several such iterations are performed to refine the multicomponent intervention.

The main problem with this approach is that it depends heavily on the non-experimental, observational analyses. As is well-known [33–35], findings that are not based on randomization are hard to replicate due to the likely presence of unknown *confounders*[¶] (e.g. the variables that affect both the receipt of a component and the outcome). As a consequence, the effects of individual components and interactions may be misinterpreted resulting in a suboptimal intervention. Collins *et al.* [36] provided a head-to-head comparison between the above approach and an experimental procedure using FFDs in an extensive simulation study. This comparison was based on a simulated model involving five components, varying levels of adherence to each component, an unknown

[¶]In the literature on FFDs, the term *confounding* often refers to *aliasing* of effects. Here we use *confounding* to mean mixing of treatment effects with effects of other variables that affect both the receipt of treatment and the outcome, and thus keep *confounding* distinct from *aliasing*.

DOI: 10.1002/sim

confounder, and a continuous outcome. In addition, the model included an antagonistic interaction between two components. The simulation results showed that the FFD-based experimental approach outperformed the traditional approach (two-arm randomized trial followed by observational analyses) in terms of various criteria, e.g. optimizing the mean outcome of the final intervention and identifying the best multicomponent intervention. Of course the relative merit of the FFD-based experimental approach depends on the degree of confounding; using observational analyses to investigate the interactions might work well when the unknown confounder is only weakly related to the receipt of the components or the outcome.

Another alternative to FFDs is to conduct a series of *dismantling* or *subtractive* trials [12] where a 'more complete' version of the multicomponent intervention is compared with a reduced version with one or more components eliminated. A close variant of this is known as the *constructive strategy* [12] or *treatment augmentation design* [37], where a base intervention is compared with an augmented version in which one or more components are added to the base intervention. Yet another alternative is known as the *comparative treatment strategy* [12], where several versions of the intervention are directly compared. For example, if there are k components under consideration, a comparative strategy would compare $(k+1)$ experimental arms: k arms, each setting a single component at the high level and the rest at the low level, plus a control arm where all components are set at the low level. The above three approaches (i.e. dismantling, constructive, and comparative strategies) sometimes come under the umbrella term of single-factor designs [24], whenever the experimental arms under comparison differ by manipulating a single factor.

Note that there are several problems with using a series of single-factor experimental designs to construct a multicomponent intervention. First, as discussed by Box *et al.* [13, pp. 510–513], the use of single-factor designs often tacitly assumes that the effect of one component is independent of the levels of other components. This is not true in general, e.g. when there is a sizeable (qualitative) interaction between the components. Thus adopting a single-factor design often implicitly assumes that there is no interaction. Because of this limitation, using a series of single-factor designs to construct a multicomponent intervention may fail to achieve the best intervention.

The second problem regarding single-factor designs arises in designing the trials, e.g. deciding which factor to add (in constructive strategy) or subtract (in dismantling strategy), or which two versions of the multicomponent interventions to compare (in comparative strategy). These decisions are often driven by theory, cost, burden, or the results of observational analyses. To the extent that the results are driven by observational dose-response analyses of the amount of treatment received, they are vulnerable to confounding bias. As a consequence, in the sequence of single-factor trials conducted to find the best multicomponent intervention, active components may be accidentally eliminated in a dismantling strategy, and less active components may be erroneously added earlier than more active components in a constructive strategy.

A third problem with single-factor designs is that they often require many more subjects than comparable factorial designs to achieve similar power [24], rendering factorial designs a more efficient choice.

To summarize, in contrast to the treatment package strategy or the single-factor designs, inference about individual components in FFDs is strictly based on randomization, and hence less vulnerable to confounding bias. Furthermore, single-factor experiments are not equipped to take care of interactions, and often have higher sample size requirement. Although some *aliasing* of effects happens in FFDs, the investigator can control this based on prior substantive knowledge (see below for more discussion on aliasing). Thus by using FFDs, one often trades uncontrolled confounding

DOI: 10.1002/sim

for controlled aliasing. Thus, full and fractional factorial experimental designs offer a gold standard for developing multicomponent interventions. In the following, we will discuss feasibility.

2.3. Feasibility of the design

When the number of components (k) is moderately large, full factorial designs may be impractical due to cost of designing and implementing too many cells, i.e. making each treatment combination work together and ensuring implementation fidelity by staff [12]. This criticism has been the main motivation behind the development of FFDs. It is possible to select an FFD with substantially fewer cells, but still estimate the main effects (and sometimes important two-way interactions) without bias and with the same precision as in a full factorial design under plausible assumptions. A full factorial design allows the estimation of every individual factorial effect, including all higher-order interactions. However, in the absence of compelling prior theory or evidence to the contrary, third- and higher-order interactions are likely negligible in size in most of the multicomponent interventions [13, 14]. FFDs sacrifice the ability to estimate some of these higher-order interactions, and, in return, enable the study to have fewer cells. The choice of interactions to be sacrificed is informed by scientific theory, past studies, and investigator's experience. The practical price paid to buy the economy offered by an FFD is that the effects of interest, such as the main effects and two-way interactions, are *aliased* with some higher-order interactions. When two or more effects are aliased, one can estimate only the sum of the aliased effects. To overcome this problem, ideally an FFD is chosen in which each 'aliased bundle' includes only one effect that is *a priori* believed to be active, with any other effects included in the bundle likely negligible in size. If this is not possible, follow-up experiments [15, 18, 38] can be conducted to settle any ambiguity about which effects are most important in the aliased bundle of effects. The above ideas were used by both *Project Quit* [8] and *Guide to Decide* [14] to design FFD trials. A technical review of aliasing and FFDs is provided in Appendix A.

The strong use of theory and investigators' experience in determining which interactions to alias in an FFD is often initially disconcerting to scientists. Note, however, that in a two-arm randomized trial of a multicomponent intervention vs control, the multicomponent intervention must be determined completely by theory and investigator's experience, and furthermore in these two-arm trials every factorial effect (main effects and interactions) is aliased with every other effect. Thus all analyses concerning individual components hinge on the use of a correct model; if the model is too simple, then finding out what each effect is estimating is often difficult or even impossible. In this regard, FFDs offer a clearly better option in that the entire aliasing pattern is under the investigator's control, and there are principled ways (e.g. follow-up experiments) to disentangle any aliased effect. Moreover in non-experimental studies (that often follows the two-group comparisons) in which often the receipt of treatment depends on adherence to or availability of certain components, staff decisions as to who to offer what treatment, etc., the resulting confounding is uncontrolled.

Often concerns about feasibility are intertwined with a perceived need to include many subjects in each cell of the design; this may occur because investigators erroneously think that comparisons between individual cells will be required. This, however, is not the case; see below for a discussion of this along with power considerations. Nonetheless there are some situations in which investigators are unable to hire sufficient staff so as to implement multiple multicomponent interventions or are unable to train the staff to implement multiple multicomponent interventions simultaneously. In these settings FFDs are not feasible.

DOI: 10.1002/sim

2.4. Inability to cross some components

To use factorial designs, one must be able to cross the components without changing dose (i.e. all combinations should be implementable). This has been a fundamental concern regarding the use of factorial designs in medication trials. In medication trials, toxicity often precludes the combined use of multiple components (e.g. drugs) unless the dosage is altered [20, 39]. That is, the combination of drug A and drug B uses lower doses of both A and B, compared with the case when either drug A or drug B is used alone. So a high level of drug A in presence (high level) of drug B does not mean the same thing as a high level of drug A in absence (low level) of drug B. In such cases, the components lose their meanings and factorial designs become inappropriate. Here we consider only those components that when crossed, retain their meaning. This includes most behavioral, delivery, or implementation components, as well as multiple medications as long as they use different biological pathways.

