Design Options for an Evaluation of Head Start Coaching

Review of Methods for Evaluating Components of Social Interventions

OPRE Report #2014-81

July 2014

Design Options for an Evaluation of Head Start Coaching

REVIEW OF METHODS FOR EVALUATING COMPONENTS OF SOCIAL INTERVENTIONS

JULY 2014

Office of Planning, Research and Evaluation
Administration for Children and Families
U.S. Department of Health and Human Services
http://www.acf.hhs.gov/programs/opre
Wendy DeCourcey, Project Officer
Christine Fortunato, Project Specialist

American Institutes for Research
1000 Thomas Jefferson Street NW
Washington, DC 20007-3835
Eboni C. Howard, Project Director
Kathryn Drummond, Project Manager

Authors
Marie-Andrée Somers, MDRC
Linda Collins, Pennsylvania State University
Michelle Maier, MDRC

OPRE Report #2014-81

Suggested Citation

Somers, M., Collins, L., Maier, M. (2014). *Review of Experimental Designs for Evaluating Component Effects in Social Interventions.* Produced by American Institutes for Research for *Head Start Professional Development: Developing Evidence for Best Practices in Coaching*. Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation.

Disclaimer

This report and the findings within were prepared under Contract #HHSP23320095626W with the Administration for Children and Families, U.S. Department of Health and Human Services. The views expressed in this publication are those of the authors and do not necessarily reflect the views or policies of the Office of Planning, Research and Evaluation, the Administration for Children and Families, or the U.S. Department of Health and Human Services. This report and other reports are available from the Office of Planning, Research and Evaluation.

CONTENTS

Acknowledgments

This report is the result of the contributions and dedicated efforts of several individuals and organizations.

We want to thank all members of the Head Start Professional Development project team for their support, guidance, and reviews of this report: Dr. Eboni Howard (AIR) is project director; Dr. Kathryn Drummond (AIR) is project manager; Dr. James Taylor (AIR) and Dr. Chrishana Lloyd (MDRC) are experts in the design and evaluation of professional development and coaching interventions.

We are also very grateful to Dr. Wendy DeCourcey and Dr. Christine Fortunato at the U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation (OPRE). Dr. DeCourcey and Dr. Fortunato provided useful guidance and constructive advice on the content and writing of this report. We also thank their OPRE colleagues for their assistance and careful review of the report.

We are also indebted to the report's expert consultants and internal reviewers—Dr. Howard Bloom, Dr. Charles Michalopoulos, Mr. Mike Fishman, Dr. Hans Bos, and Dr. Barbara Goldman—who provided very helpful feedback on earlier drafts of this report.

Finally, we want to thank the research and administrative staff at the American Institutes for Research, MDRC, Child Trends, and MEF Associates who assisted in various ways throughout this project, including contract management, information technology, database development, conferencing, financial management, quality review, editing, and report production.

Executive Summary

This report is part of a larger effort to design a study to evaluate the effect of individual coaching components in Head Start programs. The design project is guided by the following research question: *What is the effect of individual coaching components on teachers and children in the Head Start context?* The goal is to design an evaluation that will help Head Start programs, and other early childhood programs, implement stronger coaching interventions by providing them with reliable evidence on the effect of coaching components that they can use to decide which components to implement, given their local needs and budgetary constraints. The purpose of this present report is to review different experimental designs that could be used to estimate the effect of individual components within a social intervention, such as Head Start coaching.

In the research design literature, an intervention *component* is defined as any aspect, element, or feature of an intervention that can be reasonably separated out in order to study its individual effect on the outcomes of interest. For example, in Head Start and other early childhood education settings, coaching interventions consist of multiple components that are intended to improve teacher practice and classroom quality, and ultimately child outcomes. A coaching intervention may include program components related to structure or delivery (e.g., coach credentials, coach training, coach caseload, coach supervision) and components related to the content or process of coaching (e.g., use of modeling; quantity and nature of feedback to the teacher). Each component has possible values, or *levels*. A component may be "on" or "off" in an intervention, or it can take on varying levels of intensity (e.g., "low" versus "high").

Unfortunately, there is little rigorous evidence on the effect of individual intervention components. For this reason, decisions about which components to include in a social intervention such as coaching are based primarily on theory and professional experience about which components are likely to matter, rather than empirical evidence on the effect of these components. This means that social interventions may not be as effective or as cost-effective as they could be. If the effect of individual components were known *a priori*, this information could be used to design interventions that are not only more effective but also less time consuming and more economical. In order to build interventions that have maximum impact, and that are flexible to local context and needs, evaluation science needs to move towards policy experiments that test the effect of individual intervention components.

Accordingly, the goal of this report is to review potential experimental design options that could be used to estimate the effect of individual coaching components in Head Start. Five experimental designs are discussed: factorial designs, comparative treatment designs, the individual experiments design, crossover designs, and adaptive clinical trials. The differences between these designs are elucidated in terms of how well they can answer the study's research question; their sample size requirements; the number of experimental conditions that would have to be implemented; and whether interactions between components can be estimated.

The main conclusion from this review is that a factorial design is the strongest experimental design for evaluating the effect of individual intervention components, such as coaching components in the Head Start context**.** Factorial designs provide findings that are useful for policymakers and practitioners who are creating or adapting interventions in the field because they account for—and provide information on—interaction effects between components. Although evaluators often disregard factorial designs because they require many more experimental conditions than other designs, they also require a smaller sample size to statistically detect a component effect of a given magnitude. This can outweigh the disadvantage and cost of having to implement a larger number of conditions. The other four designs reviewed in this report are more suitable for different purposes—namely comparing the effect of different intervention models (as opposed to components) or estimating the effect of a single component.

The report concludes by describing several issues that need to be considered when designing a study of component effects, regardless of which experimental design is used. A unique challenge with studies of component effects is that the expected effect of a single component is likely to be smaller in magnitude than the effect of an entire intervention. This means that the total sample size needed for a study of component effects will likely be larger than the sample size needed for an evaluation of a complete intervention, and therefore it becomes especially important to use strategies to improve statistical power (e.g., use of baseline covariates, choosing a lower level of random assignment, using well-aligned and reliable outcome measures, etc.). In a study of component effects, evaluators must also decide whether or not to "fix" the levels of non-tested components, and they must gauge the study's feasibility in the field, because a test of multiple components is more operationally complex than an evaluation of a single intervention.

I. Introduction

This report is part of a larger effort to design a study to evaluate the effect of individual coaching components in Head Start programs. The design project is guided by the following research question: *What is the effect of individual coaching components on teachers and children in the Head Start context?* The goal is to design an evaluation that will help Head Start programs, and other early childhood programs, implement stronger coaching interventions by providing them with reliable evidence on the effect of coaching components that they can use to decide which components to implement, given their local needs and budgetary constraints.

As a first step in this effort, the project team identified a list of coaching components that could be systematically varied and studied in order to determine the degree to which each component improves outcomes at the program, teacher, and/or child level (Taylor et al., 2013). The next step in the project is to choose a study design that can informatively evaluate the effect of these individual components.

Accordingly, the goal of this report is to review potential experimental design options that could be used to estimate the effect of individual coaching components. Five experimental designs are discussed: factorial designs, comparative treatment designs, the individual experiments design, crossover designs, and adaptive clinical trials. The differences between these designs are elucidated in terms of how well they can answer the study's research question; their sample size requirements; the number of experimental conditions that would have to be implemented; and whether interactions between components can be estimated.

The main conclusion from this review is that a factorial design is the strongest experimental design for evaluating the effect of individual intervention components, such as coaching components in the Head Start context**.** Factorial designs provide findings that are useful for policymakers and practitioners who are creating or adapting interventions in the field because they account for—and provide information on—interaction effects between components. Although evaluators often disregard factorial designs because they require many more experimental conditions than other designs, they also require a smaller sample size to statistically detect a component effect of a given magnitude. This can outweigh the disadvantage and cost of having to implement a larger number of conditions. The other four designs reviewed in this report are more suitable for other purposes—namely comparing the effect of different intervention models (as opposed to components) or estimating the effect of a single component.

A. Rationale for a Study of Component Effects

Social interventions typically consist of multiple components that are bundled together with the goal of improving the outcomes of individuals or groups. In the evaluation literature, an intervention *component* is defined as any aspect, element, or feature of an intervention that can be reasonably separated out in order to study its individual effect on the outcomes of interest. A component can be related to the content of the services being provided; it can be a strategy that promotes compliance or adherence to an intervention strategy; or it can be an aspect of program delivery. For example, in Head Start and other early childhood education programs, coaching interventions consist of multiple components that are intended to improve teacher practice and classroom quality, and ultimately child outcomes. A coaching intervention may include program components that are related to its structure or delivery (e.g., coach credentials, coach training, coach caseload, coach supervision) and program components that are related to the content or process of coaching (e.g., use of modeling, quantity and nature of feedback to the teacher). Each component also has possible values or *levels*. A component may be "on" or "off" in an intervention—for example, a teacher assessment form (e.g., a standardized assessment tool) that may or may not be administered by the coach in in a coaching intervention. Components can also take on varying levels of intensity (e.g., minimal versus intensive, or "low" versus "high"). For example, a coaching intervention for Head Start teachers can be minimally or intensively coordinated with other professional development activities provided to teachers.

Impact evaluations of complex social interventions are now relatively common, and randomized controlled trials (RCTs) have become the gold standard for evaluating a social intervention's effects. In this type of study design, program participants are randomly assigned to either a treatment group that receives the intervention or to a control group that does not, and then the average outcomes of the treatment and control group are compared to evaluate the intervention's average effects. The growing use of RCTs to evaluate social interventions has been instrumental in providing rigorous findings to policymakers and practitioners about the effect of different interventions.

Yet unfortunately, there is relatively little strong, empirical evidence about the effect of individual intervention components (Green, Ha, and Bullock, 2010). When component effects are examined in evaluation research, it is typically done post-hoc using nonexperimental methods (Baker et al., 2011; Collins, Murphy, Nair, and Strecher, 2005). For example, after the impact of a social intervention has been estimated using an RCT, exploratory analyses are often conducted to examine whether intervention effects interacted with the implementation of particular program features.[1] The problem with this approach is that any conclusions drawn about individual intervention components are weak because they may be due to alternative explanations.

Given this lack of empirical evidence on component effects, decisions about *which* components to bundle together in a social intervention are based primarily on theory and professional experience about which elements are likely to matter. This evaluation paradigm – characterized by a narrow focus on testing whole interventions rather than individual program components – leaves evaluators, policymakers, and program developers with a "black box".

[1] For example, the Evaluation of Enhanced Academic Instruction in After-School Programs (Black, Doolittle, Zhu, Unterman, & Grossman, 2008) examined the extent to which program impacts were correlated with program implementation characteristics (e.g., number of days the after-school program was offered, attrition rates of program staff, etc.).

Although the effect of social interventions is often known, it is difficult to determine *which* particular components of an intervention are more important.

This means that social interventions may not be as effective or as cost-effective as they could be. If the effect of individual components were known *a priori*, this information could be used to design interventions that are not only more effective but also less time consuming and more economical. In order to build interventions that have maximum impact, and that are flexible to local context and needs, evaluation science needs to move towards policy experiments that test the effect of individual intervention components.

This is, in essence, the approach proposed by the multiphase optimization strategy (MOST) (Collins, Dziak, & Li, 2009; Collins et al., 2005).[2] The MOST framework is a staged approach to developing social interventions that is generating growing interest among intervention researchers, most notably in public health. The framework begins with an "optimization" experiment, the goal of which is to estimate the effect of individual intervention components on the outcomes of interest. The findings from this experiment are then used to design an optimal intervention model. The components included in this optimal model are chosen based on a pre-specified optimization criterion—for example, those that meet some minimum threshold for effect size or cost effectiveness are selected for inclusion. In a later stage, the impact of this optimal model is evaluated against a control or business as usual condition using a standard two-group RCT.

One of the most distinctive features of the MOST approach is the optimization experiment. This phase uses experimentation (as opposed to theory, practitioner experience, and minimal empirical evidence) to develop an evidence base for deciding which components to include in an intervention. MOST is informed by the *resource management principle*, which asserts that research resources should be strategically managed so that the amount of reliable information gained is maximized and the field is pushed forward at the fastest possible pace. Thus for the optimization experiment in particular, evaluators using the MOST approach must choose the experimental design that enables them to address their most important research questions, given their limited resources. The larger design project for which this report was prepared aligns with the spirit of the first phase of the MOST approach, in particular identifying a strong research design for estimating component effects.

B. Report Structure and an Illustrative Example

The remainder of the report is structured as follows. Section II begins by discussing factorial designs, which have been emphasized in research utilizing the MOST framework because they are strong and efficient designs for examining component effects (although they are usually overlooked in social interventions). Section III reviews four other experimental designs that are used to examine component effects: comparative treatment designs, individual

[2] For applications, see Caldwell et al. (2012); Collins et al. (2011).

experiments, crossover designs, and adaptive clinical trials. These types of design are most common in evaluations of social interventions and in medical research. Section IV will discuss several issues that should be considered when designing a study of component effects, regardless of the selected experimental design. Section V concludes by summarizing the key differences between the experimental designs reviewed in this report. A glossary of key terms presented in this report is provided in Appendix A.

When describing and comparing the design options in this report, we will use a simple illustrative example. As summarized in Exhibit 1, our example includes five hypothetical coaching components. In our hypothetical study, the levels of these five components will be manipulated and made to vary randomly across Head Start centers, in order to estimate the effect of the "higher" level of each component relative to the "lower" level of each component. With respect to the choice of levels, it is important to point out that some of these five components have multiple reasonable levels. For example, a component such as "coaches' use of modeling" could vary from none to any number of times per month. However, testing the effect of all of these levels would be inefficient and likely impractical, so in practice one has to choose a subset of levels to compare. As will be discussed in greater detail later (Section II-D, p.21), choosing *two* levels for each component can minimize the sample size requirements for the study.[3] The two levels of a component ("low" versus "high", "minimal" versus "intensive", or "on" versus "off"),   should be chosen to reflect levels that practitioners or policymakers would want to implement in practice, and that are feasible in a real-world context. The "low" setting, for example, could be the average level of a coaching component as currently implemented by Head Start centers, while the "high" level could be set to a level that is higher than the current level yet still practical and affordable.

In our hypothetical example, the five coaching components and their levels are the following:

1. **Staff targeted: *lead teacher* versus *teaching team***

   Coaching interventions can target their coaching to lead classroom teachers only or to the classroom teaching team (i.e., the lead teacher and assistant teacher together).

   Low level: Lead teacher only

   High level: Teaching team

2. **Delivery mode: *on-site and web-based* versus *on-site***

[3] Components can take on more than two levels, and more than two levels can be tested in an experiment. Increasing the number of levels a component takes on, however, exponentially complicates the evaluation because it increases the number of experimental conditions that need to be implemented. If there are three possible levels, for example, it may be more resource-efficient to start by testing the lowest of the three levels against the highest to make sure that there is an impact when the contrast is maximized. Later tests can then be conducted to pinpoint the level that has the greatest impact.

Coaching interventions can be delivered in a variety of ways, such as on-site (i.e., in person), online through a web-based communication system, or some combination of the two.

Low level: Half of the coaching sessions are conducted on-site and the other half are conducted online

High level: All coaching sessions are conducted on-site

3. **Exposure to modeling: *minimal* versus *intensive***

   Coaching interventions can use modeling to demonstrate positive teaching practices. This can be done minimally (where modeling takes place during the coaching session when the coach deems it necessary) or intensively (where coaches are trained to explicitly model to teachers on a monthly basis, to make sure that teachers are intentionally observing the modeling, and to debrief and reflect on the modeling).

   Low level: Minimal exposure to modeling

   High level: Intensive exposure to modeling

4. **Use of assessment tools: *none* versus *explicit use of tools and differentiation***

   Coaching interventions may or may not use global or specific formal assessment tools to identify teacher needs and then use the information resulting from the tool's use to explicitly differentiate their coaching across teachers.

   Low level: No use of tools

   High level: Explicit and standardized use of tools

5. **Supervision: *minimal* versus *intensive***

   Coaching interventions can have a coach supervision component, where coaches report to a director who conducts performance assessments and monitors coaches to ensure they deliver the coaching model as intended. The amount of time coaches spend in supervision may be minimal or more intensive.

   Low level: One hour of supervision per coach per month

   High level: One hour of supervision per coach per week

For our hypothetical example, we will assume that Head Start *centers* are randomized to experimental conditions. This means that all coaches in a given Head Start center will be assigned to implement the same set of coaching components. (Note that this is for illustrative purposes only; in practice, it might be preferable to randomize coaches instead of centers. At the end of this report in Section IV, p. 39, we will discuss factors that affect the choice of randomization unit.)

