Statistics for Social and Behavioral Sciences

Linda M. Collins · Kari C. Kugler
*Editors*

# Optimization of Behavioral, Biobehavioral, and Biomedical Interventions

## Advanced Topics

Springer

# Statistics for Social and Behavioral Sciences

**Series editor**
Stephen E. Fienberg (in memoriam)
Carnegie Mellon University
Pittsburgh, PA, USA

Statistics for Social and Behavioral Sciences (SSBS) includes monographs and advanced textbooks relating to education, psychology, sociology, political science, public policy, and law.

More information about this series at http://www.springer.com/series/3463

Linda M. Collins • Kari C. Kugler
Editors

# Optimization of Behavioral, Biobehavioral, and Biomedical Interventions

## Advanced Topics

Springer

*Editors*
Linda M. Collins
The Pennsylvania State University
The Methodology Center
University Park, PA, USA

Kari C. Kugler
The Pennsylvania State University
The Methodology Center
University Park, PA, USA

ISSN 2199-7357 ISSN 2199-7365 (electronic)
Statistics for Social and Behavioral Sciences
ISBN 978-3-319-91775-7 ISBN 978-3-319-91776-4 (eBook)
https://doi.org/10.1007/978-3-319-91776-4

Library of Congress Control Number: 2018947430



Printed on acid-free paper

This Springer imprint is published by the registered company Springer International Publishing AG part of Springer Nature.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

# Preface

Behavioral, biobehavioral, and biomedical interventions play an important role in society worldwide. These interventions are aimed at, for example, helping people quit smoking; improving reading skills in children; helping autistic children learn how to communicate verbally; improving family functioning; keeping convicted criminals who have served their time from engaging in criminal activity; treating cancer, diabetes, depression, and many other diseases and health problems; slowing the progression of heart failure; preventing the onset of drug abuse; and improving treatment regimen compliance in people living with HIV. These are just a few of many, many examples.

This book and another book, titled *Optimization of Behavioral, Biobehavioral, and Biomedical Interventions: The Multiphase Optimization Strategy (MOST)* (Collins, 2018), are companion volumes. Both are focused on MOST, an engineering-inspired framework for arriving at and then evaluating an optimized intervention. The objective is to develop an intervention that is not only effective but also efficient, economical, and scalable. MOST consists of three phases: preparation, optimization, and evaluation. Activities in the preparation phase include selection of the components that are candidates for inclusion in the intervention and development of a detailed conceptual model of the process to be intervened upon. In the preparation phase, the investigator also specifies an optimization criterion. This criterion operationalizes the goal of optimization. For example, if it has been established that to be scalable a particular intervention must cost no more than $400 per participant to implement, an appropriate optimization criterion would be "the most effective intervention that can be obtained for no more than $400 per participant in implementation costs." In the optimization phase, which occurs before an intervention is evaluated in an RCT, one or more optimization trials are conducted to gather information on the individual and combined effects of the candidate components. This information, along with the optimization criterion, forms the basis for selection of the components and component levels that make up the optimized intervention. The optimization trial may use any of a wide variety of experimental designs and approaches, depending on the type of intervention to be optimized, the precise research questions that are of interest, and the circumstances.

In the evaluation phase, the effectiveness of the optimized intervention is confirmed in a standard RCT. If the optimization criterion was appropriately specified, the resulting intervention will be immediately scalable.

Collins (2018) provides a comprehensive introduction to MOST. In addition to an overview of MOST, the book includes information on developing a conceptual model; using factorial and fractional factorial designs in optimization trials, with an entire chapter devoted to interactions; applying the resource management principle when selecting an experimental design; making decisions about selection of components and component levels based on experimental results; and numerous other topics. The book also includes a chapter introducing adaptive interventions.

Early in the process of planning a book about MOST, it became clear that a number of topics would need to be covered to arrive at a comprehensive treatment. An in-depth treatment of the process of arriving at a conceptual model was needed. Investigators wanting to conduct factorial optimization trials were asking for practical advice about implementing large and complex experiments in field settings. Intervention scientists who worked in populations with a cluster structure, such as educational researchers, had been asking whether and how they could appropriately conduct optimization trials. Clarification was needed on approaches such as the SMART experimental design and system identification experiments, and how they fit into the MOST framework. There was perennial confusion about the difference between conducting a factorial ANOVA using effect coding and using dummy coding. It seemed natural that cost-effectiveness could be a consideration in the optimization phase of MOST, but it was not apparent how. Guidance was needed on how to take advantage of the possibilities for interesting mediation analyses opened up by factorial optimization trials.

Linda Collins was not an expert in many of these topics, and it was clear that other authors would be able to do a much better job of presenting them. An edited book, with chapters written by experts in each area, was needed in addition to an authored book. Dr. Kari Kugler agreed to serve with Dr. Collins as coeditor. The editors were extremely fortunate that a number of outstanding academics agreed to contribute chapters.

## The Chapters in This Book

The first chapter, by Kugler, Wyrick, Tanner, Milroy, Chambers, Ma, Guastaferro, and Collins, describes a critically important, but too often overlooked, aspect of intervention optimization: the development of a detailed and highly specific conceptual model. Specification of a conceptual model is often a demanding and challenging task, requiring the integration of a diverse body of scientific literature and input from many members of the research team. However, it is worth the time and effort, because it is ultimately rewarding to arrive at a sophisticated conceptual model that will provide a firm conceptual foundation for the remainder of the preparation phase as well as the optimization and evaluation phases of MOST.

The chapter "Using the Multiphase Optimization Strategy (MOST) to Develop an Optimized Online STI Preventive Intervention Aimed at College Students: Description of Conceptual Model and Iterative Approach to Optimization" also introduces the idea of taking an iterative approach to optimization. In this approach, successive optimization trials are performed, with the objective of improving the intervention by revising or replacing weak or inert components and re-testing the components.

Readers who would like to use MOST in their work but are uneasy about implementation of an experiment that can be much more complex than an RCT will find that the chapter "Implementing Factorial Experiments in Real-World Settings: Lessons Learned While Engineering an Optimized Smoking Cessation Treatment" provides a wealth of valuable information. For nearly 10 years, Piper, Schlam, Fraser, Oguss, and Cook have successfully conducted factorial optimization trials in ordinary health care settings. In this chapter, they offer practical advice and lessons learned based on their extensive experience in the implementation of large factorial optimization trials in real-world field settings. Piper et al. discuss going from selection of intervention components to a workable experimental design; maintaining a high level of fidelity when conducting a complex experiment in the field; conducting random assignment with as many as 32 experimental conditions; and other considerations of particular interest to scientists who are relatively new to MOST. The chapter "Implementing Factorial Experiments in Real-World Settings: Lessons Learned While Engineering an Optimized Smoking Cessation Treatment" will be helpful to readers who would like to know how to implement large factorial experiments in field settings successfully and with few protocol deviations.

The chapter "Multilevel Factorial Designs in Intervention Development," by Nahum-Shani and Dziak, discusses design of optimization trials, statistical power, and analysis of the resulting data when there is a multilevel (also called hierarchical, cluster, or nested) structure. A multilevel structure can occur naturally when experimental subjects are grouped in schools, neighborhoods, clinics, families, or some other unit. A multilevel structure can also be induced by the experimenter, for example, if part of the experiment involves assigning individuals to some kind of group-delivered treatment. The presence of a multilevel structure has different implications for experimental design, data analysis, and statistical power depending on whether the clustering is naturally occurring or experimenter-induced and whether individuals or entire clusters are to be randomly assigned to experimental conditions. Nahum-Shani and Dziak provide a careful and comprehensive review that will help investigators decide on the best way to conduct an optimization trial when a multilevel structure must be considered. This chapter may be of particular interest to scientists developing educational or other school-based interventions.

In an adaptive intervention, the intervention content, dose, or approach can be varied across participants and across time, with the objective of achieving or maintaining a good outcome for all participants (see Chapter 8 in the companion volume). Adaptive interventions range in intensity of adaptation. In low-intensity adaptive interventions, the content, dose, or approach is varied only a few times, or adaptation occurs infrequently. In the chapter "Experimental Designs for Research

on Adaptive Interventions: Singly and Sequentially Randomized Trials," Almirall, Nahum-Shani, Wang, and Kasari discuss the design of optimization trials when the objective is optimization of a low-intensity adaptive intervention. Almirall et al. demonstrate that a variety of experimental designs can be appropriate and remind the reader that the choice of design must be based on the precise scientific questions motivating the experiment. This chapter reviews a number of experimental design alternatives, including types of singly randomized trials and sequential, multiple assignment, randomized trials (SMARTs), most of which are variations on the factorial experiment.

In contrast to low-intensity adaptive interventions, intensively adaptive interventions may vary the content, dose, or approach frequently, for example, daily or even several times per day. For example, mhealth interventions, in which the intervention is delivered via a mobile device app, are often intensively adaptive. In the chapter "Intensively Adaptive Interventions Using Control Systems Engineering: Two Illustrative Examples," Rivera, Hekler, Savage, and Downs discuss one approach to design of optimization trials when the objective is to optimize an intensively adaptive intervention. Their approach is not a variation on the factorial experiment. Instead, these authors take a control engineering perspective. From this perspective, the outcome, along with the behaviors and other factors that influence the outcome, is considered a dynamical system, and the adaptive intervention is a controller that can be used to modulate this system. Then the optimization trial is a system identification experiment, which provides the data needed to develop the controller. This chapter will appeal both to behavioral scientists considering using control engineering principles in their work and to engineers who may be interested in applying their skill set in the behavioral sciences.

Once an optimization trial has been conducted, the data need to be analyzed properly so that the results can be used in making decisions about which components and component levels will make up the optimized intervention. In the companion volume, Collins recommended using effect coding rather than dummy coding when analyzing data from a factorial optimization trial. However, dummy coding is more familiar to many behavioral scientists. In the chapter "Coding and Interpretation of Effects in Analysis of Data from a Factorial Experiment," Kugler, Dziak, and Trail compare and contrast effect coding and dummy coding of factorial experiments. They demonstrate that in most cases, effect coding and dummy coding produce different estimates of individual effects (although the omnibus $F$ will be identical). They also explain that effect coding models effects that correspond to the definitions of analysis of variance (ANOVA) main effects and interactions that appear in most statistics textbooks, whereas in general dummy coding models a different set of effects. The chapter "Coding and Interpretation of Effects in Analysis of Data from a Factorial Experiment" will clarify this important issue for data analysts and help the reader to see why effect coding is usually a better choice for analysis of data from an optimization trial.

In the companion volume, several different possible goals for optimization are discussed, with the emphasis on seeking the most effective intervention that can be obtained subject to a specified fixed upper limit on cost. However, in many

situations it may be desired to use the results of the optimization trial along with data on cost to identify the set of components and component levels that represents the most cost-effective intervention. This requires a more sophisticated approach to making decisions about selection of components and component levels; at this writing, there are still many unanswered questions about how to accomplish this. In the chapter "Optimizing the Cost-Effectiveness of a Multicomponent Intervention Using Data from a Factorial Experiment: Considerations, Open Questions, and Tradeoffs Among Multiple Outcomes," Dziak discusses issues and open research areas related to cost-effectiveness and MOST.

Optimization trials yield rich data that can form the basis for interesting and informative secondary analyses. The final chapter, "Investigating an Intervention's Causal Story: Mediation Analysis Using a Factorial Experiment and Multiple Mediators," discusses one type of secondary analysis of a factorial optimization trial, mediation analysis. The majority of readers of this book will have some familiarity with mediation analysis of data from an RCT. The purpose of such analyses is to determine which variables mediated any observed treatment effect, and thereby obtain an empirical sense of the mechanisms underlying the intervention. Mediation analysis of data from a two-arm RCT can be highly informative. However, because in an RCT the treatment is an aggregate of all the components, it is not possible to determine which variables mediate which individual components. Smith, Coffman, and Zhu review the possibilities that are opened up by mediation analysis of data when the treatment is a factorial experiment rather than a two-arm RCT. Here it is possible to model mediation of the effect of a single factor, and even to model mediation of an interaction effect! Mediation analysis of the data from a factorial optimization trial can be helpful in the optimization phase of MOST and is likely to be particularly helpful in informing the preparation phase of a subsequent cycle of MOST.

## How to Use This Book

From the beginning, the objective was that the two companion volumes would be tightly integrated. The reader will see that each book cites the other repeatedly. Moreover, the chapters in the present book assume an understanding of the material in Collins (2018), so it is a good idea to have read that book before reading this one. Each chapter in this book stands alone, and, unlike the chapters in the companion volume, it is not necessary to read them in the order they appear.

The eight chapters in this book have been presented according to roughly where they fall in the MOST process. The first two chapters discuss matters pertaining primarily to the preparation phase of MOST and the early part of the optimization phase, and the remaining chapters pertain to designing an optimization trial, conducting primary data analysis, selecting components and component levels, and conducting secondary analysis of data from an optimization trial. The material in the Smith et al. chapter could also be considered part of the preparation phase, because

mediation analyses are an excellent source of information useful in updating and refining a conceptual model. This may be done in preparation for a subsequent round of optimization aimed at further improvements to the intervention.

Intervention scientists often work in teams, and different team members may have different roles. For a scientist whose role is primarily intervention development, chapters "Using the Multiphase Optimization Strategy (MOST) to Develop an Optimized Online STI Preventive Intervention Aimed at College Students: Description of Conceptual Model and Iterative Approach to Optimization" and "Optimizing the Cost-Effectiveness of a Multicomponent Intervention Using Data from a Factorial Experiment: Considerations, Open Questions, and Tradeoffs Among Multiple Outcomes" may be of particular interest. For a team member responsible for implementation, the chapter "Implementing Factorial Experiments in Real-World Settings: Lessons Learned While Engineering an Optimized Smoking Cessation Treatment" is essential reading. The chapters "Multilevel Factorial Designs in Intervention Development," "Experimental Designs for Research on Adaptive Interventions: Singly and Sequentially Randomized Trials," and "Intensively Adaptive Interventions Using Control Systems Engineering: Two Illustrative Examples" were written to be helpful to those responsible for selecting the design of the optimization trial. Those chapters, along with chapters "Coding and Interpretation of Effects in Analysis of Data from a Factorial Experiment" and "Investigating an Intervention's Causal Story: Mediation Analysis Using a Factorial Experiment and Multiple Mediators," are likely to be interesting to a statistician, methodologist, or data analyst.

We hope you find this book and its companion helpful and that you have an opportunity to use MOST in your work.

University Park, PA
2018

Linda M. Collins
Kari C. Kugler

# Contents

# Using the Multiphase Optimization Strategy (MOST) to Develop an Optimized Online STI Preventive Intervention Aimed at College Students: Description of Conceptual Model and Iterative Approach to Optimization


**Kari C. Kugler, David L. Wyrick, Amanda E. Tanner, Jeffrey J. Milroy, Brittany Chambers, Alice Ma, Kate M. Guastaferro, and Linda M. Collins**

**Abstract** This chapter describes some aspects of an application of the multiphase optimization strategy (MOST) to optimize and evaluate itMatters, an online intervention that targets the intersection of alcohol use and sexual behaviors to reduce sexually transmitted infections (STIs) among college students. The chapter emphasizes two aspects of this application. First, we describe the development of a detailed conceptual model during the preparation phase of MOST. This conceptual model guided decisions such as the choice of outcome variables. Second, we describe an iterative approach to experimentation during the optimization phase of MOST. The objective of the iterative approach is to build a highly effective intervention by using repeated optimization trials to evaluate which intervention components meet a given criterion for effectiveness and which do not. Revisions are undertaken to improve the components that do not meet the criterion, and then a

K. C. Kugler (✉) · L. M. Collins
The Pennsylvania State University, The Methodology Center, University Park, PA, USA
e-mail: kck18@psu.edu

D. L. Wyrick · A. E. Tanner · J. J. Milroy · B. Chambers · A. Ma
Department of Public Health Education, The University of North Carolina Greensboro, Greensboro, NC, USA

K. M. Guastaferro
The Methodology Center, The Pennsylvania State University, University Park, PA, USA



L. M. Collins, K. C. Kugler (eds.), *Optimization of Behavioral, Biobehavioral, and Biomedical Interventions*, Statistics for Social and Behavioral Sciences,
https://doi.org/10.1007/978-3-319-91776-4_1

subsequent optimization trial is used to reevaluate the components. This iterative approach has the potential to enable the investigator to develop more effective, efficient, economical, and scalable interventions.

## 1 Introduction

Approximately 70% of college students are sexually experienced, yet only half of sexually active students report using a condom during their last sexual encounter, and 75% report inconsistent or no condom use (American College Health Association, 2016). Concurrent and casual sexual partnerships are also common among college students (Olmstead, Pasley, & Fincham, 2013), with one third reporting not using a condom during a penetrative hookup (Fielder & Carey, 2010); a hookup is a casual sexual encounter without the expectation of dating or a romantic relationship (Garcia, Reiber, Massey, & Merriwether, 2012). Inconsistent condom use (Trepka et al., 2008); multiple, concurrent partners (Lewis, Miguez-Burban, & Malow, 2009); and penetrative hookups (Paul, McManus, & Hayes, 2000) are all high-risk behaviors that contribute to the high prevalence of sexually transmitted infections (STIs) among college students (Kann et al., 2016). Drinking alcohol is a known risk factor for unprotected sex, particularly among college students, and, by extension, a risk factor for exposure to STIs. An extensive body of research (Scott-Sheldon et al., 2016; Shuper et al., 2010) has documented a consistently strong and positive, but also complex, relationship between alcohol use and unprotected sex (Ebel-Lam, MacDonald, Zanna, & Fong, 2009; Prause, Staley, & Finn, 2011; Shuper et al., 2010).

Numerous individual-level interventions for college students have been developed that focus separately on alcohol use (Carey, Scott-Sheldon, Elliot, Bolles, & Carey, 2009) and condom use (Scott-Sheldon, Huedo-Medina, Warren, Johnson, & Carey, 2011), but few have directly emphasized the alcohol-sex relationship. Dermen and Thomas (2011) found that a brief intervention combining alcohol risk-reduction content with HIV risk-reduction content produced effects on sexual risk behaviors (e.g., frequency of unprotected sex), but not on alcohol use frequency or intensity. Lewis and colleagues (2014) found that the use of personalized normative feedback specific to drinking in sexual situations was effective at reducing alcohol use and sexual risk behaviors (e.g., drinking alcohol prior to or during sex).

Although these studies suggest that interventions focusing on the alcohol-sex relationship show promise, more research is needed to overcome some limitations. For example, the study by Dermen and Thomas (2011) was based on a relatively small sample of predominately White students. Lewis and colleagues (2014) included only sexually active students with minimal levels of drinking behavior and focused solely on challenging normative misperceptions. It is unclear whether the findings would generalize to a more diverse population with a wider range of sexual experiences and drinking behaviors. It is also unclear whether an intervention would be more effective if it targeted other constructs beyond correcting normative misperceptions.

Thus there is a need for development of an effective STI preventive intervention that targets the intersection of alcohol use and sexual risk behaviors and is aimed at a diverse population of college students. This chapter describes an ongoing study that is attempting to accomplish this aim by developing an online intervention called itMatters. The objective of itMatters is to prevent STIs in college students by focusing on the intersection of alcohol use and sexual risk behaviors. We are applying the multiphase optimization strategy (MOST) to develop, optimize, and evaluate itMatters.

MOST is an engineering-inspired framework for building more effective, efficient, economical, and scalable interventions. MOST includes three phases: preparation, optimization, and evaluation. As part of the preparation phase, a carefully specified, theoretically driven conceptual model is established to articulate how each component that is a candidate for inclusion in the intervention is hypothesized to affect the outcome. During the optimization phase, the effectiveness of the individual intervention components is examined experimentally. Based on the information obtained via this experimentation, the components and component levels that make up the optimized intervention are selected. In the evaluation phase, the resulting optimized intervention is evaluated using a standard RCT.

MOST has been applied to develop interventions in a wide range of health areas, including school-based prevention of alcohol and drug use and HIV (Caldwell et al., 2012), drug use among NCAA athletes (Wyrick, Rulison, Fearnow-Kenney, Milroy, & Collins, 2014), smoking cessation (e.g., Baker et al., 2016), weight loss (Pellegrini, Hoffman, Collins, & Spring, 2014, 2015), and cardiology (Huffman et al., 2017). For a more detailed description of MOST, see the companion volume (Collins, 2018).

## *1.1 The Current Chapter*

The purpose of the current chapter is to describe our application of MOST to optimize and evaluate the itMatters intervention. The chapter emphasizes two aspects of this application. First, we describe the development of a detailed conceptual model during the preparation phase of MOST (see Chapter 2 in the companion volume). Second, we describe an iterative approach to experimentation during the optimization phase of MOST. The objective of the iterative approach is to build a highly effective intervention by using repeated optimization trials to evaluate which intervention components meet a given criterion for effectiveness and which do not. Revisions are undertaken to improve the components that do not meet the criterion, and then a subsequent optimization trial is used to reevaluate the components.

# 2 The Conceptual Model of the Intersection of Alcohol Use and Sexual Behaviors

## *2.1 Overview*

During the preparation phase of MOST, a carefully specified, theoretically driven conceptual model is articulated. As noted in the companion volume, the purpose of the conceptual model is to express "all of what is known or hypothesized about how the intervention under development is to intervene on the behavioral, biobehavioral, or biomedical process" (Collins, 2018, p. 64). In other words, the conceptual model forms the basis for the intervention by specifying the set of components that are candidates for inclusion in the intervention, identifying the proximal mediators that are immediate targets of each component, and outlining the causal pathways by which these candidate intervention components are intended to have an impact on the proximal and distal outcomes.

The conceptual model that forms the basis of the itMatters intervention expresses how alcohol use is hypothesized to lead to sexual risk behaviors (e.g., unprotected sex, penetrative hookups) and how this increases the risk for STIs among college students. Because examination of the intersection of alcohol use and sex has been limited primarily to laboratory studies (Davis et al., 2014; George et al., 2009; Prause et al., 2011), this conceptual model has been informed by empirical research and behavioral theory on alcohol use and sexual risk behaviors separately and together.