When some components cannot be crossed, the clinical trials literature provides some approaches. Byar *et al.* [26] discussed *incomplete factorial* designs along with analysis strategies to take care of such cases. These designs are full or fractional factorial designs, minus some unpermitted combinations. Although these designs are not balanced (see the second footnote for a definition of balance), one can still estimate many of the relevant factorial effects.

2.5. Interpretation of main effects

It is well known that the definition of the main effect, in the presence of sizeable interactions [20, 39], differs from the investigators' conceptual definition of the effect of a component. To address this criticism, here we provide precise definitions of main effects and simple effects that commonly arise in various designs for multicomponent interventions, and establish their interrelationship.

For simplicity, consider a 2×2 factorial design with two components, say A_1 and A_2, and continuous outcome Y. The presence and absence (or, high and low level) of each component is coded $+1$ and -1, respectively. Let $\mu_{(-,-)}$, $\mu_{(-,+)}$, $\mu_{(+,-)}$, and $\mu_{(+,+)}$ be the mean outcomes corresponding to the absent–absent, absent–present, present–absent, and present–present cells of the design, respectively. At the population level, the main effect of the component A_1 is defined by $\frac{1}{2}(E_1+E_2)$, where $E_1=(\mu_{(+,+)}-\mu_{(-,+)})$ and $E_2=(\mu_{(+,-)}-\mu_{(-,-)})$ are two *simple effects*, denoting the effect of A_1 when A_2 is fixed at high and low level, respectively. Thus the main effect of A_1 is defined as the average of the two simple effects E_1 and E_2, and hence can be interpreted as the effect of A_1 when half the subjects in the population are exposed to (the high level of) A_2 and the remaining half are not. On the other hand, when conceptualizing the treatment effect of A_1, an investigator usually thinks of the simple effect denoting 'the effect of A_1 in absence of A_2' [39, p. 506], i.e. E_2. In absence of interaction between the components, this mismatch does not cause a problem since the two simple effects are equal. However, the main effect could be very different from the simple effect of A_1 in presence of a sizeable interaction.

If a dismantling strategy is followed (dismantling A_1 from the full package involving both A_1 and A_2), then the effect estimated is simply $(\mu_{(+,+)}-\mu_{(-,+)})=E_1$. This effect could also be estimated if the constructive strategy is followed (augmenting A_1 to the base intervention consisting of A_2 only). Thus these alternative designs estimate simple effects rather than main effects. Finally one can imagine a *treatment package effect*, e.g. $(\mu_{(+,+)}-\mu_{(-,-)})$, which is estimated when the 'likely best' package consisting of the present or high level of all the components is compared with a

DOI: 10.1002/sim

control consisting of the absent or low levels of all the components. This does not correspond to any of the simple effects.

For three two-level components, the main effect of a component A_1 is defined as $\frac{1}{4}(E_1+E_2+E_3+E_4)$, where $E_1=(\mu_{+,+,+)}-\mu_{(-,+,+)})$, $E_2=(\mu_{(+,+,-)}-\mu_{(-,+,-)})$, $E_3=(\mu_{(+,-,+)}-\mu_{(-,-,+)})$, and $E_4=(\mu_{(+,-,-)}-\mu_{(-,-,-)})$ are the four simple effects (and also can be interpreted as the effects resulting from different dismantling or constructive trials). The most common simple effect is E_4, i.e. 'effect of A_1 in absence of other components', and is often conceptualized as the treatment effect of A_1. In general, for a setting involving k two-level components, there are 2^{k-1} simple effects that can be interpreted as effects resulting from different constructive or dismantling trials. The main effect is simply the average of these 2^{k-1} simple effects.

To more clearly understand the alternate definitions and how they differ in the presence of an interaction, consider a regression formulation. Suppose that the true data-generating model, where A_1, A_2 are coded in $0/1$, is given by

$$Y=b_0+b_1A_1+b_2A_2+b_{12}A_1A_2+\varepsilon \tag{1}$$

If we use a regression analysis with the $-1/1$ coding, e.g. we fit $\beta_0+\beta_1A_1+\beta_2A_2+\beta_{12}A_1A_2$, then we estimate the following transformed model (now A_1, A_2 are coded in $-1/1$):

$$Y=\left(b_0+\frac{b_1+b_2}{2}+\frac{b_{12}}{4}\right)+\left(\frac{b_1}{2}+\frac{b_{12}}{4}\right)A_1+\left(\frac{b_2}{2}+\frac{b_{12}}{4}\right)A_2+\left(\frac{b_{12}}{4}\right)A_1A_2+\varepsilon \tag{2}$$

The main effect of A_1 is $2\hat{\beta}_1$, which estimates the population quantity $2(b_1/2+b_{12}/4)=b_1+b_{12}/2$ (this main effect continues to be the average effect of A_1 on Y over the levels of A_2). In contrast, the two simple effects of A_1 are b_1 (effect of A_1 when A_2 is absent) and b_1+b_{12} (effect of A_1 when A_2 is present). The main effect and the effect commonly conceptualized as the treatment effect of A_1, i.e. b_1, differ by the quantity $\frac{1}{2}b_{12}$ in presence of an active interaction ($b_{12}\neq 0$). If we apply the reasoning of the *Hierarchical Ordering Principle*[‖] [18] to this setting, then in general we expect that b_{12}, if nonzero, is likely to be of smaller size than b_1 and b_2.

To summarize, when there is an interaction, the main effect has the interpretation of the average effect of A_1 on Y over the levels of other components. This is quite different from what is often conceptualized as the treatment effect of A_1, e.g. the simple effect of A_1 on Y setting other components to lower level. However the crucial point is that in screening studies, the goal is to screen components efficiently, and not to estimate either the simple effect or the main effect of a component *per se*. The important issue for screening is whether this difference in definition impinges on our ability to screen components. So in this context, the concern about the definition of main effects is actually a concern about power to screen components. We address this concern below in great detail (see the third issue below under the Power heading).

2.6. Power

Several issues lead to concerns about power when factorial designs are considered. First, investigators sometimes use factorial designs to evaluate or compare a few multicomponent interventions,

[‖]The *Hierarchical Ordering Principle* is an assumption commonly made in design of experiments in the absence of substantive theory or prior results suggesting otherwise. It states that lower-order effects are more likely to be important than higher-order effects, and effects of the same order are equally likely to be important.

DOI: 10.1002/sim

e.g. compare one cell against another cell [40], or otherwise assess simple effects. This naturally leads to a large sample size requirement since each cell (group of subjects) must be large. However to screen components, we primarily focus on main effects and sometimes also a few anticipated two-way interactions. The focus on main effects and lower-order interactions for the purpose of screening can be partially justified by the *Hierarchical Ordering Principle* [18], which says that main effects and lower-order interactions are likely to be more important than higher-order interactions. Recall that the main effect of a factor is an average of all the 2^{k-1} simple effects. Thus even though several components are studied, the total sample size required for assessing the significance of a main effect is the same as that for a two-group trial (for example in a linear model, the estimator of the main effect is proportional to the difference between the means of two groups of cells; all cells in the FFD belong to one or the other group). Furthermore, in the multiphase approach to intervention development [14, 15], ascertaining the best treatment combination is done through follow-up studies, in which one usually focuses on only a few combinations of components while holding the levels of the remaining components constant. See Section 4 for a discussion of follow-up studies.

Second, there is concern about the loss of balance and subsequent loss of power due to study dropout. In most intervention studies, patient dropout is inevitable, thus resulting in unequal cell sizes. As discussed by Montgomery *et al.* [28], this is an issue for all clinical trials rather than a criticism of factorial designs; modern-day missing data techniques will be needed in the analysis as is the case with any randomized clinical trial.