In this report, we also assume that the objective of the study is to generate findings that will help practitioners and program operators choose the most effective and/or cost-effective components for their local program. Ideally, the findings should also provide information about how the effect of these components is likely to vary when they are embedded into an existing coaching program. By extension, this means that the study must be able to (1) provide estimates of component effects that are not sensitive to the overall intervention strategy into which they are being embedded, and (2) provide information on the interactions between these components. The remainder of this report will review the extent to which different study designs are able to meet these objectives. Although the examples referenced in this report focus on coaching in Head Start, the issues and designs described are also applicable to evaluating program component effects in other types of social intervention.

Exhibit 1. Hypothetical Coaching Components Used to Illustrate the Experimental Designs

	Factor	Levels		Component
Component	Name	Lower Level	Upper Level	Type
Targeted staff	*TARGET*	Lead teacher	Teaching team (teacher and aide)	Structural
Delivery mode	*MODE*	Mix of on-site and online coaching sessions	On-site coaching sessions only	Structural
Coaches' use of modeling to demonstrate positive teaching practices	*MODELING*	Minimal, with no reflection	Intensive, with time for reflection	Process/content
Coaches' use of assessment tools to identify needs and differentiate coaching	*TOOLS*	None	Explicit	Process/content
Amount of coach supervision	*SUPER*	Minimal	Intensive	Staffing

II. Factorial Designs

This section begins by introducing the basic factorial design and demonstrating how it can be used to estimate the effect of the five components in our hypothetical scenario (that is, the effect of the upper level of each of the hypothetical components in Exhibit 1 relative to its lower level). We then describe two variants of the basic factorial design that may be more operationally feasible to implement in certain situations: fractional factorial designs and multiple factorial experiments. The section concludes by discussing several general topics and questions related to factorial designs.

A. An Introduction to Factorial Designs

We can introduce the concept of a factorial experiment by starting with a simple hypothetical example. Suppose we are interested in examining the first two intervention components in our hypothetical scenario (Exhibit 1, p.6): targeted staff and the delivery mode. In a factorial experiment, each of these components becomes an independent variable that is manipulated by the evaluator. We will call the independent variable corresponding to targeted staff *TARGET*. *TARGET* has two levels: "lead teacher only" versus "teaching team." The independent variable corresponding to the delivery mode will be called *MODE* and has two levels: "mix of online and in-person" and "on-site only." Each of these two independent variables is referred to as a *factor* in the experiment.[4]

Exhibit 2 illustrates a factorial experiment based on these two factors. As shown in the exhibit, there would be four experimental conditions, representing all four possible combinations of levels of the two factors. In this example, Head Start centers would be randomly assigned to be in one (and only one) of these experimental conditions. In centers randomly assigned to Condition 1, coaches would work only with the lead teacher and the coaching would be delivered on-site only. In contrast, in centers randomly assigned to Condition 4, coaches would work with the entire teaching team, with coaching delivered both in person and online.

Exhibit 2. Illustration of a 2x2 Factorial Experiment, Based on Two Hypothetical Components

		Staff Targeted (*TARGET*)	
		Lead teacher	Teaching team
Delivery Mode	On-site	Condition 1	Condition 2
	Mixed	Condition 3	Condition 4

Factorial designs are commonly described in terms of the number of factors in the experiment, and the number of levels in each factor. An informal notation is often used in which the number of levels of each factor is multiplied. For example, if there was one factor with three levels and one factor with two levels, the design would be called a 3×2 factorial, showing that there are 3×2=6 experimental conditions in that design. The experiment in Exhibit 2 is called a 2×2 factorial, because the first and second factors each have two levels. In statistics, a somewhat more formal and compact exponential notation is used to describe factorial experiments. Using this notation, the design in Exhibit 2 would be labeled a 2^2 factorial. This more compact notation is helpful in experiments that involve many factors (as will be evident later).

A desirable feature of the factorial design is the *balance* property. To achieve balance, it is necessary to meet two criteria: (1) each experimental condition must have the same number of subjects, and (2) each level of each factor must appear the same number of times with each level

[4] Although independent variables in a factorial experiment can have more than two levels, using more than two levels increases the sample size requirements for the study, as will be discussed later.

of every other factor. As an example of the second condition, notice that in Exhibit 2, the "lead teacher" level of *TARGET* appears once with the "On-site" level of *MODE* and once with the "mix" level of *MODE*. Similarly, the "teaching team" level of *TARGET* appears once with each level of *MODE*; the "on-site" level of *MODE* appears once with each level of *TARGET*; and the "mix" level of *MODE* appears once with each level of *TARGET*. Violations of the first condition are common and usually not very difficult to address. However, violations of the second condition are usually much more serious and can render a design more resource-intensive than it otherwise would be.[5] Balance is desirable because it minimizes the sample size needed for the study, as will be explained below.

The data from a factorial experiment are typically analyzed using an Analysis of Variance (ANOVA). There are different ways to set up an ANOVA, but in the classic approach that is covered in most statistics textbooks, two different kinds of effects are estimated. The first is called the *main effect*. The main effect of a factor is the effect of that factor averaged across all the levels of all the other factors in the experiment. In the example, the main effect of *TARGET* is the difference between the average of the conditions in which *TARGET* is set to "lead teacher" (Conditions 1 and 3) and the average of the conditions in which *TARGET* is set to "teaching team" (Conditions 2 and 4). Similarly, the main effect of *MODE* is the difference between the average of the conditions in which *MODE* is set to "mix" (Conditions 3 and 4) and the average of the conditions in which *MODE* is set to "on-site" (Conditions 1 and 2).

Main effects have three characteristics that are worth noting. First, in a factorial ANOVA, the main effect is found by comparing means based on combinations or *groups* of experimental conditions, not by directly comparing the means of individual conditions. Second, the two main effect estimates are based on different combinations of the same four experimental conditions. Thus, although the main effects are different, each is based on the entire sample *N*. (These points will be important when we discuss statistical power below.) Third, the main effect of a particular factor is defined as the effect across all the levels of all other factors in the experiment, not at any one particular level of another factor.

The second type of effect that is typically estimated in an ANOVA is the *interaction*. Two factors are said to interact when the size of the effect of one factor varies depending on the level of the other. In our example, there would be a *TARGET ×MODE* interaction if the effect of *TARGET* when *MODE* is set to "on-site" differs from the effect when *MODE* is set to "mix." If the effect of *TARGET* is the same no matter what *MODE* is set to, there is no *TARGET ×MODE* interaction. Formally, we define the two-way interaction between Component A and Component B as the effect of Component A *when Component B is set to its upper level* minus the effect of Component A when *Component B is set to its lower level*.[6] Describing how interactions are estimated is beyond the scope of this document, but interested readers can refer to Kugler, Trail,

[5] For a more detailed discussion, see Collins et al. (2009).

[6] In some fields of research, a two-way interaction effect is defined as *half* of this value. However, this alternative definition of a two-way interaction is much less useful from a policy or evaluation perspective.

Dziak, and Collins (2012) for more information. (Please note that here we are referring exclusively to interactions between factors in the experimental design.[7])

Now suppose that there are five intervention components to examine. In addition to the Head Start teaching staff targeted by the coaching and the delivery mode, the evaluator wishes to examine three other components: the use of modeling by the coach, the coach's use of tools, and the supervision of the coach. (See Exhibit 1 on page 6 for a complete list of components and levels.) This can be accomplished by using a 2^5 factorial experiment, illustrated in Exhibit 3. In the experimental design in Exhibit 3, *TARGET* and *MODE* have the same levels as before. The factor corresponding to the modeling of good teaching practices will be called *MODELING*. The factor corresponding to a coach's standardized use of assessment tools will be called *TOOLS*. Finally, the factor corresponding to coach supervision will be called *SUPER*

Many more effects can be estimated based on the design in Exhibit 3 than can be estimated based on the design in Exhibit 2. The design in Exhibit 3 can be used to estimate the main effect of each of the five factors; 10 two-way interactions; 10 three-way interactions; 5 four-way interactions; and 1 five-way interaction, for a total of 31 effects. In the multi-factor case, the concept of main effects and interactions remains essentially the same as in the two-factor case, as is shown for main effects in Exhibit 4. The main effect of *TARGET* is now found by comparing the mean of all the conditions in which *TARGET* is set to "teaching team" (Conditions 16—32) against the mean of all the conditions in which *TARGET* is set to "teacher only" (Conditions 1—16). The main effects of the other four factors are found in a similar fashion, by comparing the mean of half of the conditions against the mean of the remaining half. As shown in Exhibit 4, the conditions are "reshuffled" or regrouped to produce a unique estimate of each effect.

Evaluators who are immersed in the tradition of the standard two-group RCT sometimes have difficulty conceptualizing how effects are estimated in a factorial design. To better understand the distinction between the RCT and the factorial design, it is important to remember that their goals are fundamentally different. The goal of an RCT is to estimate the impact of an

[7] Evaluators are also often interested in investigating interactions between the experimental factors and non-manipulated variables, such as the characteristics of participants. In our hypothetical study, for example, assume that the effect of coach credentials will not be tested in the study, but that the effect of the other components may depend on (be moderated by) coach experience. For instance, it might be hypothesized that coaches with at least 10 years of experience will implement a coaching component differently than coaches with less experience. One approach to investigating moderation is to conduct the factorial experiment and to let the untested component or moderator (coach experience) vary randomly across the sample (Head Start centers). The researcher could then conduct post-hoc analyses modeling interactions between the moderator (coach experience) and the tested components in the experiment, or similarly, could estimate effects by coach subgroups defined by their experience level. Another approach is to plan to sample half of the coaches with at least 10 years of experience and half of the coaches with less than 10 years of experience, and then conduct the experiment in the usual way, essentially conducting it once for the more experienced group and once for the less experienced group. It would then be possible to conduct a power analysis and ensure that the overall sample size is large enough to provide sufficient power to detect any interactions involving coach experience. This approach is more definitive than the first approach, but it might be more costly.

entire intervention (or package of components). Subjects are randomly assigned either to a treatment condition, in which subjects receive the intervention being tested (and all components packaged within that intervention), or to a control condition, in which subjects do not receive the intervention. The outcomes of these two groups of subjects are then directly compared to evaluate the impact of the intervention. Viewing the factorial design through the lens of an RCT can hinder the evaluator's understanding of the factorial experiment because the latter design is conceptually different. The objective of a factorial experiment is to estimate the main effect of individual intervention components and their interactions, and therefore two individual

Exhibit 3. Illustration of a 2^5 Factorial Design, Based on Five Hypothetical Components

Experimental Condition	Tested Components				
	Targeted Staff (*TARGET*)	Delivery Mode (*MODE*)	Use of Modeling (*MODELING*)	Assessment Tools (*TOOLS*)	Supervision of Coach (*SUPER*)
1	Lead teacher	On-site	Minimal	None	Minimal
2	Lead teacher	On-site	Minimal	None	Intensive
3	Lead teacher	On-site	Minimal	Explicit	Minimal
4	Lead teacher	On-site	Minimal	Explicit	Intensive
5	Lead teacher	On-site	Intensive	None	Minimal
6	Lead teacher	On-site	Intensive	None	Intensive
7	Lead teacher	On-site	Intensive	Explicit	Minimal
8	Lead teacher	On-site	Intensive	Explicit	Intensive
9	Lead teacher	Mix	Minimal	None	Minimal
10	Lead teacher	Mix	Minimal	None	Intensive
11	Lead teacher	Mix	Minimal	Explicit	Minimal
12	Lead teacher	Mix	Minimal	Explicit	Intensive
13	Lead teacher	Mix	Intensive	None	Minimal
14	Lead teacher	Mix	Intensive	None	Intensive
15	Lead teacher	Mix	Intensive	Explicit	Minimal
16	Teaching team	Mix	Intensive	Explicit	Intensive
17	Teaching team	On-site	Minimal	None	Minimal
18	Teaching team	On-site	Minimal	None	Intensive
19	Teaching team	On-site	Minimal	Explicit	Minimal
20	Teaching team	On-site	Minimal	Explicit	Intensive
21	Teaching team	On-site	Intensive	None	Minimal
22	Teaching team	On-site	Intensive	None	Intensive
23	Teaching team	On-site	Intensive	Explicit	Minimal
24	Teaching team	On-site	Intensive	Explicit	Intensive
25	Teaching team	Mix	Minimal	None	Minimal
26	Teaching team	Mix	Minimal	None	Intensive
27	Teaching team	Mix	Minimal	Explicit	Minimal
28	Teaching team	Mix	Minimal	Explicit	Intensive
29	Teaching team	Mix	Intensive	None	Minimal
30	Teaching team	Mix	Intensive	None	Intensive
31	Teaching team	Mix	Intensive	Explicit	Minimal
32	Teaching team	Mix	Intensive	Explicit	Intensive

Note. Shading denotes the upper level of the tested component; unshaded cells represent the lower level of the component.

Exhibit 4. How Main Effects Are Estimated in the 2^5 Factorial Experiment

	Compare Subjects in the Following Two Sets of Experimental Conditions:	
Main Effect of:	Set A	Set B
Targeted Staff (*TARGET*)	1–15	16–32
Delivery Mode (*MODE*)	1–8, 17–24	9–16, 25–32
Use of Modeling (*MODELING*)	1–4, 9–12, 17–20, 25–27	5–8, 13–16, 21–24, 29–32
Use of Tools (*TOOLS*)	1–2, 5–6, 9–10, 13–14, 17–18, 21–22, 25–26, 29–30	3–4, 7–8, 11–12, 15–16, 19–20, 23–24, 27–28, 31–32
Coach Supervision (*SUPER*)	Odd numbers	Even numbers

Note. The numbers in this table refer to the experimental conditions in Exhibit 3.

experimental conditions are not directly compared to each other. Instead, *groups* of conditions are compared, as illustrated in Exhibit 4.[8] Thus, each effect estimate is based on a unique combination or grouping of all experimental conditions.

This leads to an important property of the factorial experiment—each main effect is estimated using the entire sample *N*. This means that the statistical power of a factorial experiment is determined by the total sample size *N*, and *not* the sample size per experimental condition *n*. To demonstrate this, we return to the experiment in Exhibit 2 (p.7), which examines two coaching components and has four experimental conditions. Suppose that a total sample size *N* of 160 Head Start centers provides the desired level of statistical power in this hypothetical example. With this sample size, each of the experimental conditions will have a per-condition sample size (*n*) of 40 (=160/4). However, because each effect estimate is based on a combination of all the experimental conditions, the entire sample of *N* centers is used for each effect estimate.

Now consider the experiment in Exhibit 3 (p.11). This one has many more experimental conditions—32 compared to the 4 in Exhibit 2. At first glance, it may seem that because it has so many more experimental conditions, it must also require many more subjects. However, perhaps surprisingly, this is not necessarily the case. By referring to Exhibit 4, one can see that the main effect of *MODELING* is based on all 32 experimental conditions, and thus based on the entire study sample *N*. The same logic applies for the main effect of *COACH* and the main effect of *SUPER*.[9] This means that for a given sample size *N*, the larger factorial design in Exhibit 3 will achieve the same statistical power as the smaller factorial design in Exhibit 2, even though the sample size for each experimental condition is smaller in the former design (for a sample size of 160, *n* is 5 compared to 40).

[8] For a more detailed discussion, see Collins, Dziak, Kugler, and Trail (submitted).

[9] These figures are approximate; when the number of factors is increased, the error degrees of freedom are reduced, which can result in a very small decrease in power that can be dealt with by a minimal increase in sample size.

It is important to make two central points here. First, it is quite common for factorial experiments to be adequately powered, despite having per-condition *n*s that are small relative to those that would be required in an RCT. This is because, all else being equal, the power for a factorial experiment is determined by the overall *N*, not the per-condition *n.* Second, factorial experiments make extremely efficient use of experimental subjects. In this hypothetical example, an investigation of five intervention components would require essentially the same *N* as an investigation of only two intervention components. In fact, when considering a factorial experiment, it may be helpful to remember that it is often possible to increase the number of factors, and thereby increase the scientific yield of the experiment, without increasing the sample size by more than a minimal amount. Of course, there is a trade-off; every time a factor is added to the design, it becomes necessary to implement many more experimental conditions. Adding factors also increases the severity of the multiple hypothesis testing problem—more hypothesis tests will be conducted (one for each factor at minimum), which increases the risk of concluding that a component is effective when in fact it is not. (Multiple hypothesis testing is discussed at greater length in Section IV, p. 43).

Exhibit 5. Example of a Study Using a Factorial Design: The HealthWise Study

Intervention: *HealthWise South Africa: Life Skills for Adolescents (HW)* is an evidence-based substance use and sexual risk prevention program that emphasizes the positive use of leisure time. The purpose of the study is to evaluate the effect of three factors hypothesized to affect the quality and fidelity of HW implementation.