The itMatters conceptual model is depicted in Fig. 1. The purpose of Fig. 1 is to provide a visual representation of how the intervention components are hypothesized to prevent alcohol-related sexual risk behaviors and, ultimately, STIs. As the figure suggests, a conceptual model is similar to a logic model but goes a step further by detailing the mechanisms by which each intervention component is expected to effect change in the primary outcome(s) (see Chapter 2 in the companion volume for more detail).

Before examining Fig. 1 in more detail, it is necessary to define two terms. The first is protective behavioral strategies (PBS). In this case, PBS are approaches an individual uses to reduce the potentially negative consequences associated with alcohol-related sexual risk behaviors (Treloar, Martens, & McCarthy, 2015). Examples of PBS include limiting alcohol intake; using a condom, including making sure they are readily available and that the skills needed to use them properly have been acquired; designating a friend to step in if an individual appears headed for excessive alcohol use or an unintended sexual encounter; and proactively sharing sexual boundaries with a partner.

The second term to be defined is myopic effects. Myopic effects are cognitive effects of alcohol that affect an individual's appraisal of sex potential and risk (Sevincer & Oettingen, 2014). In particular, alcohol use leads to cognitive impairment that can affect decision-making and lead to a higher probability of risk-taking. Alcohol myopia theory (Sevincer & Oettingen, 2014) helps explain this.

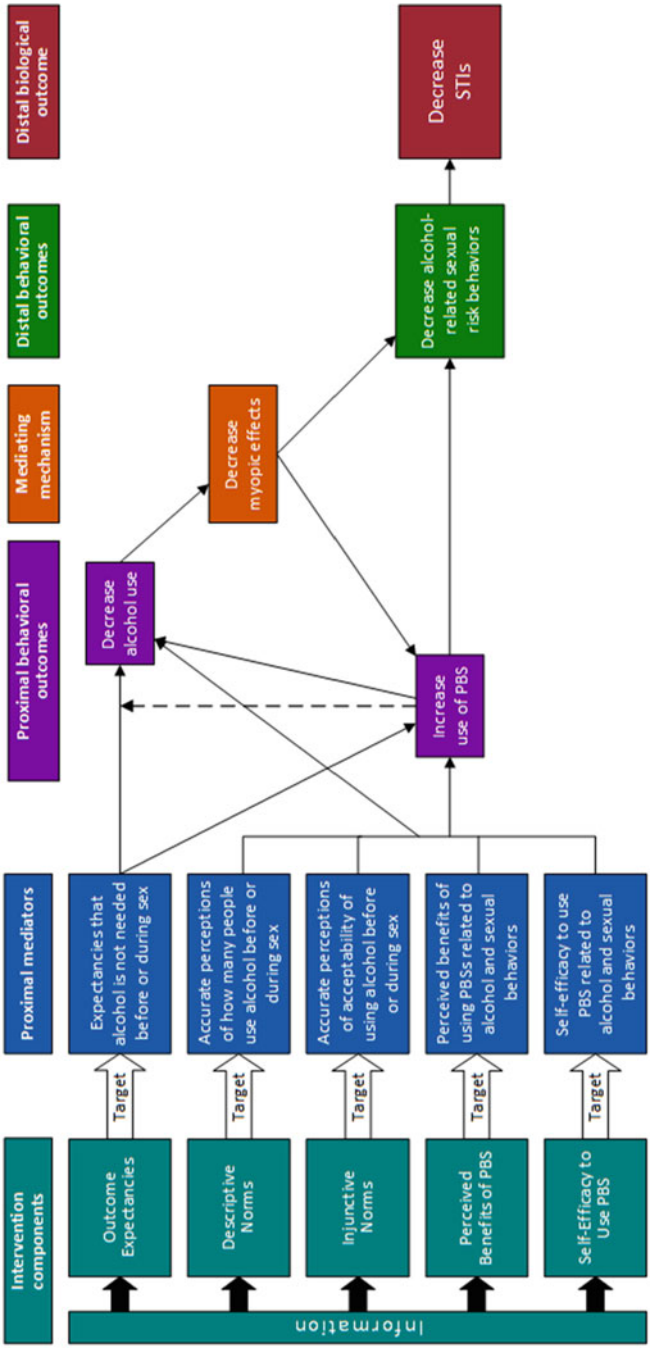


**Fig. 1** Conceptual model for a behavioral intervention to reduce sexually transmitted infections (STIs) among college students. PBS stands for protective behavioral strategies

Alcohol myopia theory posits that alcohol increases a person's concentration on the immediate situation (e.g., enjoyment), limits higher-level cognitive functioning, and reduces attention on more distant events or cues (e.g., reducing the risk of unprotected sex); these effects are intensified as the quantity or dose of alcohol increases (Dry, Burns, Nettelbeck, Farquharson, & White, 2012). As suggested by this theory, the mechanism of how alcohol use affects sexual behaviors is further influenced by several factors, such as primary (e.g., sex potential) and secondary (e.g., STI risk) appraisals, which are anticipated to influence alcohol-related sexual behaviors directly (Purdie et al., 2011) and indirectly through PBS strategies (Abbey, Saenz, & Buck, 2005).

Examining Fig. 1 from left to right shows how each component targets a particular putative proximal mediator (henceforth termed proximal mediator). These proximal mediators, in turn, affect their respective proximal behavioral outcomes: they reduce alcohol use and increase the use of PBS. A decrease in alcohol use leads to a decrease in myopic effects, which decreases the likelihood of engaging in alcohol-related sexual risk behaviors directly and indirectly by increasing the likelihood of using PBS. Increased use of PBS leads to a decrease in the likelihood of engaging in alcohol-related sexual risk behavior.

Figure 1 is not a structural equation modeling diagram, although it resembles one in some ways. One important difference is that Fig. 1 is meant to convey the rationale for the intervention, not provide a summary of how data would be analyzed. For this reason, the figure does not contain an arrow representing every anticipated nonzero regression coefficient. Another difference is that some of the boxes represent an increase or decrease in a variable. This is not always a feature of figures representing conceptual models. We used this approach here to avoid complicating the figure with negative signs on some paths.

## *2.2 Intervention Components*

Figure 1 shows six components. One component, information, is represented by a bar on the far left of the figure to indicate that information is considered a necessary foundation for the other components. The information component will not be examined experimentally during the optimization phase. Because this material is foundational to the remaining components, an a priori decision has been made to include it in the intervention. All experimental subjects will be provided with the information component. The remaining five components are candidates for inclusion in itMatters and therefore will be examined experimentally. These are listed in the left-hand area of Fig. 1. Each is labeled with a brief name of the proximal mediator it targets: outcome expectancies, descriptive norms, injunctive norms, perceived benefits of PBS, and self-efficacy to use PBS. The arrow labeled Target indicates the immediate target of each component. (Note that even though each component is connected by an arrow only to the mediator it directly targets, a component may have an effect on other mediators. As mentioned previously,

the purpose of Fig. 1 is to depict the reason why a component is a candidate for inclusion in the intervention, not to show every possible nonzero path.) Figure 1 specifies the hypothesized causal pathways of the effect of each of these intervention components on the proximal behavioral outcomes (i.e., alcohol use, use of PBS), the distal behavioral outcome (i.e., alcohol-related sexual risk behaviors), and the distal biological outcome, STIs, via the proximal mediators. A detailed description of the pathways is provided below. First, we review each candidate component.

### 2.2.1 Outcome Expectancies

Informed by expectancy theory (Jones, Corbin, & Fromme, 2001), this component challenges positive expectancies related to alcohol use before or during sex, such as expectancies that using alcohol will increase the likelihood of engaging in sex (Davis et al., 2010). Thus, the component is designed to convince participants that no, or at most limited, alcohol use is needed before or during sex. Outcome expectancies are consistently associated with behavioral outcomes, with positive expectancies associated with an increased likelihood of alcohol consumption (Davis et al., 2010) and a decrease in PBS (Logan, Koo, Kilmer, Blayney, & Lewis, 2015). There is a notable moderating effect by the use of PBS.

Grazioli and colleagues (2015) found that the association between expectancies and alcohol use was weaker for students with a high use of PBS (e.g., predetermined strategies to limit or stop drinking). This implies an interaction between PBS and the outcome expectancies proximal mediator. This is represented by a dashed line in Fig. 1 running from the box representing PBS to the line representing the relation between outcome expectancies and alcohol use.

### 2.2.2 Descriptive Norms and Injunctive Norms

Two different types of social norms, descriptive (perceived prevalence of a behavior; social norms theory (Berkowitz, 2004)) and injunctive (perceived peer approval of a behavior (Ajzen, 1991)), are positively associated with alcohol use (Reid & Carey, 2015) and inversely associated with PBSs (Lewis, Rees, Logan, Kaysen, & Kilmer, 2010). For example, perceptions that participating in sexual behaviors while under the influence of alcohol is prevalent (descriptive norms) and is approved of by one's peers (injunctive norms) increase the likelihood of using alcohol and decrease the use of PBS. However, a recent study by Lewis and colleagues (2014) found that, although college students tend to underestimate the prevalence of protective behaviors (e.g., condom use) and overestimate the prevalence of risk behaviors (e.g., drinking prior to sex), only norms pertaining to the overestimation of sexual risk behaviors (i.e., descriptive norms) are related to actual behavior.

### 2.2.3 Perceived Benefits

According to the health belief model (Rosenstock, 1990), perceived benefits of using PBS to reduce the negative consequences of using alcohol or having sex are expected to have an impact on both alcohol use and use of PBS. Although there is evidence to support the idea that perceived benefits of using PBS reduce alcohol consumption (Pearson, 2013), there is less empirical evidence that perceived benefits of using PBS lead to an actual increase in use of PBS. Thus, inclusion of this component is more supported theoretically than empirically. Nevertheless, we hypothesize that increasing perceived benefits of using PBS will lead to decreased alcohol use and increased use of PBS.

### 2.2.4 Self-Efficacy to Use PBS

This component is designed to increase self-efficacy to use PBS. Self-efficacy (Bandura, 1977) for using PBS such as limiting alcohol intake or planning to discuss sexual boundaries with a partner when intoxicated is expected to decrease alcohol use (Pearson, Prince, & Bravo, 2017) and increase the use of PBS. In a study specifically about alcohol use, Ehret, Ghaidarov, and LaBrie (2013) found that drinking refusal self-efficacy was associated with decreased weekly alcohol use. In a study focused on sexual risk behaviors, Nesoff, Dunkle, and Lang (2016) found that condom use negotiation skills were positively associated with condom use.

## *2.3 Pathways from the Intervention Components to Alcohol-Related Sexual Risk Behaviors and STIs*

Figure 1 shows that each component targets one of the proximal mediators. In turn, each mediator is hypothesized to produce an effect on the proximal behavioral outcomes, that is, a reduction in alcohol use and an increase in the use of PBS. Two main pathways lead to a reduction in alcohol-related sexual risk behaviors and a reduction in STIs, each associated with one of the proximal behavioral outcomes. Reducing alcohol use leads to a decrease in myopic effects and also to an increase in the use of PBS, both of which lead to a decrease in alcohol-related sexual risk behaviors. The increased use of PBS can itself decrease alcohol use (and then the subsequent pathway to alcohol-related sexual risk behaviors can be followed as described above), and it can directly decrease the use of alcohol-related sexual risk behaviors. For example, even without reducing alcohol intake, a student could be sure to use a condom as appropriate in any sexual encounter. In this case, even though the sexual encounter may be alcohol-related, its risk is greatly reduced.

## *2.4 Potential Moderators*

We do not make specific hypotheses about moderators, but we note a number of potential moderators that will be examined in data analysis. They are not included in the conceptual model because we chose to develop a model focusing specifically on how the candidate intervention components are hypothesized to have an effect on the proximal mediators, proximal behavioral outcomes, and distal biological outcomes. However, we include a description of potential moderators here for completeness.

### 2.4.1 Gender

Gender differences in alcohol use and sexual behaviors are well documented. For example, males report higher participation in heavy episodic drinking in the past 30 days (American College Health Association, 2016) and typically report higher levels of specific sexual behaviors than females (e.g., more lifetime sexual partners; (Chandra, Copen, & Mosher, 2013)). Males and females report comparable numbers of hookups, yet males report more penetrative behaviors during a hookup (i.e., vaginal and anal sex) than females. Fisher (2009) suggests that this may be due to reporting bias that is a by-product of the social desirability of penetrative behaviors among males. Further, in a study by Kirmani and Suman (2010), males reported more positive norms for engaging in sexual behaviors and more positive alcohol- and sex-related expectancies than females. Although these differences are important, they are not expected to moderate either component effects or the effects of mediators. The itMatters intervention components have been developed to work equally well for both males and females. Although we hypothesize that there will be no gender-by-component interactions, we will explore this interaction empirically.

### 2.4.2 Race/Ethnicity

Notable race/ethnicity differences have been observed in alcohol use and sexual risk behaviors. Although African American/Black and Hispanic/Latino students typically report lower alcohol use than White students (Paves, Pedersen, Hummer, & Labrie, 2012), Black and Latino students report more unprotected sex and more partners than White students (Randolph, Torres, Gore-Felton, Lloyd, & McGarvey, 2009) and carry a disproportionate STI burden (Centers for Disease Control and Prevention, 2013). This disparity is often attributed to the gender ratio (more females than males on campus) and available sex partner pools on college campuses, particularly at historically black colleges and universities (Ferguson, Quinn, Eng, & Sandelowski, 2006; Jennings et al., 2010). Thus the association between self-efficacy and use of PBS might be weaker for Black students than White students:

Black women may perceive less power to negotiate condom use and discuss safer sex openly due to fear of losing a male partner to another woman.

#### 2.4.3 Other Individual-Level

Another possible moderator is individual differences in sexual sensation-seeking. Individuals who are high in sexual sensation-seeking are expected to experience a more negative impact of the effects of alcohol (e.g., reduced condom use and increased penetrative hookups) than those who are lower in sexual sensation-seeking (Hendershot, Stoner, George, & Norris, 2007). Relational factors may moderate the association between the use of PBS and alcohol-related sexual risk behaviors. For example, the association between negotiating condom use and using a condom is weaker for individuals in a committed relationship than those in a casual relationship (Brown & Vanable, 2007) and weaker when a sexual partner is 3 or more years older compared to less than 3 years older (Ford, Sohn, & Lepkowski, 2001). In addition, the association is weaker if there is a reliance on hormonal contraception versus a barrier method such as a condom (Bailey, Fleming, Catalano, Haggerty, & Manhart, 2012).

#### 2.4.4 Environmental

There is less scientific literature on environmental moderators between the proximal mediators and distal behaviors among college students. However, we hypothesize the association between accurate perceptions of descriptive norms, and behavioral outcomes is weaker when a Greek system exists on campus than not, and we hypothesize that the association between having a plan to use condoms (a PBS) and using condoms is stronger if free condoms are available on campus (Reeves, Ickes, & Mark, 2016).

## 3 Optimizing itMatters

### *3.1 Overview of the Iterative Approach to Optimization*

The goal of the current study is to build an effective and efficient STI preventive intervention. By effective intervention, we mean an intervention that has been empirically demonstrated to decrease alcohol-related sexual risk behaviors and, ultimately, STIs. By efficient intervention, we mean an intervention that is made up exclusively of components that have empirically detectable effects on the proximal mediators. In other words, we plan to use the all active components optimization criterion (see Chapter 2 in the companion volume).

As mentioned above, this application of MOST is using an iterative approach to optimization. An iterative approach involves conducting more than one, in this case two, separate and sequential optimization trials. The research plan calls for us to proceed as follows: Use the first experiment to determine which of the five candidate components described previously have an empirically detectable effect. Any components that do not have an empirically detectable effect are then revised, with the objective of improving their effectiveness. Next, conduct a second optimization trial to evaluate the new set of five components, made up of the components found to be acceptable in the first optimization trial plus the newly revised components. After this second experiment, construct the optimized itMatters intervention using the all active components optimization criterion; in other words, the optimized intervention will consist of all the components that had empirically detectable effects. Finally, proceed to the evaluation phase of MOST and confirm by means of an RCT that the optimized intervention has a statistically significant and clinically meaningful effect.

At the time of this writing, the first optimization trial has been completed, and the second is in the field. Below we describe the general strategy we used for identifying which components require revision and for revising those components in preparation for the second experiment.

## *3.2 Criteria for Determining Whether a Component Has an Empirically Detectable Effect*

Whether each component has an empirically detectable effect will be established in the optimization trial, described below. To our knowledge there are no currently established standards of what constitutes an effective intervention component. We specified a priori that a component will be deemed effective if the results of the optimization trial indicate that it achieves a main effect of $d \geq 0.15$ in the anticipated direction. This is what we consider the minimum clinically significant effect size for a component, and it reflects the notion that in an efficient intervention, every component should have an effect that is at least small by Cohen's rule of thumb (Cohen, 1988). We will also examine interactions between components, though based on the conceptual model, we do not anticipate any large interactions.

We recognize that if there are interactions between components, the combined effect of the components will be different from what would be expected based on the main effects. In particular, if any interactions are primarily antagonistic, this combined effect will be less than what would be expected based on the main effects. Nevertheless, if we are able to arrive at a set of five components that achieve the stated minimum effect size, we expect that the resulting intervention package will achieve an effect size in the $d = 0.35$–$0.5$ range. This would exceed the effects of existing interventions aimed at alcohol use (Scott-Sheldon et al., 2016; Tanner-Smith & Lipsey, 2015) and condom use (Scott-Sheldon et al., 2011).

More about how decision-making can be carried out based on the results of a factorial experiment can be found in Chapter 7 in the companion volume.

## 3.3 Design of the Optimization Trials

As noted above, in this application of MOST, the purpose of the optimization trials is to determine which of the candidate components achieve a main effect of $d \geq 0.15$. The resource management principle (see Chapter 1 in the companion volume) states that the most appropriate experimental design for the optimization trial is one that addresses the key research questions while making the best use of the resources available.

A factorial design was selected for the optimization trials for three reasons. First, a factorial experiment provides the necessary scientific information because it separates component effects, enabling estimation of the main effect of each candidate component (five in the current study). Second, the factorial experiment is the only design that will enable us to examine interactions between components. For example, the results of the factorial experiment will address the question of whether, contrary to our conceptual model, the effect of the expectancies component varies depending on whether a participant is provided with the self-efficacy component. Third, a factorial experiment is a highly efficient way to examine multiple intervention components. To achieve the same statistical power for tests of component effects, a factorial experiment requires substantially fewer participants than alternative approaches, such as conducting individual experiments on each component (Collins, Dziak, & Li, 2009). (For more about factorial experiments, see Chapters 3, 4, 5, and 6 in the companion volume.)

Each of the two factorial experiments to be conducted in the optimization phase uses the same experimental design. There are five factors—a factor corresponding to each component except the information component. Each factor has two levels: no, in which the component is not provided to the participant, and yes, in which the component is provided. A factorial experiment including five two-level factors requires $2^5 = 32$ experimental conditions. Table 1 shows the names assigned to each factor and the 32 conditions in this experiment. Note that all of the participants receive the information component. For example, a participant randomly assigned to experimental condition #8 receives, in addition to the information component, the injunctive norms (*INORM* = yes), perceived benefits (*BENEFITS* = yes), and self-efficacy candidate components (*SELFEFF* = yes). By contrast, a participant randomized to experimental condition #32 receives all of the components.

We considered conducting a $2^{5-1}$ fractional factorial design (see Chapter 5 in the companion volume), which would have cut the number of experimental conditions in half, to 16. This would have meant that, as in all incomplete factorial designs, there would have been aliasing (combining) of effects. In this case, each main effect would have been aliased with a four-way interaction, and each two-way interaction would have been aliased with a three-way interaction. Aliasing can be an acceptable

**Table 1** Experimental conditions in factorial design

| Experimental Condition Number | | FACTORS | | | | |
|---|---|---|---|---|---|---|
| | Information component | *EXP* | *DNORM* | *INORM* | *BENEFITS* | *SELFEFF* |
| 1 | Yes | No | No | No | No | No |
| 2 | Yes | No | No | No | No | Yes |
| 3 | Yes | No | No | No | Yes | No |
| 4 | Yes | No | No | No | Yes | Yes |
| 5 | Yes | No | No | Yes | No | No |
| 6 | Yes | No | No | Yes | No | Yes |
| 7 | Yes | No | No | Yes | Yes | No |
| 8 | Yes | No | No | Yes | Yes | Yes |
| 9 | Yes | No | Yes | No | No | No |
| 10 | Yes | No | Yes | No | No | Yes |
| 11 | Yes | No | Yes | No | Yes | No |
| 12 | Yes | No | Yes | No | Yes | Yes |
| 13 | Yes | No | Yes | Yes | No | No |
| 14 | Yes | No | Yes | Yes | No | Yes |
| 15 | Yes | No | Yes | Yes | Yes | No |
| 16 | Yes | No | Yes | Yes | Yes | Yes |
| 17 | Yes | Yes | No | No | No | No |
| 18 | Yes | Yes | No | No | No | Yes |
| 19 | Yes | Yes | No | No | Yes | No |
| 20 | Yes | Yes | No | No | Yes | Yes |
| 21 | Yes | Yes | No | Yes | No | No |
| 22 | Yes | Yes | No | Yes | No | Yes |
| 23 | Yes | Yes | No | Yes | Yes | No |
| 24 | Yes | Yes | No | Yes | Yes | Yes |
| 25 | Yes | Yes | Yes | No | No | No |
| 26 | Yes | Yes | Yes | No | No | Yes |
| 27 | Yes | Yes | Yes | No | Yes | No |
| 28 | Yes | Yes | Yes | No | Yes | Yes |
| 29 | Yes | Yes | Yes | Yes | No | No |
| 30 | Yes | Yes | Yes | Yes | No | Yes |
| 31 | Yes | Yes | Yes | Yes | Yes | No |
| 32 | Yes | Yes | Yes | Yes | Yes | Yes |

price to pay for greatly increased efficiency (for example, see (Piper et al., 2016)). However, ultimately we decided that in this case, the increase in efficiency would not be great enough to compensate for aliasing. Because all of the candidate components were to be delivered online, the additional resources required to conduct 32 as compared to 16 experimental conditions would have been minimal and, in our view, did not justify the use of a fractional factorial design.