A third issue related to power is how one should formulate the test statistics to detect the effects of treatment components in a screening study. Note that in a screening study the goal is to screen out inactive components, and not to estimate either a simple effect or a main effect *per se*. Below we show that even when the data are generated using non-zero simple effects, often the power to detect the resulting main effect is higher than the power to detect the original simple effect. Hence in a screening study, formulating the test statistics based on main effects is in general better than formulating test statistics based on simple effects. To discuss this consider again the 2×2 factorial design with two components, say A_1 and A_2, r subjects per cell, and the continuous outcome Y. The true data-generating model is specified in terms of simple effects, which is consistent with an investigator's conceptualization. Thus, the true data-generating model is given by (1) in which the components A_1, A_2 are coded in 0/1. In the following, we show that by basing the test statistic on main effects, we can in general screen non-zero simple effects with greater power.

For simplicity, assume $\mathrm{Var}(\varepsilon)=\sigma^2$ is known (and homogeneous across cells). If a linear regression model with 0/1 coding is used as in Piantadosi [39, pp. 508–509], then the following model is fit:

$$\beta_0+\beta_1 A_1+\beta_2 A_2+\beta_{12}A_1A_2 \tag{3}$$

In 0/1 coding, β_1, the coefficient of A_1, is a simple effect representing the comparison of the $(1,0)$ cell with the $(0,0)$ cell, i.e. $\beta_1=\mu_{(1,0)}-\mu_{(0,0)}=b_1$, where $\mu_{(1,0)}$ is the population mean of Y in the $(1,0)$ cell, and so on. Now β_1 is estimated by $\hat{\beta}_1=\bar{Y}_{(1,0)}-\bar{Y}_{(0,0)}$, where $\bar{Y}_{(1,0)}$ is the sample mean of Y in the $(1,0)$ cell, and so on. Clearly, $E(\hat{\beta}_1)=b_1$ and

$$\mathrm{Var}(\hat{\beta}_1)=\mathrm{Var}(\bar{Y}_{(1,0)})+\mathrm{Var}(\bar{Y}_{(0,0)})=\frac{\sigma^2}{r}+\frac{\sigma^2}{r}=\frac{2\sigma^2}{r}$$

DOI: 10.1002/sim

So the signal-to-noise ratio (SNR) governing the power to screen A_1 with 0/1 coding in the analysis model (i.e. basing the test statistics on simple effects) is

$$\mathrm{SNR}_{[0/1]}=\frac{|E(\hat{\beta}_1)|}{\sqrt{\mathrm{Var}(\hat{\beta}_1)}}=\frac{|b_1|\sqrt{r}}{\sqrt{2}\sigma}$$

On the other hand, if we use the analysis model (3) with $-1/1$ coding, it follows that

$$\begin{aligned}\beta_1 &= \tfrac{1}{4}[(\mu_{(+,+)}-\mu_{(-,+)})+(\mu_{(+,-)}-\mu_{(-,-)})]\\ &= \tfrac{1}{2}\times(\text{the main effect of } A_1)\\ &= \tfrac{1}{2}\times(\text{the average of two simple effects})\end{aligned}$$

and is estimated by the sample version $\hat{\beta}_1$ (where μ is replaced by $\bar{Y}$). Then,

$$E(\hat{\beta}_1)=\beta_1=\left(\frac{b_1}{2}+\frac{b_{12}}{4}\right)$$

$$\mathrm{Var}(\hat{\beta}_1)=\frac{1}{4}\times\frac{1}{2}\times(\text{variance of an estimated simple effect})=\frac{1}{4}\times\frac{1}{2}\times\frac{2\sigma^2}{r}=\frac{\sigma^2}{4r}$$

So the SNR governing the power to screen A_1 with $-1/1$ coding in the analysis model (i.e. basing the test statistics on main effects) is

$$\mathrm{SNR}_{[-1/1]}=\frac{|E(\hat{\beta}_1)|}{\sqrt{\mathrm{Var}(\hat{\beta}_1)}}=\left|b_1+\frac{b_{12}}{2}\right|\frac{\sqrt{r}}{\sigma}$$

A measure of relative efficiency of the two coding schemes (equivalently, two ways of forming the test statistics) in screening A_1 is given by

$$\eta=\frac{\mathrm{SNR}_{[-1/1]}}{\mathrm{SNR}_{[0/1]}}=\sqrt{2}\left|b_1+\frac{b_{12}}{2}\right|\Big/|b_1|=\sqrt{2}\left|1+\frac{b_{12}}{2b_1}\right|$$

In absence of an interaction (i.e. $b_{12}=0$), $\eta=\sqrt{2}>1$, and hence the $-1/1$ coding gives higher power for screening components. In case of synergistic interaction (i.e. b_1 and b_{12} are of same sign), η is even larger, so the $-1/1$ coding gives higher power. Even in the case of antagonistic interaction (i.e. b_1 and b_{12} are of opposite sign), the $-1/1$ coding gives higher power in screening components (i.e. $\eta>1$) if $b_1<0$ and $0<b_{12}<-(2+\sqrt{2})b_1$, or if $b_1>0$ and $0>b_{12}>-(2-\sqrt{2})b_1$. If we have $k(\geqslant 2)$ components in a factorial experiment, and there may be a two-way but no higher-order interaction in the true data-generating model, then the relative efficiency of the two coding schemes (measured by η) increases with k. A verification of this appears in Appendix B.

To illustrate the power implications of basing the test statistics on main effects rather than simple effects in a regression analysis, we consider a small simulation study with the data-generating model $Y|A_1, A_2\sim\mathrm{N}(\mu=b_0+b_1A_1+b_2A_2+b_{12}A_1A_2, \sigma=1)$, where A_1, A_2 are coded in 0/1. That is, the data-generating model is specified in terms of simple effects (as is usually conceptualized by an investigator). The coefficients b_1, b_2 are set according to Cohen's [41] small or medium effect size (i.e. $b_1=b_2=0.2, 0.5$). The coefficient b_{12} of the interaction term is varied: $b_{12}=b_1$,

DOI: 10.1002/sim

Table I. Power to screen A_1 in absence and presence of an interaction.

			$b_1=0.2$		$b_1=0.5$	
n	Interaction size (b_{12})	Interaction type	Analysis model in 0/1 coding	Analysis model in $-1/1$ coding	Analysis model in 0/1 coding	Analysis model in $-1/1$ coding
100	Same ($=b_1$)	Synergistic	0.1030	0.2910	0.4150	0.9550
	Half ($=\frac{b_1}{2}$)	Synergistic	0.1030	0.2290	0.4150	0.8730
	Absent ($=0$)	None	0.1030	0.1720	0.4150	0.6830
	Half ($=-\frac{b_1}{2}$)	Antagonistic	0.1030	0.1110	0.4150	0.4420
	Same ($=-b_1$)	Antagonistic	0.1030	0.0820	0.4150	0.2290
200	Same ($=b_1$)	Synergistic	0.1690	0.5440	0.6920	1.0000
	Half ($=\frac{b_1}{2}$)	Synergistic	0.1690	0.3940	0.6920	0.9870
	Absent ($=0$)	None	0.1690	0.2840	0.6920	0.9430
	Half ($=-\frac{b_1}{2}$)	Antagonistic	0.1690	0.1720	0.6920	0.7510
	Same ($=-b_1$)	Antagonistic	0.1690	0.1040	0.6920	0.3940
500	Same ($=b_1$)	Synergistic	0.3460	0.9210	0.9740	1.0000
	Half ($=\frac{b_1}{2}$)	Synergistic	0.3460	0.8040	0.9740	1.0000
	Absent ($=0$)	None	0.3460	0.6050	0.9740	1.0000
	Half ($=-\frac{b_1}{2}$)	Antagonistic	0.3460	0.3730	0.9740	0.9870
	Same ($=-b_1$)	Antagonistic	0.3460	0.1890	0.9740	0.8040

$b_1/2$, 0, $-b_1/2$, $-b_1$ (i.e. same size and sign as b_1, half the size of and same sign as b_1, absent, half the size of b_1 but of opposite sign, same size as b_1 but of opposite sign). A 0.05 level of significance is used throughout, while varying the sample size: $n=100, 200, 500$. The goal of this simulation is to illustrate that even when the data-generating model is specified in terms of simple effects, basing the test statistics on main effects (using $-1/1$ coding) leads to higher power in most settings than basing the test statistics on simple effects (using 0/1 coding). Note that the SNRs of the coding schemes govern the corresponding powers. In the following, we consider the power to screen A_1 both in presence and absence of an interaction term A_1A_2 (synergistic as well as antagonistic). Table I contains a Monte Carlo estimate (using 1000 iterations) of the power for screening A_1 under different scenarios.