Experimental Design: The effect of the three components was examined in a 2^3 factorial experiment. The three components (factors) in the experiment were: enhanced teacher training (“standard” versus “enhanced”); teacher structure, support, and supervision (“not provided” versus “provided”); and enhanced school climate (“not provided” versus “provided”).

Target Population and Sample Size: Fifty-six schools in the Cape Town area were randomly assigned to one of the eight experimental conditions.

Outcomes of Interest: Outcomes of interest included adherence to the intervention delivery protocol and quality of delivery.

Findings: This experiment is currently in the field so there are no results as of yet.

Further Reading: Caldwell, L. L., Smith, E. A., et al. (2012). Translational research in South Africa: Evaluating implementation quality using a factorial design. *Journal of Research and Practice in Children's Services, 41*(2): 119-136.

The idea of conducting factorial experiments to examine the effect of intervention components in field settings is still relatively new, but studies are beginning to emerge. An example of a factorial experiment in an educational setting is the HealthWise study (see Exhibit 5 for a summary). The purpose of the study was to evaluate three strategies for improving the implementation fidelity of HealthWise, a school-based drug abuse and HIV prevention program for children in South Africa.[10] A 2^3 factorial design was used to evaluate the effect of these three strategies—with 56 schools randomized to eight experimental conditions—demonstrating that it is feasible to use this design in field settings.

B. Fractional Factorial Designs

As discussed, factorial experiments make very efficient use of experimental subjects, but they are costly in a different way: they can require implementation of many experimental conditions and combinations of components. A comparison of Exhibits 2 and 3, in the previous section, shows that as the number of factors increases, the number of experimental conditions increases rapidly. Although a 2^5 factorial may require no more subjects than a 2^2 factorial, it requires eight times as many experimental conditions.

The *fractional factorial design* is a variation on the factorial design that can be more economical. In a fractional factorial design, only some of the experimental conditions—a fraction of the conditions, which is where the name "fractional factorial" comes from—are implemented.

Given a particular number of factors, there are usually several different fractional factorial designs from which to choose. When five dichotomous factors are to be examined, for example, the evaluator can choose a design with 16 conditions, which is called a half fraction (because 16 is half of the conditions required for a complete factorial), or a design with eight conditions, which is called a quarter fraction. Exhibit 6 shows a half fraction design that could be used to examine the five intervention components in our hypothetical example. The design in Exhibit 6 is referred to as a 2^{5-1}. This notation conveys the following information: (1) the corresponding complete factorial is a 2^5; (2) this fractional factorial design is a half fraction because it involves 2^{-1}, or ½, of the conditions in the complete factorial; and (3) the fractional factorial design has $2^{5-1} = 2^4 = 16$ experimental conditions.

An important feature of a fractional factorial design is that it has exactly the same overall sample size requirements as a complete factorial design. For a given sample size, the per condition *n* is larger in a fractional factorial experiment, because the overall *N* is distributed across fewer experimental conditions. However, the statistical power of a fractional factorial design is the same as in a complete factorial experiment. This is because a fractional factorial experiment estimates the main effect of each factor using the entire study sample, like a complete factorial experiment.

A related feature of fractional factorial designs is that they preserve the balance property. For example, examination of Exhibit 6 shows that the "on-site" level of *MODE* appears exactly four times at the "lead teacher" level of *TARGET*, and exactly four times at the "teaching team" level of *TARGET*. This holds for every level of every factor at every level of every other factor.

[10] See Caldwell et al. (2012).

Because fractional factorial designs preserve the balance property, they are as efficient as complete factorial designs in their use of subjects, but they require fewer experimental conditions to be implemented.

Exhibit 6. Illustration of a $2^{5\text{-}1}$ Fractional Factorial Experiment, Based on Five Hypothetical Components

Experimental Condition		Tested Components				
In this design	In 2^5 design (Exhibit 3)	Targeted Staff (*TARGET*)	Delivery Mode (*MODE*)	Use of Modeling (*MODELING*)	Use of Assessment Tools (*TOOLS*)	Supervision of Coach (*SUPER*)
1	2	Lead teacher	On-site	Minimal	None	Intensive
2	3	Lead teacher	On-site	Minimal	Explicit	Minimal
3	5	Lead teacher	On-site	Intensive	None	Minimal
4	8	Lead teacher	On-site	Intensive	Explicit	Intensive
5	9	Lead teacher	Mix	Minimal	None	Minimal
6	12	Lead teacher	Mix	Minimal	Explicit	Intensive
7	14	Lead teacher	Mix	Intensive	None	Intensive
8	15	Lead teacher	Mix	Intensive	Explicit	Minimal
9	17	Teaching team	On-site	Minimal	None	Minimal
10	20	Teaching team	On-site	Minimal	Explicit	Intensive
11	22	Teaching team	On-site	Intensive	None	Intensive
12	23	Teaching team	On-site	Intensive	Explicit	Minimal
13	26	Teaching team	Mix	Minimal	None	Intensive
14	27	Teaching team	Mix	Minimal	Explicit	Minimal
15	29	Teaching team	Mix	Intensive	None	Minimal
16	32	Teaching team	Mix	Intensive	Explicit	Intensive

Note. Shading denotes the upper level of the tested component; unshaded cells represent the lower level of the component. The fractional factorial design shown here is a Resolution V design, which means that main effects are only aliased with four-way interactions and higher (main effects are never aliased with two-way or three-way interactions). In addition, two-way interactions are never aliased with each other or with main effects.

Therefore, the real economy of the fractional factorial design comes from the reduction in the number of experimental conditions. This means that fractional factorial designs are worth considering when, as is often the case in intervention science, implementing and monitoring many experimental conditions would be excessively resource intensive or logistically difficult. Assuming each factor has two levels, fractional factorial designs require half or fewer of the experimental conditions required by a corresponding complete factorial. In other words, fractional factorial designs can potentially cut the study costs associated with implementing experimental conditions by half or more, while keeping costs associated with subjects the same.

Although fractional factorial designs are now just emerging in the social sciences, they have been used routinely for decades in fields such as engineering. Fractional factorial experiments have also been used successfully in health field settings, most notably to develop interventions in the smoking cessation area. One study, described in Strecher et al. (2008), involves the use of a 2^{5-1} fractional factorial experiment to examine components of an internet-delivered smoking cessation intervention. Other examples in the smoking cessation literature are Collins et al. (2011) and McClure et al. (2012). It may be helpful to note that these articles report experiments involving between 8 and 32 experimental conditions, all of which were implemented or in the process of being implemented in field settings. This provides evidence that, with careful planning and the appropriate staff, it is possible to conduct studies with many experimental conditions.

Of course, there are trade-offs that must be taken into account when using a fractional factorial design (rather than a complete factorial design). It was mentioned earlier that data from a 2^5 factorial design can be used to estimate up to 31 main effects and interactions. In general, if there are *k* experimental conditions in a factorial experiment, *k*-1 main and interaction effects can be estimated. In a fractional factorial design, *k* is smaller than it would be in the corresponding complete factorial, so fewer effects can be estimated. This happens because some effect estimates are combined or bundled together and the bundled effects can no longer be disentangled from each other. We will refer to this as *aliasing*, which is a term used in engineering. Fortunately, for every balanced fractional factorial design, the effects that are aliased together are known and documented.

To demonstrate the concept of aliasing, we can turn back to our hypothetical example. Exhibit 7 shows the aliasing structure of the fractional factorial design in Exhibit 6 (p.15). As seen in this table, each main effect in the fractional factorial design is aliased (bundled) with one four-way interaction, and each two-way interaction is aliased with one three-way interaction. For example, the main effect of *TARGET* is aliased with the *MODE×MODELING×TOOLS×SUPER* interaction. This means that if this experiment was conducted and the data analyzed, there would be a single estimate for the combination of the main effect of *TARGET* and the *MODE×MODELING×TOOLS×SUPER* interaction. It would not be possible to disentangle these two effects from each other without additional experimentation.

Aliasing occurs whenever conditions are removed from a factorial experiment, and therefore is present not only in fractional factorial experiments, but also in other study designs including the ones reviewed in the next section of this report.[11] In fractional factorial designs, the fraction reveals the number of effects that are aliased together. The design in Exhibit 6 is a half fraction, and therefore each effect is aliased with one other effect; in other words, the effects are estimated in bundles of two. In designs that are quarter fractions, each effect is aliased with three other effects; in designs that are eighth fractions, each effect is aliased with seven other effects; and so on.

[11] See Collins et al. (2009) for a further discussion on aliasing.

Exhibit 7. Aliasing in the 2^{5-1} Design Based on the Five Hypothetical Components

This effect…	Is aliased with this effect…
Main effects	
TARGET	*MODE×MODELING×TOOLS×SUPER*
MODE	*TARGET×MODELING×TOOLS×SUPER*
MODELING	*TARGET×MODE×TOOLS×SUPER*
TOOLS	*TARGET×MODE×MODELING×SUPER*
SUPER	*TARGET×MODE×MODELING×TOOLS*
Two-way interactions	
TARGET×MODE	*MODELING×TOOLS×SUPER*
TARGET×MODELING	*MODE×TOOLS×SUPER*
TARGET×TOOLS	*MODE×MODELING×SUPER*
TARGET×SUPER	*MODE×MODELING×TOOLS*
MODE×MODELING	*TARGET×TOOLS×SUPER*
MODE×TOOLS	*TARGET×MODELING×SUPER*
MODE×SUPER	*TARGET×MODELING×TOOLS*
MODELING×TOOLS	*TARGET×MODE×SUPER*
MODELING×SUPER	*TARGET×MODE×TOOLS*
TOOLS×SUPER	*TARGET×MODE×MODELING*

Note. This table shows the aliasing structure in the fractional factorial design in Exhibit 5.

Fractional factorial designs are often referred to as having a particular resolution. The resolution of a design is a convenient shorthand way of describing its aliasing. By convention, resolution is expressed as a Roman numeral. In a Resolution IV design, for example, the main effects are aliased with only three-way interactions and higher; in a Resolution V design, the main effects are aliased with only four-way interactions and higher; and so on. The design in Exhibit 6 is Resolution V, which means that main effects are never aliased with two-way or three-way interactions. In addition, two-way interactions are never aliased with each other or with main effects. In general, higher resolution designs are more desirable, but they often require more experimental conditions and therefore can be more expensive.

Fractional factorial designs can and should be selected strategically. An appropriate fractional factorial design for any given study is one that aliases effects of scientific interest with effects that are expected to be negligible in size. In general, this means aliasing main effects and lower-order interactions that are of scientific interest with higher-order interactions that are expected to be negligible (by using this phrase, we mean not specifically predicted to be sizeable by theory or prior research).

For example, if the design in Exhibit 6 were used, we would have to assume that the interactions in the right-hand column of Exhibit 7 are negligible in size. Suppose that we want to estimate the effect of *MODE*. In this design, the main effect of *MODE* is aliased with the four-way interaction between the other four components, *TARGET*×*MODELING*×*TOOLS*×*SUPER*. Therefore, to interpret the estimated effect of *MODE* as a main effect, one would have to assume that the four-way interaction between the other components is negligible in size. Similarly, assume that the interaction between *TARGET* and *MODE* is of scientific interest. This two-way interaction is aliased with the three-way interaction between the other three components. Therefore, it would not be possible for the evaluator to estimate the *TARGET*×*MODE* interaction without assuming that *MODELING*×*TOOLS*×*SUPER* is negligible in size. If these are reasonable assumptions, then this design is acceptable. However, if higher-order interactions are of scientific interest, or are predicted *a priori* to be large, fractional factorial designs are usually not appropriate and a complete factorial design may be preferable.

Selecting a fractional factorial design is not intuitive, but it is not difficult either. For most researchers, with the possible exception of scientists who have had extensive training in experimental design, it is not possible to look at a complete factorial design (like the one in Exhibit 3, p.11) and to figure out which experimental conditions to remove and which to retain to arrive at a balanced fractional factorial design with desirable properties, an acceptable aliasing structure, and the highest resolution. Fortunately, software is available to help evaluators select a fractional factorial design (see Appendix C for a demonstration of how to use Proc FACTEX in SAS to obtain the experiment shown in Exhibit 5).

In a real-world context where resources are scarce, such as in Head Start settings, evaluators must consider the opportunities and costs of using a fractional factorial design versus a complete factorial design. If resources were unlimited, any evaluator would prefer to conduct a complete factorial experiment and thereby eliminate the aliasing of effects. However, more often the reality is that an evaluator wants to examine a particular set of intervention components, but does not have sufficient resources to conduct a complete factorial experiment. The evaluator is then faced with a clear choice between two very different ways of managing research resources. Option 1 is to conduct a complete factorial experiment on a subset of the components. This approach implicitly assumes that there are enough higher-order interactions that are large enough to render a fractional factorial design inadvisable. The evaluator can only hope that funding can be obtained in the future to support another experiment on the remaining components. Option 2 is to consider a fractional factorial experiment that will enable investigation of the complete set of components, but will result in the aliasing of some effects. This approach explicitly assumes that the higher-order interactions are negligible in size.

As can be seen here, there is a trade-off between the amount of scientific information obtained and the cost of the study. Option 1 will produce estimates of main effects and interactions that are not aliased, but there is a scientific cost, namely that fewer components can be examined. The risk is that if the assumption that many higher-order interactions are substantially large is incorrect, resources will have been wasted estimating them. These resources could have been devoted instead to investigating more intervention components in a fractional factorial design. Option 2 provides the opportunity to obtain estimates of more main effects and interactions, but the estimates will involve aliasing. The risk is that if some of the assumptions about higher-order interactions being negligible in size are incorrect, some of the scientific conclusions based on the fractional factorial experiment may be incorrect. However, if the assumptions are approximately correct, considerably more scientific information will have been gained. It is up to the evaluator to judge which approach is right for a given situation.

C. Conducting Multiple Factorial Experiments

One alternative to conducting a 2^5 factorial experiment or a 2^{5-1} fractional factorial experiment would be to divide the set of intervention components that are to be examined into subsets, and to conduct multiple complete factorial experiments—that is, a separate complete factorial experiment on each subset. We could, for example, decide to examine the first two components—*TARGET* and *MODE*—using a 2x2 factorial, and to examine the other three components—*MODELING*, *TOOLS*, and *SUPER*—in a separate 2^3 factorial.

We can call this the multiple complete experiments (MCE) approach and compare it to the 2^5 factorial experiment (Exhibit 3) and the 2^{5-1} fractional factorial experiment (Exhibit 6). The MCE approach would require a total of 12 experimental conditions—fewer than the 16 required by the fractional factorial design.[12] On the other hand, an MCE approach requires a new sample of subjects for each experiment. In this hypothetical example, there are two experiments and this would require roughly twice the number of experimental subjects as either the complete or fractional factorial experiments.

One way to look at the MCE approach is to view the complete factorial experiment as the starting point (Exhibit 3). Conditions are then removed to arrive at the complete factorial experiments to be conducted. Assume that when one experiment is conducted, the factors that are not included in that experiment are arbitrarily set to the lower level. Exhibit 8 indicates which conditions would be selected out of the 2^5 factorial design for implementation in the MCE approach. Note that Condition 1 would be implemented twice. We stated above that removing conditions from a complete factorial experiment results in aliasing. Thus, the MCE approach always results in aliasing. The set of conditions in Exhibit 8 does not represent a balanced fractional factorial design, but rather what Collins et al. (2009) call an incomplete factorial. In incomplete factorial designs, it is usually not immediately clear which effects are aliased, although the aliasing can be determined.[13] Often the aliasing in incomplete factorials is less desirable than that offered by a balanced fractional factorial design. Before deciding to opt for

[12] There is a 2^{5-2} fractional factorial design that would require only eight experimental conditions, but it is a Resolution III, and therefore is probably not suitable for most behavioral science applications.

[13] This is because incomplete factorials have not been tabled by statisticians and are not offered as an alternative by software like Proc FACTEX.

any MCE approach, it is important to determine what aliasing will result from the strategy under consideration. Collins et al. (2009) demonstrate how to do this.

Exhibit 8. Illustration of a 2^5 Factorial Experiment Implemented as Two Experiments,[a] Based on Five Hypothetical Components

Condition in 2^5 experiment (Exhibit 3)	Tested Components				
	Targeted Staff (*TARGET*)	Delivery Mode (*MODE*)	Use of Modeling (*MODELING*)	Use of Assessment Tools (*TOOLS*)	Supervision of Coach (*SUPER*)
1[b]	Lead teacher	On-site	Minimal	None	Minimal
2	Lead teacher	On-site	Minimal	None	Intensive
3	Lead teacher	On-site	Minimal	Explicit	Minimal
4	Lead teacher	On-site	Minimal	Explicit	Intensive
5	Lead teacher	On-site	Intensive	None	Minimal
6	Lead teacher	On-site	Intensive	None	Intensive
7	Lead teacher	On-site	Intensive	Explicit	Minimal
8	Lead teacher	On-site	Intensive	Explicit	Intensive
9	Lead teacher	Mix	Minimal	None	Minimal
17	Teaching team	On-site	Minimal	None	Minimal
25	Teaching team	Mix	Minimal	None	Minimal

Note. Shading denotes the upper level of the tested component; unshaded cells represent the lower level of the component.