## *3.4 Subjects and Measures*

### 3.4.1 Subjects

Subjects for both optimization trials were freshmen students at several 4-year, coeducational, public universities in the United States. The universities varied in characteristics such as size, location, and ethnic composition, providing a diverse sample.

### 3.4.2 Outcome Measures

The outcome measures for the optimization trials are drawn directly from the conceptual model. As shown in Fig. 1, each component targeted a proximal mediator—expectancies that alcohol is not needed before or during sex, accurate perceptions of how many people use alcohol before or during sex, accurate perceptions of acceptability of using alcohol before or during sex, perceived benefits of using PBS related to alcohol and sex behaviors, and self-efficacy to use PBS related to alcohol and sex behaviors. Measures of these mediators were the outcomes used for making decisions about whether a particular component needs revision. We used proximal mediators as outcomes, instead of the proximal behavioral outcomes of sexual risk behavior in the first experiment, because the proximal mediators could be measured sooner. This helped provide enough time for us to analyze the data, determine which components are working, and make the necessary revisions before conducting the subsequent experiment. We plan to use proximal mediators as outcomes for the second experiment as well as to allow us the time to prepare the itMatters-optimized intervention package for an evaluation by means of an RCT within 1 year.

Our conceptual model justifies this approach. According to the conceptual model, the proximal mediators ultimately affect alcohol-related sexual risk behaviors, which in turn are hypothesized to affect STI incidence. Thus if each component has an effect on its target mediator, this indicates that the optimized intervention package can be expected to have the desired effect on the proximal behavioral outcomes (i.e., alcohol use, use of PBS), the distal behavioral outcome (i.e., alcohol-related sexual risk behaviors), and the distal biological outcome, STIs. As will be discussed below, for the RCT to be conducted in the evaluation phase of MOST, a measure of alcohol-

related sexual risk behavior will be used as the primary outcome, and alcohol use and use of PBS will be secondary outcomes.

### *3.5 Revision of Components*

As noted above, this is an ongoing study, with the first of two optimization trials completed. This section describes very briefly how results from the first experiment were used to identify which components needed revision. A regression approach to ANOVA was used to determine which intervention components were satisfactory (i.e., achieved an effect of $d \geq 0.15$) or needed revision (i.e., achieved an effect of $d \leq 0.15$). Preliminary analyses suggest that only the descriptive and injunctive norms components were satisfactory; the remaining three components required revisions. We made some initial revisions based on feedback from student focus groups. We then asked several outside experts to give us a fresh perspective on the revised components. We specifically asked them to identify content that was missing and/or unclear. We revised the components based on their feedback. These components will be evaluated in a second factorial experiment.

### *3.6 Secondary Analysis of Data from the Optimization Trials*

Above we reviewed a number of variables that could be moderators, including gender, race/ethnicity, other individual-level variables, and certain environmental variables. Secondary analyses will examine whether any of these variables moderate the effects of any of the five candidate components. We consider any analyses involving moderating effects to be exploratory and acknowledge that we may not have power to detect moderation effects. However, we expect these analyses to be helpful in generating hypotheses that can be evaluated in subsequent studies that will be powered for that purpose.

## 4 Evaluating itMatters

After the two rounds of optimization trials, we will know which of the candidate components achieved the desired effectiveness on the proximal mediators and will be included in the optimized intervention. The next step will be to evaluate whether the optimized intervention is more effective than a suitable control using a two-arm RCT. The primary outcome for this phase will be the proximal behavioral outcome, alcohol-related sexual risk behaviors. This is the primary outcome specified in the conceptual model, but measuring this outcome requires following students for a longer period of time. We will use the necessary time to measure this outcome in

the RCT. If shown to be effective, the optimized intervention will then be released and made available to other universities interested in reducing alcohol-related sexual behaviors and ultimately STIs among college students.

# 5 Discussion

## *5.1 The Conceptual Model*

The first objective of this chapter is to describe the conceptual model for itMatters, an online STI prevention intervention among college students. The development of a conceptual model is a critical part of the preparation phase of MOST. In order to create the conceptual model (Fig. 1), we drew heavily on theory and relevant literature on alcohol use and sexual risk behaviors separately and laboratory studies and the few interventions that have specifically targeted the intersection of alcohol and sex.

The conceptual model is critical in MOST for several reasons. First, during intervention development, the conceptual model has served as a powerful reminder to the research team that it is essential to retain focus on the *intersection* of alcohol and sex, rather than to develop an intervention that is a disjointed amalgam of interventions focused separately on alcohol and sexual behaviors. Second, the proximal mediating variables (i.e., expectancies, descriptive norms, injunctive norms, perceived benefits, and self-efficacy) were clearly identified in the conceptual model as the intervention targets, which guided the content of the components. Third, as discussed above, the conceptual model pointed the way to selection of short-term outcomes for making decisions about the effectiveness of a given component.

Development of a conceptual model can itself be an iterative process. During the course of this project, we have revisited the conceptual model (and the literature) numerous times as new literature emerged and as we refined our understanding of the mechanisms by which alcohol use influences sexual risk behaviors. The research team drafted more than 20 versions of the conceptual model, stopping only when we felt it accurately represented the current empirical literature, scientific theory, and our own ideas about mediating pathways by which proximal cognitive factors influence behavior and ultimately STIs. Figure 1 represents our current thinking.

## *5.2 The Iterative Approach to Optimization*

The second objective of this chapter is to describe the iterative approach to optimization used in the current study. Using sequential optimization trials in an iterative fashion provides an opportunity to improve the effectiveness of individual candidate components before making final decisions about whether or not to include them in the optimized intervention. In the current study, we are conducting two

optimization trials, but more than two could be used if resources permit. To conduct two or more optimization trials within a single optimization phase of MOST, it is necessary to have access to enough subjects and sufficient time to conduct multiple experiments.

Because we are conducting this study in university settings, we have a new set of freshmen every year, so we have access to sufficient subjects. Using measures of mediators as short-term outcomes afforded us enough time to conduct two optimization trials in 2 years. However, when mediators are used as short-term outcomes, the test of effectiveness is less definitive than it would be if the outcome of ultimate interest were used. In this case it is particularly important to confirm the effectiveness of the optimized intervention by means of an RCT using the outcome of ultimate interest.

In this application of MOST, we used the all active components optimization criterion, which means we were primarily interested in achieving effectiveness and efficiency. As discussed in the companion volume, economy and scalability may also be important in other settings. For example, in this application we could have used the iterative approach to develop the most effective intervention we could obtain that could be completed within some upper limit on time, say 30 min. This might have improved both the economy (expressed in terms of participant time) and scalability of the intervention. Thus the iterative approach can be used with the objective of improving effectiveness, efficiency, economy, or scalability.

The iterative approach may not be feasible in situations where subjects, for example, clinical subjects, must be recruited over a period of time or offered generous compensation for their participation. It also may not be feasible for interventions that take a long time to deliver or where the outcome of interest is far in the future and it is not desirable to use measures of mediators as short-term outcomes. But where resources permit its use, the iterative approach has considerable appeal, because it has the potential to systematically and incrementally strengthen the effect of an intervention before it is evaluated in an RCT. We are hopeful that the use of this approach will enable us to develop an online intervention that approaches the effectiveness of comparable traditional implementer-led interventions, and we believe it could improve the public health impact of many behavioral, biobehavioral, and biomedical interventions.

**Acknowledgments** Preparation of this chapter was supported by Award Number R01 AA022931 from the National Institute on Alcohol Abuse and Alcoholism and Award Number P50 DA039838 from the National Institute on Drug Abuse. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors thank Amanda Applegate for editorial assistance.

## References

Abbey, A., Saenz, C., & Buck, P. O. (2005). The cumulative effects of acute alcohol consumption, individual differences and situational perceptions on sexual decision making. *Journal of Studies on Alcohol and Drugs, 66*(1), 82.

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes, 50*, 179–211.

American College Health Association. (2016). *American College Health Association-National College Health Assessment II: Undergraduate Students Reference Group Data Report Fall 2015*. Retrieved from Hanover, MD.

Bailey, J. A., Fleming, C. B., Catalano, R. F., Haggerty, K. P., & Manhart, L. E. (2012). Romantic relationship characteristics and alcohol use: Longitudinal associations with dual method contraception use. *Journal of Adolescent Health, 50*(5), 450–455.

Baker, T. B., Collins, L. M., Mermelstein, R., Piper, M. E., Schlam, T. R., Cook, J. W., … Fiore, M. C. (2016). Enhancing the effectiveness of smoking treatment research: Conceptual bases and progress. *Addiction, 111*(1), 107–116. https://doi.org/10.1111/add.13154

Bandura, A. (1977). *Social learning theory*. Englewood Cliffs, NJ: Prentice Hall.

Berkowitz, A. (2004). Applications of social norms theory to other health and social justice issues. In H. W. Perkins (Ed.), *The social norms approach to preventing school and college age substance abuse: A handbook for educators, counselors, and clinicians* (pp. 259–279). San Francisco, CA: Jossey-Bass.

Brown, J. L., & Vanable, P. A. (2007). Alcohol use, partner type, and risky sexual behavior among college students: Findings from an event-level study. *Addictive Behaviors, 32*(12), 2940–2952.

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.* Paper presented at the Child & Youth Care Forum, *41*, 119. https://doi.org/10.1007/s10566-011-9164-4

Carey, K. B., Scott-Sheldon, L. A. J., Elliot, A. J., Bolles, J. R., & Carey, M. P. (2009). Computer-delivered interventions to reduce college student drinking: A meta-analysis. *Addiction, 104*(11), 1807–1819.

Centers for Disease Control and Prevention. (2013). CDC fact sheet: Incidence, prevalence, and cost of sexually transmitted infections in the United States. CDC.

Chandra, A., Copen, C. E., & Mosher, W. D. (2013). Sexual behavior, sexual attraction, and sexual identity in the United States: Data from the 2006–2010 National Survey of Family Growth. In *International handbook on the demography of sexuality* (pp. 45–66). Dordrecht: Springer. https://doi.org/10.1007/978-94-007-5512-3_4

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: LEA.

Collins, L. M. (2018). *Optimization of behavioral, biobehavioral, and biomedical interventions: The Multiphase Optimization Strategy (MOST)*. New York, NY: Springer.

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. https://doi.org/10.1037/a0015826

Davis, K. C., Masters, N. T., Eakins, D., Danube, C. L., George, W. H., Norris, J., & Heiman, J. R. (2014). Alcohol intoxication and condom use self-efficacy effects on women's condom use intentions. *Addictive Behaviors, 39*(1), 153–158.

Davis, K. C., Norris, J., Hessler, D. M., Zawacki, T., Morrison, D. M., & George, W. H. (2010). College women's sexual decision making: Cognitive mediation of alcohol expectancy effects. *Journal of American College Health, 58*(5), 481–489. https://doi.org/10.1080/07448481003599112

Dermen, K. H., & Thomas, S. N. (2011). Randomized controlled trial of brief interventions to reduce college students' drinking and risky sex. *Psychology of Addictive Behaviors, 25*, 583–594.

Dry, M. J., Burns, N. R., Nettelbeck, T., Farquharson, A. L., & White, J. M. (2012). Dose-related effects of alcohol on cognitive functioning. *PLoS One, 7*(11), e50977.

Ebel-Lam, A. P., MacDonald, T. K., Zanna, M. P., & Fong, G. T. (2009). An experimental investigation of the interactive effects of alcohol and sexual arousal on intentions to have unprotected sex. *Basic and Applied Social Psychology, 31*, 226–233.

Ehret, P. J., Ghaidarov, T. M., & LaBrie, J. W. (2013). Can you say no? Examining the relationship between drinking refusal self-efficacy and protective behavioral strategy use on alcohol outcomes. *Addictive Behaviors, 38*(4), 1898–1904.

Ferguson, Y. O., Quinn, S. C., Eng, E., & Sandelowski, M. (2006). The gender ratio imbalance and its relationship to risk of HIV/AIDS among African American women at historically black colleges and universities. *AIDS Care, 18*(4), 323–331. https://doi.org/10.1080/09540120500162122

Fielder, R. L., & Carey, M. P. (2010). Predictors and consequences of sexual "hookups" among college students: A short-term prospective study. *Archives of Sexual Behavior, 39*(5), 1105–1119.

Fisher, T. D. (2009). The impact of socially conveyed norms on the reporting of sexual behavior and attitudes by men and women. *Journal of Experimental Social Psychology, 45*(3), 567–572.

Ford, K., Sohn, W., & Lepkowski, J. (2001). Characteristics of adolescents' sexual partners and their association with use of condoms and other contraceptive methods. *Family Planning Perspectives, 33*(3), 100–132.

Garcia, J. R., Reiber, C., Massey, S. G., & Merriwether, A. M. (2012). Sexual hookup culture: A review. *Review of General Psychology, 16*(2), 161.

George, W. H., Davis, K. C., Norris, J., Heiman, J. R., Stoner, S. A., Schacht, R. L., . . . Kajumulo, K. F. (2009). Indirect effects of acute alcohol intoxication on sexual risk-taking: The roles of subjective and physiological sexual arousal. *Archives of Sexual Behavior, 38*(4), 498–513.

Grazioli, V. S., Lewis, M. A., Garberson, L. A., Fossos-Wong, N., Lee, C. M., & Larimer, M. E. (2015). Alcohol expectancies and alcohol outcomes: Effects of the use of protective behavioral strategies. *Journal of Studies on Alcohol and Drugs, 76*(3), 452.

Hendershot, C. S., Stoner, S. A., George, W. H., & Norris, J. (2007). Alcohol use, expectancies, and sexual sensation seeking as correlates of HIV risk behavior in heterosexual young adults. *Psychology of Addictive Behaviors, 21*(3), 365–372. https://doi.org/10.1037/0893-164x.21.3.365

Huffman, J. C., Albanese, A. M., Campbell, K. A., Celano, C. M., Millstein, R. A., Mastromauro, C. A., . . . Collins, L. M. (2017). The positive emotions after acute coronary events behavioral health intervention: Design, rationale, and preliminary feasibility of a factorial design study. *Clinical Trials, 14*(2), 128–139.

Jennings, J. M., Taylor, R., Iannacchione, V. G., Rogers, S. M., Chung, S.-E., Huettner, S., & Ellen, J. M. (2010). The available pool of sex partners and risk for a current bacterial sexually transmitted infection. *Annals of Epidemiology, 20*(7), 532–538.

Jones, B. T., Corbin, W., & Fromme, K. (2001). A review of expectancy theory and alcohol consumption. *Addiction, 96*(1), 57–72.

Kann, L., McManus, T., Harris, W. A., Shanklin, S. L., Flint, K. H., Hawkins, J., . . . Zaza, S. (2016). Youth risk behavior surveillance—United States, 2015. *MMWR Surveillance Summaries, 65*(6), 1–174. https://doi.org/10.15585/mmwr.ss6506a1

Kirmani, M. N., & Suman, L. N. (2010). Gender differences in alcohol related attitudes and expectancies among college students. *Journal of the Indian Academy of Applied Psychology, 36*(1), 19–24.

Lewis, J. E., Miguez-Burban, M.-J., & Malow, R. M. (2009). HIV risk behavior among college students in the United States. *College Student Journal, 43*(2), 475–491.

Lewis, M. A., Patrick, M. E., Litt, D. M., Atkins, D. C., Kim, T., Blayney, J. A., . . . Larimer, M. E. (2014). Randomized controlled trial of a web-delivered personalized normative feedback intervention to reduce alcohol-related risky sexual behavior among college students. *Journal of Consulting and Clinical Psychology, 82*(3), 429.

Lewis, M. A., Rees, M., Logan, D. E., Kaysen, D. L., & Kilmer, J. R. (2010). Use of drinking protective behavioral strategies in association to sex-related alcohol negative consequences: The mediating role of alcohol consumption. *Psychology of Addictive Behaviors, 24*(2), 229.

Logan, D. E., Koo, K. H., Kilmer, J. R., Blayney, J. A., & Lewis, M. A. (2015). Use of drinking protective behavioral strategies and sexual perceptions and behaviors in US college students. *Journal of Sex Research, 52*(5), 558–569.

Nesoff, E. D., Dunkle, K., & Lang, D. (2016). The impact of condom use negotiation self-efficacy and partnership patterns on consistent condom use among college-educated women. *Health Education & Behavior, 43*(1), 61–67.

Olmstead, S. B., Pasley, K., & Fincham, F. D. (2013). Hooking up and penetrative hookups: Correlates that differentiate college men. *Archives of Sexual Behavior, 42*(4), 573–583.

Paul, E. L., McManus, B., & Hayes, A. (2000). "Hookups": Characteristics and correlates of college students' spontaneous and anonymous sexual experiences. *Journal of Sex Research, 37*(1), 76–88.

Paves, A. P., Pedersen, E. R., Hummer, J. F., & Labrie, J. W. (2012). Prevalence, social contexts, and risks for prepartying among ethnically diverse college students. *Addictive Behaviors, 37*(7), 803–810. https://doi.org/10.1016/j.addbeh.2012.03.003

Pearson, M. R. (2013). Use of alcohol protective behavioral strategies among college students: A critical review. *Clinical Psychology Review, 33*(8), 1025–1040.

Pearson, M. R., Prince, M. A., & Bravo, A. J. (2017). Moderators of the effects of alcohol protective behavioral strategies: Three attempts of replication and extension. *Substance Use & Misuse, 52*(7), 939–949.

Pellegrini, C. A., Hoffman, S. A., Collins, L. M., & Spring, B. (2014). Optimization of remotely delivered intensive lifestyle treatment for obesity using the Multiphase Optimization Strategy: Opt-IN study protocol. *Contemporary Clinical Trials, 38*, 251–259.

Pellegrini, C. A., Hoffman, S. A., Collins, L. M., & Spring, B. (2015). Corrigendum to "Optimization of remotely delivered intensive lifestyle treatment for obesity using the Multiphase Optimization Strategy: Opt-IN study protocol". *Contemporary Clinical Trials, 38*(2014), 251–259. *Contemp Clin Trials, 45*(Pt B), 468–469. https://doi.org/10.1016/j.cct.2015.09.001

Piper, M. E., Fiore, M. C., Smith, S. S., Fraser, D., Bolt, D. M., Collins, L. M., … Jorenby, D. E. (2016). Identifying effective intervention components for smoking cessation: A factorial screening experiment. *Addiction, 111*(1), 129–141.

Prause, N., Staley, C., & Finn, P. (2011). The effects of acute ethanol consumption on sexual response and sexual risk-taking intent. *Archives of Sexual Behavior, 40*(2), 1–12. https://doi.org/10.1007/s10508-010-9718-9

Purdie, M. P., Norris, J., Davis, K. C., Zawacki, T., Morrison, D. M., George, W. H., & Kiekel, P. A. (2011). The effects of acute alcohol intoxication, partner risk level, and general intention to have unprotected sex on women's sexual decision making with a new partner. *Experimental and Clinical Psychopharmacology, 19*(5), 378.

Randolph, M. E., Torres, H., Gore-Felton, C., Lloyd, B., & McGarvey, E. L. (2009). Alcohol use and sexual risk behavior among college students: Understanding gender and ethnic differences. *The American Journal of Drug and Alcohol Abuse, 35*(2), 80–84. https://doi.org/10.1080/00952990802585422

Reeves, B., Ickes, M. J., & Mark, K. P. (2016). Gender differences and condom-associated embarrassment in the acquisition of purchased versus free condoms among college students. *American Journal of Sexuality Education, 11*(1), 61–75.

Reid, A. E., & Carey, K. B. (2015). Interventions to reduce college student drinking: State of the evidence for mechanisms of behavior change. *Clinical Psychology Review, 40*, 213–224.

Rosenstock, I. M. (1990). The health belief model: Explaining health behavior through expectancies. In K. Glanz, F. Lewis, & B. Rimer (Eds.), *Health behavior and health education*. San Francisco, CA: Jossey-Bass.

Scott-Sheldon, L. A., Carey, K. B., Cunningham, K., Johnson, B. T., Carey, M. P., & Team, M. R. (2016). Alcohol use predicts sexual decision-making: A systematic review and meta-analysis of the experimental literature. *AIDS and Behavior, 20*(1), 19–39.

Scott-Sheldon, L. A. J., Huedo-Medina, T. B., Warren, M. R., Johnson, B. T., & Carey, M. P. (2011). Efficacy of behavioral interventions to increase condom use and reduce sexually transmitted infections: A meta-analysis, 1991 to 2010. *JAIDS: Journal of Acquired Immune Deficiency Syndromes, 58*, 489.

Sevincer, A. T., & Oettingen, G. (2014). Alcohol myopia and goal commitment. *Frontiers in Psychology, 5*, 169.

Shuper, P. A., Neuman, M., Kanteres, F., Baliunas, D., Joharchi, N., & Rehm, J. (2010). Causal considerations on alcohol and HIV/AIDS—a systematic review. *Alcohol and Alcoholism, 45*(2), 159–166. https://doi.org/10.1093/alcalc/agp091

Tanner-Smith, E. E., & Lipsey, M. W. (2015). Brief alcohol interventions for adolescents and young adults: A systematic review and meta-analysis. *Journal of Substance Abuse Treatment, 51*, 1–18.

Treloar, H., Martens, M. P., & McCarthy, D. M. (2015). The protective behavioral strategies scale-20: Improved content validity of the serious harm reduction subscale. *Psychological Assessment, 27*(1), 340.

Trepka, M. J., Kim, S., Pekovic, V., Zamor, P., Velez, E., & Gabaroni, M. V. (2008). High-risk sexual behavior among students of a minority-serving university in a community with a high HIV/AIDS prevalence. *Journal of American College Health, 57*(1), 77–84.

Wyrick, D. L., Rulison, K. L., Fearnow-Kenney, M., Milroy, J. J., & Collins, L. M. (2014). Moving beyond the treatment package approach to developing behavioral interventions: Addressing questions that arose during an application of the Multiphase Optimization Strategy (MOST). *Translational Behavioral Medicine, 4*(3), 252–259.