Note that in Table I, the power to screen A_1 is higher in general when the analysis model is coded in $-1/1$ compared with when the analysis model is coded in 0/1 (e.g. comparing the 4th vs 5th column, and comparing the 6th vs 7th column), except when the interaction is of same size and opposite in sign as the simple effect of A_1 (as expected from the above discussion). However according to the *Hierarchical Ordering Principle* [18], interactions are usually of smaller order of magnitude than the main effects (absent strong scientific theory to the contrary), and hence this is a fairly unlikely scenario. A secondary point to note is that when the data-analysis model uses the $-1/1$ coding, there is a decrease in power to screen A_1 as the interaction term b_{12} decreases from highly synergistic to highly antagonistic (moving down the 5th and 7th columns). However, when the data-analysis model uses the 0/1 coding, the power for screening A_1 is independent of the size of the interaction term b_{12} (moving down the 4th and 6th columns). But the decrease in power in the 5th and 7th columns due to interaction often does not pose a serious threat (as compared

DOI: 10.1002/sim

with the loss of power from using 0/1 coding) if the goal is to screen components, since in most settings $-1/1$ coding gives better power anyway.

A fourth issue related to power is the power to detect interactions. Factorial designs are often criticized on the ground that the power to detect an interaction is much lower than the power to detect a main effect of the same size [28, 39]. However, it is also recognized that factorial designs are the only experimental designs that can systematically investigate interactions. To overcome the low power for detecting interactions in a confirmatory (not screening) trial, the general recommendation in the literature [20] is that if an interaction is strongly anticipated based on the investigator's prior knowledge, the study should be powered with a larger sample size. When criticizing factorial designs on the ground of low power for interactions in the present context of screening trials for developing multicomponent interventions, it is useful to consider the pros and cons of the possible alternatives. The natural alternative is to conduct non-experimental analyses using treatment adherence or other post-randomization outcomes as doses or factor levels from a randomized trial or to use observational data sets. As discussed previously, the relative merit of FFDs over the above strategy depends on the degree of confounding in the data. The crux is that the low power to detect interactions in a factorial design can be offset by its ability to perform valid estimation and inference, and its ability to control (by design) aliasing in a principled manner, in comparison with observational analyses.

3. OPERATIONALIZATION OF SCREENING TRIALS

This section provides an example of how screening trials can be operationalized using FFDs. The choice of an appropriate FFD is often governed by prior knowledge regarding the intervention to be developed. To move forward, two definitions are useful. An FFD is completely characterized by its *defining relation* [18], a rule from which the aliasing pattern of the FFD can be obtained. Moreover, FFDs are sometimes categorized by their *resolution*. Loosely speaking, the higher the resolution, the better the design. Resolution IV and resolution V designs are considered here. In particular, in a resolution V design, main effects are aliased with 4-way (or higher order) interactions, and 2-way interactions are aliased with 3-way (or higher order) interactions. Likewise in a resolution IV design, main effects are aliased with 3-way (or higher order) interactions, and 2-way interactions are aliased with other 2-ways (or higher order). Typically resolution V designs are better than resolution IV designs, but resolution V designs require more cells. Hence, sometimes investigators have to use resolution IV designs due to cost and feasibility constraints. Further review of the *defining relation* and *resolution* are given in the Appendix A. In the following, we first discuss the screening design used in the *Project Quit* study. Next, we discuss a general approach to construct screening designs (e.g. appropriate FFDs).

3.1. Screening design in the project quit study

Denote the six components of the *Project Quit* study, e.g. depth of *outcome expectations*, depth of *efficacy expectations*, depth of *success stories*, personalization of *message source*, mode of message framing, and exposure schedule by A_1, A_2, A_3, A_4, A_5, and A_6, respectively. In this study, prior knowledge suggested that the interactions between *outcome expectations* and *efficacy expectations* (A_1A_2), *outcome expectations* and *success stories* (A_1A_3), *outcome expectations* and message framing (A_1A_5), and *efficacy expectations* and message framing (A_2A_5) were likely active

DOI: 10.1002/sim

(let us call them *anticipated* interactions), and that all other interactions should be negligibly small in size. So a design was constructed such that one could estimate the A_1A_2, A_1A_3, A_1A_5, and A_2A_5 interactions, assuming all others to be small. Owing to cost constraints, 16 cells were used in the design. So the design used was a 16-cell FFD with the defining relation

$$I = A_1A_2A_4A_5 = A_1A_3A_4A_6 = A_2A_3A_5A_6 \tag{4}$$

This is a resolution IV design where some of the 2-way interactions are aliased with other 2-way interactions. The anticipated 2-way interactions are listed on the left-hand side of the following aliasing equations (obtained from the defining relation (4)):

$$A_1A_2 = A_4A_5$$

$$A_1A_3 = A_4A_6$$

$$A_1A_5 = A_2A_4$$

$$A_2A_5 = A_1A_4$$

Note that the anticipated interactions were aliased with other 2-way interactions that were considered negligible, and hence were estimable without bias. The defining relation $I = A_1A_2A_3A_5 = A_1A_3A_4A_6 = A_2A_4A_5A_6$ was 'cleverly' chosen to accomplish this goal. Of course, the investigator's assumption about the interactions could be wrong. But one can verify any critical working assumptions made in the screening study using follow-up studies [15].

3.2. Screening design construction in general

As a starting point we assume that regardless of the number of components studied, the number of cells used can be at most 16 (equal to the number of cells used in the *Project Quit* study). Of course this number can vary from one setting to another. If four or fewer components are to be studied, a full factorial design can be used. If five components, say $A_1, \ldots, A_5$, are to be studied, then one should use the resolution V FFD with the defining relation $I = A_1A_2A_3A_4A_5$ (this is the case in the *Guide to Decide* project described in [14]). If six components, say $A_1, \ldots, A_6$, are to be studied, resolution IV designs are generally recommended. If prior knowledge suggests a few anticipated 2-way interactions, an FFD can be chosen carefully so that the anticipated 2-way interactions are not aliased with each other (this consideration often drives the construction of the design). Assuming the unanticipated interactions to be negligible, this ensures that each anticipated interaction can be estimated without bias. When there is only one anticipated interaction, any 16-cell resolution IV FFD can be used. However, for two or more anticipated interactions, choices are limited. Software (e.g. SAS PROC FACTEX, JMP, Minitab) can be used to generate the designs in such cases (they provide one possible design that satisfies the constraints of resolution and/or anticipated interactions, instead of giving the complete list of possible designs). For two or three anticipated interactions, the complete set of recommended designs are given in Table II.

3.3. Power and sample size in screening trials

In a screening trial using a factorial design, the power calculation used to size the trial focuses on main effects of each component. Thus, the power calculation is similar to that of a two-arm randomized trial in that the two levels of a single component (averaged over the levels of

DOI: 10.1002/sim

Table II. Recommended resolution IV FFDs under varying anticipated interactions.