[a] The first experiment would be a 2x2 experiment of *TARGET* and *MODE* (conditions 1, 9,17, and 25) and the second experiment would be a 2x2x2 experiment of *MODELING*, *TOOLS*, and *SUPER* (conditions 1-8)

[b] Condition 1 would be implemented in both experiments for a total of two implementations.

D. Additional Considerations for Factorial Experiments

This section discusses four topics that often arise when designing factorial experiments: how to power the experiment, whether one can use more than two levels for each component, how to code effects for the analysis, and how to use the findings from a factorial experiment for decision making.

1. Sample size requirements for a factorial experiment

The general approach for determining the necessary sample size for a factorial experiment (complete or fractional) is the same as for a standard two-group RCT—that is, identify the desired alpha and an expected effect size, and then determine the sample size necessary to statistically detect an effect size of that magnitude. A quick and fairly accurate estimate of the required *N* can be obtained by identifying the smallest main effect to be detected, and then determining the *N* required for a *t*-test of that effect at the chosen level of power and statistical significance. Any other effects of equivalent or larger size will be associated with at least the desired level of power with the resulting *N*.[14] As discussed earlier, the sample size for a complete factorial experiment and a fractional factorial experiment are the same.

If groups of individuals are to be randomly assigned (e.g., all coaches in a Head Start center, as in our hypothetical scenario), the power analysis must account for the clustering of outcomes. Again, this is done in the same way as an RCT. However, a special challenge with factorial experiments is that if subjects are clustered in relatively few groups (e.g., many coaches per center and relatively few centers), there may not be enough clusters to populate all of the experimental conditions in a large factorial experiment. In this case, consideration may be given to reducing the number of experimental conditions by selecting a fractional factorial design.[15] The choice of random assignment level is discussed later in this report.

Like an RCT, the statistical power for a factorial design is also maximized when the sample is balanced (equal) across experimental conditions. That is, the per-condition *n* should be the same across experimental conditions. When the sample size varies across conditions, it is relatively easy to adjust for this statistically with little loss of statistical power. Nevertheless, the closer an experiment can come to perfectly equal experimental conditions (for a given study sample *N*), the smaller the effect that can be detected statistically for a given total number of subjects, so it is useful to maintain this as an ideal and to try to come as close as possible to achieving it.[16]

In summary, power calculations to determine the required sample size for a factorial design are conducted essentially the same way as in a standard two-group RCT. Moreover, these calculations are affected by the same types of design considerations (clustering, balance, and so on).

However, it is also important to point out that a unique challenge with studies of component effects is that the expected effect of a single component is likely to be smaller in magnitude than the effect of an entire intervention. This means that the total sample size needed for a study of component effect will likely be larger than the sample size needed for an evaluation of a complete intervention. Strategies for dealing with this challenge are discussed in Section IV.

2. Factors with more than two levels

As mentioned previously, factorial experiments can have more than two levels per factor. However, for experiments aimed at selecting components for inclusion in social interventions, using only two levels per factor is recommended whenever possible. The primary reason for this recommendation—which the majority of factorial experiments in engineering and related fields

[14] A more precise power analysis can be conducted using commercial software such as Proc POWER in SAS.
[15] See Dziak, Nahum-Shani, and Collins (2012) for a discussion of statistical power for cluster randomized experiments.
[16] For a more detailed discussion, see Collins et al. (2009).

follow—is that experiments with dichotomous factors use experimental subjects in the most economical way.

Suppose the small 2^2 factorial experiment in Exhibit 2 is sufficiently powered with N centers. We showed that several factors can be added to the experiment without necessarily requiring an increase in the total sample size N to maintain statistical power. However, adding a level to a factor does require an increase in N. For example, assume that in the simple example in Exhibit 2 (p.7), the sample size N needed to detect an effect of a given magnitude is 160 centers. Now suppose the evaluator decides to add a third level, "online only," to the *MODE* factor. To maintain the ability to detect an effect of the same magnitude as in Exhibit 2, the comparison between any two levels must be based on N=160 centers. In the original experiment (Exhibit 2), each level of *MODE* had 80 centers, so the comparison of "on-site" versus "mix" is based on a total sample N of 160. If a third level is added, another 80 centers must be added to maintain the same level of power for the "online only" versus "mix" comparison. This would increase the total sample size for the study to 240 centers (=160+80), an increase of 50 percent. The key point is that if there are more than two levels, it is no longer true that the entire study sample is used to estimate each factor's main effect, because there are now two relevant comparisons for each factor. Factors with two levels are more economical in terms of sample size.

In addition to their sample size benefits, another important advantage of factorial experiments involving two-level factors is that they simplify the choice of a fractional factorial design (i.e., which subset of conditions to retain in the experiment) and later statistical analysis of either a complete or fractional factorial. Two-level designs are also inherently more elegant and produce results that are more easily understood by policymakers and other non-researchers. Of course, there are times when it may be necessary to use a factor with more than two levels, but careful thought should be given to alternatives before making this choice.

3. Coding of effects in an ANOVA

When setting up the analysis of a factorial experiment, there are two ways in which the levels of a factor can be coded. Two approaches to coding are in wide use—effect coding and dummy coding (here we assume that each factor has two levels):

- In *effect coding*, which is used routinely in engineering and related fields, the levels of each factor are represented by 1 or -1 (or $–a$ and $+a$, where a is some constant).
- In *dummy coding*, which is most often used in the social sciences, the levels of each factor are represented by 0 or 1.

Although the two approaches to coding produce the same overall test of the model (omnibus F), the estimates of individual effects and the associated hypothesis tests are usually different. Effect coding produces main effects and interactions that are interpreted using the classical definitions discussed earlier (Section A, p.8; see also definitions in Appendix A). Dummy coded effects are interpreted differently from the classical definitions; in fact, they are not main effects and interactions and should not be referred to using these terms (although they usually are). The use of effect coding is recommended for experiments that are conducted for the purpose of selecting intervention components. For a technical discussion of the difference between effect coding and dummy coding, see Kugler et al. (2012).

4. Using the results of a factorial experiment

Suppose the experiment in Exhibit 3 was conducted and the main effect of each factor—and interactions between these factors—has been estimated. How might practitioners and policymakers use these findings to make decisions about which intervention components to use? Although the analysis of a factorial experiment is straightforward (based on an ANOVA), interpreting the findings and using them for decision-making purposes requires a more nuanced approach.

In order to make optimal decisions about which intervention components to implement, practitioners and policymakers should specify *a priori* a set of decision-making criteria, based on their priorities and constraints. For example, one strategy would be to decide on a minimum acceptable effect size for a main effect—either in absolute terms or by dollars spent—and then tentatively select components that meet that threshold. These tentative decisions could then be reconsidered in light of any sizeable interactions between components.[17] For example, if there is evidence that a selected component performs considerably better when a particular second component is present, that second component might be considered for selection even if its main effect is not large enough to meet the threshold. Tentative decisions could also be reconsidered in light of the statistical significance level (the size of the p-value). However, as noted earlier, the effect of components is harder to detect statistically due to their smaller expected magnitude, so relying on the usual standards of statistical significance may be too strict a rule for decision-making purposes. Issues related to the statistical significance level for hypothesis testing will be further discussed in Section IV.

Regardless of which criterion is used, one can see from these examples that optimal decision-making requires careful thought and consideration of effect sizes, costs, and interactions. In order to improve practitioners' utilization of the results, evaluators could start by convening a group of practitioners to learn about their main resource constraints and program priorities, and then suggest several optimal models based on the decision-making criteria most likely to be relevant to the field. Because some practitioners might want to "construct" their own intervention, evaluators could also provide them with guidance about how to choose a decision-making criterion and how to select components that will meet this criterion.

E. Summary

Factorial experiments are most appropriate when the research agenda calls for an investigation into the effect of several independent variables, and interactions between those

[17] This approach to decision-making is outlined in Collins et al. (accepted pending minor revisions).

variables, for the purpose of making decisions about which intervention components to select. Unlike an RCT, factorial experiments are not intended to provide a direct comparison of experimental conditions. Thus, in general, they do not provide a definitive answer to the question of which single combination of components (i.e., which intervention model) is best. Instead, they provide the evaluator with an efficient and cost-effective means of identifying individual components that are likely to be effective. This approach can be an excellent way to move intervention science forward, and to build a coherent body of knowledge about what works and what does not work in a particular intervention area. In addition to their rigorous scientific contributions, factorial experiments also make very efficient use of the study sample when estimating effects.

One potential drawback of factorial designs is that they usually require implementation of many experimental conditions. To lessen this burden, a fractional factorial experiment can be used instead; these designs provide the same efficient use of experimental subjects, but require implementation of half or fewer experimental conditions. Nevertheless, the number of experimental conditions may be more than intervention scientists are accustomed to handling logistically. (See Appendix B for a reading list that includes factorial experiments that were successfully implemented in field settings.)

Another important requirement of factorial experiments is that all combinations of factors and factor levels are plausible and implementable. If a combination of factors and levels is logically inconsistent or somehow toxic, it usually is not possible to conduct a factorial experiment with that set of factors and levels.

III. Other Experimental Design Options

In this section, we review four other experimental designs that could, in theory, also be used to estimate component effects. However, as we will demonstrate, these designs are less suitable than factorial designs for examining the effect of multiple intervention components. First, it is not possible to examine interactions between program components with these designs, rendering their findings less useful for decision-making purposes. In addition, the designs reviewed in this section require a larger study sample than a factorial design to statistically detect an effect of a given magnitude.

Throughout this section, we will refer to Exhibit 9 below, which summarizes the key differences between the factorial design and the alternative designs, assuming that *k* components are to be tested. We will also refer to Exhibit 10 (p.25), which illustrates these differences more concretely by comparing the sample size needed to detect a component effect of 0.20 for each type of design, based on our five-component hypothetical example.

Exhibit 9. Comparison of Experimental Design Options for an Evaluation of *k* Components

Design	Number of Experimental Conditions	Sample Size	Interactions That Can Be Estimated
Factorial Designs			
Complete factorial	2^k	*N*	All
Fractional factorial	2^{k-a}	*N*	Selected subset
Other Design Options			
Comparative treatment (CT)	$k + 1$	$(k + 1)N/2$	None
Individual experiments	$2k$	kN	None
Crossover design	$k!$	Less than a CT design, assuming no carryover effects	None
Adaptive trial	$k + 1$	Less than a CT design, depends on the size of treatment effects relative to each other	None

Note. k = number of tested components, N = sample size for a factorial design, and a = number of times by which the fractional factorial halves the number of original experimental conditions or $(1/2)^a$ (e.g., a one-half fractional factorial design would be $a = 1$). $k! = k \text{ x } (k\text{-}1) \text{ x } (k\text{-}2) \text{ x } (k\text{-}3) \text{ x } (\ldots) \text{ x } 1$.

Exhibit 10. Number of Experimental Conditions and Sample Size Needed to Detect an Effect Size of 0.20 in the Five-Component Hypothetical Scenario

Design	Number of Experimental Conditions	Number of Centers Needed to Detect an Effect Size ≥ 0.20
Factorial Designs		
Complete factorial	32	220
Fractional factorial (1/2)	16	220
Other Design Options		
Comparative treatment	6	660
Individual experiments	10	1100

Note. These calculations are based on the following assumptions: random assignment of centers to experimental conditions (to simplify the calculations, we assume that there is one coach per center so that this is equivalent to randomly assigning coaches); power of 80 percent; alpha level of 5 percent, four lead teachers per coach (or equivalently, per center), a center-level intra-class correlation of 17 percent in teacher outcomes, and baseline measures that explain 35 percent of the variation between centers and 20 percent of the variation between teachers.

A. Comparative Treatment Designs

Comparative treatment (CT) designs—also known as multigroup or multiarm experiments—are typically used to compare the effect of different interventions or approaches.[18] For example, the Head Start CARES[19] study is using a CT design to test the impact of three different classroom interventions aimed at improving early childhood outcomes in the Head Start setting (Lloyd & Modlin, 2012). In this study, Head Start centers were randomly assigned to one of four experimental conditions (one control group representing business as usual and three different interventions). This design makes it possible to estimate the impact of each of the three interventions against the control condition, as well as the differential impact of the three interventions relative to each other. Appendix D provides more information on the design of the study.

A CT design can also be used to test the effect of adding a component to a given intervention in order to examine whether this addition increases the intervention's impact.[20] This approach was used in a study of reading professional development strategies for second-grade teachers (Garet et al., 2008). The study randomized second-grade teachers to one of three groups: (1) a control condition representing business as usual professional development in the study districts; (2) a teacher institute series focused on reading principles and how children learn to read; or (3) the same institute series plus in-school coaching that focused on how to integrate this knowledge into teaching practice. This experimental design made it possible to estimate the impact of the basic intervention (the institute series) on teacher practice and children's reading skills, and to examine whether adding coaching to the basic intervention further increased its impact. Appendix D provides more information on the study design.

However, the CT design is less informative when the goal of the study is to provide the information needed to select a subset of intervention components from a larger set. To demonstrate this point, suppose that we want to use a CT design to evaluate the effect of the five components in our hypothetical Head Start coaching scenario (Exhibit 1, p.6). Two types of CT design could be used for this purpose.

In the first type of CT design (Exhibit 11), Head Start centers could be randomly assigned to one of six experimental conditions. Centers in the first condition would receive a basic intervention where the components are set at the lower level of each component: the early childhood coach works with the lead classroom teacher only, coaching is delivered using a mix of on-site and online coaching sessions, the coach uses only minimal modeling of good practice, the coach does not use assessment tools for identification and differentiation, and the coach is minimally supervised. In the other five experimental conditions, Head Start centers would receive an intervention that differs from the basic intervention with respect to one of the five tested components. The effect of changing a given component could then be estimated by

[18] See Collins et al. (2009) for a further discussion of CT designs.

[19] CARES = Classroom-based Approaches and Resources for Emotion and Social Skill Promotion.

[20] Similarly, a CT design could also be used to test the effect of removing one component from an intervention.

comparing the outcomes of centers in Condition 1 (the basic intervention) to the outcomes of centers in the experimental group where that component is set to its upper level. For example, the effect of coaches using assessment tools more explicitly—relative to not using them—would be estimated by comparing centers assigned to Condition 5 to centers assigned to Condition 1.

Exhibit 11. Illustration of a Basic Comparative Treatment Design, Based on Five Hypothetical Components

Experimental Condition	Tested Components				
	Targeted Staff (*TARGET*)	Delivery Mode (*MODE*)	Use of Modeling (*MODELING*)	Use of Assessment Tools (*TOOLS*)	Supervision of Coach (*SUPER*)
1 (base model)	Lead teacher	Mix	Minimal	None	Minimal
2	Teaching team	Mix	Minimal	None	Minimal
3	Lead teacher	On-site	Minimal	None	Minimal
4	Lead teacher	Mix	Intensive	None	Minimal
5	Lead teacher	Mix	Minimal	Explicit	Minimal
6	Lead teacher	Mix	Minimal	None	Intensive
	Effect of targeting: 2 versus 1	*Effect of mode: 3 versus 1*	*Effect of modeling: 4 versus 1*	*Effect of tool use: 5 versus 1*	*Effect of supervision: 6 versus 1*

Note. Shading denotes the upper level of the tested component; unshaded cells represent the lower level of the component.

In the second type of CT design (Exhibit 12), centers could be assigned to experimental conditions that differ from each other in an additive manner. In this version, Condition 1 is still the same basic intervention model; however, the other five conditions differ by one component from each other, instead of differing from the basic intervention. In this design, for example, the effect of more explicit use of assessment tools is estimated by comparing the outcomes of centers in Condition 5 to Condition 4 (instead of comparing Condition 5 and Condition 1, as in Exhibit 11).

Comparing these two examples, one can see that the findings from a CT design represent the effect of a given component *when the other tested components are set to a particular level*. For instance, in Exhibit 11 we can estimate the effect of more explicit use of assessment tools in the specific context where the other four coaching components are set to their lower levels (Condition 1). In Exhibit 12, we can estimate the effect of more explicit tool use when the other four coaching components are set to the levels in Condition 4. If there are interaction effects between the components being tested, the effect of a component (such as the use of assessment tools) will depend on the levels of the other components. By extension, the two CT designs will yield different estimates of the effect of more explicit tool use. Neither result is incorrect; they simply have different interpretations.