# Implementing Factorial Experiments in Real-World Settings: Lessons Learned While Engineering an Optimized Smoking Cessation Treatment


**Megan E. Piper, Tanya R. Schlam, David Fraser, Madeline Oguss, and Jessica W. Cook**

**Abstract** Treatment development research is challenging. The multiphase optimization strategy (MOST) is a new, efficient method to engineer effective treatment packages. MOST helps researchers understand intervention component effects and how these components can best be combined into an optimized treatment package and then evaluated. However, researchers may be discouraged by the perceived challenges of optimization trials. Our research team has considerable experience implementing optimization trials, including implementing three factorial experiments in primary care clinics simultaneously ($N \approx 1700$). This chapter provides evidence for the feasibility of conducting large factorial experiments in real-world settings and shares strategies for successfully carrying out this type of research. We present practical information on implementation considerations, including treatment factor selection, factor integration and implementation, fidelity in implementation, randomization, data collection, considering the participant perspective, and reporting the results.

M. E. Piper (✉) · T. R. Schlam · D. Fraser · M. Oguss
Center for Tobacco Research and Intervention, Department of Medicine, School of Medicine and Public Health, University of Wisconsin, Madison, WI, USA
e-mail: mep@ctri.wisc.edu

J. W. Cook
Center for Tobacco Research and Intervention, Department of Medicine, School of Medicine and Public Health, University of Wisconsin, Madison, WI, USA

William S. Middleton Memorial Veterans Hospital, Madison, WI, USA



L. M. Collins, K. C. Kugler (eds.), *Optimization of Behavioral, Biobehavioral, and Biomedical Interventions*, Statistics for Social and Behavioral Sciences,
https://doi.org/10.1007/978-3-319-91776-4_2

## 1 Introduction

> The way to build a complex system that works is to begin with a very simple system that works. Kevin Kelly, American Editor/Publisher (Kelly, 1994)

Smoking is the leading preventable cause of death and disease in the United States, killing almost half a million people each year (U.S. Department of Health and Human Services, 2014). Tobacco researchers have worked for decades to develop and disseminate evidence-based treatments to help smokers quit. However, even with such treatments, the vast majority of smokers who try to quit ultimately return to smoking (Fiore et al., 2008). To address this issue, our research team at the University of Wisconsin Center for Tobacco Research and Intervention (UW-CTRI) sought to build more effective treatment packages more efficiently and decided to try a novel approach to smoking treatment development – the multiphase optimization strategy (MOST; Collins, 2018, companion volume, Collins et al., 2011; Collins, Kugler, & Gwadz, 2016). Briefly, in MOST, promising behavioral and pharmacologic intervention components are examined in an optimization trial (e.g., a factorial experiment). The most promising (e.g., effective, scalable) components are then combined into an optimized treatment package consisting only of components shown to be effective and to work well together; the optimized treatment package is then evaluated in a randomized clinical trial (RCT), and the treatment package is disseminated.

Over the last 9 years, in conjunction with collaborators at the Pennsylvania State University and the University of Illinois, Chicago, we have worked to develop optimized smoking cessation treatments that can be effectively and efficiently delivered in primary care settings. To achieve that goal, we have completed five optimization trials and one randomized controlled trial (RCT) that completed a full cycle of MOST. Our optimization trials have primarily been factorial screening experiments; we have also conducted one fractional factorial screening experiment (Collins, companion volume) and are currently conducting a sequential, multiple assignment, randomized trial (SMART; Almirall, Nahum-Shani, Wang, & Kasari, this volume). Both fractional factorial experiments and SMARTs are special cases of the factorial experiment. The term screening experiment refers to a factorial or fractional factorial experiment whose primary purpose is to screen out inactive or underperforming components and/or to choose the collection of components that will be most effective given constraints on time or money. In one center grant, we recruited almost 1700 participants into three factorial experiments implemented in primary care clinics; we implemented more than 80 different experimental conditions simultaneously and with fidelity. Informed by the results from the first round of optimization trials, we are currently conducting another factorial screening experiment and a SMART. The goal of this chapter is to provide evidence for the feasibility of conducting large factorial experiments in real-world settings and to share strategies and important considerations for successfully implementing such experiments. Some of these strategies are also relevant to implementing RCTs

and may be familiar, but implementing factorial experiments also involves unique considerations and modifying standard strategies.

## 2 Tobacco Treatment Optimization Trials

We began our research with a center grant from the National Cancer Institute (NCI; 2009–2014). The goal of this center grant was to identify promising intervention components to treat all smokers seen in primary care, including those who are not ready to quit. Based on behavior change theory and prior literature, different intervention components were designed to (1) motivate smokers who were initially unwilling to quit to make quit attempts sooner and increase the success of those attempts and (2) help smokers who were willing to quit to become smoke-free. All patients attending primary care visits were asked about their smoking status by clinic staff. Identified smokers were invited to participate in a study where they would receive treatment to either cut down on their smoking or quit. Interested patients were then referred to our research office via the electronic health record (Fraser, Christiansen, Adsit, Baker, & Fiore, 2013; Piper et al., 2013).

Patients who smoked who were eligible and interested in reducing their smoking, but not in quitting, were assigned to Motivation Study 1 (Cook et al., 2016; see Table 1). Participants were randomized to 1 of 16 treatment conditions as part of a 4-factor screening experiment (see Table 2). The intervention components were designed to reduce smoking and promote future cessation. Smoking patients who were interested in quitting were randomized to one of two cessation screening experiments. Cessation Study 1 (Piper, Fiore, et al., 2016; see Table 1) was a fractional factorial experiment that examined the effects of six intervention components focused on the preparation and cessation phases of treatment, targeting treatment mechanisms relevant to the opportunities presenting in the 3 weeks prequit and 2 weeks postquit, respectively (Baker et al., 2011; see Table 3). Cessation Study 2 (Schlam et al., 2016; see Table 1) was a five-factor experiment with 32 conditions (see Table 4) that screened components designed to target challenges to medication adherence and challenges in the maintenance phase. Thus, this center grant allowed us to conduct three factorial optimization trials and screen 15 potential intervention components in primary care clinics.

The findings from this research were used to engineer an abstinence-optimized treatment package that we then evaluated in an RCT as part of a subsequent NCI-funded center grant (2014–2019). This cessation RCT – designed to evaluate the effectiveness of the optimized treatment package – represented the completion of one cycle of MOST. We found that the MOST-engineered treatment package doubled abstinence rates at all time points (through 52 weeks after the target quit day), compared to a recommended usual care treatment (Piper et al., in press).

As part of this 2014–2019 center grant, we are also conducting a second optimization trial of intervention components for smokers initially unwilling to quit, in an effort to more effectively increase the percentage who eventually make

**Table 1** Different optimization trials conducted by UW-CTRI

| Study name | Design | Factors | Sample size | Grant | References |
|---|---|---|---|---|---|
| Motivation Study 1 | $2^4$ factorial | 1. Nicotine patch vs. none<br>2. Nicotine gum vs. none<br>3. Behavioral reduction counseling vs. none<br>4. Motivational interviewing strategies vs. none | 517 | NCI P50 (2009–2014) | Cook et al., (2016) |
| Cessation Study 1 | $2^{5-1}$ fractional factorial | 1. Prequit nicotine patch vs. none<br>2. Prequit nicotine gum vs. none<br>3. Prequit counseling vs. none<br>4. Intensive cessation in-person counseling vs. minimal<br>5. Intensive cessation telephone counseling vs. minimal<br>6. 16 vs. 8 weeks of combination NRT (patch + gum) | 637 | NCI P50 (2009–2014) | Piper, Fiore, et al. (2016)<br>Piper, Cook, et al. (2016)<br>Piper, Schlam et al. (2016) |
| Cessation Study 2 | $2^5$ factorial | 1. 26 vs. 8 weeks of nicotine patch + nicotine gum<br>2. Maintenance phone counseling vs. none<br>3. Medication adherence counseling vs. none<br>4. Automated medication adherence calls vs. none<br>5. Electronic medication monitoring with feedback and counseling vs. electronic medication monitoring alone | 544 | NCI P50 (2009–2014) | Schlam et al. (2016) |
| Motivation Study 2 | $2^4 \times$ factorial | 1. Nicotine mini-lozenge vs. none<br>2. Behavioral reduction counseling vs. none<br>3. Brief motivation counseling vs. none<br>4. Behavioral activation counseling vs. none | Target N = 512 | NCI P01 (2014–2019) | |

| | | | | | |
|---|---|---|---|---|---|
| Testing relapse recovery intervention components study | SMART | First randomization (one of three relapse recovery conditions):<br>1. Control<br>2. Behavioral smoking reduction counseling and nicotine mini-lozenges<br>3. Recycling counseling<br>Second randomization if they choose to make a new aided quit attempt (in a $2^2$ factorial design):<br>1. Supportive counseling vs. brief information<br>2. Skill training counseling vs. brief information | Target<br>N = 1200 | NCI P01<br>(2014–2019) | |
| Quitline Study | $2^3$ factorial | 1. 6 weeks of NRT vs. 2 weeks of NRT<br>2. Nicotine patch vs. patch plus nicotine gum<br>3. Standard four-call counseling vs. four-call counseling plus medication adherence counseling | 987 | NCI K05<br>(2008–2014)<br>and NCI<br>RC1<br>(2008–2012) | Smith et al. (2013) |
| Smokefree.gov study | $2^4$ factorial | 1. Access to Smokefree.gov vs. a "lite" website<br>2. Telephone quitline counseling vs. none<br>3. A smoking cessation brochure vs. a lite brochure<br>4. Motivational e-mail messages vs. none<br>5. Nicotine mini-lozenge vs. none | 1034 | NCI K05<br>(2008–2014)<br>and NCI<br>ARRA<br>funding | Fraser et al. (2014) |

*UW-CTRI* University of Wisconsin Center for Tobacco Research and Intervention, *NRT* nicotine replacement therapy, *SMART* sequential, multiple assignment, randomized trial

**Table 2** The factors and experimental conditions for Motivation Study 1

| Conditions | Nicotine patch | Nicotine gum | Behavioral reduction counseling | Motivational interviewing strategies |
|---|---|---|---|---|
| 1 | Patch | Gum | Yes | MI |
| 2 | Patch | Gum | Yes | No MI |
| 3 | Patch | Gum | No | MI |
| 4 | Patch | Gum | No | No MI |
| 5 | Patch | None | Yes | MI |
| 6 | Patch | None | Yes | No MI |
| 7 | Patch | None | No | MI |
| 8 | Patch | None | No | No MI |
| 9 | None | Gum | Yes | MI |
| 10 | None | Gum | Yes | No MI |
| 11 | None | Gum | No | MI |
| 12 | None | Gum | No | No MI |
| 13 | None | None | Yes | MI |
| 14 | None | None | Yes | No MI |
| 15 | None | None | No | MI |
| 16 | None | None | No | No MI |

*MI* motivational interviewing

quit attempts and the success of those attempts (Motivation Study 2). In that trial, participants are randomized to one of two levels on four different factors aimed at reducing smoking (see Table 1). The third experiment in this center grant screens intervention components designed to help smokers who have tried to quit and relapsed (Testing Relapse Recovery Intervention Components Study). This experiment uses a SMART design – an innovative approach aimed at optimizing adaptive treatments. In a SMART design, randomization occurs at more than one stage, with randomization at a later stage based on response to treatment at an earlier stage. This optimization trial tests adaptive intervention components at two stages: when a smoker relapses and when a smoker decides to make a new quit attempt. Participants initially receive a usual care cessation treatment. If they relapse, they are randomized to one of three Relapse Recovery conditions (see Table 1). If participants not in the control condition choose to make a new aided quit attempt, they are randomly assigned to one of four treatment conditions (in a $2 \times 2$ factorial design) screening two cessation intervention components. We have also conducted other factorial experiments detailed in Table 1.

In the subsequent sections of this chapter, we will share how we implemented these optimization trials, the challenges we encountered, and the lessons we learned. While we do not profess to have all the answers, we will discuss our approach to addressing the complexities of implementing factorial designs. Specifically, this chapter will address treatment factor selection, factor integration and implementation, fidelity in implementation, randomization, data collection, considering the participant perspective, and reporting results of factorial experiments.

**Table 3** The factors and experimental conditions for Cessation Study 1

| Condition | Prequit nicotine patch | Prequit nicotine gum | Prequit counseling | In-person cessation counseling | Phone cessation counseling | Postquit nicotine patch + gum duration |
|---|---|---|---|---|---|---|
| 1 | Patch | Gum | Intensive | Minimal | Minimal | 8 weeks |
| 2 | Patch | Gum | Intensive | Minimal | Intensive | 16 weeks |
| 3 | Patch | Gum | Intensive | Intensive | Minimal | 16 weeks |
| 4 | Patch | Gum | Intensive | Intensive | Intensive | 8 weeks |
| 5 | Patch | Gum | None | Minimal | Minimal | 16 weeks |
| 6 | Patch | Gum | None | Minimal | Intensive | 8 weeks |
| 7 | Patch | Gum | None | Intensive | Minimal | 8 weeks |
| 8 | Patch | Gum | None | Intensive | Intensive | 16 weeks |
| 9 | Patch | None | Intensive | Minimal | Minimal | 16 weeks |
| 10 | Patch | None | Intensive | Minimal | Intensive | 8 weeks |
| 11 | Patch | None | Intensive | Intensive | Minimal | 8 weeks |
| 12 | Patch | None | Intensive | Intensive | Intensive | 16 weeks |
| 13 | Patch | None | None | Minimal | Minimal | 8 weeks |
| 14 | Patch | None | None | Minimal | Intensive | 16 weeks |
| 15 | Patch | None | None | Intensive | Minimal | 16 weeks |
| 16 | Patch | None | None | Intensive | Intensive | 8 weeks |
| 17 | None | Gum | Intensive | Minimal | Minimal | 16 weeks |
| 18 | None | Gum | Intensive | Minimal | Intensive | 8 weeks |
| 19 | None | Gum | Intensive | Intensive | Minimal | 8 weeks |
| 20 | None | Gum | Intensive | Intensive | Intensive | 16 weeks |
| 21 | None | Gum | None | Minimal | Minimal | 8 weeks |
| 22 | None | Gum | None | Minimal | Intensive | 16 weeks |
| 23 | None | Gum | None | Intensive | Minimal | 16 weeks |
| 24 | None | Gum | None | Intensive | Intensive | 8 weeks |
| 25 | None | None | Intensive | Minimal | Minimal | 8 weeks |
| 26 | None | None | Intensive | Minimal | Intensive | 16 weeks |
| 27 | None | None | Intensive | Intensive | Minimal | 16 weeks |
| 28 | None | None | Intensive | Intensive | Intensive | 8 weeks |
| 29 | None | None | None | Minimal | Minimal | 16 weeks |
| 30 | None | None | None | Minimal | Intensive | 8 weeks |
| 31 | None | None | None | Intensive | Minimal | 8 weeks |
| 32 | None | None | None | Intensive | Intensive | 16 weeks |

## 3 Factor Selection

In a factorial experiment, researchers select possible levels of an intervention component, and together these levels make up a factor (a manipulated independent variable in the experiment). Some factors have an intervention component turned on or off (e.g., nicotine gum vs. none), whereas in others the levels represent a more intensive versus a less intensive version of an intervention component (e.g.,

**Table 4** The factors and experimental conditions for Cessation Study 2

| | Cessation interventions | | Adherence interventions | | |
|---|---|---|---|---|---|
| Condition | Nicotine patch + gum duration | Maintenance phone counseling | Medication adherence counseling | Automated medication adherence calls | Medication monitoring feedback |
| 1 | 26 weeks | Intensive | MAC | Calls | Feedback |
| 2 | 26 weeks | Intensive | MAC | Calls | No feedback |
| 3 | 26 weeks | Intensive | MAC | None | Feedback |
| 4 | 26 weeks | Intensive | MAC | None | No feedback |
| 5 | 26 weeks | Intensive | None | Calls | Feedback |
| 6 | 26 weeks | Intensive | None | Calls | No feedback |
| 7 | 26 weeks | Intensive | None | None | Feedback |
| 8 | 26 weeks | Intensive | None | None | No feedback |
| 9 | 26 weeks | None | MAC | Calls | Feedback |
| 10 | 26 weeks | None | MAC | Calls | No feedback |
| 11 | 26 weeks | None | MAC | None | Feedback |
| 12 | 26 weeks | None | MAC | None | No feedback |
| 13 | 26 weeks | None | None | Calls | Feedback |
| 14 | 26 weeks | None | None | Calls | No feedback |
| 15 | 26 weeks | None | None | None | Feedback |
| 16 | 26 weeks | None | None | None | No feedback |
| 17 | 8 weeks | Intensive | MAC | Calls | Feedback |
| 18 | 8 weeks | Intensive | MAC | Calls | No feedback |
| 19 | 8 weeks | Intensive | MAC | None | Feedback |
| 20 | 8 weeks | Intensive | MAC | None | No feedback |
| 21 | 8 weeks | Intensive | None | Calls | Feedback |
| 22 | 8 weeks | Intensive | None | Calls | No feedback |
| 23 | 8 weeks | Intensive | None | None | Feedback |
| 24 | 8 weeks | Intensive | None | None | No feedback |
| 25 | 8 weeks | None | MAC | Calls | Feedback |
| 26 | 8 weeks | None | MAC | Calls | No feedback |
| 27 | 8 weeks | None | MAC | None | Feedback |
| 28 | 8 weeks | None | MAC | None | No feedback |
| 29 | 8 weeks | None | None | Calls | Feedback |
| 30 | 8 weeks | None | None | Calls | No feedback |
| 31 | 8 weeks | None | None | None | Feedback |
| 32 | 8 weeks | None | None | None | No feedback |

*MAC* medication adherence counseling

26 vs. 8 weeks of medication). Selecting which intervention components to test in a factor, according to MOST, should be based on theory, an explicit conceptual model, and, most importantly, extant research supporting promising intervention components (Collins et al., 2011, 2016; Collins, Murphy, Nair, & Strecher, 2005; see Chapter 2 of the companion volume, Collins, 2018). There are also practical considerations to take into account when selecting components and component

levels for an experiment, including selection of robust intervention components with appropriate comparison levels, common treatments/constant components, and the similarity of the intervention components being tested.

## *3.1 Selection of Robust Intervention Components with Appropriate Comparison Levels*

One of the strengths of factorial designs is they enable researchers to tease apart the main effects of each intervention component tested and determine which intervention components are and are not exerting effects on the primary outcome. Therefore, selecting components and appropriate comparison levels is key. It is important to include comparison levels that would be appropriate to include in the final optimized treatment package (i.e., both levels would be feasible and appropriate to implement), and it is important to ensure that the intervention and the comparison level address the research question. Finally, in order to have the best probability of observing an effect, if there is one, while maintaining a reasonable sample size, the comparison levels of the factors must be selected to produce a clinically meaningful effect size. The difference in outcome between the on vs. off (or of more intensive vs. less intensive) levels needs to be sufficient so that the main effect of on vs. off produces a measurable, ideally statistically significant (however that is defined), independent effect. This is particularly important because, in designs with more than two factors, the majority of the participants receive additional intervention components. For instance, in the Cessation Studies 1 and 2, we included a factor that examined medication duration (8 vs. 16 weeks of combination nicotine replacement therapy (NRT) in Cessation Study 1 and 8 vs. 26 weeks of combination NRT in Cessation Study 2). Although we did not find significant differences in 26-week abstinence rates in the 8- vs. 16-week comparison, 26 weeks of combination NRT yielded higher abstinence rates than 8 weeks (Piper, Fiore, et al., 2016; Schlam et al., 2016). This was an important empirical question: Do smokers need 26 weeks of medication, or does 16 weeks sufficiently improve upon the effects of 8 weeks of medication? However, if we only had one study and wanted to know whether extended medication was better than the standard 8 weeks of medication, we would have wanted to ensure that our comparison was strong enough to exert an effect (e.g., 26 vs. 8 weeks of medication). In other words, the target intervention component (the on condition) needs to be robust, and the comparison level (the off condition) needs to be appropriate, such as the current clinical practice; the comparison should not be a straw man nor should it be so robust that it won't yield a main effect of the intervention component. For example, if a factor compared three counseling sessions (on) vs. two counseling sessions (off), one might need a very large sample size to achieve sufficient statistical power to detect the main effect. In summary, it is important when powering a factorial design to choose the effect size carefully and to select factor levels that are robust enough to deliver that effect.

In addition, it is important to ensure that either level of a factor (i.e., the lower and higher levels) could be included in the optimized treatment package. If it turns out that 16 weeks of combination NRT is not more effective than 8 weeks, then the lower level of 8 weeks of medication would be the level identified as appropriate for the optimized treatment package. Therefore, the lower level and the higher level need to be appropriate and feasible for future implementation. That is, by including 8 weeks of NRT as the lower level, one is committing to providing at least 8 weeks of NRT to all participants in the eventual RCT and scaled intervention.

## *3.2 Common Treatments or Constant Components*

When conducting a factorial experiment, it is also important to consider whether to use a common treatment (called a constant component in Chapter 3 in the companion volume). For instance, in Cessation Studies 1 and 2, all participants received a common treatment of 8 weeks of combination NRT, a very effective smoking cessation treatment (Cahill, Stevens, Perera, & Lancaster, 2013) that exceeded standard of care. This design likely made it more difficult to demonstrate a main effect of intervention components that were added to this strong common treatment. When the lower level of a factor already has a considerable effect, or when the common treatment is strong, the on conditions need to be strong enough to produce detectable main effects (or, alternatively, a larger sample size is needed for adequate power to detect the smaller effect size).

## *3.3 Similarity of the Intervention Components*

Another practical consideration when selecting intervention components is ensuring that two intervention components are not so similar that they could be considered equivalent. For example, researchers should consider whether the desired goal is to compare two different approaches to addressing one specific treatment mechanism or whether each intervention component addresses an independent treatment mechanism. Redundant intervention components included as distinct factors that address the same treatment mechanism may produce an antagonistic interaction (see Chapter 4 in the companion volume). In Cessation Study 1, we tested an in-person counseling factor and a phone counseling factor, each with intensive vs. minimal levels; both forms of counseling were delivered during the first 2 weeks after the target quit day and covered similar topics. We wanted to determine whether one modality was better than the other (in-person vs. phone) and whether they exerted additive effects (i.e., was more counseling contact better than less as reported in Fiore et al., 2008). However, by treating the in-person and phone counseling as independent factors, we ended up with an antagonistic interaction: receiving both treatments, via different modalities, did not improve cessation rates on average

above receiving just one or the other. Clearly identifying the scientific question can help protect against this problem. If the goal is to compare two intervention components that are similar (i.e., they target the same mechanism), but only one component would be included in the final treatment package, then consider making each component a level of a single factor (e.g., in-person counseling vs. phone counseling), rather than two separate factors (in-person counseling vs. none; phone counseling vs. none). Using this approach, no participant would get both in-person and phone counseling.