Case	Anticipated interactions of the form	Recommended designs (defining relations)
1	A_1A_2, A_3A_4 (no component shared)	$I=A_1A_2A_3A_5=A_1A_3A_4A_6=A_2A_4A_5A_6$ $I=A_1A_2A_3A_5=A_2A_3A_4A_6=A_1A_4A_5A_6$ $I=A_1A_2A_4A_5=A_1A_3A_4A_6=A_2A_3A_5A_6$ $I=A_1A_2A_4A_5=A_2A_3A_4A_6=A_1A_3A_5A_6$ $I=A_1A_2A_3A_6=A_1A_3A_4A_5=A_2A_4A_5A_6$ $I=A_1A_2A_3A_6=A_2A_3A_4A_5=A_1A_4A_5A_6$ $I=A_1A_2A_4A_6=A_1A_3A_4A_5=A_2A_3A_5A_6$ $I=A_1A_2A_4A_6=A_2A_3A_4A_5=A_1A_3A_5A_6$
2	A_1A_2, A_1A_3 (one component shared)	$I=A_1A_2A_4A_5=A_1A_3A_4A_6=A_2A_3A_5A_6$ $I=A_1A_2A_4A_6=A_1A_3A_4A_5=A_2A_3A_5A_6$ $I=A_1A_2A_4A_5=A_1A_3A_5A_6=A_2A_3A_4A_6$ $I=A_1A_2A_5A_6=A_1A_3A_4A_5=A_2A_3A_4A_6$ $I=A_1A_2A_4A_6=A_1A_3A_5A_6=A_2A_3A_4A_5$ $I=A_1A_2A_5A_6=A_1A_3A_4A_6=A_2A_3A_4A_5$
3	A_1A_2, A_3A_4, A_5A_6	Same as case 1
4	A_1A_2, A_1A_3, A_4A_5	$I=A_1A_2A_5A_6=A_1A_3A_4A_6=A_2A_3A_4A_5$ $I=A_1A_2A_4A_6=A_1A_3A_5A_6=A_2A_3A_4A_5$
5	A_1A_2, A_1A_3, A_2A_4	$I=A_1A_2A_5A_6=A_1A_3A_4A_5=A_2A_3A_4A_6$ $I=A_1A_2A_5A_6=A_1A_3A_4A_6=A_2A_3A_4A_5$
6	A_1A_2, A_1A_3, A_1A_4	$I=A_1A_2A_5A_6=A_1A_3A_4A_5=A_2A_3A_4A_6$
7	A_1A_2, A_1A_3, A_2A_3	Same as case 2

all other components) serve as the two arms. Below we provide the power calculation for the *Project Quit* study as an example. For *Project Quit*, the planned initial recruitment size was 2000; this number was chosen to achieve a total sample size of 1500 for the analysis, anticipating a 75 per cent response rate at the 6-month follow-up. Assuming no differential attrition across cells, this meant roughly 750 subjects per level of each intervention component. The primary outcome was binary, e.g. 7-day point-prevalence smoking cessation at the 6-month follow-up. So the power analysis involved binomial calculations (using a normal approximation) assuming a baseline average cessation rate of 10 per cent found in a previous study [42]. For each main effect, the sample size of 750 per level provides approximately 80 per cent power for detecting a 4.5 per cent difference in cessation rates. The same power characteristics exist for each of the six components. Note that to achieve the same power to detect the same difference in cessation rates, one would need the same sample size in a usual two-arm study (so the sample size requirement is not increased by using a factorial design). The formula for calculating power in the present set-up is given by

$$\Phi\left(\frac{\sqrt{\frac{n}{2}}|\Delta|-z_{\alpha/2}\sqrt{2p(1-p)}}{\sqrt{p(1-p)+(p+\Delta)(1-p-\Delta)}}\right)$$

DOI: 10.1002/sim

where n is the total sample size, p is the baseline cessation rate, Δ is the change in cessation rate to be detected, α is the Type I error, $z_{\alpha/2}$ is the upper $100(\alpha/2)$ per cent cutoff point of a standard normal distribution, and Φ is the standard normal distribution function.

3.4. Additional practical considerations regarding study duration and cost

A primary advantage of using factorial designs in a screening study lies in its efficiency, i.e. its ability to answer several screening questions (regarding multiple intervention components) quickly from a single study. The use of an FFD-based approach in *Project Quit* was motivated by the concern that advances in communication technologies were moving well beyond the understanding of message content, presentation, and delivery principles in the field of smoking cessation. Investigators of this study realized that research using the field's most widely used designs (e.g. randomized trials with a small number of groups) [42–45] would take years to assess even a few basic questions. By the time these findings would be disseminated, the technology and target populations would likely be changed (e.g. become more sophisticated in their understanding of a communications channel), and consequently the field would continue to lag behind. Thus in the context of this concern, the FFD-based multiphase approach provided a huge benefit by offering a shorter total study duration to answer so many questions compared with the alternative designs.

There are two kinds of cost associated with designing multicomponent intervention trials, e.g. (1) cost associated with sample size requirement and (2) cost of designing and implementing different cells. We have already discussed that the sample size requirement does not go up by using an FFD. The only additional cost of designing an FFD over a two-group trial is the cost of designing and delivering too many versions of the intervention that might limit the applicability of FFDs in certain settings. In case of *Project Quit*, the intervention was delivered entirely through the Internet. So the delivery of 16 versions of the multicomponent intervention did not cost additional staff time and training over and above the cost of software programming to generate the different versions, which turned out to be manageable. See Collins *et al.* [24] for a detailed comparison of FFDs with single-factor designs (dismantling, constructive and comparative trials) from a resource management perspective.

3.5. Screening analysis

The screening analysis uses a linear model (in case of continuous outcome) or a generalized linear model (in case of binary or categorical outcome). A few considerations to be made during the analysis are:

1. The level of significance α for testing the effects in the screening study might be set higher than 0.05 to achieve greater power for detecting effects. α can be viewed as a tuning parameter of the procedure. One possible choice is to use $\alpha = 0.1$ for the main effects and *anticipated* two-way interactions, and a Bonferroni-corrected 0.1 level for the *unanticipated* interactions.
2. As an alternative (or augmentation) to performing significance tests at the screening study, one can rank-order the absolute values of the test statistic corresponding to the factorial effects (or equivalently p-values) and move to follow-up studies with the largest m. Then this m becomes a tuning parameter of the procedure. This approach should work better in case all individual effects are small, but together they produce some effect (significance test often accepts the null hypothesis of no effect in such cases, and hence performs poorly). To be resistant to the noise in the data, one may choose to rank-order only the main effects

DOI: 10.1002/sim

and *anticipated* interactions. This strategy with $m=3$ was followed in the simulation study described in [36].

Examples of the screening analysis in the *Guide to Decide* and *Project Quit* studies can be found in [8, 14]. Based on the screening analysis of the *Project Quit* study, the investigators decided to move to the follow-up study with the components having the highest two p-values (e.g. *success stories* and *message source*). Furthermore, since three of the components (*outcome expectations*, *efficacy expectations*, and *success stories*) were set at levels corresponding to high depth of tailoring vs low depth of tailoring, the investigators considered a regression of overall depth of tailoring (over all components) and found that as the depth of tailoring increased, the smoking cessation rate increased. Hence the investigators decided to use a high depth of tailoring in the follow-up study.

4. FOLLOW-UP STUDIES

In the process of developing a multicomponent intervention, an investigator often conducts follow-up studies involving the *significant*** factorial effects from the screening study to fine-tune the results, e.g. finding the best level (or dose) of a significant component, which is either continuous or has more than two levels by a dose-response experiment (where the subjects are randomized to ethically acceptable doses of the component), or de-aliasing significant aliased interactions by a smaller factorial experiment. In this section, first we provide a few hypothetical examples (of varying level of complexity) of follow-up studies to provide some general intuition, and then briefly describe the follow-up phase of the *Project Quit* study.