Exhibit 12. Illustration of a Constructive Comparative Treatment Design, Based on Hypothetical Components

Experimental Condition	Tested Components				
	Targeted Staff (*TARGET*)	Delivery Mode (*MODE*)	Use of Modeling (*MODELING*)	Use of Assessment Tools (*TOOLS*)	Supervision of Coach (*SUPER*)
1 (base model)	Lead teacher	Mix	Minimal	None	Minimal
2	Teaching team	Mix	Minimal	None	Minimal
3	Teaching team	On-site	Minimal	None	Minimal
4	Teaching team	On-site	Intensive	None	Minimal
5	Teaching team	On-site	Intensive	Explicit	Minimal
6	Teaching team	On-site	Intensive	Explicit	Intensive
	Effect of targeting: *2 versus 1*	*Effect of mode:* *3 versus 2*	*Effect of modeling:* *4 versus 3*	*Effect of tool use:* *5 versus 4*	*Effect of supervision:* *6 versus 5*

Note. Shading denotes the upper level of the tested component; unshaded cells represent the lower level of the component.

The design-specific nature of findings from a CT design can make them more difficult to use for decision-making purposes. Turning again to our example, assume that, based on the CT design in Exhibit 11, we find that more intensive coach supervision does *not* have an effect when the other components are set to their low level. In practice, however, supervision of the coach might be important if the other components were implemented at a more intensive level. For instance, intensive supervision of the coach could be necessary if modeling were used more frequently or if some coaching sessions were delivered online. (In other words, there is an interaction between coach supervision and some of the other components.) If the hypothetical findings from the design in Exhibit 12 were published, Head Start centers might decide to give their coaches minimal supervision, when in fact they should be supervising them more intensively when using more complex coaching delivery models. Problematically, a CT design does not make it possible to determine the extent to which estimated component effects are sensitive to the levels of the other components (that is, whether there are interactions between components).

In contrast, a factorial experiment provides information about the main effect of a component across all possible levels of the other components. In addition, a factorial design also makes it possible to examine interactions between these components to better understand the extent to which component effects are sensitive to the levels of the other components. These types of findings are useful for a policymaker or practitioner who is thinking about which intervention components to implement or how to design an intervention package. With a CT design, however, there is no way to determine whether a given component would still be effective if it were used in combination with different levels of the other components, and for this reason the findings from a CT design are more difficult to interpret and utilize for decision making.[21]

One reason why evaluators may consider using a CT design for a study of component effects—even though its findings are less useful for this purpose—is that the CT design has fewer experimental conditions than the factorial design. Exhibit 9 shows that the number of experimental conditions needed to evaluate the effect of k components is 2^k for a complete factorial design, and k+1 for a comparative treatment design. More concretely, in our hypothetical five-component scenario, the CT design has six experimental conditions while a complete factorial design requires 32 conditions (or 16 conditions for a half fractional design). This might lead evaluators to think that the CT design is less costly and more operationally feasible than a factorial design.

However, many researchers do not realize that the CT design requires a much larger sample than the factorial design, which increases the cost and complexity of the study and may outweigh the advantage of having fewer experimental conditions. In a CT design, component effects are estimated by comparing experimental conditions directly; for example, the effect of explicit tool usage for the CT design in Exhibit 11 is estimated by comparing centers in Conditions 5 and 1. In contrast, when using a factorial design, component effects are estimated by comparing groups of conditions; experimental conditions are never directly compared to each other. This is an important distinction between the two designs: in a factorial design, the entire sample is used to estimate each component effect, while in a CT design only a subset of the sample is used to estimate the effect of a component.

This means that the CT design requires a larger sample than the factorial design to detect an effect of a given magnitude. As shown in Exhibit 9, the sample size needed to detect a component effect of a given size is (k+1)/2 times larger for a CT design than for a factorial design. This means that in our hypothetical five-component scenario, the CT design requires three times more sample members than the factorial design to detect the same effect size. Exhibit 10 illustrates this point even more clearly—as seen in this exhibit, the sample size would need to be 660 centers if using a CT design versus 220 for a factorial design.

In summary, the CT design is less suitable for a study of component effects and less useful from a policy perspective, because its findings about the effect of a particular component are specific to the levels at which the other components are set. In the best case scenario, using a CT design would require recruitment of a larger study sample than necessary; in the worst case scenario, the results from this design may lead practitioners to make the wrong decision about whether or not to embed a component into their current intervention strategy. The CT design is best used for its intended purposes—to compare different intervention models (like in the Head Start CARES study) or to examine the incremental effect of one component (like in the reading

[21] More technically, in a CT design, the main effect of the components is aliased with all interactions (Collins et al., 2009).

PD study). For these two purposes, interaction effects are much less relevant and therefore the CT design is appropriate.

B. Individual Experiments Design

The individual experiments (IE) design is an evaluation strategy that can be used when it is not operationally feasible to implement an experiment with multiple experimental conditions, like the CT design. In essence, the IE design breaks the CT design into multiple, two-group "mini-experiments." For example, had the Head Start CARES study (discussed on page 26) used an IE design, there would have been three two-group experiments. The first experiment would have tested the first instructional model against a control group (business as usual); the second experiment would have tested the second instructional model against a control group; and the third experiment would have tested the third instructional model against a control group. These three experiments could be implemented at the same time, but more likely they would be implemented in sequence. One advantage of the IE design, therefore, is that it allows experiments to be rolled out over time when short-term resources are scarce.[22]

Exhibit 13 below shows the IE design version of the CT design in Exhibit 11. This IE design and its CT design counterpart provide the same estimates of component effects, which are subject to the same interpretation limitations. As a result, findings from the IE design—as with the CT design—are less useful when the goal of the study is to compare the effect of two or more components.

The IE design also has three additional drawbacks beyond those of the CT design. First, while there is only one control group in the CT design, there are multiple control groups receiving the same set of services (one control group per experiment) in an IE design. This means that the IE design has more experimental conditions in total than a CT design. For an experiment of *k* components, there are 2*k* experimental conditions, whereas for a CT design there are only *k*+1 conditions (see Exhibit 9).

Second, the IE design has the largest sample requirement of all the designs reviewed in this report because different participants are used in each experiment. As shown in Exhibit 9, the sample size needed to detect a component effect of a given magnitude is *k* times larger for an IE design than for a factorial design. Thus, in our five-component scenario, the sample size for the IE design would be five times larger than the sample size required for a factorial design (a sample size of 1100 centers to detect an effect of 0.20, as shown in Exhibit 10).

Third, although each experiment can provide internally valid estimates of a component's effect, the estimated effects are difficult to compare across experiments because each experiment uses a different sample of study participants. For example, assume that the estimated effect size of coach supervision is 0.50 from Experiment 5, and the estimated effect size of delivery mode is 0.30 from Experiment 2. The relative magnitude of these two effects could reflect their true

[22] See Collins et al. (2009) for a further discussion of IE designs.

relative impact, but it could also be due to differences in study population across the two experiments. Therefore, evaluators have to be especially careful about making the study samples as similar as possible across experiments when using an IE design.

Exhibit 13. Illustration of an Individual Experiments Approach, Based on Five Hypothetical Components

Experiment	Intervention	Tested Components				
		Targeted Staff (*TARGET*)	Delivery Mode (*MODE*)	Use of Modeling (*MODELING*)	Use of Assessment Tools (*TOOLS*)	Supervision of Coach (*SUPER*)
A	Control (base model)	Lead teacher	Mix	Minimal	None	Minimal
	Treatment	Teaching team	Mix	Minimal	None	Minimal
B	Control (base model)	Lead teacher	Mix	Minimal	None	Minimal
	Treatment	Lead teacher	On-site	Minimal	None	Minimal
C	Control (base model)	Lead teacher	Mix	Minimal	None	Minimal
	Treatment	Lead teacher	Mix	Intensive	None	Minimal
D	Control (base model)	Lead teacher	Mix	Minimal	None	Minimal
	Treatment	Lead teacher	Mix	Minimal	Explicit	Minimal
E	Control (base model)	Lead teacher	Mix	Minimal	None	Minimal
	Treatment	Lead teacher	Mix	Minimal	None	Intensive
		Effect of target: T versus C in Exper. A	*Effect of mode: T versus C in Exper. B*	*Effect of modeling: T versus C in Exper. C*	*Effect of tool use: T versus C in Exper. D*	*Effect of supervision: T versus C in Exper. E*

Note. Shading denotes the upper level of the tested component; unshaded cells represent the lower level of the component.

C. Crossover or Switching Designs

Crossover designs, also known as switching designs, are used in medical research. In this type of design, subjects are randomly assigned to all possible sequences of the treatment options. As a simple example, assume that an evaluator wants to test and compare the effect of two different therapies (A and B). In a crossover design, one group of patients would be randomly assigned to receive Treatment A then Treatment B, while the other group would be assigned to receive Treatment B then Treatment A. (This is called an AB/BA design.) The effect of each treatment can then be estimated by comparing each patient's outcomes under Treatment A to the patient's own outcomes under Treatment B and then averaging across all patients. Each treatment's impact relative to a control condition can also be estimated by comparing a patient's outcomes under a given treatment to their outcomes at baseline (prior to the receipt of any treatment).[23]

In theory, one could use a crossover design to estimate the effect of intervention components. Exhibit 14 below shows a simple crossover design for estimating the effect of three program components. Each experimental condition represents a different sequence in which a given component is turned "on" or "high." With three components, there would be a total of six (=3!) experimental conditions, representing every possible sequence of the three components.[24]

Exhibit 14. Illustration of a Crossover Design with Three Components (A, B, and C)

Experimental	Sequence of Components Received by Subjects in Each Experimental Condition		
	Round 1	Round 2	Round 3
1	Component A	Component B	Component C
2	Component A	Component C	Component B
3	Component B	Component A	Component C
4	Component B	Component C	Component A
5	Component C	Component A	Component B
6	Component C	Component B	Component A

Note. Each cell indicates which component is turned "on" or set to "high" in a given round (other components are set to off or low in that round).

The advantage of using a crossover design is that under certain assumptions, it provides effect estimates with relatively greater precision than a CT design or even a factorial design, which means that it is possible to detect a smaller effect for a given sample size. This is due to the fact that each patient receives all treatments or components, and this makes it possible to control for random variation across patients or subjects in the analysis (e.g., by controlling for patient fixed effects). Thus, this design is more sample-efficient than a CT design because the entire sample is used to estimate the effect of each component.

However, crossover designs have several features that make them more difficult to use for evaluating social interventions such as coaching (as opposed to medical treatments). First, crossover designs can require many experimental conditions. For example, as shown in Exhibit 9 (p.25), a crossover design that evaluated five interventions or components would have 120 (=5!) experimental conditions.[25] Second, data collection for the crossover design is more costly. The design requires multiple rounds of measurement, which means that survey tools and other instruments have to be administered multiple times. Third, the impact of a social intervention (or a component) can take longer to manifest itself, which means that the duration of each round

[23] See Friedman, Furberg, and DeMets (1998) for further discussion of crossover designs.
[24] 3! = 3x2x1.
[25] 5! = 5x4x3x2x1. This explains why we do not show a crossover design based on our five-component scenario.

(and the study as a whole) would be longer than when evaluating a medical treatment. Fourth, crossover designs assume that there is no carryover effect—that is, the effect of a treatment must disappear once the patient ceases to receive it in the next round. Yet, social interventions are likely to have carryover effects.[26] For all of these reasons, the crossover design is unlikely to be a feasible option for evaluating the effect of components in a social intervention such as coaching.

More importantly, crossover designs, like CT designs, are better suited to comparing the effect of different *interventions* as a package as opposed to individual component effects. In the design illustrated in Exhibit 14 (p.32), one can see that component effects are estimated by allowing each component to vary in turn (much like the CT design in Exhibit 11, p.27). This means that this crossover design, if used, would provide estimates of a component's effect when all other components are set to their low level. For reasons already discussed in the context of CT designs, such context-specific results are less useful for policymakers and practitioners.

D. Adaptive Clinical Trials

Though not yet commonly used, adaptive clinical trials have been garnering increasing amounts of attention in medical research.[27] The experimental conditions in an adaptive trial are similar to those in a comparative treatment (CT) design—there is a control condition and one or more treatment conditions receiving different therapeutic regimens. However, an adaptive design differs from CT designs with respect to random assignment and analysis. In an adaptive design, sample intake and random assignment (RA) to experimental conditions happen on a rolling basis, and assignment probabilities (RA ratios) are updated at each round of random assignment:

i. The first round of subjects is randomly assigned to treatment conditions (therapies) based on equal probabilities (RA ratios).

ii. Data on outcomes is collected as soon as it is feasible to expect impacts. The relative effect of the treatment(s) is estimated, overall and by subgroups.

iii. The RA probabilities are then updated based on these results, with higher RA ratios given to treatments that are more effective based on the findings. When new subjects are recruited into the study, they are given a higher probability of being assigned to treatments that are more effective for individuals with their particular characteristics (and treatments that are uniformly ineffective might even be dropped).

iv. The process then cycles repeatedly through Steps (ii) and (iii).

An adaptive design is constantly updating information about effects and then applying this information to new subjects as they enter the study. This process can improve the precision

[26] In our hypothetical example, one strategy for eliminating the carryover effect would be to randomize Head Start centers to experimental conditions, but to recruit a new group of teachers in these centers in each phase. However, if each center has few teachers or little staff turnover, recruiting new teachers could be challenging. Also, this would increase the sample size requirements for the study, thereby eliminating the main advantage of the crossover design.

[27] For an overview, see Scott and Baker (2007).

impact estimates for the more effective treatments, thereby making it possible to detect smaller effects for a given sample size. This design is also fairer to newer participants because they have a higher probability of getting the treatment(s) that appears to be the most effective. Finally, it can simplify sample intake because participant recruitment can be spread across multiple rounds.[28]

In theory, this approach could be used instead of a classic CT design to compare social interventions. However, the multiple rounds of assignment/measurement that characterize an adaptive design might be challenging to implement in practice because social institutions (such as Head Start centers) operate on set schedules (such as school years), making it hard to recruit participants mid-year. In addition, impacts on social outcomes take longer to appear than impacts on medical or physical outcomes, and therefore there would need to be a substantial amount of time between random assignment and measurement rounds. Measurement is also more challenging in social institutions because measures are more complex and time consuming (requiring surveys, observation, testing, etc.). All of these factors would result in more costly data collection and a longer study time frame, were an adaptive design used.

It is also important to remember that adaptive designs—like CT designs—have so far been used to compare and evaluate different *interventions* and treatment regimens as opposed to gathering information on the effect of *components*. In future, it is possible that the adaptive approach could be extended to evaluations of component effects; however, at present, the analytics and applicability of adaptive designs to factorial experiments is not well understood. Therefore adaptive trials currently have the same limitation as a CT design—they are less suitable for studies of component effects because the interpretation of estimated effects is too narrow to be policy-relevant or practically useful.

IV. Important Design Issues in a Study of Component Effects

When designing a study of component effects, several issues need to be carefully considered by evaluators, regardless of which experimental design is used. Although some of these issues are also relevant to evaluations of social interventions, they are exponentially more complex in component evaluations, and therefore they have to be dealt with more strategically. This section discusses three broad topics: strategies for dealing with the fact that component studies are likely to yield effects that are small in magnitude; deciding whether or not to "fix" the levels of non-tested components; and using a pilot phase to assess the feasibility of the chosen study design.

[28] This approach is Bayesian, in that information about probabilities and impact estimates is constantly being updated as new information becomes available. Therefore, it requires that data are analyzed using Bayesian methods. See Berry (2006) for a discussion of Bayesian versus frequentist statistics in the analysis of clinical trials.

A. Smaller Expected Effects

An important part of any evaluation is to determine the sample size that is needed to detect an impact of meaningful magnitude. The *minimum detectable effect* (MDE) is a useful concept for making this decision. Formally, the MDE is the smallest true impact (on the outcome of interest) that can be detected with a reasonable degree of power (for example, 80 percent) for a given level of statistical significance (usually 5 percent). A related concept, the *minimum detectable effect size* (MDES), is the MDE scaled as an effect size—obtained by dividing the MDE by the standard deviation of the outcome measure. The MDES is more useful than the MDE when the outcome of interest is measured in a scale whose units do not have a readily interpretable meaning (for example, standardized test scores and behavioral composite scales). In the paragraphs below, we refer to the MDES for simplicity, but the discussion also applies to the MDE.

The most important determinant of the MDES is the sample size. The sample size and the MDES are inversely related: the greater the number of participants in the study, the smaller the estimated impact can be to conclude that it is statistically significant. Conversely, the smaller the true expected impact of an intervention or component, the larger the sample size needs to be to conclude that its estimated effect is statistically significant.