## 4 Factor Integration and Implementation

Once the intervention components and their appropriate comparisons have been selected, the components need to be delivered in a compatible way, even if the components' mechanisms of action are quite different. Given that the goal is to develop an optimized integrated treatment package from these components, it is essential that intervention components do not contradict one another, operate via antagonistic mechanisms, or require incompatible delivery methods. For example, in Motivation Study 1, we tested behavioral reduction counseling and motivational interviewing. The theoretical rationales behind these two different treatments were quite distinct; behavioral reduction is a behavioral, directive approach that involves setting reduction goals, whereas motivational interviewing is a nondirective, client-centered counseling technique for eliciting behavior change by helping clients to explore and resolve ambivalence. Although these treatments likely targeted different potential mechanisms, it was important to ensure the protocols were presented in a compatible manner to participants randomized to the on levels of both factors (e.g., the more participant-focused motivational interviewing was provided first, followed by the more directive behavioral reduction counseling).

Another key design consideration involves participant burden. How intense is the treatment, and what is the participant burden, if all factors are set to on or if all are set to off? Components and levels need to be selected to ensure that if any combination of levels (including all on or all off) is selected, the condition is feasible to deliver. If all factors are set to off, is the participant getting any treatment? Is this ethical? If all factors are set to on, is the burden beyond what a participant may be willing to endure? We found that if participants do not receive any meaningful treatment to help them quit smoking, it can compromise participant retention. This finding highlighted the need for there to be either a constant component or for at least some of the factors to designate lower levels that include evidence-based treatment to ensure that all participants got at least some treatment (although, as noted previously, the constant component or lower level should not be so potent that the effects of the higher or on level will likely not be detectible). We also learned that when all levels were set to on, some participants experienced high burden: (1) long in-person counseling sessions (e.g., 45 minutes in the Relapse Recovery Study), (2) long phone calls consisting of assessment plus three types of counseling (50-minute

calls; Motivation Study 2), and (3) four counseling sessions in the first week of the quit attempt (Cessation Study 1). The amount of time and attention required for this much counseling content was difficult for some participants and may have resulted in failure to sustain engagement or to benefit from the intervention. To the extent that this engendered treatment fatigue, it likely lowered the "dose" of treatment and lowered the effect size of the individual treatment components among this subgroup of participants. The benefit of a factorial design, however, is that when considering how many factors of each type (e.g., counseling) to include in a study, researchers can prioritize the factors and perhaps choose less burdensome factors so that those included in the experiment are delivered in an adequate dose to all participants.

Finally, when integrating factors, the timing of intervention component delivery needs to be considered. If a factor is not implemented until later in the experiment (e.g., after the target quit day or after an early smoking relapse), then fewer participants may receive that treatment due to study attrition. For example, in Cessation Study 2, maintenance counseling calls did not begin until 3 weeks after the target quit day. This timing may make it harder to detect an effect of this intervention component because some participants will be unwilling or unavailable to receive the treatment because they are discouraged about being able quit and have dropped out of the study. A similar pattern may occur if an intervention component delivered early in treatment negatively affects treatment engagement, but this is difficult to predict in advance. This type of effect appeared to occur in the Smokefree.gov study (Fraser et al., 2014); we observed that sending multiple daily emails starting immediately after enrollment was associated with lower use of the website, which interfered with the beneficial effect of the website.

## 5 Implementing Factors with Fidelity

We identified four key elements to implementing treatment with fidelity – appropriate protocols, staff training, a rigorous system to guide study activities, and fidelity checks. These are important in all research, but when using complex factorial designs, they are critical.

### *5.1 Protocols*

Implementation of each intervention component as intended requires a detailed protocol for each component and a clear plan for integrating various components. In other words, protocols are required for all levels of each factor, as are protocols for implementing multiple factors set to the on level. In addition, there needs to be a clear, consistent temporal order of intervention components. If two factors are on, in what order should they be provided? Should one factor be provided first, based on its purported mechanism? As noted above, in Motivation Study 1, we

kept the sequencing of the counseling intervention components consistent across participants with both components on (e.g., motivational interviewing counseling always preceded behavioral reduction counseling in hopes that the mechanisms would have the least interference).

Once the implementation order has been determined, pragmatic adjustments may be necessary for certain combinations of intervention components. For instance, the timing of intervention components that co-occur should be considered. If two intervention components are scheduled to occur on the same day and that might result in excessive burden, then can the timing be adjusted to maintain fidelity to the treatment yet minimize participant burden? In Cessation Study 1, we tested in-person and phone counseling designed to support participants in the days surrounding the target quit day. Both factors included a contact on the quit day. However, it did not make sense to have both a quit day visit and a phone call. Therefore, for participants randomized to the ON level of both the in-person and phone counseling factors, we moved the quit day call to 1 day after the quit day to keep the combined protocol as consistent as possible with the standard protocol, yet also make clinical sense. If we had decided to include both components in our optimized treatment package, this is how we would have implemented the optimized treatment. Having appropriate protocols to guide the implementation of each intervention component, including the implementation of certain components in concert, increases the likelihood that each intervention component will be implemented as intended, providing the most rigorous test of the intervention.

## *5.2 Staff Training*

As with any study, staff training is critical to implementing treatment with fidelity. Even with a large number of treatment conditions, staff training and study design can ensure that implementation with fidelity is feasible for staff. In our studies, bachelor-level staff serve as health counselors. When preparing to launch a study, our health counselors (HCs) are first trained on individual protocols for each level of each factor (2–5 factors). This approach ensures the HCs are able to administer each of the intervention components with fidelity. We then train HCs to administer combinations of intervention components. Transitions from one intervention component to the next are scripted in the counseling protocols, with the order of implementation specified by the study database. Thus, in our 5-factor experiment (Cessation Study 2), we did not train staff to administer 32 different treatment conditions. Rather, we thoroughly trained HCs in each individual treatment component protocol and then in any necessary transitions, so they were able to easily use the protocols in all possible combinations. HCs were required to demonstrate proficiency in each intervention component and in a few key combinations before treating participants.

### 5.3 Rigorous Systems

Our ability to focus on training HCs to administer just five intervention protocols (rather than 32) is due to the use of a rigorous database with embedded protocols to guide study activities. This system allows researchers to ensure that participants receive all the intervention components to which they are assigned and only the intervention components to which they are assigned. Because the components are independent of one another, they can be clearly defined and specified so that HCs know exactly who gets what and at which study contact each intervention component should be carried out. Our study databases also include a scheduling program that tracks when research contacts occur, including when visits or calls need to be scheduled based on the participants' condition. The length and timing of appointments vary by condition; this information is populated in the database by condition, making scheduling relatively straightforward for the HC.

For our studies, we use an Access/SQL Server database (other relational databases could also perform this function) that allows researchers to create condition-specific templates, randomize participants, and then immediately create a participant's treatment and assessment schedule based on the templates and randomized condition. Figure 1 provides an example of a participant's study appointment schedule. It includes the ideal date to schedule the appointment (typically based on either the date of consent or the target quit date), as well as start and end windows that tell HCs the range of dates in which they can schedule the appointment and not have it fall "out of window." The database also contains condition-specific preparation lists for each appointment (not shown) so that HCs know what they need to do to prepare for the upcoming visit or phone

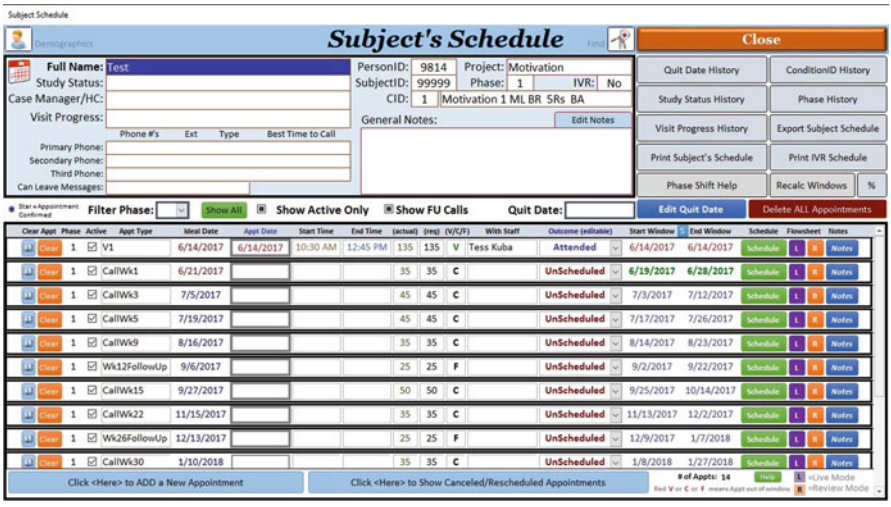


**Fig. 1** Condition 1 participant schedule

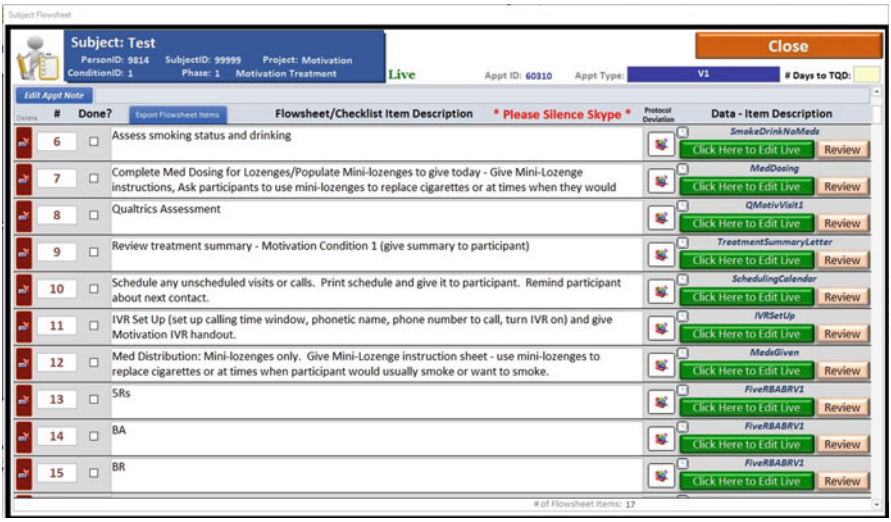


**Fig. 2** Condition 1 participant flowsheet (Note. In Condition 1, all factors are on. At this visit, the health counselor distributes medication and provides three types of counseling. There are separate lines in the flowsheet for each counseling type (5Rs motivation counseling, behavioral activation [BA], and behavioral reduction [BR]); this allows the counselor to see what counseling needs to be provided at a glance and to document that they provided all three types of counseling. However, the counseling protocol provided in the link (the green box) is the combined counseling guide customized for Condition 1 with appropriate transitions between the three types of counseling)

call, including what medication (type, dose, and amount) needs to be prepared. For each appointment (in-person or phone), the database provides HCs with a procedure checklist or "flowsheet" (see Fig. 2). HCs proceed through the flowsheet, implementing each appointment activity, in order, for that appointment. For each relevant flowsheet item, the database has an electronic link to the appropriate online assessment (e.g., Visit 2 assessment) or treatment tailored to each participant's specific condition and appointment. For example, there is a link to the medication distribution instructions for Visit 4 or a link to the counseling protocol for that participant for that contact (e.g., Visit 1 5Rs motivation counseling, behavioral activation, and behavioral reduction protocols with appropriate transitions between the types of counseling). As shown in Fig. 2, HCs simply click on the green "Click Here to Edit Live" button to the right of the Qualtrics assessment item, opening up the Visit 1 assessment for participants randomized to Condition 1.

In sum, when participants provide consent, they are randomized to an experimental condition, and the database creates a study schedule for them based on their experimental condition and corresponding flowsheets for each contact that guide the delivery of treatments and assessments (i.e., the database ensures that HCs are prompted to deliver all participants in Condition 14 identical Condition 14 interventions and assessments and nothing else). It is important to note that flowsheets and scheduling could be tracked using paper forms and/or Excel

worksheets (e.g., Pellegrini, Hoffman, Collins, & Spring, 2014). This method would likely be more cumbersome than using a database but would likely still achieve the goal of ensuring that what happens to each participant, and when, is clearly spelled out for each experimental condition.

## *5.4 Fidelity Checks*

Fidelity checks are also critical for ensuring that participants are receiving the intervention components consistent with their assigned experimental condition and, conversely, are not receiving intervention components to which they are not assigned. For our research, we audio-record counseling sessions to provide supervision and feedback. We also score the recordings for counseling content to ensure the appropriate content is being covered (i.e., we confirm each highlighted topic is addressed), and no non-protocol content is addressed (see Fig. 3). If an experiment includes more than one counseling intervention component, it is important to verify that there is no bleeding over of counseling content between intervention components. Further, implementing a factorial experiment with fidelity requires ensuring that each level of each factor is implemented the same way irrespective of the other levels of the other factors (except in the case of planned exceptions, such as those addressed above regarding avoiding scheduling two study contacts on the same day).

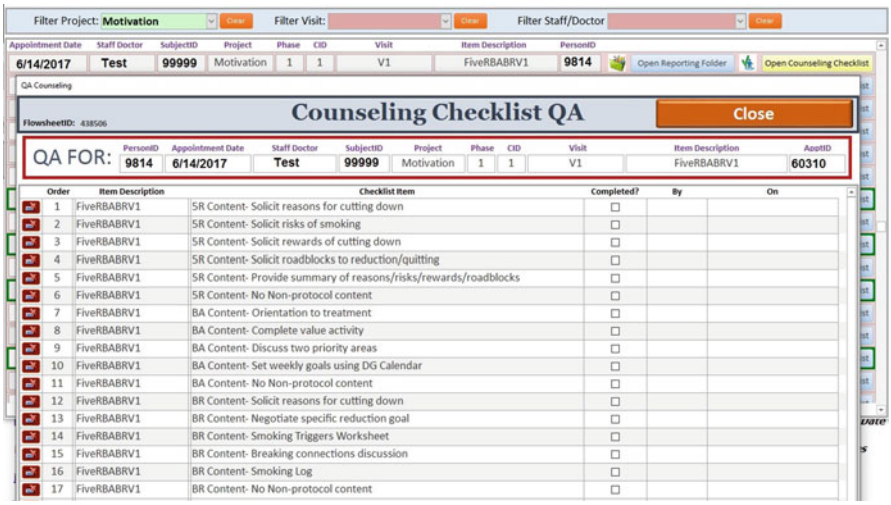


**Fig. 3** Condition 1 participants' fidelity checklist for Visit 1 counseling (5Rs motivation counseling, behavioral reduction, and behavioral activation)

While implementing factorial experiments with fidelity is a challenge, it is definitely possible. When beginning this line of research, we found it helpful to recruit slowly for the study at first and to enroll "pilot" participants. This allowed us to get feedback from staff on all aspects of implementation, from participant understanding of the consent procedures to the ability of the database to keep staff on track regarding which treatments should and should not to be delivered. This also allowed us to talk to participants about their experiences and get their feedback. These processes allowed us to address unforeseen problems and feel confident when we officially launched the study.

## 6 Randomization

Randomization for factorial experiments, as for RCTs, is essential because it strengthens the inferences that can be made about the effects of treatment on outcome. This is, in part, because randomization reduces the impact of individual difference variables; participants randomized to an experimental condition should have equal representation in terms of gender, age, race, etc. However, randomization is more likely to fail (i.e., experimental conditions will not have equal representation of baseline characteristics) when individual experimental condition sample sizes are small (Saint-Mont, 2015). We opted to use stratification/blocking variables (e.g., gender, healthcare clinic) to minimize failures of randomization. We found that when using stratified randomization for a factorial study, overall condition allocation initially looked somewhat uneven. However, distribution across all 8, 16, or 32 conditions tended to even out over time (e.g., after there were about 10 participants per condition). We carefully track the condition allocation by block and stratum to ensure that the randomization routines work as intended.

The complexity of implementing large-scale factorial designs (with 4–5 factors resulting in 16–32 different intervention combinations) led us to randomize participants prior to their first treatment visit and prior to them providing informed consent. In RCTs, treatment condition should not be known by the staff or participants prior to providing consent to ensure that participants do not differentially decline, based on their treatment assignment. However, we found that informing staff of the participants' treatment condition prior to consent is important in our factorial experiments in order to know how long to schedule the initial visit, which may vary considerably by experimental condition when lengthy intervention components are assigned. Knowing the treatment assignment also ensures staff have adequate time to prepare the appropriate treatment (e.g., review counseling protocols, have appropriate handouts, prepare medication with proper dosing based on information collected during screening). It is critical that staff are not rushed so there is no increased risk of providing the wrong treatment. We do not, however, reveal the treatment condition to participants prior to consent. An alternative solution would be to randomize participants after consent but not provide treatment until a subsequent visit. This rigor should be weighed against the burden on participants

of an additional study contact and the likelihood that some participants will drop out after providing written consent and never attend the first treatment contact if these procedures are not completed in the same visit.

## 7 Data Collection

As with any study, in a factorial experiment, it is important to consider who will conduct the assessments (i.e., will they be conducted by the treatment provider or at another time by an independent assessor), as well as the duration and timing of assessments. As for who conducts the assessments, the most rigorous assessment approach would be to have nontreatment providers blind to treatment condition conduct the assessments. However, this can be a challenge in a factorial design where some participants may have many more treatment contacts than others. To minimize participant burden and reduce the overall number of contacts, the treatment provider can conduct both assessments (especially intervention-specific assessments) and treatment. This is typically how we have conducted assessments in our factorial experiments. During in-person visits, participants complete most assessments independently using a web-based survey on a tablet the treatment provider does not see. During follow-up calls, nontreatment providers collect outcome data. However, during treatment phone contacts that involve assessment, HCs collect assessments first and then provide the treatment in an effort to minimize demand characteristics. This approach has two unfortunate side effects: the treatment provider collecting the data potentially biases the answers and calls which involve assessment followed by counseling can become excessively long and burdensome.

By the time the assessment is complete (especially if participants elaborate on their answers during the assessment), some participants may have little energy left to devote to the counseling. This problem is particularly challenging during certain phone contacts. We have found that if the HC clearly explains that topics that arise during the assessment can be explored in more detail during the counseling, this can reduce the duration of the assessment and mitigate fatigue. The design and administration of assessment – as it relates to both burden and possible conflict with treatment delivery – are important to consider during protocol development. Generally, we have tried to limit the assessment burden during sessions where lengthier counseling treatments are being delivered and to administer the majority of assessments during in-person visits when assessments can be completed via web-based surveys, which are faster than answering questions over the phone.

In factorial designs, assessments related to specific intervention components (e.g., medication adherence, mechanisms of action, outcomes) need to be scheduled appropriately for the participants receiving those components. Therefore, researchers need to ensure that they have clear flowsheets that specify the timing of each assessment for each condition, especially if there are different versions of assessments for different conditions or at different times (e.g., we asked about

medication use late in treatment only of those randomized to extended medication). As discussed above, our database has assessment links embedded in participants' flowsheets so that the staff can click on the link and go directly to the assessment appropriate for that treatment condition and that study contact.

Variability in treatment and assessment contacts in factorial designs can make it challenging to assess constructs consistently. For instance, in Cessation Study 2, at Week 16 only half of the participants attended an in-person treatment visit, but all participants received an assessment phone call (the call was scheduled before the visit for those receiving both). We therefore did not obtain in-person biochemical verification of abstinence from any participants at Week 16; rather, in the Week 16 phone call, we assessed self-reported abstinence from everyone because that was the measure we could collect consistently from all participants regardless of whether they had an in-person visit. Regardless of condition differences, it is important that all major outcomes be assessed at an appointment that is identical for all participants (either an in-person visit or a call) and that all participants are scheduled to attend (e.g., a 26-week assessment call for all participants).

## 8 The Participant Perspective

Factorial experiments are an efficient way to address many research questions, but they are more complicated to implement than randomized controlled trials. However, participants do not need to understand or experience this complexity. Participants need to understand the different treatments they might receive in order to provide informed consent, but they do not need to understand the complexities involved in creating and keeping track of 32 sets of flowsheets and checking for treatment fidelity. Rather, participants should experience their assigned treatment condition as a single, seamless treatment package. When we inform participants about our studies, we describe each level of each factor, including potential risks, and explain that they will get a combination of these intervention components (e.g., we tell them "you may get one or more of the following treatments"). This approach ensures that participants are providing informed consent to receive all possible intervention components and combinations. We also inform participants that the study will have a range of study contacts (e.g., 3–5 study visits, up to 4 coaching calls, and 3 follow-up calls). To ensure that participants experience the study as seamless, once they are randomized, we provide them with condition-specific treatment summary letters that list exactly which intervention components they will receive as part of the study and when they will have study contacts.

It is important to note that participants may elect not to engage or continue with a specific intervention component (e.g., take medication, attend counseling sessions), even after consenting to receive all possible intervention components and combinations. Because each factor is independent, participants should still receive the other components, as appropriate, based on their experimental condition. Therefore, if a participant were no longer able to use medication due to an adverse

event, the participant would still receive the counseling intervention components. Note that discontinuation of certain constant components may affect the relevance of certain intervention components. For example, if a participant in Cessation Study 2 decided to discontinue the constant component of cessation medication, intervention components designed to increase medication adherence would be irrelevant. In such cases, it is important to have infrastructure for documenting intervention discontinuations and protocol deviations (e.g., noting that an irrelevant treatment was discontinued). Despite changes to individual treatment regimens, all participants should be included in analyses, using the intent-to-treat principle, based on their assigned condition, regardless of whether they engaged in each intervention component. In fact, one of the benefits of a factorial design is the ability to assess overall engagement in each individual component and how such engagement is related to the outcomes of interest.