4.1. Hypothetical examples

In the following examples, for simplicity, we assume that there are six components in the study, e.g. $A_1, \ldots, A_6$, out of which only A_1 is a 3-level component (say, high, medium, low levels—only high and low levels are studied at the screening trial) and the rest are binary (high and low). High values of the outcome are preferred. A 16-cell resolution IV FFD is used as the screening design (see Section 3 for details). We assume throughout that three-way (or higher-order) interactions are negligible in size compared with the noise in the data; hence, even though main effects are aliased with three-way interactions, we assign the estimated effect to the main effect.

Example 1
Suppose the significant effects along with their signs based on screening analysis are

$$A_1(+), A_2(+), A_3(-), A_5(-), A_2A_3 = A_4A_5(-)$$

where the aliased interaction A_2A_3 is unanticipated, but A_4A_5 is anticipated. So the investigator may dismiss A_2A_3 as a possible effect and assign the observed effect entirely to A_4A_5. Since the main effect of A_4 is insignificant, the main effect of A_5 is negative, and the A_4A_5 interaction is negative; it is reasonable to set A_4 at its high level and A_5 at its low level to maximize the mean outcome. In addition, from the signs of the estimated main effects of A_2 and A_3 (and ignoring the unanticipated A_2A_3 interaction), A_2 and A_3 should be set at their high and low levels, respectively.

**Throughout this section, we use the term *significant* loosely to mean any effects that come out important according to the screening analysis strategy outlined in Section 3.

DOI: 10.1002/sim

Since A_6 is insignificant, it should be set at low level. Hence the follow-up study might be a 2-group trial varying A_1 at its medium and high levels (since its main effect is positive), setting A_2, A_3, A_4, A_5, and A_6 at high, low, high, low, and low level, respectively. If A_4 is an expensive or particularly burdensome component, then it may be worthwhile to affirm that A_4 is not significant yet its interaction with A_5 is. In that case, the follow-up study can be a 8-group trial where the two levels of A_1 (high/medium) are crossed with two levels of A_4 and A_5 each. In all the groups, A_2, A_3, and A_6 should be set at high, low, and low level, respectively.

Example 2
Suppose the significant effects along with their signs based on screening analysis are

$$A_1(+), A_3(+), A_1A_3 = A_2A_6(-)$$

where aliased interaction A_1A_3 is anticipated, but A_2A_6 is unanticipated. As before, we dismiss A_2A_6 based on prior considerations and assign the observed effect to A_1A_3. The follow-up study could be a 6-group trial crossing three levels of A_1 with two levels of A_3. In all the groups, the levels of A_2, A_4, A_5, and A_6 should be set at the low level.

Example 3
Suppose the significant effects along with their signs based on screening analysis are

$$A_1(+), A_2(+), A_3(+), A_5(-), A_2A_3 = A_4A_5(-)$$

where both the interactions A_2A_3 and A_4A_5 involved in the aliased bundle are unanticipated. Since the main effect of A_4 is insignificant, the main effect of A_5 is negative, and the aliased A_4A_5 interaction is negative (even though we are not sure if the observed effect is really due to A_4A_5); one reasonable step would be to set A_4 at its high level (provided the high level of A_4 is not very expensive or burdensome) and A_5 at its low level (note that our decision about the optimum levels of A_4 and A_5 would be the same when A_4A_5 effect is really negative as when A_4A_5 is null). In addition, we would set A_6 to the low level. If there is a concern about the potential A_2A_3 interaction, then the follow-up study could be a 8-group trial, where medium and high levels of A_1 are crossed with the two levels of A_2 and A_3 each to form the 8 groups (setting A_4, A_5, and A_6 at high, low, and low levels, respectively).

4.2. Follow-up study design of Project Quit

An alternative to the follow-up studies outlined above is provided by *Project Quit* study, in which all components were two-level (hence a dose-response experiment was unnecessary) and no (unanticipated) aliased interaction were found to be significant (hence no de-aliasing experiment was necessary). The investigators decided to study different aspects (not studied in the screening trial) of the two important components (e.g. *success stories* and *message source*). The decision was to vary *message source* at two levels (high/low) of additional personalization, and to vary *success stories* in terms of the archetype (language and picture) of the hypothetical character in the story at three levels (e.g. a rebel, care-giver, or self-made character). Two new two-level components, e.g. *order* (of appearance on the web site: *success stories* first vs *health advice* first) and *email quit status request* (yes/no) were added to the follow-up study. Subjects randomized to the 'yes' level of *email quit status request* were contacted by the study staff at regular intervals about their quit status. The follow-up study consisted of 25 groups in total: 24 groups from the

DOI: 10.1002/sim

$2 \times 3 \times 2 \times 2$ factorial structure of the above four components, plus a control group. In all groups, the original components from the screening trial not studied in the follow-up study were set as follows: deeply-tailored *efficacy expectation* and *outcome expectation* messages, gain framing, and multiple exposures. All three levels of the *success stories* were also deeply tailored. The control group received the best intervention according to the results of screening study (e.g. highly personalized source at the first session only, deeply tailored story with fixed archetype as in the screening study, deeply tailored *efficacy expectation* and *outcome expectation* messages, gain framing, and multiple exposures)—they did not receive any email about their quit status.

5. DISCUSSION

Multicomponent interventions are becoming increasingly common in health sciences. In this paper, we have addressed the criticisms and misconceptions regarding the use of full and FFDs (e.g. attractive alternatives to and feasibility of such designs, inability to cross components, interpretation of main effects, and concerns about power) in the context of screening studies to develop multicomponent interventions. Other issues regarding the use of factorial designs, as discussed by Couper *et al.* [46], are slower recruitment rate (since subjects need to meet the inclusion criteria for all the components) and potential lower compliance (due to a more complicated treatment protocol) than single-component trials. However, these are common to any studies of multicomponent interventions, and not problems specific to factorial designs.

We provided some examples of follow-up studies that often need to be conducted (e.g. to de-alias significant aliased interactions) after the completion of the screening study. Further strategies for conducting follow-up studies can be found, for example, in [18, 38]. In addition, in case there is at least one component with more than two levels (e.g. a continuous component), dose-response experiments [47, 48] where subjects are randomized to ethical doses should be used to find the optimal dose of these components. Operationalizing a wider variety of follow-up studies needs more targeted future research.

In our discussion of FFDs, we assumed that third- and higher-order interactions are negligible [13, 14]. This is not a binding constraint. Suppose prior knowledge suggests that interactions up to order 3 involving a certain component are likely important, whereas even two-way interactions involving some other components are negligible. One can still use a carefully chosen FFD [18].

One setting in which factorial designs are not well suited is when the main effects of all the individual components are weak, but there are some high-order interactions in the data-generating model that produce a strong effect on the outcome (i.e. a setting where the *Hierarchical Ordering Principle* is violated). Another important caveat regarding the use of factorial designs for developing multicomponent interventions is the presence of nested components (e.g. levels of component B are nested within the levels of component A). Generalization of the usual factorial designs called *nested factorials* [49, 50] can incorporate nested components. Analysis of such designs can employ mixed-effects models [51]. A somewhat similar issue is when some intervention components are applied most naturally in a grouped setting. For example, some intervention components are provided to all patients at a clinic [52] or to all children in a classroom or school [53]. Development of an experimental framework tailored to such settings is an avenue for future research. To conclude, FFDs provide a powerful tool for conducting screening studies to aid in the development of multicomponent interventions.