It follows that evaluators must think carefully about the effect size that the study should be able to detect (i.e., the target or expected effect size) because this will determine the required sample size for the study. Evaluators can use several approaches to identify a target effect size. One approach is to use effect sizes from prior evaluations of interventions similar to the one being tested. For example, if a particular type of professional development (PD) approach has been shown to produce an estimated effect size of 0.15, a new study that aims to evaluate a similar PD approach could choose a sample size that will allow it to detect a target effect size of 0.15. Another approach is to choose the smallest impact deemed policy-relevant or cost-effective as the target effect size. For example, in studies conducted by the U.S. Department of Education (ED), an effect size of 0.20 is usually considered a policy-relevant impact on student achievement, and therefore many evaluations funded by ED use this as the target effect size.[29]

Identifying a target effect size (and a sample size to detect this effect) is more complicated in an evaluation of component effects. Prior empirical research on the effects of intervention components is scarce, and the policy relevance of components is difficult to gauge in isolation. However, what we do know is that the impact of an individual intervention *component* is likely to be smaller than the impact of an entire multi-component *intervention.* By extension, the target effect size used for sample size determination in a component evaluation should be smaller than for the evaluation of an entire intervention, which in turn implies that the study sample for a component evaluation will be larger (all else equal).

[29] See Bloom, Hill, Black, and Lipsey (2008) for a general discussion of how to think about expected effect sizes in the context of K-12 education.

Increasing the study sample is costly, however, so evaluators should consider ways in which they can reduce the sample size requirements for the component study. Logically, this can be accomplished by strategically choosing design elements that will either increase the expected (target) effect size of the components, or that will reduce the MDES for a given sample size. Specifically:

- The expected effect size can be increased by carefully choosing the components to be tested, the levels of these components, the outcome of interest, and the measurement of these outcomes.
- The MDES for a given sample size can be reduced by choosing a different significance level, by blocking random assignment, by collecting data on baseline measures or pretests, and by changing the level of random assignment.

By choosing design elements that will increase the expected effect size or reduce the MDES, an evaluator can make the sample size requirements for the study more manageable. These design elements—and how they affect the sample size—are discussed in greater detail in the remainder of this section.[30]

In the discussion that follows, we assume that the sample size will be chosen based on the expected effect size for an intervention component. In theory, one could also choose a sample size based on the expected effect size for an interaction between components. However, this is unlikely to be a desirable strategy in practice because: (1) choosing a target effect size for an interaction is even more difficult than choosing a target effect size for a component because almost nothing is known about interaction effects; (2) the statistical power for an interaction effect is less than for a component effect; for example, if a component (main) effect of 0.20 can be statistically detected given the sample size, a two-way interaction effect would have to be twice as large (i.e., 0.40) to be detected;[31] and (3) component effects are typically of greater scientific interest, whereas interaction effects are most useful as a secondary source of information to help evaluators interpret component effects.

1. The Choice of Components and Their Levels

To maximize their target or expected effect size, the components tested in the study should have the potential to affect participants’ outcomes. Tested components should have reasonably large expected impacts (relative to non-tested components) based on prior experimental or quasi-experimental evaluations and/or practitioners’ experience. To maximize their expected effect, the components chosen for testing should also be components that can be implemented with a high degree of fidelity during the study’s time frame.

[30] These design elements are important for all types of evaluation—not just studies of component effects—but they are especially important in component evaluations, due to the fact that intrinsically smaller effects are to be expected.

[31] This is due to the fact that when estimating a two-way interaction, the sample must be split into subgroups based on the levels of one of the components.

In some cases, expected effect sizes can also be increased by bundling two components together. If two components are interrelated—meaning that, in practice, they would either both be implemented or neither would be implemented— then these two components could be packaged together and tested as one meta-component, the expected effect size of which would be greater than for each component individually. For example, in our hypothetical five-component scenario, assume that a coach's use of assessment tools requires that he or she be closely supervised. This implies that the combination of "explicit assessment tool use" and "minimal coach supervision" might never be implemented in practice in the field. In this situation, the components "coach's use of assessment tools" and "coach supervision" could be bundled into one component, and the two levels for this new component would be "typical assessment tool use and minimal coach supervision" versus "explicit assessment tool use with intensive coach supervision."[32]

In addition to carefully choosing the components, it is also important to carefully choose their levels because this also affects the expected effect size of the components. As explained in Section I, choosing two levels for each component—a level that is "off" or "low" versus a level that is "on" or "high"—represents the most efficient use of resources. Although in theory one could construct a factorial or other design with more than two levels, this would exponentially increase the sample size requirements for the study, and it would increase the number of experimental conditions that need to be implemented (as discussed in Section II). Therefore, if a component has multiple feasible levels—and if little is known about the effect of the component at any of these levels—the most resource-efficient strategy is to test the lowest feasible level against the highest feasible level. The rationale here is that testing the effect of an intermediary level of the component is less important when the effect of the component at its highest level is not yet known.[33]

When specifying the two levels of a component, it is important to remember that the greater the contrast between the lower and upper level, the greater the expected effect size. Ideally, evaluators should maximize the service contrast between the two levels, while still making sure that both levels are feasible in a real-world context. The "off" or "low" level of each component should offer the lowest possible amount of the component, but should not be so low that the level would never be implemented in practice. At the other end of the spectrum, the "on" or "high" level should be as service-intensive as possible while still being feasible to implement in the field. For example, in our hypothetical five-component scenario, the lower level of the

[32] The disadvantage of bundling components is that the individual effect of the original components is no longer estimable. For example, if the level of coach supervision interacts with other components' effects (and not only the "data usage" component), it might be preferable to have it remain as an individual component in order to estimate its independent effect and its interaction with other components. Evaluators should carefully consider all of these issues if components are to be bundled.

[33] The resource management principle states that evaluators should choose the levels that provide the most useful information given the cost of the study. Based on this principle, a sound strategy is to start by establishing that the high level of a component (versus the low level of a component) improves participants' outcomes, so that resources are not wasted looking at an intermediate level if there is no high-low difference. If the high versus low level of a component is shown to have an effect, then the effect of more moderate levels could be tested in a future study.

coach supervision component could be set at the minimum amount of coach supervision used in practice, while the upper level could be set at the highest number of hours that is affordable in Head Start centers.

2. The Choice of Outcomes and Measures

In order to maximize the expected effect size for the components in the study, evaluators should also consider choosing a primary outcome that is more proximal in the theory of action and that is known to be associated with longer-term outcomes. In general, impacts on shorter-term (proximal) outcomes are larger than for longer-term (distal) outcomes. For example, in our five-component hypothetical scenario, the primary outcome could be teacher practice and/or child-teacher interactions, as opposed to children's academic or behavioral outcomes. The disadvantage of this approach is that the effect of the components on the longer-term outcome of interest (in this case, child outcomes) would not be evaluated. On the other hand, if the theory of action is logical, intervention components (such as coaching components) must affect short-term outcomes (teacher practice) in order to have an effect on longer term outcomes (child outcomes). Therefore, focusing on proximal outcomes in a component evaluation makes rational sense as a strategy for making the sample size requirements more reasonable.[34]

On a related point, evaluators must also decide the extent to which the outcome of interest should be aligned with the components. The outcome of interest could be a general latent measure targeted by the components ("use of effective teaching practices") or the specific behaviors and/or knowledge targeted by the components ("use of scaffolding").

There are several trade-offs between these two options. Expected effects on specific behaviors will probably be larger because they are more closely aligned with the components, which would in turn reduce the sample size requirements for the study. However, this option would likely require having to develop new measures for the purposes of the evaluation, which would be costly and time consuming to pilot and may not be as valid and reliable as one would wish.[35] Conversely, the expected effect of the components on a more *general outcome* might be smaller (and the sample size requirements larger), but it might be possible to use an existing measure that is already being used in the field. This would reduce the cost of data collection, and could also provide findings that are more policy relevant and reliable. For example, the Office of Head Start (OHS) uses the Classroom Assessment Scoring System or CLASS (Pianta, La Paro, & Hamre, 2008) to measure program quality, which includes examining teacher practices for

[34] If the component evaluation is being conducted as part of the screening experiment of the MOST approach (see Section I), impacts on longer-term outcomes could be explored in the second phase of MOST (i.e., the confirmatory experiment that evaluates the impact of the optimal model). Of course, with this approach, one would only know the effect of the optimal intervention on longer-term outcomes (as opposed to the effect of the components).

[35] Measures should be valid and reliable. Validity means that the measure is correctly capturing the latent outcome that it is intended to measure. Reliability refers to the ability to provide an accurate measurement of the outcome of interest (one with little measurement error). Reliability is important because the standard error of the impact estimate (and the MDE) is smaller for more reliable measures, all else equal. Reliability is also important because a measure cannot be valid unless it is reliable.

monitoring and accountability purposes. Expected impacts on the CLASS might be smaller than on a study-developed measure, but the evaluation results might be more policy relevant given that the CLASS is used in the field.

Somewhere between these two extremes, an alternative approach would be to choose a primary outcome that represents a subdomain of behaviors that is reasonably well aligned with the intervention components, and measured by existing assessments. For example, the CLASS has subtests for three aspects of teacher practice: emotional support, instructional support, and class organization. Using a particular subdomain of the CLASS as the primary outcome – for example "instructional support" – would reduce data collection costs (since the CLASS has already been developed) while also ensuring that the outcome is suitably specific. This approach would also make it possible to use a different outcome measure for each tested component. In our hypothetical scenario, for example, one might expect coaches' use of modeling to have a larger impact on the "instructional support" subdomain of the CLASS, while the target of the coaching (lead teacher versus the entire teaching staff) might have a larger effect on classroom organization.[36] In this situation, it might make sense to let the primary (subdomain) outcome differ across components. This would increase the alignment between the components and the outcome measures, and therefore the expected effect size. The Teacher Behavior Rating Scale (TBRS) is another example of a validated measure used in the field that has subtest scores for teacher practice, in this case by content area (mathematics, language and literacy).[37]

3. The Unit of Randomization

In any experiment, evaluators must choose a level (unit) of random assignment: who, or what, should be randomly assigned to the conditions in the experimental design? This is an important decision because it has implications for the study's sample size requirements.[38]

For the purposes of this discussion, assume that the intervention or the components are group-administered. In a group-administered intervention, a *provider* (such as a coach) delivers a set of services to a group of recipients (such as Head Start teachers). Many educational interventions are group-administered. Notable examples in K-12 education evaluations are whole-school reforms, whereby a school provides services that are intended to affect the outcomes of all students at the school.

When an intervention is group-administered, an obvious choice for the randomization unit is the provider of the intervention. For example, in our hypothetical scenario, the provider of the intervention is the coach, so the simplest strategy is to randomize coaches to experimental conditions. A coach would then be tasked with administering (to his or her teachers) the combination of components in the condition to which the coach was assigned.

[36] If different subdomain measures were to be used for each component – and these subdomain measures are on different scales – then the magnitude of estimated effects could be compared across components based on a standardized metric such as the effect size.
[37] For more information on the TBRS, see Landry (2007).
[38] See Bloom (2005) for a discussion of random assignment levels.

Although one might be tempted to randomize recipients rather than providers, randomizing recipients in a group-administered intervention can increase the operational complexity of the evaluation and muddy the interpretation of the evaluation findings. To illustrate this point, consider what would happen if Head Start teachers were randomized to experimental conditions in our hypothetical example. As already noted, a coach can work with multiple teachers. If teachers were randomized to conditions, a coach would have to work with teachers assigned to receive different sets of coaching components. In other words, a coach might be assigned to multiple experimental conditions. This means that the coach would have to administer a different set of coaching components depending on which teacher they were working with. This would be problematic for three reasons. First, some components cannot be switched on and off by the coach; for example, a coach cannot receive different levels of supervision. Second, it would be difficult in practice for a coach to remember to use different components with different teachers (and also for evaluators to monitor whether coaches are doing so). This in turn would reduce the service contrast between the experimental conditions, and therefore the expected effects. Third, the estimated effect of the components would be difficult to interpret; because coaches are not randomized, the observed effect of a component could be confounded with the effect of coach quality (i.e., the intrinsic ability of the coaches implementing that component). Weiss (2010) discusses these issues in the context of K-12 education. For all of these reasons, randomizing service providers is preferable in a group-administered intervention.[39]

Alternatively, if spillover can occur, then researchers might have to consider randomizing groups of providers rather than the providers themselves. A spillover effect—also called control group contamination—happens when subjects assigned to a particular experimental condition are also exposed to the treatment in a *different* experimental condition. Spillover effects are most common when evaluating interventions that provide information that can be shared across individuals. In our five-component scenario, for example, there may be scheduled shared planning time in which Head Start teachers discuss their practices and knowledge. A teacher working with Coach A (who has been randomly assigned to administer a specific combination of coaching components) might share his or her new knowledge with a teacher working with Coach B (who has been tasked with administering a different combination of components). This would reduce the contrast in services received by different experimental groups, which would in turn reduce the magnitude of the effects that one would expect to see in the evaluation. By extension, the sample size for the study would have to be increased to detect this smaller effect. In a component evaluation, this is an especially important concern because component effects are expected to be relatively small even in the absence of spillover.

Fortunately, spillover can be reduced by randomizing groups of providers, where groups are defined in such a way as to minimize social interactions between providers (or recipients who

[39] One could also randomize teachers to coaches to prevent centers from assigning particular teachers to the coach who will be using a particular set of practices, since such non-random selection by centers would compromise random assignment.

work with these providers). In our hypothetical scenario, for example, it might be preferable to randomize Head Start centers (i.e., all coaches at a Head Start center) instead of the coaches themselves.[40] Assuming that each coach works at only one Head Start center, the randomization of centers would preserve the service contrast between experimental conditions and the expected effect size.[41]

However, randomizing at a higher level (randomizing groups) also has an important disadvantage—it increases the MDES. All else equal, randomizing groups decreases the precision of estimated impacts, which in turn makes it harder to detect an effect of a given magnitude for a given sample size (or in other words, it increases the MDES). In our hypothetical example, the MDES would be larger if Head Start centers were randomly assigned to experimental conditions, as opposed to directly randomizing the coaches in those centers.[42] A larger MDES is undesirable because it means that estimated effects have to be larger for evaluators to conclude that they are statistically significant.

In summary, if spillover is a possibility, then the challenge for evaluators is to decide whether the degree of spillover is larger enough to warrant group-level randomization. As a practical strategy, one could compare the sample size requirements for group-level randomization versus provider-level randomization, based on different assumptions about the range and magnitude of spillover. If anticipated spillover is small or moderate, then it might still be preferable to randomize providers rather than groups.[43]

Finally, it is important to note that when an intervention (or a component) is group-administered, the experiment is by definition a *cluster* randomized experiment, regardless of whether providers or groups of providers are randomized. This is because the unit of analysis is the recipient (the teacher in our hypothetical example), and recipients are clustered within providers (coaches). The analysis of the experiment will have to account for this two-level clustering, typically by using a hierarchical or random-effects analysis. If groups of providers are randomized, then the analysis would have to account for three levels of clustering (*recipients* within *providers* within *groups of providers*). Bloom (2005) provides information on the analysis of cluster-randomized experiments. As noted in Section II, factorial designs can also accommodate cluster randomization.

[40] Note that if there is only one coach per center, randomizing centers is the same as randomizing coaches, and there is no risk of spillover across teachers at the center because all teachers are coached by the same person.

[41] Randomizing centers would only be useful if each coach works at only one Head Start center (i.e., coaches are nested within centers). If coaches work across centers (i.e., coaches are cross-classified across centers), randomizing centers would not work because a coach could be implicitly assigned to implement multiple experimental conditions. This is problematic because a coach would have to use different sets of coaching components depending on which center they were working at on a given day. This would be difficult for a coach to maintain, and would likely reduce the service contrast across experimental conditions.

[42] For a discussion of statistical power in this context see Bloom (2005); Dziak et al. (2012).

[43] If spillover is large on some components but not others, one could randomize the former set of components at the group level, and randomize the latter set of components at the recipient level. This is sometimes called a split-plot design and it can be used with a complete factorial design (Clarke & Kempson, 1997).