As mentioned above, participant burden is an important consideration from both the clinical and scientific perspective. To minimize participant burden, it is important to consider the amount and spacing of all types of contacts (assessment, reminder calls, treatment). If the research procedures are so demanding that participants do not engage in the treatment (or do not engage fully), this compromises the ability to draw firm conclusions about the effects of the intervention component (although the argument could be made that lack of engagement is an important outcome). The number and timing of treatment contacts in the factorial experiment should be examined with the "frequency of contact" burden in mind. To determine our contact schedule for participants, we first develop a list of ideal contact dates, with corresponding windows surrounding each ideal contact date (showing the range of acceptable dates in which to complete that contact) to provide adequate opportunity to collect data or provide treatment at our desired time points as well as ensuring breaks between contacts. Figure 1 shows the ideal date and the start and end window dates surrounding the ideal date of each appointment.

## 9 Reporting Results

In a reporting and publishing world built for RCTs, factorial designs are sometimes a round peg trying to fit in a square hole. We have encountered a few challenges and developed solutions when reporting results in both the National Clinical Trial Registry and in journals.

A factorial screening experiment funded by the National Institutes of Health in the United States is considered a clinical trial and needs to be registered in clinicaltrials.gov (see Smokefree.gov study: NCT01342523 or Motivation Study 1: NCT01122238). However, clinicaltrials.gov is designed for reporting RCTs. We have developed strategies for registering our factorial trials. First, each level of each factor should be entered as a treatment (e.g., for Cessation Study 1, treatments would include minimal in-person counseling, intensive in-person counseling, 8 weeks of patch, and 16 weeks of patch). Second, all combinations of treatment (e.g., every

experimental condition) must be entered as if they were separate conditions (i.e., separate "study arms"); select which intervention components are offered in each particular condition. This produces a rather large matrix (depending on the number of factors). Third, some results (e.g., number recruited, withdrawals, abstinence rates) are reported at the condition/study arm level (e.g., Condition 1: N = 12, withdrawals = 2, percent abstinent = 34%). Other results (e.g., adverse events) are reported for each level of each factor (e.g., the adverse events for 8 weeks of patch vs. the adverse events for 16 weeks of patch); if there are participants randomized to not receive medication (as in Motivation Study 1 for those with both nicotine patch and gum turned off), these can be combined in a single group. Finally, the main outcomes are entered in relation to the primary study aims (e.g., smoking abstinence) for each of the study's primary comparisons (i.e., on vs. off). It is important to note that the professionals at clinicaltrials.gov will provide consultation as needed.

Another substantial reporting challenge for factorial clinical trials involves reporting participant flow in consort diagrams and outcome tables. Our first published factorial experiment, the Quitline Study (see Table 1), was a $2^3$ factorial design. The outcome paper (Smith et al., 2013) included a consort diagram with factors branching off one by one until each condition was represented with its individual results (e.g., 8 conditions with 8 abstinence rates). For our subsequent factorial studies (Cook et al., 2016; Fraser et al., 2014; Piper, Fiore, et al., 2016; Schlam et al., 2016), consort diagrams and outcomes were reported based on the factors (on vs. off; subjects were thus counted multiple times). We prefer the latter approach because factorial studies evaluate outcomes based on main effects (i.e., comparisons of on vs. off); therefore, it is important to examine participant flow in each on vs. off level. This approach also provides a cleaner presentation of results, especially when there are more than three factors.

Finally, translating the complexity of factorial designs into a written manuscript, within the word limits imposed by most journals, is a challenge. The design is somewhat more complex and requires descriptions not only of the design but also of each level of each factor. In addition, the need to report main effects and interaction effects increases the number of primary outcome results to report, plus there may be additional relevant analyses to report. One approach for investigators to consider is publishing the study methods in a separate paper and referring to this paper in subsequent outcome papers.

## 10 Conclusions

Factorial experiments permit efficient evaluation of multiple intervention components as part of the MOST approach to engineering optimized treatment packages. While there are complexities to implementing factorial experiments, they are certainly feasible, as we have demonstrated in our simultaneous screening of 15 potential intervention components with 80 different experimental conditions

within a single NCI center grant. In this chapter, we reviewed some of the implementation complexities, including implementation considerations to address during study design, factor integration, fidelity, timing and implementation of assessments, randomization, issues to consider regarding the participant perspective, and reporting results. We reviewed various implementation challenges and shared our approach to addressing these issues.

One final note of encouragement to researchers is that while there may be additional considerations in the initial implementation of factorial experiments and in reporting the results, once the protocols are written and the appropriate systems are in place, factorial designs can run as smoothly as RCTs while providing much more detailed and nuanced information.

## References

Almirall, D., Nahum-Shani, I., Wang, L., & Kasari, C. (2018). Experimental designs for research on adaptive interventions: Singly and sequentially randomized trials. In L. M. Collins & K. C. Kugler (Eds.), *Optimization of behavioral, biobehavioral, and biomedical interventions: The multiphase optimization strategy (MOST)*. New York, NY: Springer.

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*(2), 192–207.

Cahill, K., Stevens, S., Perera, R., & Lancaster, T. (2013). Pharmacological interventions for smoking cessation: An overview and network meta-analysis. *Cochrane Database of Systematic Reviews, 5*, CD009329. https://doi.org/10.1002/14651858.CD009329.pub2

Collins, L. M. (2018). *Optimization of behavioral, biobehavioral, and biomedical interventions: The multiphase optimization strategy (MOST)*. New York, NY: Springer.

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 effective tobacco use interventions. *Annals of Behavioral Medicine, 41*(2), 208–226. https://doi.org/10.1007/s12160-010-9253-x

Collins, L. M., Kugler, K. C., & Gwadz, M. V. (2016). Optimization of multicomponent behavioral and biobehavioral interventions for the prevention and treatment of HIV/AIDS. *AIDS and Behavior, 20*(Suppl 1), S197–S214. https://doi.org/10.1007/s10461-015-1145-4

Collins, L. M., Murphy, S. A., Nair, V. N., & Strecher, V. J. (2005). A strategy for optimizing and evaluating behavioral interventions. *Annals of Behavior Medicine, 30*(1), 65–73. https://doi.org/10.1207/s15324796abm3001_8

Cook, J. W., Collins, L. M., Fiore, M. C., Smith, S. S., Fraser, D., Bolt, D. M., … Mermelstein, R. (2016). Comparative effectiveness of motivation phase intervention components for use with smokers unwilling to quit: A factorial screening experiment. *Addiction, 111*(1), 117–128. https://doi.org/10.1111/add.13161

Fiore, M. C., Jaen, C. R., Baker, T. B., Bailey, W. C., Benowitz, N., Curry, S. J., … Wewers, M. E. (2008). Treating tobacco use and dependence: 2008 update. Retrieved from http://bphc.hrsa.gov/buckets/treatingtobacco.pdf

Fraser, D., Christiansen, B. A., Adsit, R., Baker, T. B., & Fiore, M. C. (2013). Electronic health records as a tool for recruitment of participants' clinical effectiveness research: Lessons learned from tobacco cessation. *Translational Behavioral Medicine, 3*(3), 244–252. https://doi.org/10.1007/s13142-012-0143-6

Fraser, D., Kobinsky, K., Smith, S. S., Kramer, J., Theobald, W. E., & Baker, T. B. (2014). Five population-based interventions for smoking cessation: A MOST trial. *Translational Behavioral Medicine, 4*(4), 382–390. https://doi.org/10.1007/s13142-014-0278-8

Kelly, K. (1994). Out of control: the new biology of machines, social systems and the economic world, 393. Retrieved from http://kk.org/mt-files/books-mt/ooc-mf.pdf

Pellegrini, C. A., Hoffman, S. A., Collins, L. M., & Spring, B. (2014). Optimization of remotely delivered intensive lifestyle treatment for obesity using the multiphase optimization strategy: Opt-IN study protocol. *Contemporary Clinical Trials, 38*, 251–259.

Piper, M. E., Baker, T. B., Mermelstein, R., Collins, L. M., Fraser, D. L., Jorenby, D. E., ... Fiore, M. C. (2013). Recruiting and engaging smokers in treatment in a primary care setting: Developing a chronic care model implemented through a modified electronic health record. *Translational Behavioral Medicine, 3*(3), 253–263. https://doi.org/10.1007/s13142-012-0178-8

Piper, M. E., Cook, J. W., Schlam, T. R., Jorenby, D. E., Smith, S. S., Collins, L. M., ... Baker, T. B. (in press). A randomized controlled trial of an optimized smoking treatment delivered in primary care. *Annals of Behavioral Medicine*.

Piper, M. E., Cook, J. W., Schlam, T. R., Smith, S. S., Bolt, D. M., Collins, L. M., ... Baker, T. B. (2016). Toward the development of precision smoking cessation treatment II: Proximal effects of smoking cessation intervention components on putative mechanisms of action. *Drug and Alcohol Dependence, 171*, 50–58 PMCID: PMC5262527.

Piper, M. E., Fiore, M. C., Smith, S. S., Fraser, D., Bolt, D. M., Collins, L. M., ... Baker, T. B. (2016). Identifying effective intervention components for smoking cessation: A factorial screening experiment. *Addiction, 111*(1), 129–141. https://doi.org/10.1111/add.13162

Piper, M. E., Schlam, T. R., Cook, J. W., Smith, S. S., Bolt, D. M., Loh, W.-Y., ... Baker, T. B. (2016). Toward precision smoking cessation treatment I: Moderator results from a factorial experiment. *Drug and Alcohol Dependence, 171*, 59–65 PMCID: PMC5263119.

Saint-Mont, U. (2015). Randomization does not help much, comparability does. *PLoS One, 10*(7), e0132102. https://doi.org/10.1371/journal.pone.0132102

Schlam, T. R., Fiore, M. C., Smith, S. S., Fraser, D., Bolt, D. M., Collins, L. M., ... Baker, T. B. (2016). Comparative effectiveness of intervention components for producing long-term abstinence from smoking: A factorial screening experiment. *Addiction, 111*(1), 142–155. https://doi.org/10.1111/add.13153

Smith, S. S., Keller, P. A., Kobinsky, K. H., Baker, T. B., Fraser, D. L., Bush, T., ... Fiore, M. C. (2013). Enhancing tobacco quitline effectiveness: Identifying a superior pharmacotherapy adjuvant. *Nicotine & Tobacco Research, 15*(3), 718–728. https://doi.org/10.1093/ntr/nts186

U.S. Department of Health and Human Services. (2014). The health consequences of smoking: 50 years of progress. A report of the Surgeon General. Retrieved from http://www.surgeongeneral.gov/library/reports/50-years-of-progress/full-report.pdf

# Multilevel Factorial Designs in Intervention Development


**Inbal Nahum-Shani and John J. Dziak**

**Abstract** Factorial designs are one of the many useful experimental tools that can be used to inform the construction of multicomponent behavioral, biobehavioral, and biomedical interventions. Clustering presents various challenges to investigators aiming to implement such designs. Clustering means that some or all individuals are nested in higher-level social or administrative units (e.g., schools, therapy groups). These multilevel settings generate dependency in data within clusters because individuals in one cluster tend to be more similar to each other than to individuals in other clusters. Such dependency has implications for the design of the factorial experiment, the model used to analyze the data, and the power for detecting the effects of interest. In this chapter, we discuss five classes of multilevel factorial designs that vary in terms of the nature of clustering (i.e., the process by which individuals become clustered or the reason why they are considered to be clustered), as well as the randomization scheme employed (i.e., whether randomization to experimental conditions is done at the individual level, the cluster level, or both). For each of the five classes, we discuss the scientific motivation for employing the multilevel factorial design, provide a model for analyzing data arising from employing a multilevel factorial design of this class, and offer formulas that investigators can use to calculate the expected power. Design considerations are also discussed with respect to each class. Our goal is to provide a comprehensive review to help investigators select the most suitable design given their scientific questions, target population, and available resources.

Inbal Nahum-Shani and John J. Dziak have contributed equally to this chapter.

I. Nahum-Shani (✉)
Institute for Social Research, The University of Michigan, Ann Arbor, MI, USA
e-mail: inbal@umich.edu

J. J. Dziak
The Methodology Center, The Pennsylvania State University, University Park, PA, USA



L. M. Collins, K. C. Kugler (eds.), *Optimization of Behavioral, Biobehavioral, and Biomedical Interventions*, Statistics for Social and Behavioral Sciences,
https://doi.org/10.1007/978-3-319-91776-4_3

## 1 Introduction

Factorial designs are a useful experimental tool in the construction of multicomponent behavioral, biobehavioral, and biomedical interventions (MBIs). These designs can be used in the optimization phase of the multiphase optimization strategy (MOST) to screen out multiple candidate intervention components simultaneously and inform their inclusion in an intervention package (see Collins, 2018, which is a companion to this volume). One challenge in implementing factorial experiments relates to the multilevel[1] nature of the targeted population. Specifically, in many cases the target population involves lower-level units (such as students, employees, or patients) that are clustered, namely, nested in higher-level social or administrative units (e.g., schools, organizations, therapy groups). This generates dependency in data within clusters because individuals in one cluster tend to be more similar to each other than to individuals in other clusters.

Clustering has implications for statistical power for detecting treatment effects. In addition to standard considerations that are relevant in any experiment (e.g., sample size, effect size, selected type I error rate), investigators planning an experiment with clustered individuals also have to take into account the number of clusters ($J$), the number of individuals within each cluster ($n$), and the intraclass correlation ($ICC$). The ICC reflects the degree of dependence in the response (i.e., outcome) among individuals within clusters. A large ICC can result in reduced power, particularly if $J$ is small or $n$ is large. However, the implications of these elements on power also depend on the nature of clustering in the experiment, as well as the randomization scheme employed.

By *the nature of the clustering*, we refer to the process by which individuals become clustered or the reason why they are considered to be clustered; there are two main reasons. First, individuals might belong to pre-existing units, such as clinics, hospitals, schools, or organizations. Here, we use the term pre-existing clusters (PECs) to describe units that exist prior to experimentation. Second, clusters might be generated as part of the experimentation itself. This can occur when individuals are independent prior to experimentation, but dependence within clusters of individuals is generated over the course of the experiment. This dependence may be generated for reasons that are practical (e.g., each of several therapists delivers the intervention to a subset of individuals), scientific (e.g., investigators seek to facilitate therapeutic group processes and capitalize on social reinforcers), or both. Such experimentally induced clusters (EICs) can be generated for all study participants (i.e., full EIC) or for only a subset of them (partial EIC).

[1] We would like to point out that in this chapter, the term *level* is used in two different ways. Here the term level is used to refer to a level of nesting in data with a cluster structure. Later, the term level will also be used to refer to one of the values that can be taken on by a manipulated factor in an experiment. We considered using a different term in one of these contexts but ultimately decided to remain consistent with the existing scientific literature. We made an effort to write this chapter so that the context makes the meaning clear.

The *randomization scheme* can also differ from one study to another. Specifically, randomization to experimental conditions can be done at the individual level (i.e., individuals are randomized), the cluster level (i.e., clusters are randomized), or both. The options available here depend on the nature of clustering as well as other considerations, which we discuss in detail below.

In this chapter, we discuss five classes of multilevel factorial designs that vary in terms of the nature of clustering, as well as the randomization scheme employed. The first involves PECs, with individuals within each cluster being randomized to experimental conditions. We label this class *within-PEC factorial experiments*. The second class also involves PECs, but clusters are randomized to experimental conditions. We label this class *between-PEC factorial experiments*. The third class can be conceptualized as a hybrid of a within-PEC and a between-PEC factorial: the unit of assignment is the individual (rather than cluster), but each individual has only a subset of possible experimental conditions to which s/he can be assigned, with the specific subset determined by membership in PECs. We label this class *hybrid-PEC factorial experiments*. The fourth class involves EICs that are generated for all individuals. Here, either individuals or clusters can be randomized to experimental conditions. We label this class *full-EIC factorial experiments*. The fifth class also involves EICs, but here clusters are generated for only a subset of the individuals. Hence, clusters of individuals are randomized to some experimental conditions, and single individuals are randomized to others. We label this class *partial-EIC factorial experiments*. These classes are summarized in Table 1.

Existing literature (e.g., Dziak, Nahum-Shani, & Collins, 2012; Nahum-Shani, Dziak, & Collins, 2017) provides some information on various forms of multilevel factorial designs, including power planning resources. Our goal in this chapter is to provide a comprehensive review to help investigators select the most suitable design given their scientific questions, target population, and available resources. Hence, for each of the five classes above, we begin by providing an example to clarify the scientific motivation for employing such multilevel factorial design and the settings in which this class would be most relevant. We then provide a model for analyzing data arising from employing a multilevel factorial design of this class. Based on these models, we then provide formulas that investigators can use to calculate the expected power for each class of multilevel factorial designs. The power formulas provided in this chapter are simplified versions of existing power planning resources; we re-expressed the original formulas so that they would include parameters that an investigator can relatively easily elicit, such as correlations rather than variance components. Finally, for each class we discuss design considerations, namely, strategies investigators can employ when planning a multilevel factorial design to make the experiment as informative and efficient as possible. We focus somewhat more on the hybrid-PEC factorial experiments than on the other four classes because the planning and analysis of hybrid-PEC factorial experiments has not yet been explored elsewhere to our knowledge.

For simplicity our discussion focuses on complete factorial designs. However, fractional factorial designs can also be employed with all five classes, and the modeling principles and power planning resources are relevant to both the complete

**Table 1** Summary of five classes of multilevel factorial designs

| | Type of clustering | | Randomization scheme | |
|---|---|---|---|---|
| | PECs or EIC? | Full or partial clustering? | Unit randomly assigned to experimental conditions | Procedures for randomizing individuals or clusters |
| Within PEC | PEC | Full | Individuals; clusters are complete blocks | Within-cluster assignment: one or more individuals within each cluster are randomly assigned to each condition |
| Between PEC | PEC | Full | Clusters | Between-cluster assignment: whole clusters are randomly assigned to conditions |
| Hybrid PEC | PEC | Full | Individuals; clusters are incomplete blocks | Hybrid between within- and between-cluster assignment. Individuals within each cluster are randomly assigned to only a subset of the experimental conditions (i.e., a packet) |
| Full EIC | EIC | Full | Clusters (after assigning individuals to clusters) | Individuals are either (a) assigned to clusters, which are then assigned to conditions, or (a) assigned to conditions and then to clusters within each condition |
| Partial EIC | EIC | Partial | Individuals or clusters, depending on the condition | Individuals are assigned to conditions and are assigned to clusters only within those conditions that involve clustering |

Notes: *PEC* pre-existing clusters, *EIC* experimentally induced clusters

and the fractional factorial case (see Dziak et al. 2012; Nahum-Shani, Dziak, et al. 2017). Further, fractional factorial designs play an important role in our discussion of hybrid-PEC factorials. Also for simplicity, we restrict the discussion to designs with only three factors of theoretical interest, which we label $X_1$, $X_2$, and $X_3$. Of course, the number of factors can be higher; see Collins, Dziak, and Li (2009) and the companion volume (Collins, 2018) for a detailed discussion of how this may affect sample size requirements. We assume the factors are dichotomous, with two levels[2] each, which are labeled on (experimental) and off (control) for convenience.

[2]Note that here, the term level refers to a value that can be taken on by an experimental factor.

**Table 2** Experimental conditions in the 2 × 2 × 2 factorial examples

| Experimental condition | $X_1$ | $X_2$ | $X_3$ | Example with blocked within-PEC | |
|---|---|---|---|---|---|
| | | | | Clinic Strategy I | Clinic Strategy II |
| 1 | Off (−1) | Off (−1) | Off (−1) | A | A |
| 2 | Off (−1) | Off (−1) | On (+1) | A | B |
| 3 | Off (−1) | On (+1) | Off (−1) | A | B |
| 4 | Off (−1) | On (+1) | On (+1) | A | A |
| 5 | On (+1) | Off (−1) | Off (−1) | B | B |
| 6 | On (+1) | Off (−1) | On (+1) | B | A |
| 7 | On (+1) | On (+1) | Off (−1) | B | A |
| 8 | On (+1) | On (+1) | On (+1) | B | B |

Notes: *PEC* pre-existing clusters

The number of levels per factor can be higher, but this would have large implications for sample size planning, which we do not explore here. Further, the levels of the dichotomous factors could instead be labeled as low and high or any other dichotomy, instead of on or off. A complete factorial design in this setting will result in $2 \times 2 \times 2 = 8$ experimental conditions (see Table 2). Throughout, we assume that the levels of the factors are effect-coded, where the off level is coded −1 and the on level is coded +1; see Kugler, Dziak, and Trail (2018) in the current volume for a detailed discussion of coding in factorial designs. We also assume that the outcome response $Y$ is normally distributed and that the effects of the factors on $Y$ are being modeled, using a pretest $P$ as a covariate.

## 2 Classes of Multilevel Factorial Designs

### *2.1 Within-PEC Factorials*

These factorial experiments involve clusters that exist *prior to experimentation* and a randomization scheme in which individuals are randomized to experimental conditions *within* each cluster. Hence, different members of the same cluster are independently assigned to different conditions (see Dziak et al. 2012 for published examples).

**Example** Suppose an investigator wishes to develop an intervention program to improve the emotional well-being of high-school students. Assume the outcome of interest, denoted $Y$, is some measure of emotional well-being, with higher values representing a more desirable outcome. The investigator would like to address three scientific questions, namely, whether the targeted outcome would be improved by including each of the following three components: (1) weekly videos that provide stress-coping training (Video), (2) access to an interactive website (Website), and (3) access to a mobile app that facilitates daily self-monitoring of mood and provides supportive messages (App). Hence, the three factors of theoretical interest are *VIDEO* ($X_1$), *WEBSITE* ($X_2$), and *APP* ($X_3$). To address the three questions noted

above, high-school students within each school are randomized to the resulting eight experimental conditions.