DOI: 10.1002/sim

APPENDIX A: DEFINING RELATION AND RESOLUTION OF AN FFD

The *defining relation* of an FFD specifies the aliasing. Suppose a study involving five components, say $A_1, \ldots, A_5$, is restricted to 16 cells (as in the *Guide to Decide* study described in [14]). Then a $\frac{1}{2}$ fraction of the 2^5 full factorial design should be used. With 16 cells, one can construct a full factorial with four components, say with $A_1, \ldots, A_4$. The strategy is to alias the fifth component, say A_5, with the 4-way interaction $A_1A_2A_3A_4$. This means, the column (in the design matrix) of A_5 is identical to that of the element-wise product of the columns of A_1, A_2, A_3, and A_4, i.e. $A_5 = A_1A_2A_3A_4$. Note that all the elements in any of the columns are either $+1$ or -1. So element-wise product of any column with itself leads to the identity column, say I (with all its entries $+1$). In particular, $A_5A_5 = I$. Multiplying both sides of the equation $A_5 = A_1A_2A_3A_4$ by A_5 gives

$$I = A_1A_2A_3A_4A_5 \tag{A1}$$

The condition (A1) completely specifies the aliasing pattern of the 2^{5-1} FFD under consideration, and hence called its *defining relation*. The alias of any factorial effect can be found by multiplying both sides of (A1) by that effect and then using the facts that $A_jI = A_j$ and $A_jA_j = I$ for all j. The word $A_1A_2A_3A_4A_5$ is called the *defining word*. The length (i.e. number of elements) of the *defining word* is called the *resolution* of the design. So the design specified by (A1) is a resolution V design. In a resolution V design, the main effects are aliased with 4-way interactions and the 2-way interactions are aliased with 3-way interactions. In a setting where the third- or higher-order interactions are negligible, resolution V FFDs are almost as good as the full factorials in that the main effects and 2-way interactions are estimable without bias.

However due to cost and feasibility constraints, one often has to use smaller (than $\frac{1}{2}$) fraction of full factorial designs, leading to lower resolution. The *Project Quit* study described before used a resolution IV FFD. In the following, we illustrate resolution IV designs with an example. Suppose there are six components, say $A_1, \ldots, A_6$, in a study that is restricted to 16 cells (as in *Project Quit*). This means constructing a $\frac{1}{4}$ fraction of the 2^6 ($=64$ cells) full factorial design. With 16 cells, one can construct a full factorial with four components, say with $A_1, \ldots, A_4$. Now, the strategy is to make the columns of the remaining two components A_5 and A_6 identical to some higher-order interactions. One such choice is to set $A_5 = A_1A_3A_4$ and $A_6 = A_2A_3A_4$. Using the same rules as before, one gets $I = A_1A_3A_4A_5$ from the first aliasing relation, and $I = A_2A_3A_4A_6$ from the second aliasing relation. Multiplying these two, a third equation $I = A_1A_2A_5A_6$ follows. Thus the defining relation of this FFD is

$$I = A_1A_3A_4A_5 = A_2A_3A_4A_6 = A_1A_2A_5A_6 \tag{A2}$$

By definition, (A2) is a resolution IV design, since the length of each defining word is 4. In a resolution IV design, the main effects are aliased with 3-way or higher-order interactions, but the 2-way interactions are aliased with other 2-ways.

APPENDIX B: RELATIVE EFFICIENCY OF CODING SCHEMES

Here we show that if there are $k(\geqslant 2)$ components in a factorial experiment, and there may be a two-way but no higher-order interaction in the true data-generating model, then the relative efficiency of the two coding schemes (measured by η) increases with k.

DOI: 10.1002/sim

Note that in 0/1 coding, regardless of the total number of components (k), β_1, the coefficient of A_1, is a simple effect given by $\mu_{(1,0,\ldots,0)} - \mu_{(0,\ldots,0)}$; the corresponding estimator is $\hat{\beta}_1 = \bar{Y}_{(1,0,\ldots,0)} - \bar{Y}_{(0,\ldots,0)}$, where $\bar{Y}_{(1,0,\ldots,0)}$ is the sample mean of Y in the $(1,0,\ldots,0)$ cell, and so on. It follows that $E(\hat{\beta}_1) = b_1$ and $\mathrm{Var}(\hat{\beta}_1) = 2\sigma^2/r$ (since $\hat{\beta}_1$ is a comparison of two cells each of size r). So with 0/1 coding, the SNR governing the power to detect A_1 in the 2^k design is same as that in a 2×2 design considered before, e.g.

$$\mathrm{SNR}_{[0/1]} = \frac{|E(\hat{\beta}_1)|}{\sqrt{\mathrm{Var}(\hat{\beta}_1)}} = \frac{|b_1|\sqrt{r}}{\sqrt{2}\sigma}$$

In $-1/1$ coding, $\beta_1 = \frac{1}{2} \times$ (the main effect of A_1) $= \frac{1}{2} \times$ (the average of 2^{k-1} simple effects), which is estimated by its sample version $\hat{\beta}_1$ where all the μ's in the expression of simple effects are replaced by the corresponding $\bar{Y}$'s. In case there is a two-way interaction (b_{12}) but no higher-order interaction in the true data-generating model, $E(\hat{\beta}_1) = \beta_1 = (b_1/2 + b_{12}/4)$,

$$\mathrm{Var}(\hat{\beta}_1) = \frac{1}{4} \times \frac{1}{2^{k-1}} \times (\text{variance of an estimated simple effect}) = \frac{1}{4} \times \frac{1}{2^{k-1}} \times \frac{2\sigma^2}{r} = \frac{\sigma^2}{2^k r}$$

and

$$\mathrm{SNR}_{[-1/1]} = \frac{|E(\hat{\beta}_1)|}{\sqrt{\mathrm{Var}(\hat{\beta}_1)}} = 2^{(k/2-1)} \left| b_1 + \frac{b_{12}}{2} \right| \frac{\sqrt{r}}{\sigma}$$

Thus,

$$\eta = \frac{\mathrm{SNR}_{[-1/1]}}{\mathrm{SNR}_{[0/1]}} = 2^{(k-1)/2} \left| 1 + \frac{b_{12}}{2b_1} \right|$$

is an increasing function of k.

ACKNOWLEDGEMENTS

We acknowledge support for this project from National Institutes of Health grants RO1 MH080015, P50 DA10075, and P50 CA101451.

REFERENCES

1. Golin C, Earp J, Tien H, Stewart P, Porter C, Howie L. A 2-arm, randomized, controlled trial of a motivational interviewing-based intervention to improve adherence to antiretroviral therapy (ART) among patients failing or initiating ART. *Journal of Acquired Immune Deficiency Syndrome* 2006; **42**:42–51.
2. Cuffe M. The patient with cardiovascular disease: treatment strategies for preventing major events. *Clinical Cardiology* 2006; **29**(II):4–12.
3. Williams J, Gerrity M, Holsinger T, Dobscha S, Gaynes B, Dietrich A. Systematic review of multifaceted interventions to improve depression care. *General Hospital Psychiatry* 2007; **29**:91–116.
4. Paul G, Smith S, Whitford D, O'Kelly F, O'Dowd T. Development of a complex intervention to test the effectiveness of peer support in type 2 diabetes. *BMC Health Services Research* 2007; **7**:136.
5. Riggs N, Elfenbaum P, Pentz M. Parent program component analysis in a drug abuse prevention trial. *Journal of Adolescent Health* 2006; **39**:66–72.