4. The Blocking of Random Assignment

By blocking random assignment, evaluators can reduce the sample size that is needed to detect an effect of a given magnitude. Blocking simply entails randomizing providers or groups of providers *by strata* defined by geography or some other characteristic.[44] To maximize the benefits of blocking, it is important to define the strata based on characteristics across which there is a substantial variation in the outcome of interest. For example, in our hypothetical scenario, teacher practices are likely to vary across Head Start centers and grantees, so one could randomize coaches *by center* (assuming that the number of coaches per center is greater or equal to the number of experimental conditions in the chosen design, which is probably unlikely), or one could randomize Head Start coaches or centers by grantee. Blocking is desirable because it increases the precision of estimated effects, which in turn reduces the sample size needed to detect a particular target effect size (the MDES).[45]

However, one disadvantage of blocking is that it can reduce the external validity of the results, especially in an evaluation with many experimental conditions (such as a component evaluation). For example, assume that a half fractional factorial design is chosen for our five-component scenario, such that there are 16 experimental conditions. This means that in each block, there would have to be at least 16 random assignment units (coaches or centers, depending on the level of random assignment). Therefore, sites recruited for the study would have to be larger centers and/or grantees, and consequently the results from the evaluation might only be applicable to larger grantees/centers. To mitigate this limitation, one could randomize coaches or centers by *groups* of similar grantees or centers (e.g., combine two or more small grantees with similar characteristics into a block, and randomize centers within this combined block). This strategy would make it possible to recruit smaller sites and improver the external validity of the findings.

5. Using Covariates to Improve Precision

Another strategy that can be used instead of, or in addition to, blocking to reduce the sample size requirements for the study is to collect data on subjects' characteristics at baseline, prior to the start of the intervention(s). Like blocking, using baseline characteristics as control variables in the analysis increases the precision of estimated effects, which in turn reduces the sample size that is needed to detect a particular target effect size (the MDES). Baseline or pre-intervention measures of the outcome of interest are especially useful in this respect. In almost all program areas, the most predictive covariate of an outcome of interest is a pretest of that outcome (i.e., an earlier pre-intervention value of the outcome).[46] For example, in our

[44] Blocking random assignment decreases the standard error of the impact estimate, and therefore the MDES, for a given sample size (Bloom, 2006).

[45] Blocking can also be useful for program operations because random assignment (and program implementation) can happen as soon as a grantee is recruited into the study, without having to wait for the full study sample to be recruited.

[46] Controlling for covariates in the impact model improves the precision of the impact estimates, which decreases the MDES for a given sample size (Bloom, 2006).

hypothetical scenario, one could collect data on teacher practices at baseline before the coaching components are administered, and, in fact, the sample size requirements presented in Exhibit 10 assume that such baseline data would be available. If this were not the case, the sample size for the factorial design would increase from 220 to 252 centers.[47]

However, there is a trade-off that accompanies the use of baseline measures. Although it reduces the sample size needed to detect an effect of a given size (thereby reducing study costs), it also increases data collection costs. Evaluators should use simulations to determine whether the additional costs of data collection outweigh the cost savings of a smaller study sample. [48]

6. The Statistical Significance Level

Recall that the MDES is the smallest true impact that can be detected with a reasonable degree of power for a given level of statistical significance (Type I error). Two types of error should be of concern to evaluators:

- The Type I error rate (or the statistical significance level) is the probability of mistakenly concluding that an ineffective component is effective.
- The Type II error rate is the probability of mistakenly concluding that an effective component is ineffective. A study's power is the likelihood of correctly concluding that an effective component is effective (= 1 – the Type II error rate).

The tradition in scientific hypothesis testing is to fix the Type I error rate at 5 percent or smaller and then devote resources to achieving a Type II error rate of 20 percent or less (equivalently, power of 80 percent or greater).

However, one could consider relaxing these standard levels. For example, one could relax the standard 5 percent Type I error rate, and instead choose a statistical significance level that achieves a given level of power, say 80 percent. In our hypothetical scenario, assume that component effects are evaluated using a factorial design and that 100 Head Start centers are randomly assigned to experimental conditions. In this situation, the power to detect a target effect size of 0.25 would be approximately 65 percent, based on the standard Type I error of 5 percent. In other words, the probability of correctly concluding that an effective component is effective is only 65 percent. However, one could increase the power to 80 percent by setting the Type I error rate to 13 percent instead.[49]

[47] This number was obtained by setting the within-center R^2 to 0.

[48] Evaluators may also want to consider another reason for using baseline measures: they can be used to ensure that random assignment "worked" and also to describe the sample of participants at the start of the study.

[49] For these calculations, we make the following assumptions about parameters affecting the standard error of impact estimates: centers are randomly assigned to experimental conditions, but there is only one coach per center, so this is equivalent to randomizing coaches to experimental conditions; there are 4 teachers per center (or equivalently, coach); the center-level intra-class correlation in teacher practice is 17 percent, and teacher characteristics and baseline measures explain 35 percent of the variation between centers and 20 percent of the variation between teachers.

Increasing the Type I error may seem unorthodox, especially in light of the multiple hypothesis testing problem that arises in a study of component effects. Specifically, in a component evaluation, many hypothesis tests are conducted (at minimum, one per component), which increases the risk of concluding that a component is effective when in fact it is not. That is, it increases the probability of a Type I error relative to a study where only one impact is being estimated. From this perspective, the Type I error for sample size calculations should be decreased *below* the standard 5 percent level.

Yet it is important to bear in mind that the goal of a component evaluation is to help practitioners and policymakers develop stronger interventions. In other words, the intrinsic focus is on sound *decision-making* rather than formal hypothesis testing. From this perspective, reducing the Type II error rate (the likelihood that practitioners will not use a component that is effective) may be more important than reducing the Type I error rate (the likelihood that practitioners will use a component that is not effective).[50] Viewed in this light, adjustments for multiple hypothesis testing (which is related to the Type I error rate) are less crucial. It is also worth noting that when the study design is a factorial experiment with an approximately balanced sample, component effects are uncorrelated when effect coding is used, which alleviates the multiple hypothesis testing problem.

If p-values and the Type I error *do* matter, then of course one could make adjustments to the Type I error to account for multiple hypothesis testing. This would require a larger sample size and would therefore increase the complexity and cost of the study.

B. Untested Components: Fixed or Allowed to Vary Randomly?

Social interventions consist of many distinct components, some of which are strong candidates for an evaluation of component effects because they have the potential to affect outcomes and can be varied in an experiment. Some components, however, may be more difficult for evaluators to vary, or of lesser scientific interest. Such components are probably not good candidates for planned variation in an experimental study.

For components that will not (or cannot) be tested in the evaluation, an important question is whether their levels/content should be fixed by evaluators, as opposed to being allowed to vary naturally (and randomly) across sites. For example, in our hypothetical five-component study, coach credentials are *not* included in the list of tested components, yet a coach’s education and experience could potentially have an effect on teacher outcomes. If coach credentials are not tested, then evaluators are faced with two options—to fix coach credentials or to let this component vary naturally. To fix the component, the evaluation team could recruit only Head Start centers whose coaches have some set level of education and experience (for example, a bachelor’s degree and five years of experience). Alternatively, coach credentials could be allowed to vary randomly: any interested Head Start center could participate in the

[50] This is especially true if the study of component effects is going to be followed by another experiment (a two-group RCT) to confirm that an intervention package composed of the “best” components has an impact.

study, regardless of the credentials of its coaches. The advantage of using the latter strategy is that site recruitment would be less complicated and the study findings would be applicable to coaches of all credentials. Letting coach credentials vary randomly would also make it possible to examine whether coaching credentials moderate the effect of the tested components (see next section). On the other hand, the disadvantage of letting an untested component vary randomly is that it can reduce the precision of estimated effects and therefore increase the sample size requirements.

Evaluators need to weigh the trade-off between recruitment efforts, desired external validity, and sample size requirements. Depending on the nature of the untested components, there may also be a middle ground in which evaluators let the untested components vary naturally, but within a more limited band. For example, in our hypothetical scenario, the study could recruit Head Start centers with coaches who have between five and 10 years of experience.

C. The Importance of a Pilot Phase

Prior to the impact evaluation of a social intervention, a pilot phase is often used to help evaluators understand and address various issues related to study implementation and operations (e.g., intervention fidelity, likelihood of spillover, ability to monitor and maintain the service contrast, reliability of measures, site cooperation). These issues are even more salient in a component evaluation, because there are many more experimental conditions to implement and monitor, requiring more buy-in from study sites. Therefore, a pilot phase could be especially informative when undertaking an evaluation of component effects.

A pilot phase would also allow evaluators to gather more information on the factors discussed earlier in the context of choosing the sample size. For example, as already noted, it is important to choose components that show potential for having an impact on outcomes based on prior research, and that can be implemented with fidelity. However, if prior research is too scant to identify strong components—or if implementation fidelity could be variable—then a pilot phase could be used to identify the strongest components and their likely effect sizes, based on a strong quasi-experimental design. A pilot phase would also help evaluators establish the highest feasible upper level of the components to be tested, thereby allowing the study to maximize the service contrast.

V. Conclusion

The main conclusion from this review is that a factorial design is usually the most appropriate design for a study of component effects, for several reasons. First, factorial experiments provide estimates of a component's effect across all levels of the other components, which is useful information for policymakers and practitioners who are looking for components whose effect is robust across different intervention settings. Second, factorial designs make it possible to examine whether a component's effect is sensitive to the levels of the other components being tested (that is, whether there are interaction effects). Such findings are especially useful for the purposes of constructing or refining local interventions. The factorial design is also very sample-efficient because the entire sample is used to estimate each component effect.[51] Although a complete factorial design requires a large number of experimental conditions, a fractional factorial design can be used to reduce this burden. The fractional factorial design requires the same sample size as a complete factorial, but it has fewer experimental conditions and is therefore more feasible to implement and less costly.

In contrast, the other designs reviewed in this report—such as the comparative treatment (CT) design—are less suitable for a study of component effects. Although the CT design requires fewer experimental conditions than a factorial design, the sample size needed to detect an effect of a given magnitude is larger. Another limitation of the CT design is that its findings are very specific—they represent the effect of a given component when the other tested components are set to a particular level. Interactions between components cannot be examined, so it is not possible to understand how the effect of a component might be sensitive to how the other components are implemented. For this reason, CT designs are best used in situations where teasing out interaction effects is less relevant—that is, to directly compare the effects of different intervention models, or to evaluate the effect of adding or removing one component from an existing intervention.

The other designs reviewed in this report—the individual experiments (IE) design, the crossover design, and adaptive trials—have practical limitations that make them unlikely to be used to evaluate component effects in practice. The IE design is even more problematic than the CT design for estimating component effects, because it has more experimental conditions *and* it requires the largest study sample of all the designs discussed in this report. A crossover design would have even more experimental conditions than a factorial design, as well as higher data collection costs due to repeated rounds of assessment. An adaptive trial would also have high data collection costs, and similar to the CT design, it would provide estimates of component effects whose interpretation is less useful for decision-making purposes.

Despite these limitations, there may still be a tendency for evaluators to want to use the CT design for studies of component effects, because it requires fewer experimental conditions than a factorial design and is therefore less operationally complex. However, what is not well understood by some evaluators is that the CT design requires a much larger sample size to detect a component effect of a given magnitude, as clearly demonstrated by Exhibit 10 (p. 25). The additional cost of recruiting and serving more study participants may outweigh the cost savings of implementing fewer experimental conditions.

In addition to cost, it is also important to consider the scientific contributions of the study. The resource management principle states that evaluators should choose the design that provides the most useful information given its cost. Based on this principle, evaluators should weigh the cost differential between study design options against the quality of the information

[51] For these reasons, factorial designs are the design most often used in the screening experiment of the first phase of MOST approach described earlier in this report.

provided by each design. If a factorial design is more expensive than a CT design, evaluators and funders need to decide whether the cost savings of the CT design is worth the risk of obtaining findings about component effects that may be less useful in field settings. In the long run, gaining reliable and robust information on coaching components could result in a more efficient use of public resources.

VI. References

Baker, T. B., Mermelstein, R. J., Collins, L. M., Piper, M. E., Jorenby, D. E., Smith, S. S., . . . Fiore, M. C. (2011). New methods for tobacco dependence treatment research. *Annals of Behavioral Medicine, 41*, 192-207.

Berry, D. A. (2006). Bayesian clinical trials. *Nature Reviews, 5*, 27-36.

Black, A. R., Doolittle, F., Zhu, P., Unterman, R., & Grossman, J. B. (2008). *The evaluation of enhanced academic instruction in after-school programs: Findings after the first year of implementation (NCEE 2008-4021)*. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Bloom, H. S. (2005). Randomizing groups to evaluate place-based programs. In H. S. Bloom (Ed.), *Learning more from social experiments: Evolving analytical approaches* (pp. 115-172). New York: Russell Sage.

Bloom, H. S. (2006). *The core analytics of randomized experiments for social research*. New York, NY: MDRC.

Bloom, H. S., Hill, C., Black, A. R., & Lipsey, M. W. (2008). *Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions*. New York, NY: MDRC.

Caldwell, L. L., Smith, E. A., Collins, L. M., Graham, J. W., Lai, M., Wegner, L., . . . Jacobs, J. (2012). Translational research in South Africa: Evaluating implementation quality using a factorial design. *Journal of Research and Practice in Children's Services, 41*(2), 119-136.

Clarke, G. M., & Kempson, R. E. (1997). *Design and analysis of experiments*. New York NY: John Wiley & Sons.

Collins, L. M., Baker, T. B., Mermelstein, R. J., Piper, M. E., Jorenby, D. E., Smith, S. S., . . . Fiore, M. C. (2011). The multiphase optimization strategy for engineering effectice tobacco use interventions. *Annals of Behavioral Medicine*, *41*(2), 208-226.

Collins, L. M., Dziak, J. J., Kugler, K. C., & Trail, J. B. (submitted). Investigating the effectiveness of individual treatment components: The surprising efficiency of factorial experiments.

Collins, L. M., Dziak, J. J., & Li, R. (2009). Design of experiments with multiple independent variables: A resource management perspective on complete and reduced factorial designs. *Psychological Methods, 14*(3), 202-224.

Collins, L. M., Murphy, S. A., Nair, V. N., & Strecher, V. J. (2005). A strategy for optimizing and evaluating behavioral interventions. *Annals of Behavioral Medicine*, *30*(1), 65-73.

Collins, L. M., Trail, J. B., Kugler, K. C., Baker, T. B., Piper, M. E., & Mermelstein, R. J. (accepted pending minor revisions). Evaluating individual intervention components: Making decisions based on the results of a factorial component screening experiment. *Translational Behavioral Medicine*.

Dziak, J. J., Nahum-Shani, I., & Collins, L. M. (2012). Multilevel factorial experiments for developing behavioral interventions: Power, sample size, and resource considerations. *Psychological Methods, 17*, 153-175.

Friedman, L., Furberg, C., & DeMets, D. (1998). *Fundamentals of clinical trials*. New York, NY: Spring Verlag.

Garet, M. S., Cronen, S., Eaton, M., Kurki, A., Ludwig, M., & Jones, W. (2008). *The impact of two professional development interventions on early reading instruction and achievement (NCEE 2008-4030)*. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Green, D. P., Ha, S. E., & Bullock, J. G. (2010). Enough already about “black box” experiments: Studying mediation is more difficult than most scholars suppose. *The ANNALS of the American Academy of Political and Social Science, 628*(1), 200-208.

Kugler, K. C., Trail, J. B., Dziak, J. J., & Collins, L. M. (2012). *Effect coding versus dummy coding in analysis of data from factorial experiments (Technical Report 12-120)*: The Methodology Center, Pennsylvania State University.

Landry, S. (2007). *Teacher Behavior Rating Scale TBRS*. Houston, TX: Center for Improving the Readiness of Children for Learning and Education.

Lloyd, C. M., & Modlin, E. L. (2012). *Coaching as a key component in teachers' professional development: Improving classroom practices in Head Start settings*. Washington, D.C.: Office of Planning, Research, and Evaluation: Administration for Children and Families (http://www.mdrc.org/project/head-start-cares-project#featured_content).

McClure, J. B., Derry, H., Riggs, K. R., Westbrook, E. W., St. John, J., Shortreed, S. M., . . . An, L. C. (2012). Questions about quitting (Q(2)): Design and methods of a Multiphase Optimization Strategy (MOST) randomized screening experiment for an online, motivational smoking cessation intervention. *Contemporary Clinical Trials, 33*(5), 1094-1102.

Pianta, R., La Paro, K., & Hamre, B. K. (2008). *Classroom Assessment Scoring System*. Baltimore MD: Paul H. Brookes.

Scott, C. T., & Baker, M. (2007). Overhauling clinical trials. *Nature Biotechnology, 25*(3), 287-292.

Strecher, V. J., McClure, J. B., Alexander, G. W., Chakraborty, B., Nair, V. N., Konkel, J. M., . . . Pomerleau, O. F. (2008). Web-based smoking cessation programs: Results of a randomized trial. *American Journal of Preventive Medicine, 34*(373-381).