**Model** Assuming there are $j = 1, \ldots, J$ clusters, with $n_j$ individuals in the $j$th cluster, and denoting the response for individual $i$ within cluster $j$ as $Y_{ij}$, a model for a within-PEC factorial can take on the following form:

$$\begin{aligned} Y_{ij} = {} & \gamma_0 + \gamma_P P_{ij} + \\ & \gamma_1 X_{1ij} + \gamma_2 X_{2ij} + \gamma_3 X_{3ij} + \\ & \gamma_4 X_{1ij} X_{2ij} + \gamma_5 X_{1ij} X_{3ij} + \gamma_6 X_{2ij} X_{3ij} + \\ & \gamma_7 X_{1ij} X_{2ij} X_{3ij} + \\ & u_j + e_{ij}. \end{aligned} \tag{1}$$

Here the $e_{ij}$, which represent a combination of random individual variability and measurement error, are independent $N(0, \sigma^2)$; and the $u_j$, which represent the random cluster effects, are independent $N(0, \tau^2)$. The regression coefficient $\gamma_P$ represents the effect of pretest on posttest; this effect can be omitted if pretest is unavailable. $\gamma_1, \ldots, \gamma_K$ are the regression coefficients for the $K$ effects of interest, including main effects and interactions. Because we use effect coding, the *main effect* of each factor can be expressed as two times the coefficient representing the effect of this factor. For example, the average effect of offering Video (versus not offering Video) would be expressed by $2\gamma_1$, where the multiplier 2 comes from the difference between effect codes for the on and off levels, that is, $(+1)-(-1) = 2$. Notice that these coefficients represent effects at the individual level. For example, $\gamma_1$ represents one-half the expected difference between *individuals* assigned to the on level of Video and those assigned to the off level of Video.

If the investigator wishes to assume that one or more interactions are negligible, the term representing that particular interaction can be omitted. For simplicity, similar to Dziak and colleagues (2012), Model (1) assumes no cluster-by-treatment interactions. This assumption could be relaxed by including cluster-by-treatment interaction terms; for details, see Raudenbush and Liu's (2000) description for a single-factor design.

**Power** For all models provided in this chapter, the main effects and interactions of interest can be tested using significance tests for the corresponding regression coefficients (e.g., $\gamma_1$ for the main effect of *VIDEO* ($X_1$) and $\gamma_4$ for the interaction of *VIDEO* ($X_1$) and *WEBSITE* ($X_2$)). The power for such a significance test can be estimated as the probability that a variable with a noncentral $F$ distribution with noncentrality parameter $\lambda$ exceeds the critical value for the test (see Appendix 1 for more details). In general, $\lambda$ equals $\gamma^2/\mathrm{Var}(\widehat{\gamma})$, where $\gamma$ is the true value of the coefficient being tested and $\mathrm{Var}(\widehat{\gamma})$ is the sampling variance (squared standard error) of its maximum likelihood estimate for samples of the size proposed.

The appropriate variances for a within-PEC factorial were derived in Dziak et al. (2012), and the resulting noncentrality parameter formula is shown in Table 3. As

**Table 3** Noncentrality parameters and degrees of freedom for calculating power for different designs, in terms of the regression coefficient $\gamma$ for a main effect or interaction of interest

| Factorial design | Model | Noncentrality parameter λ | Denominator degrees of freedom υ |
|---|---|---|---|
| 1. Independent | | $\frac{N}{(1-\rho_{\text{pre,post}})}\frac{\gamma^2}{\sigma_Y^2}$ | $N-p$ |
| 2. Within PEC | (1) | $\frac{N}{(1-\rho_{\text{pre,post}})}\frac{\gamma^2}{\sigma_Y^2}$ | In theory $N-p$, but $(N-J)-p$ might be more accurate |
| 3. Between PEC | (2) | $\frac{N}{2(1-\rho_{\text{pre,post}})\left(1+n\frac{\rho_{\text{change}}}{1-\rho_{\text{change}}}\right)(1-\rho_{\text{post}})}\frac{\gamma^2}{\sigma_Y^2}$ | $J-p$ |
| 4. Hybrid PEC | (1) | $\frac{N}{(1-\rho_{\text{pre,post}})}\frac{\gamma^2}{\sigma_Y^2}$ | $N-p$ |
| 5. Full EIC | (2) | $\frac{N}{2(1-\rho_{\text{pre,post}})+n\frac{\rho_{\text{post}}}{1-\rho_{\text{post}}}}\frac{\gamma^2}{\sigma_Y^2}$ | $J-p$ |

(continued)

**Table 3** (continued)

| Factorial design | Model | Noncentrality parameter λ | Denominator degrees of freedom υ |
|---|---|---|---|
| 6. Partial EIC | (3) | $\frac{1}{\frac{\rho_{\text{post}}}{4J_1\left(1-\rho_{\text{post}}\right)}+\frac{1-\rho^2_{\text{pre,post}}}{4J_1 n}+\frac{1-\rho^2_{\text{pre,post}}}{4J_0}}\frac{\gamma^2}{\sigma^2_Y}$ $=\frac{N}{2n\frac{\rho_{\text{post}}}{\left(1-\rho_{\text{post}}\right)}+\frac{1}{2}\frac{\left(1-\rho^2_{\text{pre,post}}\right)}{2}+\frac{1}{2}\frac{1-\rho^2_{\text{pre,post}}}{2}}\frac{\gamma^2}{\sigma^2_Y}$ under equal allocation of subjects to clustered and unclustered conditions (so that $J_1 = N/2n$, $J_0 = N/2$) | Conservative estimate: $J_1 - p$. When analyzing data, use a Satterthwaite approximation for degrees of freedom |

Notes: *PEC* pre-existing clusters, *EIC* experimentally induced clusters. λ represents the overall amount of evidence against $H_0$. υ represents the number of independent sources by which this evidence is provided, $\gamma$ is the true regression coefficient, $\sigma^2_Y$ is the overall posttest outcome variance adjusting for treatment but not for cluster (it is the variance in the posttest response that is not explained by treatment effect, this expression enables to easily re-express the formulas in terms of a standardized effect size as we do below), $N$ is the total sample size, $n$ is the cluster size, and $p$ is the number of regression coefficients. $\rho_{\text{pre, post}}$ is the within-person correlation between pretest response and posttest response, and $\rho_{\text{post}}$ is the posttest ICC. To generalize any of these formulas to a situation without posttest, set $\rho_{\text{pre, post}} = 0$. For designs 2 through 5, $J$ is the number of clusters

In Design 3, $\rho_{\text{change}}$ is the within-cluster ICC for change scores. If no estimate for $\rho_{\text{change}}$ is available, it may be reasonable to assume that $\rho_{\text{change}} = \frac{1}{2}\rho_{\text{post}}$ as in the simulation section of Dziak et al. (2012). It is assumed that the pretest ICC and the posttest ICC are equal, so both are designated by $\rho_{\text{post}}$. Further, in Design 3, the cluster size may vary widely because clusters are pre-existing; this can be accounted for by replacing the $n$ in the denominator by $\left(1 + CV_n^2\right)\overline{n}$, where $\overline{n}$ is the expected average cluster size and where $CV_n$ is the coefficient of variability of cluster sizes (see Dziak et al. 2012). In Design 6, $J_1$ is the number of clusters, and $J_0$ is the number of unclustered individuals. The formula for Design 6 assumes equal error variances between clustered and unclustered conditions, but this assumption can be relaxed, as described in detail by Nahum-Shani, Dziak, et al. (2017)

shown, the formula for λ in the case of within-PEC factorials is similar to the standard formula used in the case of independent randomization (i.e., no clustering). Specifically, λ will increase to the extent that the squared true effect coefficient as a ratio of the total error variance ($\gamma^2/\sigma_Y^2$) increases. For convenience, in Table 4 we also re-expressed the formulas in terms of $d = 2\gamma/\sigma_Y$ to clarify the connection to Cohen's (1988) standardized effect size for the main effect of a given component. Here, $d$ is the ratio of the expected mean difference in the response (adjusted for pretest) between the two levels of a given factor to the overall within-condition standard deviation ($\sigma_Y$). Hence, as in the case of independent randomizations, power will increase to the extent that the standardized effect size is larger.

The noncentrality parameter λ will also increase to the extent that the within-person correlation between pretest response and posttest response ($\rho_{\text{pre, post}}$) and the total sample size ($N$) are larger. This is illustrated in Fig. 1, which provides the expected power (based on the formula for Design 2 in Table 3) for varying levels of $\rho_{\text{pre, post}}$ and sample size. This figure emphasizes the added value of having a relatively large $\rho_{\text{pre, post}}$. For example, with $\rho_{\text{pre, post}} = 0.75$ , a researcher may achieve adequate power (0.80) even with relatively small sample size. Of course, this feature is not unique to within-PEC factorials or to multilevel factorials more generally. As discussed by Dziak et al. (2012), if the probabilities of assignment to experimental conditions in the within-PEC factorial case are the same across all clusters, and if there are no cluster-by-treatment interactions, then the implications of clustering on power for within-cluster randomized designs are negligible. Hence, for power planning purposes, the data can be conceptualized as having only one level rather than two, namely, as if the individuals were independent.

The critical value for the $F$-test is determined by the alpha level of the test, as well as the degrees of freedom, which are 1 for the numerator, representing one coefficient that is being tested and some value υ for the denominator. For example, the critical value is about 3.90 for a test with an alpha level of 0.05 and with 150 denominator degrees of freedom. The denominator degrees of freedom υ depends on the design and the sample size. As can be seen in Table 1, for a within-PEC factorial, υ relies heavily on the total number of individuals rather than the number of clusters. Software packages such as SAS or R can quickly compute the power given λ and υ; sample code for doing this is provided in Appendix 2.

Throughout, our power formulas assume no random slopes for cluster – that is, no cluster-by-treatment interactions. For example, we assume that Video does not work better for some schools than for others. If this assumption does not hold, the formulas here may be too optimistic. Formulas for a single-factor within-PEC experiment with cluster-by-treatment interactions are given by Raudenbush and Liu (2000). Further methodological research on cluster-by-treatment interactions in various factorial designs may be warranted.

**Design Considerations** When considering a within-PEC factorial, investigators should take into account the feasibility of assigning different individuals within the same cluster to different conditions, as well as the risk for contamination. Here, risk for contamination refers to the chances that an intervention component intended for

**Table 4** Noncentrality parameters and degrees of freedom for calculating power for different designs, in terms of the Cohen effect size $d$ for a main effect of interest

| Factorial design | Model | In terms of $d = 2\gamma/\sigma_Y$ | Denominator degrees of freedom $\upsilon$ |
|---|---|---|---|
| 1. Independent | | $\frac{Nd^2}{4\left(1-\rho_{\text{pre,post}}\right)}$ | $N-p$ |
| 2. Within PEC | (1) | $\frac{Nd^2}{4\left(1-\rho_{\text{pre,post}}\right)}$ | In theory $N-p$, but $(N-J)-p$ might be more accurate |
| 3. Between PEC | (2) | $\frac{Nd^2}{8\left(1-\rho_{\text{pre,post}}\right)\left(1+n\frac{\rho_{\text{change}}}{1-\rho_{\text{change}}}\right)\left(1-\rho_{\text{post}}\right)}$ | $J-p$ |
| 4. Hybrid PEC | (1) | $\frac{Nd^2}{4\left(1-\rho_{\text{pre,post}}\right)}$ | $N-p$ |
| 5. Full EIC | (2) | $\frac{Nd^2}{8\left(1-\rho_{\text{pre,post}}\right)+4n\frac{\rho_{\text{post}}}{1-\rho_{\text{post}}}}$ | $J-p$ |
| 6. Partial EIC | (3) | $\frac{d^2}{\frac{\rho_{\text{post}}}{J_1\left(1-\rho_{\text{post}}\right)}+\frac{1-\rho^2_{\text{pre,post}}}{J_1 n}+\frac{1-\rho^2_{\text{pre,post}}}{J_0}} = \frac{Nd^2}{8n\frac{\rho_{\text{post}}}{\left(1-\rho_{\text{post}}\right)}+\left(1-\rho^2_{\text{pre,post}}\right)+1-\rho^2_{\text{pre,post}}}$ under equal allocation of subjects to clustered and unclustered conditions (so that $J_1 = N/2n$, $J_0 = N/2$) | Conservative estimate: $J_1 - p$. When analyzing data, use a Satterthwaite approximation for degrees of freedom |

Notes: *PEC* pre-existing clusters, *EIC* experimentally induced clusters. The notation and assumptions used in this table are analogous to those in Table 3. Here, the noncentrality parameters are re-expressed in terms of $d = 2\gamma/\sigma_Y$, namely, the standardized effect size of a main effect of interest

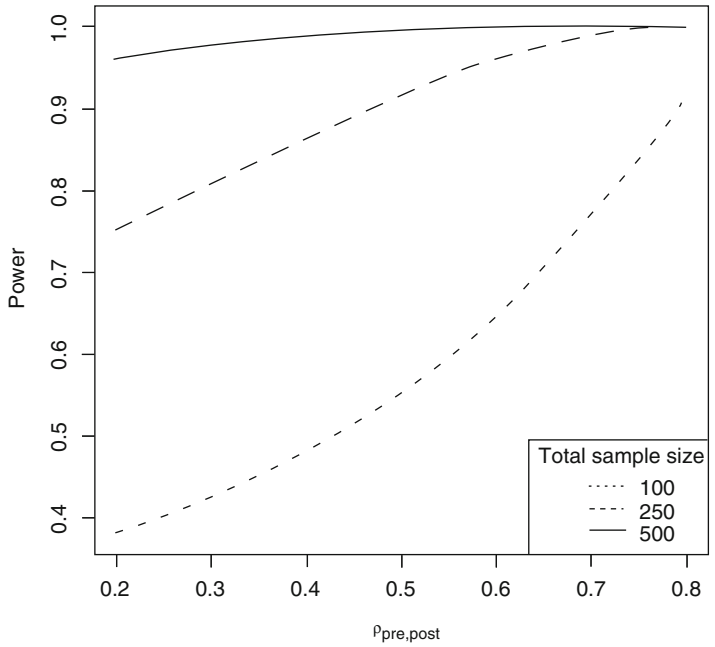


**Fig. 1** Expected power for within-PEC factorials. (Notes: *PEC* pre-existing clusters. This figure shows the calculated power for a test of a main effect or interaction in Model (1) for a within-PEC factorial design. This means we assume a linear model with normally distributed responses, no cluster-by-treatment interaction, and a fitted model that contains three dichotomous factors and all interactions included (hence 9 regression coefficients including the pretest). Note that the number of parameters has only a modest effect on power for a given parameter. Balance is also assumed, namely, that each cluster has the same number of individuals in each condition, although this may not be exactly possible in practice. To illustrate the link between power and the within-person correlation between pretest and posttest responses ($\rho_{\text{pre, post}}$), it is assumed that there are 10 clusters with 10, 25, or 50 members each and no dropout and that the true value of the standardized regression coefficient is .15 (hence a Cohen's *d* of 0.3 if a main effect is being tested))

individuals in one experimental condition will be received by individuals in another experimental condition to which this intervention component was not intended to be provided. As we discuss below, if assigning different individuals within the same cluster to different experimental conditions is not feasible for logical, practical, financial, or ethical reasons, such as in a setting where the risk for contamination is high, a between-PEC approach (described below) should be considered.

The random assignment of individuals to experimental conditions in a within-PEC factorial can be done in several ways. The simplest way is to assign each individual independently without regard for their cluster membership. While rather straightforward, this approach entails a risk that some clusters might be highly unbalanced on one or more components. The most extreme case of imbalance would arise if a cluster is missing a level of a component. As an example, consider a case where all of the students in a particular cluster receive the off level of

Video. This particular cluster would provide no information at all about the effect of that component. This potential problem can be resolved by treating each PEC as a "block" for randomization purposes: that is, to conduct the randomization in a restricted way so that each cluster (i.e., "block") contains at least one member assigned to each experimental condition. In other words, each cluster is a replication of the entire experiment. In the context of our example, each school should ideally have at least one student assigned to each of the eight experimental conditions resulting from the $2 \times 2 \times 2$ factorial design, and the number of students from each cluster in each condition should be as close to equal as possible.

## *2.2 Between-PEC Factorials*

We now turn our attention to between-PEC factorials. Like within-PEC factorials, these factorial experiments involve clusters that exist prior to experimentation. However, between-PEC factorials use a randomization scheme in which clusters, rather than individuals, are randomized to experimental conditions (see Dziak et al. 2012 for published examples).

**Example** Suppose again that an investigator wishes to develop an intervention program to improve the emotional well-being of high-school students. Although the outcome and three factors of theoretical interest remain the same, assume that now the investigator is concerned about contamination. For example, suppose the investigator has reasons to expect that students may share the videos with other students in their school. In this case, randomizing students within a school to the eight experimental conditions entails high risk that conditions involving the off level of Video will be contaminated.

Besides contamination, there are other logical, practical, financial, and ethical reasons to consider randomizing clusters rather than individuals within a cluster. For example, the nature of one or more intervention components may require administering them at the cluster level. This is the case in curriculum reforms or educational programs that are designed to be implemented at the school level. This is also the case when intervention components are designed to affect the individual by changing social processes at the cluster level, such as interventions that are designed to facilitate prosocial and pro-academic school climates, where school climate is defined as "the shared beliefs, values, and attitudes that shape interactions between the students, teachers, and administrators" (Mitchell, Bradshaw, & Leaf, 2010, p. 23). Another example concerns intervention components that involve training professionals (e.g., healthcare providers, teachers) who support or treat many other individuals (e.g., patients, students). This makes individual-level randomization infeasible, as it is usually impossible to train professionals for treating one participant and then untrain them for the next. In some cases, randomizing clusters is preferable for ethical reasons. This includes settings where it is considered unethical (e.g., from the point of view of fairness and equity) to offer a

particular component to some individuals and not to others within the same cluster (Moerbeek & Teerenstra, 2015). The rationale for electing to randomize clusters rather than individuals has been discussed in detail in literature on single-factor between-PEC trials (e.g., Donner & Klar, 2000; Eldridge, Ashby, Feder, Rudnicka, & Ukoumunne, 2004; Murray, 1998; Taljaard et al. 2009; Weijer et al. 2011).

**Model** Assuming, as before, that there are $j = 1, \ldots, J$ clusters, with $n_j$ individuals in the $j$th cluster, and denoting the response for individual $i$ within cluster $j$ as $Y_{ij}$, a model for a between-PEC factorial can take on the following form:

$$\begin{aligned} Y_{ij} = {} & \gamma_0 + \gamma_P P_{ij} + \\ & \gamma_1 X_{1j} + \gamma_2 X_{2j} + \gamma_3 X_{3j} + \\ & \gamma_4 X_{1j} X_{2j} + \gamma_5 X_{1j} X_{3j} + \gamma_6 X_{2j} X_{3j} + \\ & \gamma_7 X_{1j} X_{2j} X_{3j} + \\ & u_j + e_{ij} \end{aligned} \tag{2}$$

where, as before, the $e_{ij}$ and the $u_j$ are independent $N(0, \sigma^2)$ and $N(0, \tau^2)$, respectively. Notice that the only difference between Models (2) and (1) is the subscript for the $X$ variables, which are written here as $X_{kj}$ rather than $X_{kij}$ because factors in Model (2) cannot vary within a cluster. This has implications for the interpretation of the regression coefficients $\gamma_1, \ldots, \gamma_K$ (as well as for the power formula, as we explain below). In Model (1), $\gamma_1, \ldots, \gamma_K$ represent individual-level effects, but in Model (2) they represent cluster-level effects. For example, $\gamma_1$ represents (half) the expected difference between *clusters* assigned to the on level of *VIDEO* and those assigned to the off level of *VIDEO*.

Note that in Dziak et al. (2012), the subscripts for the regression coefficients also differed between the model for within-PEC and the model for between-PEC. This is because the authors used formal multilevel notation, where the subscripts of the regression coefficients represent levels of nesting. For example, a coefficient labeled here as $\gamma_k$ was labeled there as $\gamma_{k0}$ for within-PEC factorials and $\gamma_{0k}$ for between-PEC factorials, to emphasize that it represented an individual-level contrast in the within-PEC case and a cluster-level contrast for the between-PEC case. Here, we elect to use the same subscripts for the regression coefficients in Models (1) and (2) because algebraically Model (2) is merely a special case of (1) where all $X_{ij}$ within a cluster are set to the same value $X_j$. Hence, the same mixed model can be applied in both cases.

**Power** Based on the work of Dziak et al. (2012), Table 3 presents a simplified formula for calculating the noncentrality parameter $\lambda$, as well as the denominator degrees of freedom $\upsilon$, when planning power for a between-PEC factorial. Note that although Model (2) is suitable for analyzing data from a between-PEC factorial, the power formula in Table 3 is based on a more elaborate repeated measures model. This is because planning power for between-cluster designs requires specifying how the cluster effect on the pretest relates to the cluster effect on the posttest (see Murray, 1998 for more details). The parameters of Model (2) cannot be used

to specify this because the pretest response is treated as a covariate. A repeated measures model, on the other hand, treats the pretest response as a separate time point in a multilevel repeated measures design. Hence it enables the researcher to specify how cluster effect on pretest response and cluster effect on posttest response are related (see Dziak et al. 2012 for a more detailed explanation). The use of a repeated measures model, rather than the pretest-adjusted model, to guide the development of the power formula explains why the power formula in Table 3 for between-PEC factorials seems different from other formulas in that table. For example, unlike the other formulas in Table 3, the formula for between-PEC factorials involves the within-cluster ICC of the change score ($\rho_{\text{change}}$), which represents the extent to which members of a given cluster are similar in how their response has changed from pretest to posttest.

Similar to the within-PEC scenario, the noncentrality parameter λ in the formula for between-PEC factorials will increase to the extent that the within-person correlation between pretest and posttest response ($\rho_{\text{pre, post}}$), the total sample size ($N$), and the standardized effect size ($d$) increase. However, in the between-PEC scenario, λ is also a function of the within-cluster ICC for change scores ($\rho_{\text{change}}$) and the posttest ICC ($\rho_{\text{post}}$), which represents the extent to which members of a given cluster are similar in their posttest response. Note that for simplicity, this formula is based on the assumption that the posttest ICC is the same as the pretest ICC.