DOI: 10.1002/sim

6. Allore H, Tinettia M, Gill T, Peduzzi P. Experimental designs for multicomponent interventions among persons with multifactorial geriatric syndromes. *Clinical Trials* 2005; **2**(1):13–21.
7. Bluford D, Sherry B, Scanlon K. Interventions to prevent or treat obesity in preschool children: a review of evaluated programs. *Obesity* 2007; **15**:1356–1372.
8. Strecher V, McClure J, Alexander G, Chakraborty B, Nair V, Konkel J, Greene S, Collins L, Carlier C, Wiese C, Little R, Pomerleau C, Pomerleau O. Web-based smoking cessation components and tailoring depth: results of a randomized trial. *American Journal of Preventive Medicine* 2008; **34**(5):373–381.
9. Campbell M, Fitzpatrick R, Haines A, Kinmonth A, Sandercock P, Spiegelhalter D, Tyrer P. Framework for design and evaluation of complex interventions to improve health. *British Medical Journal* 2000; **321**:694–696.
10. Friedli K, King M. Psychological treatments and their evaluation. *International Review of Psychiatry* 1998; **10**:123–126.
11. Stephenson J, Imrie J. Why do we need randomised controlled trials to assess behavioural interventions? *British Medical Journal* 1998; **316**:611–613.
12. West S, Aiken L, Todd M. Probing the effects of individual components in multiple component prevention programs. *American Journal of Community Psychology* 1993; **21**(5):571–605.
13. Box G, Hunter W, Hunter J. *Statistics for Experimenters*: *An Introduction to Design*, *Data Analysis*, *and Model Building*. Wiley: New York, 1978.
14. Nair V, Strecher V, Fagerlin A, Ubel P, Resnicow K, Murphy S, Little R, Chakraborty B, Zhang A. Screening experiments and fractional factorial designs in behavioral intervention research. *American Journal of Public Health* 2008; **98**:1354–1359.
15. Collins L, Murphy S, Nair V, Strecher V. A strategy for optimizing and evaluating behavioral interventions. *Annals of Behavioral Medicine* 2005; **30**:65–73.
16. Yates F. The design and analysis of factorial experiments. *Imperial Bureau of Soil Sciences—Technical Communication No. 35*, Harpenden, 1937.
17. Fisher R. *The Design of Experiments* (3rd edn). Oliver & Boyd: Edinburgh, 1942.
18. Wu C, Hamada M. *Experiments*: *Planning*, *Analysis*, *and Parameter Design Optimization*. Wiley: New York, 2000.
19. Curhan R. The effects of merchandising and temporary promotional activities on the sales of fresh fruits and vegetables in supermarkets. *Journal of Marketing Research* 1974; **11**(3):286–294.
20. Byar D, Piantadosi S. Factorial designs for randomized clinical trials. *Cancer Treatment Reports* 1985; **69**: 1055–1063.
21. Brittain E, Wittes J. Factorial designs in clinical trials: the effects of non-compliance and subadditivity. *Statistics in Medicine* 1989; **8**:161–171.
22. Byar D. Factorial and reciprocal control designs (with Discussion). *Statistics in Medicine* 1990; **9**:55–64.
23. West S, Aiken L. Toward understanding individual effects in multicomponent prevention programs: design and analysis strategies. In *The Science of Prevention*: *Methodological Advances from Alcohol and Substance Use Research*, Bryant K, Windle M, West S (eds). American Psychological Association: Washington, DC, 1997.
24. Collins L, Dziak J, Li R. Design of experiments with multiple independent variables: a resource management perspective on complete and reduced factorial designs. *Psychological Methods* 2009; **14**.
25. Box G, Hunter J. The 2^{k-p} fractional factorial designs. *Technometrics* 1961; **3**:311–351 and 449–458.
26. Byar D, Herzberg A, Tan W. Incomplete factorial designs for randomized clinical trials. *Statistics in Medicine* 1993; **12**:1629–1641.
27. White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 1980; **48**:817–838.
28. Montgomery A, Peters T, Little P. Design, analysis and presentation of factorial randomized controlled trials. *BMC Medical Research Methodology* 2003; **3**(26):14633287.
29. Kazdin A. The evaluation of psychotherapy: research design and methodology. In *Handbook of Psychotherapy and Behavior Change* (3rd edn), Garfield S, Bergin A (eds). Wiley: New York, 1986; 23–68.
30. Tinetti M, Baker D, McAvary G, Claus E, Garrett P, Gottschalk M, Koch M, Trainor K, Horwitz R. A multifactorial intervention to reduce the risk of falling among elderly people living in the community. *New England Journal of Medicine* 1994; **331**:821–827.
31. Tinetti M, McAvary G, Claus E. Does multiple risk factor reduction explain the reduction in fall rate in the Yale FICSIT trial? *American Journal of Epidemiology* 1996; **144**:389–399.
32. Wolchik S, West S, Westover S, Sandler I, Martin A, Lustig J, Tein J, Fisher J. The children of divorce intervention project: outcome evaluation of an empirically based parenting program. *American Journal of Community Psychology* 1993; **21**:293–331.

DOI: 10.1002/sim

33. Holland P. Statistics and causal inference. *Journal of the American Statistical Association* 1986; **81**:945–970.
34. Rosenbaum P, Rubin D. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 1984; **79**:516–524.
35. Rubin D. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 1974; **66**:688–701.
36. Collins L, Chakraborty B, Murphy S, Strecher V. Comparison of a phased experimental approach and a single randomized clinical trial for developing multicomponent behavioral interventions. *Clinical Trials* 2009; **6**(1):5–15.
37. Hosking J, Cisler R, Couper D, Gastfriend D, Kivlahan D, Anton R. Design and analysis of trials of combination therapies. *Journal of Studies on Alcohol* 2005; **S15**:34–42.
38. Meyer R, Steinberg D, Box G. Follow-up designs to resolve confounding in multifactor experiments. *Technometrics* 1996; **38**:303–332.
39. Piantadosi S. *Clinical Trials*: *A Methodologic Perspective*. Wiley: New York, 2005.
40. Green S, Liu P, O'Sullivan J. Factorial design considerations. *Journal of Clinical Oncology* 2002; **20**(16): 3424–3430.
41. Cohen J. *Statistical Power for the Behavioral Sciences* (2nd edn). Erlbaum: Hillsdale, NJ, 1988.
42. Dijkstra A, DeVries H, Roijackers J, Van Breukelen G. Tailored interventions to communicate stage-matched information to smokers in different motivational stages. *Journal of Consulting and Clinical Psychology* 1998; **66**(3):549–557.
43. Strecher V. Computer-tailored smoking cessation materials: a review and discussion. *Patient Education and Counselling* 1999; **36**:107–117.
44. Strecher V, Shiffman S, West R. Randomized controlled trial of a web-based computer-tailored smoking cessation program as a supplement to nicotine patch therapy. *Addiction* 2005; **100**:682–688.
45. Swartz L, Noell J, Schroeder S, Ary D. A randomized control study of a fully automated internet based smoking cessation programme. *Tobacco Control* 2006; **15**:7–12.
46. Couper D, Hosking J, Cisler R, Gastfriend D, Kivlahan D. Factorial designs in clinical trials: options for combination treatment studies. *Journal of Studies on Alcohol* 2005; **S15**:24–32.
47. Box G, Draper N. *Empirical Model-building and Response Surfaces*. Wiley: New York, 1987.
48. Myers R, Montgomery D. *Response Surface Methodology*. Wiley: New York, 1995.
49. Smith J, Beverly J. The use and analysis of staggered nested factorial designs. *Journal of Quality Technology* 1981; **13**:166–173.
50. Ankerman B, Aviles A, Pinheiro J. Optimal designs for mixed-effects models with two random nested factors. *Statistica Sinica* 2003; **13**:385–401.
51. Searle S, Casella G, McCulloch C. *Variance Components*. Wiley: New York, 2002.
52. Daunica A, Smitha S, Branka E, Penfield R. Classroom-based cognitive-behavioral intervention to prevent aggression: efficacy and social validity. *Journal of School Psychology* 2006; **44**(2):123–139.
53. Flay B, Collins L. Historical review of school-based randomized trials for evaluating problem behavior prevention programs. *Annals of the American Academy of Political and Social Science* 2005; **599**:115–146.

DOI: 10.1002/sim