Taylor, J. E., Lloyd, C. M., Tout, K., Powell, D., Zaslow, M., Agnamba, L. A., . . . Farber, J. (2013). *Head Start professional development: Developing the evidence for best practices in coaching - Review of coaching frameworks, components, and outcomes* Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Weiss, M. J. (2010). *The implications of teacher selection and teacher effects in some education experiments (MDRC Methodological Working Paper)*. New York, NY: MDRC.

Appendix A. Glossary of Key Terms

- **Adaptive clinical trial.** A two-group or multiarm experiment in which the random assignment of subjects happens on a rolling basis, and where assignment probabilities are based on estimated treatment effects at the time at which a subject is assigned (with a subject having a higher probability of being assigned to a treatment that is more effective). This design requires frequent measurement of participant outcomes, and Bayesian analytical techniques. At this point in time, the use of adaptive trials with factorial designs is not well understood and they are mainly used in conjunction with a two-group experiment or a comparative treatment (CT) design. If it were used in a study of component effects, the adaptive trial would provide an estimate of a component's effect at a specified level of the other components; it would not provide estimates of interactions between components.

- **Aliasing.** Two effects are aliased when they cannot be estimated separately; instead, they must be estimated as a bundle. One can also think of aliasing as purposeful confounding. It is important to choose an experimental design that does not alias or bundle together effects that are of primary scientific interest. Ideally, each main effect (and other effects of primary interest) should be aliased only with effects or higher-order interactions that are expected to be small or zero.

- **Comparative treatment (CT) design.** This design is also called a multigroup or multiarm experiment. A CT design is typically used to test the differential impact of two or more interventions. For example, the differential impact of Intervention A and Intervention B could be examined by randomly assigning one group of participants to receive Intervention A, one group of participants to receive Intervention B, and the remainder of participants to the control group. A CT design can also be used to test component effects relative to an existing intervention. In this situation, one group is assigned to receive a particular intervention, while the other participants are assigned to experimental conditions representing variants of the intervention that differ from the original intervention with respect to a given component. To test the effect of *k* components, one would need a CT design with *k*+1 experimental conditions. If it were used in a study of component effects, this design would provide an estimate of a component's effect at a specified level of the other components; it would not provide estimates of interactions between components.

- **Component.** Any aspect of an intervention that can be reasonably separated out in order to study its effect on the outcomes of interest. A component can be a specific element, feature, or dimension of an intervention, or it can be a bundle of intervention elements that cannot function independently of each other.

- **Component levels.** The possible *values* of a component. Intrinsically, a component could feasibly take on a wide range of levels. However, for the purposes of an experiment, the evaluator must choose a finite number of levels to examine—usually two levels because this is most resource efficient. These two levels could be "yes/no" or "low/high." The levels should be set so that there is a strong contrast in the services received by participants at each level, while also making sure that both levels are feasible in a real-world context.

- **Crossover or switching design.** An experimental design in which subjects are assigned to all possible sequences of the K treatments (or components) being to be tested. The treatment (or component) received by a subject changes in each of *k* rounds; participant outcomes are assessed at the end of each round. This design has *k*! experimental conditions. This design assumes that there are no carryover effects, i.e. that the effect of a treatment (or component) disappears once it is withdrawn from the subject. If it were used in a study of component effects, the crossover design would provide an estimate of a component's effect at a specified level of the other components; it would not provide estimates of interactions between components.

- **Factorial design – complete (CF).** An experimental design used to estimate the effect of intervention components. The experimental conditions in this design are all possible combinations of the levels of the *k* components of interest—resulting in 2^k treatment conditions, assuming that each component has two levels. This design provides an estimate of a component's effect averaged across all possible levels of the other components, as well as estimates of interactions between components.

- **Factorial design – fractional (FF).** A factorial design whereby a *carefully selected* subset (fraction) of treatment conditions from the complete factorial (CF) design are retained in the experiment. A 1/2 factorial design, for example, retains half of the treatment conditions in the complete factorial design. The subset of conditions retained in the design are the ones that maximize the resolution of the design (minimize the amount of aliasing). An FF design is described by the number of times (*a*) that the original number of experimental conditions is reduced by half. For example, for a half fractional design, *a* = 1 or $(1/2)^a$. Assuming that each component has two levels, there are $2^{k\text{-a}}$ treatment conditions, where *k* is the number of components. The FF design provides an estimate of a component's effect averaged across a subset of all possible levels of the other components, as well as estimates of a subset of interactions between components.

- **Individual experiments (IE) design.** A design consisting of multiple "mini-experiments"—one experiment for each component whose effect is being tested. In a given experiment, part of the study sample for that experiment is assigned to receive a particular intervention (control group), while the remainder of the sample receives an intervention that differs from the base intervention by one component (treatment group).

The experiments can be conducted simultaneously or sequentially. If it were used in a study of component effects, this design would provide an estimate of a component's effect at a specified level of the other components; it would not provide estimates of interactions between components.

- **Interaction effect.** There is an interaction effect when the effect of a component differs depending on the level at which the other intervention components are set. Specifically, we define the two-way interaction between Component A and Component B as the effect of Component A *when Component B is set to its upper level* minus the effect of Component A when *Component B is set to its lower level*.

- **Minimum detectable effect (MDE).** The smallest true impact (on a particular outcome) that can be detected with a reasonable degree of power (for example, 80 percent) for a given level of statistical significance (usually a Type I error rate of 5 percent for a two-tailed test).

- **Minimum detectable effect size (MDES).** The MDE scaled as an effect size. This is obtained by dividing the MDE by the standard deviation of the outcome measure.

- **Power.** The likelihood of correctly concluding that an effective component or intervention is effective. This is equal to 1 – Type II error.

- **Randomized Control Trial (RCT):** A study design in which program participants are randomly assigned to one of two groups: (1) a treatment group that receives the social intervention or (2) a control group that does not receive the intervention. The impact of the intervention is estimated by comparing the average outcomes of individuals in the treatment to the average outcomes of individuals in the control group.

- **Resolution.** A design's resolution describes the amount of aliasing between main effects and interactions. Resolution is designated by a Roman numeral, usually III, IV, V or VI. A design's resolution increases as main effects and two-way interactions become increasingly free of aliasing with higher-order interactions. For example, in a design that is Resolution III, none of the main effects are aliased with each other, but they are aliased with interactions. In a Resolution IV design, main effects are not aliased with each other, nor aliased with any two-way interactions—only higher-order interactions. Ideally, evaluators should choose a design that is Resolution IV or higher, allowing them to estimate main effects as well as any important two-way interactions.

- **Type I error.** The probability of mistakenly concluding that an ineffective intervention or component is effective.

- **Type II error.** The probability of mistakenly concluding that an effective intervention or component is ineffective.

Appendix B. Suggested Reading List

1. Rationale for Factorial Experiments

Chakraborty, B., Collins, L. M., Strecher, V., & Murphy, S. A. (2009). Developing multicomponent interventions using fractional factorial designs. *Statistics in Medicine, 28,* 2687-2708. PMCID: PMC2746448.

Explains fractional factorial experiments for an audience of biostatisticians.

Collins, L. M., Dziak, J. J., Kugler, K. C., & Trail, J. B. (submitted). *Investigating the effectiveness of individual treatment components: The surprising efficiency of factorial experiments*.

Explains factorial experiments for those trained primarily in the RCT.

Collins, L. M., Dziak, J. J., & Li, R. (2009). Design of experiments with multiple independent variables: A resource management perspective on complete and reduced factorial designs. *Psychological Methods, 14*(3), 202-224. PMCID: PMC2796056.

Compares and contrasts individual experiments, comparative treatment (and similar) designs, factorial designs, and fractional factorial designs. It includes a brief conceptual tutorial on fractional factorial designs. It is aimed primarily at social and behavioral scientists.

Dziak, J. J., Nahum-Shani, I., & Collins, L. M. (2012). Multilevel factorial experiments for developing behavioral interventions: Power, sample size, and resource considerations. *Psychological Methods*. Advance online publication. doi: 10.1037/a0026972 PMCID: PMC3351535.

Discusses when it is feasible to conduct a factorial experiment when cluster randomization (e.g. assigning schools to experimental conditions) is necessary.

Nair, V., Strecher, V., Fagerlin, A., Ubel, P., Resnicow, K., Murphy, S. A., Little, R., Chakraborty, B., & Zhang, A. (2008). Screening experiments and the use of fractional factorial designs in behavioral intervention research. *American Journal of Public Health, 98,* 1354-1359. PMCID: PMC244645.

Discusses use of fractional factorial experiments in public health research.

2. Multiphase Optimization Strategy (MOST)[52]

Baker, T. B., Mermelstein, R. J., Collins, L. M., Piper, M. E., Jorenby, D. E., Smith, S. S., Schlam, T. R. Cook, J. W., & Fiore, M. C. (2011). New methods for tobacco dependence treatment research. *Annals of Behavioral Medicine, 41,* 192-207. PMCID: PMC3073306

and

Collins, L. M., Baker, T. B., Mermelstein, R. J., Piper, M. E., Jorenby, D. E., Smith, S. S., Schlam, T. R., Cook, J. W., & Fiore, M. C. (2011). The multiphase optimization strategy for engineering effective tobacco use interventions. *Annals of Behavioral Medicine, 41,* 208-226. PMCID: PMC3053423

These are companion articles. Together they are currently the most comprehensive introduction to MOST.

Caldwell, L. L., Smith, E. A., Collins, L. M., Graham, J. W., Lai, M., Wegner, L., Vergnani, T., Matthews, C., & Jacobs, J. (2012). Translational research in South Africa: Evaluating implementation quality using a factorial design. *Child and Youth Care Forum, 41*, 119-136.

Describes an example of a factorial experiment in a school setting.

Collins, L. M., Trail, J. B., Kugler, K. C., Baker, T. B., Piper, M. E., & Mermelstein, R. J. (accepted pending minor revisions). Evaluating individual intervention components: Making decisions based on the results of a factorial component screening experiment. *Translational Behavioral Medicine*.

Outlines a procedure for making decisions about component selection based on the results of a factorial or fractional factorial experiment.

McClure, J. B., Derry, H., Riggs, K. R., Westbrook, E. W., St. John, J., Shortreed, S. M., Bogart, A., & An, L. C. (2012). Questions about quitting (Q(2)): Design and methods of a Multiphase Optimization Strategy (MOST) randomized screening experiment for an online, motivational smoking cessation intervention. *Contemporary Clinical Trials, 33*(5), 1094-1102. PMCID: PMC3408878

Describes an application of MOST to development of an internet-delivered smoking cessation intervention.

Strecher, V. J., McClure, J. B., Alexander, G. W., Chakraborty, B., Nair, V. N., Konkel, J. M., Greene, S. M., Collins, L. M., Carlier, C. C., Wiese, C. J., Little, R. J., Pomerleau, C. S., & Pomerleau, O. F. (2008). Web-based smoking cessation programs: Results of a randomized trial. *American Journal of Preventive Medicine*, *34*, 373-381. PMCID: PMC2697448

An early application of a fractional factorial design in a behavioral study. This concerned development of an internet-delivered smoking cessation intervention.

[52] As described in Section I, *MOST* is an engineering-inspired approach to building, optimizing, and evaluating multicomponent behavioral interventions. The readings in this section describe MOST and the use of factorial designs to build and optimize behavioral interventions.

Appendix C. Fractional Factorial Designs: Software

Here we offer an illustration using Proc FACTEX in SAS. Comparable routines are available in Minitab, R, and other software.

There are several alternative ways to select a fractional factorial design; these are reviewed briefly in Collins et al. (2009). Here is one example:

```
proc factex;
title COACHING STUDY FIVE FACTORS;
factors TARGET MODE MODEL TOOLS SUPER;
SIZE DESIGN=16;
MODEL ESTIMATE=(TARGET--SUPER);
EXAMINE ALIASING(5);
RUN;
```

We can break this down line by line:

- `factors TARGET MODE MODEL TOOLS SUPER;` This lists the independent variables, that is, factors, that are to be included in the experiment. The default is that each factor has two levels.
- `SIZE DESIGN=16;` This specifies that a design with 16 experimental conditions is desired. Instead of requesting a design with a particular number of experimental conditions, it is possible to request a design of a particular resolution.
- `MODEL ESTIMATE=(TARGET--SUPER);` The MODEL command enables the user to specify two types of effects. ESTIMATE is for listing the effects that are of primary scientific interest and should be aliased only with effects expected to be negligible in size. Here we have listed only the main effects, but we could have listed any other effects, such as two-way interactions. We also could have listed effects in the NONNEGLIGIBLE category. These are effects that are not of scientific interest but that may be sizeable and therefore should not be aliased with effects in the ESTIMATE category. Any effects not designated ESTIMATE or NONNEGLIGIBLE are assumed to be negligible and therefore candidates for aliasing with effects in the ESTIMATE list.
- `EXAMINE ALIASING(5);` This requests a list of which effects are aliased with which up to the five-way interaction.

The output from Proc FACTEX will include the resolution of the resulting design. Upon request, the effect codes for the design can be provided. The design produced by this SAS code is the one in Exhibit 6 (2^{5-1}, or 16 experimental conditions). The aliasing structure of the design is shown in Exhibit 7.

Appendix D. Examples of Studies Based on a Comparative Treatment Design

Example #1: Head Start CARES Study

Interventions: The Head Start CARES demonstration is a national research project sponsored by the Office of Head Start and the Office of Planning, Research, and Evaluation in the Administration for Children and Families. This study was a randomized control trial testing the effects of three theoretically distinct, evidence-based social-emotional program enhancements in Head Start settings (Incredible Years Teacher Training Program, Preschool PATHS, and Tools of the Mind). The program enhancements trained lead and assistant teachers on classroom strategies that ranged from the delivery of standard classroom management procedures to a less common set of play-based activities designed to support self-regulation.

Target Population and Sample Size: The study was implemented in 104 centers in 17 Head Start grantees/delegate agencies from urban, suburban, and rural areas across the country. Grantees/delegate agencies were excluded from the sample for a number of reasons, including if they were already systematically implementing a social-emotional curriculum or only had 3-year-old classrooms. Teachers in 307 classrooms participated in the study, which included over 2,880 4-year-olds and 960 3-year-olds.

Experimental Design: Using a group-based randomized design, grantee/delegate agencies were stratified by region of the country, racial/ethnic composition of child enrollment, and urbanity of the location. Each Head Start center within a block (strata) was randomly assigned one of the three different social-emotional enhancements or to a business as usual comparison group. This design allowed an examination of the impact of each social-emotional program enhancement in comparison to business as usual.

Outcomes of Interest: Teacher reports on children's social-emotional and academic skills; direct assessments of children's social-emotional and academic skills; parent surveys of children's social-emotional skills; classroom observations of teachers' practice; and teachers' demographic and psychosocial characteristics.

Findings: No published findings yet.

Further Reading: Lloyd, C. M., and Modlin, E. L. (2012). *Coaching as a key component in teachers' professional development: Improving classroom practices in Head Start settings*. Washington, D.C.: Office of Planning, Research, and Evaluation: Administration for Children and Families. (http://www.mdrc.org/project/head-start-cares-project#featured_content)

Example #2: Professional Development Interventions in Reading Instruction Study

Interventions: The study examines the impact of year-long professional development (PD) interventions aimed at improving teachers' knowledge of reading principles and reading instruction. The PD intervention consisted of an institute series for teachers that focused on reading principles and how children learn to read; the institute series began in the summer and continued through the school year. The study also examined the effect of the same institute series *plus* in-school coaching that focused on showing teachers how to integrate this knowledge into teaching.

Target Population and Sample Size: The study was implemented in 90 schools in six school districts (a total of 270 second-grade teachers). Schools selected for the study were high-poverty urban or urban fringe public elementary schools where fewer than half the students were designated as English language learners (ELLs).

Experimental Design: In each district, equal numbers of schools were randomly assigned to receive the teacher institute series, the institute series *plus* in-school coaching, or a business as usual group that only received the usual PD offered by the district (30 schools per group). This design made it possible to determine the impact of each of the institute series relative to business-as-usual—and also to determine the incremental impact of coaching above and beyond the institute series.

Outcomes of Interest: Teachers' knowledge of scientifically based reading instruction based on a standardized test; three teacher practices targeted by the PD and measured using classroom observations (explicit instruction in reading, use of guided student practice in reading, and differentiated instruction to address students' diverse needs); and student reading test scores as measured using district tests.

Findings: Both PD interventions produced positive impacts on teachers' knowledge of reading instruction and on explicit instruction in reading. Unfortunately, these impacts had disappeared by the end of the year following the PD intervention. Impacts on student achievement were not statistically significant. The incremental effects of coaching (relative to just the institute series) were not statistically significant.

Further Reading: Garet, M. S., Cronen, S., Eaton, M., Kurki, A., Ludwig, M., Jones, W., et al. (2008). *The impact of two professional development interventions on early reading instruction and achievement* (NCEE 2008-4030). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.