Figure 2a provides the expected power (based on the formula for Design 3 in Table 3) for a varying number of clusters, as well as varying values of $\rho_{\text{post}}$ and $\rho_{\text{change}}$. This figure shows that more clusters will be needed to achieve adequate power when the within-cluster ICC for change scores is large than when it is small. Interestingly, given a fixed within-cluster ICC for change scores, the number of clusters needed to achieve an adequate level of power is only slightly affected by the posttest ICC. One intuition for why the within-cluster ICC of the change score might be more important for power than the ICC of the posttest is that some of the variance in posttest response is shared with the pretest, and this variance is removed when taking into account the change in response from pretest to posttest (see Dziak et al. 2012).

Also, notice that in between-PEC factorials, the denominator degrees of freedom (υ) are calculated based on the number of clusters ($J$) rather than the total number of individuals ($N$). This is one reason why power in between-PEC experiments is more heavily influenced by the total number of clusters than by the number of individuals within a cluster ($n$). This feature is illustrated in Fig. 2b, which provides the expected power (based on the formula for Design 3 in Table 3) for varying values of $N$, $J$, and $n$.

**Design Considerations** In between-PEC factorials, all members of a given cluster are assigned to the same experimental condition, so that clusters are nested within conditions. To preserve the balance property of a factorial design, namely, to ensure that each level of each factor contains about the same number of clusters, it is recommended that clusters be assigned to conditions in a restricted manner. For example, if there are 20 schools in the hypothetical scenario described above, the investigator could first randomly assign schools in a way that ensures that there are

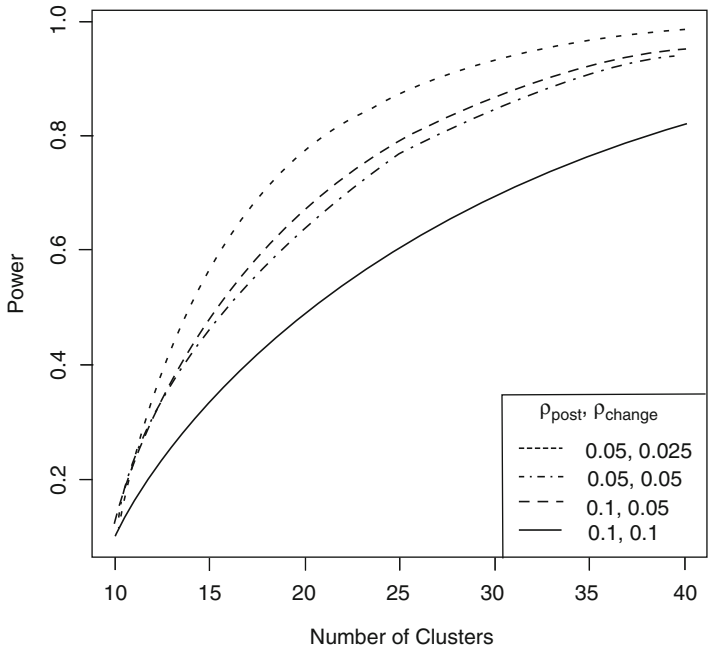


**Fig. 2** (**a**) Expected power for between-PEC factorials: fixed cluster size, varying ICCs. (Notes: *PEC* pre-existing clusters. This figure shows the calculated power for a test of a main effect or interaction in Model (2) for a between-PEC multilevel factorial design. Model (2) assumes normally distributed responses and no cluster-by-treatment interaction. Here we further assume that the fitted model has three dichotomous factors and all interactions included (hence 9 regression coefficients including the pretest), although the number of parameters has only a modest effect on power for a given parameter. Balance is also assumed, namely, that each condition receives the same number of clusters and each cluster has the same number of individuals, although this may not be exactly possible in practice. To illustrate the link between the number of clusters and power, it is assumed that the within-person correlation between pretest and posttest response ($\rho_{\text{pre, post}}$) is .65 and that the true value of the standardized regression coefficient is .15 (hence a Cohen's *d* of 0.3 if a main effect is being tested). The cluster size is set to 20, and various possible posttest ICC ($\rho_{\text{post}}$) and change ICC ($\rho_{\text{change}}$) are being compared).

two schools in each of the eight experimental conditions and then randomize the remaining four schools to any condition in a way that ensures that no condition receives more than one of them. In this example, each condition will contain either two or three clusters, so that conditions will not differ in size by more than one cluster.

In some situations, there may be important cluster-level demographic covariates (such as whether a school is public or private, urban or rural, or has received a particular kind of intervention before). Because the number of clusters is often modest, the distribution of such a covariate may easily be somewhat imbalanced between treatment levels on an assigned factor, even though the assignment is

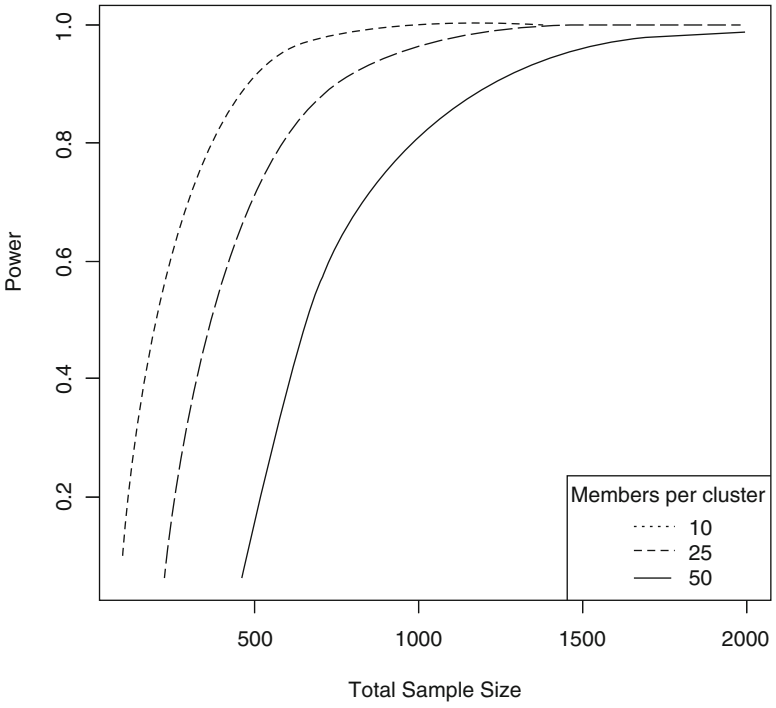


**Fig. 2** (continued) (**b**) Expected power for between-PEC factorials: fixed ICCs, varying cluster size. (Notes: *PEC* pre-existing clusters. This figure shows the calculated power for a test of a main effect or interaction in Model (2) for a between-PEC multilevel factorial design. Model (2) assumes normally distributed responses and no cluster-by-treatment interaction. Here we further assume that the fitted model has three dichotomous factors and all interactions included (hence 9 regression coefficients including the pretest), although the number of parameters has only a modest effect on power for a given parameter. Balance is also assumed, namely, that each condition receives the same number of clusters and each cluster has the same number of individuals, although this may not be exactly possible in practice. To illustrate the link between the total sample size, number of clusters, number of members per cluster, and power, it is assumed that the within-person correlation between pretest and posttest response ($\rho_{\text{pre, post}}$) is .65 and that the true value of the standardized regression coefficient is .15 (hence a Cohen's *d* of 0.3 if a main effect is being tested). The posttest ICC ($\rho_{\text{post}}$) is set to .10, the change-score ICC ($\rho_{\text{change}}$) is set to .05, and various possible cluster sizes are being compared for every given total sample size)

random. That is, the covariate may have a non-negligible sample correlation with a factor, even though the factors are randomized independently. This may not always be avoidable, especially if there are many covariates of interest. However, if there is severe imbalance on a covariate of theoretical interest, then the randomization should probably be repeated. Obviously, this must be done before administering treatment, so it would be important for these cluster-level covariates to be available to the person performing the randomization. A more advanced statistical way of dealing with the risk of poor randomizations in experiments is described by Morgan and Rubin (2015). However, because they described their method primarily in

the case of a single-factor experiment with unclustered participants, some further methodological work is needed before their method would be convenient for between-PEC factorial experiments.

## *2.3 Hybrid-PEC Factorials*

The discussion above of factorial experiments with PECs focuses on scenarios in which the randomization scheme involves randomizing either individuals (within-PEC) or clusters (between-PEC) to experimental conditions. However, in some factorial experiments with PEC, the randomization scheme lies somewhere between these two approaches. Specifically, these PEC factorials involve a randomization scheme in which individuals within a cluster are randomized to only a subset of the experimental conditions. Here, different members of the same cluster are independently assigned to different conditions, but these conditions might differ from those conditions to which members of another cluster may be assigned. As noted earlier, this class of multilevel factorials can be conceptualized as a hybrid between a within-PEC factorial, in which the members of a given cluster can receive any of the eight conditions, and a between-PEC factorial, in which the members of a given cluster must all receive the same condition. Here, a member of a given cluster has only a subset of possible experimental conditions to which s/he can be assigned, with the specific subset determined at the cluster level. The unit of assignment is still the individual, not the cluster, but the possible conditions are limited for individuals within any given cluster.

**Example** Suppose that now the investigator wishes to develop an intervention for improving the coping skills of adolescents with mental illness in primary care practices. Unlike the previous examples, the goal here is to build an intervention program that will be delivered or managed *by local clinic staff*. Assume that now the outcome of interest, $Y$, is a reverse-coded measure of symptom severity, so that as before, higher values represent a more desirable outcome. As before, three components are of theoretical interest; however, now these components are (1) in-person biweekly coping skills training (In-Person), (2) brief phone coaching sessions that focus on addressing barriers to implementing adaptive coping strategies (Phone), and (3) weekly email contact with parent(s) focusing on ways parents could be most helpful with their adolescents' coping skills training (Parent). Hence, the corresponding factors are *IN-PERSON* ($X_1$), *PHONE* ($X_2$), and *PARENT* ($X_3$).

In this hypothetical scenario, local staff at each clinic must be trained and supervised by study staff to implement each experimental condition. Suppose that local staff at each clinic have the capacity (in terms of time and available resources) to implement at most four experimental conditions simultaneously with a reasonable degree of fidelity to the protocol. Further, suppose that resources are not available to retrain the local staff so that they can implement four conditions first and then implement the other four. Hence, the investigator decides to divide the

eight experimental conditions (perhaps randomly) into two "packets," Packet A and Packet B; each packet includes four experimental conditions, and each clinic will be assigned only one packet, either A or B. In other words, each clinic will conduct four and only four of the eight conditions.

**Model** Although in this scenario the possible conditions to which an individual can be randomized are limited by cluster membership, the design is still a within-PEC factorial, and Model (1) can still be used to analyze the resulting data. It might seem as though the packets of experimental conditions to which members of a given cluster can be randomized ought to be considered as an additional level of nesting that needs to be included in the model. However, the selected packets do not generate dependency in the responses of individuals in the same manner that cluster membership does. Clusters assigned to a given packet are similar only in that their members have access to the same subsets of conditions; hence, *conditional on the experimental condition*, members of the different clusters are not necessarily statistically dependent simply because their clusters receive the same packet. As such, the packets do not introduce another source of variation beyond that generated by the experimental conditions. Rather, the packets are merely another form of restricted randomization, similar to the restrictions used to guarantee that the members in a cluster (for an ordinary within-PEC design) or the clusters in the sample (for a between-PEC design) are allocated in a roughly balanced way. As in the case of a standard within-PEC factorial, it is conceivable that the presence of one treatment component in a cluster might affect the outcomes of even those cluster members who, by design, are not supposed to receive the component; however, this would be a case of contamination, and if it is considered likely to occur, then a between-PEC design might be more suitable. In classic experimental design terminology, a within-PEC experiment without packets uses "complete blocks," and a within-PEC experiment with packets uses "incomplete blocks" (see, e.g., Kuehl, 2000).

In summary, it is reasonable to analyze data resulting from a hybrid-PEC factorial by using the same model proposed above (Model 1) for standard within-PEC factorials. First, as discussed above, in a hybrid setting, it is not necessary to conceptualize and model the packets as another level of analysis. Second, as we explain in detail below under design considerations, when the possible conditions to which an individual can be randomized are limited by cluster membership, it is sometimes necessary to assume that certain cluster-by-treatment interactions are negligible. This assumption is compatible with Model (1), which already assumes no cluster-by-treatment interactions.

**Power** Assuming no cluster-by-treatment interactions, and assuming the number of conditions within each packet is large enough to ensure balance across clusters for main effects, power calculations for main effects can reasonably be done in the same way as for an ordinary within-PEC design. That is, the same formulas for calculating the noncentrality parameter $\lambda$ and the denominator degrees of freedom $\upsilon$ can be employed for within-PEC factorials, whether or not the possible conditions

to which individuals can be randomized are limited to packets determined by cluster membership. However, as described below, power calculations for interactions can be more complicated in the hybrid-PEC case, because it is possible that not all of the clusters will provide as much relevant information for estimating a given interaction.

**Design Considerations** As described earlier, in the case of a hybrid-PEC factorial, each cluster is assigned to only a subset ("packet") of conditions. Here, the randomization involves two stages: first, randomly determine which clusters get each of the packets and second, randomly determine which cluster members get which experimental condition from the packet available to them. The packets themselves should be constructed prior to the first stage in a careful and systematic way so as to make the overall experiment as informative and efficient as possible.

To clarify this, assume there are only two clinics in the example above, labeled clinic A and clinic B. Thus, there must also be two packets, one to be assigned to each clinic. Note that this assumption is made for illustrative purposes alone and does not imply that a design with only two clusters is recommended. One arbitrary way to assign the conditions would be to assign conditions (1) through (4) as a packet to clinic A and conditions (5) through (8) as a packet to clinic B. This is shown as Strategy I in Table 2. This strategy has a serious limitation: patients in clinic A are only assigned to conditions in which the In-Person component ($X_1$) is set to off, whereas patients in clinic B are only assigned to conditions in which In-Person is set to on. Thus, the effect of cluster membership is *aliased* (confounded by design), with the main effect of In-Person. To make this concept more concrete, assume the investigator finds that the difference in mean outcome $Y$ between individuals in conditions 5 through 8 and individuals in conditions 1 through 4 is two units on the scale of measurement being used, in favor of individuals in conditions 1 through 4. Suppose that this is large enough to be statistically and practically significant. What should the investigator conclude based on this information? Unfortunately, the investigator cannot make any definitive conclusions. It is unclear whether the difference of two units indicates that providing In-Person sessions (the on level of $X_1$) is on average more effective than not providing In-Person sessions (the off level of $X_1$) or that members of clinic B do better on average compared to members of clinic A for some other reason unrelated to the treatment factors. Thus, the main effect of $X_1$ and the effect of clinic membership are confounded – they cannot be distinguished when using Strategy I.

The main problem with Strategy I is not the fact that it involves aliasing per se but rather the nature of the aliasing employed. Specifically, Strategy I aliases the main effect of $X_1$ with cluster membership. The underlying assumption is that cluster membership has a negligible effect and hence can be confounded with another effect that is of considerable theoretical interest. However, in many cases this assumption might be unreasonable. For example, for some reasons intrinsic to the clinic, such as more qualified staff, better facilities, or a more supportive management, members of one clinic might do better on average than members of another. This would seriously bias the estimate of the main effect if Strategy I were to be used. Fortunately, much more informative strategies can be found by using ideas from the fractional

factorial experimental design literature, which focuses on making experiments more practically feasible yet still informative.

Conceptualizing cluster membership as a factor in a factorial design makes it easy to see that using cluster membership to limit the possible conditions to which an individual can be randomized can cause cluster membership to be confounded with some other effects. Specifically, in our hypothetical example, conceptualizing cluster membership as another factor in the design, a full factorial would include $2^4$ experimental conditions (because there are four factors with two levels each, three of them are the treatment factors and the remaining one is the cluster membership factor). However, the design implemented in Strategy I is a $2^{4-1}$ fractional factorial (see Collins (2018) companion volume), including half of the conditions a full factorial would include in our hypothetical scenario. Selecting only a subset of the conditions, a full factorial would include aliasing (i.e., confounding) – the inability to estimate certain effects separately. However, just like in a fractional factorial experiment, it would also be possible to make a hybrid-PEC factorial more informative by aliasing cluster membership with a high-level interaction that is of little theoretical interest and is assumed to be negligible, rather than with a main effect. To clarify this, consider Strategy II in Table 2, where the first packet includes conditions 1, 4, 6, and 7; and the second packet includes conditions 2, 3, 5, and 8. This strategy was selected using an experimental design procedure in SAS called PROC FACTEX; details are provided in the appendix. In this carefully constructed strategy, none of the main effects of theoretical interest, and none of the two-way interactions, are confounded with clinic membership. Instead, the main effects are only confounded with three-way interactions involving two assigned factors and the cluster membership factor (e.g., $X_1 \times X_2 \times CLINIC$). Additionally, two-way interactions between each pairs of factors are aliased with an interaction between clinic membership and the remaining third factor. For example, the two-way interaction between *IN-PERSON* ($X_1$) and *PHONE* ($X_2$) is aliased with the interaction between *PARENT* ($X_3$) and clinic membership. Finally, the three-way interaction between all three factors is aliased with the main effect of clinic membership itself.

Overall, Strategy II allows the cluster random effect (namely, the effect of the clinic), as well as the fixed main effect of each of the three factors, to be estimated subject to more reasonable assumptions. For example, to estimate the main effects of the three factors without bias, it is only necessary for three-way interactions involving two assigned factors and the cluster membership factor (e.g., $X_1 \times X_2 \times CLINIC$) to be negligible. It is often more reasonable to assume that higher-order interactions, such as $X_1 \times X_2 \times CLINIC$, are negligible than to assume that lower-order effects, such as the effect of the clinic itself, are negligible (see Wu & Hamada, 2000). Hence, Strategy II would offer more useful and interpretable information compared to Strategy I, because of its more reasonable aliasing structure.

In the discussion above, we assume that there are only two clinics available. However, there will often be more than two clusters available for experimentation. This provides additional design options, which, if used wisely, will allow the

experimenter to estimate the effects of scientific interest with less aliasing, more power for estimated effects, or both.

For convenience, suppose that there are ten clusters. One approach would be to replicate Strategy II five times, namely, to use the same two packets, but assign five clinics to each packet, instead of one to each packet. Recall that doing so does not create another level of nesting in the model. Precisely which five clusters get the first packet and which get the second packet should be random, but the packets are the same as those that were found to be optimal in the two-cluster case; that is, the two packets each include the same subset of conditions as the corresponding packet in the two-cluster example. Of course, this randomization should be restricted so that five clusters receive each packet. Although it would be possible to assign seven clusters to one packet and three clusters to the other, this would not provide as much statistical power and precision because the sample sizes receiving different conditions would be highly unbalanced. If there are a total of eleven clinics instead of ten, the assignment would have to be slightly imbalanced, whereby six clinics would be assigned to one packet and five to the other, but this imbalance would be much less severe. That is, all else being equal, it is best to assign clusters to packets in a manner that provides as much balance as possible for the comparisons of primary theoretical interest.

However, simply repeating Strategy II several times has a remaining limitation: it does very little to resolve the confounding between cluster membership and the three-way interaction. Because there are ten instead of two clusters, there is now some information for distinguishing the two formerly aliased effects but only a very small amount. Within any given cluster, there is still no information for distinguishing the two effects, so inference for this interaction can only be done at the between-clusters level, for which there are very few degrees of freedom. Recall that ten clusters is a very small sample size for a between-clusters comparison, although it may be ample for within-clusters comparisons. In practice, the three-way interaction can now be tested but will have very low power; this is called *partial confounding*.

Instead of repeating Strategy II five times, another strategy is to sample randomly among all possible packets. For example, there are $\binom{8}{4} = 70$ possible packets of four conditions that could be assigned to any given clinic, so the investigator would randomly assign each clinic to one of the 70 packets. Alternatively, to ensure balance, the investigator might randomly choose the packets for half of the clinics and then assign the complementary packets to the other half. For example, if one clinic received cells 3, 4, 5, and 8, then another should receive 1, 2, 6, and 7. Because it explores a more diverse range of combinations of experimental conditions, a random strategy may have less aliasing than a repetitive strategy.

Although sampling randomly among all possible packets might perform very well by chance, it also might perform very poorly by chance. It can be shown that among the 70 possible packets of 4 experimental conditions, only 20 are balanced across all 3 factors (i.e., for each of the factors, the number of conditions in which the factor is set to on equals the number of conditions in which it is set to off) and

only 2 out of them – specifically, the packets mentioned in Strategy II – are balanced across the 3 factors and their 2-way interactions. Therefore, if it is reasonable to ignore three-way interactions, then simply replicating Strategy II for as many pairs of clusters as are available is likely to offer more efficiency than a more complicated but poorly chosen approach. Especially if there are relatively few clusters, it is quite possible that a random arrangement would cause aliasing, or at least strong imbalance and low power, for an important effect of interest. Therefore, it is not recommended to create the packets randomly; or at least, if this is done, simulations should be performed after selecting the random packets in order to ensure that there is no serious limitation in the set of packets chosen.

Another strategy would be to combine the careful planning of Strategy II with the diversity of the random strategy. The concept of "foldover" in the literature on fractional factorial design suggests that the aliasing in a particular fractional factorial design can be overcome by doing a new experiment with carefully chosen conditions that complement the information in the original experiment. In the current context, this idea means that the aliasing in a particular pair of packets can be overcome by adding more packets and that it can be overcome most efficiently by choosing these packets in a careful way. Following a procedure outlined in the documentation for PROC FACTEX (see the "Replicated Blocked Design with Partial Confounding" example, SAS Institute, 2011, pp. 679–682), it can be shown that there is a particular set of four packets that would allow all of the main effects and interactions to be tested. Specifically, in terms of the cells listed in Table 2, the packets would be {1, 4, 6, 7}, {2, 3, 5, 8}, {3, 4, 5, 6}, and {1, 2, 7, 8}. The first pair of packets is the same as in Strategy II. If only these first two packets were available, then the three-factor interaction would be aliased with cluster effects, but nothing else would be aliased. The remaining pair of packets, if it were used alone, would allow the three-way interaction to be estimated but would alias one of the two-way interactions (e.g., *IN-PERSON* × *PHONE*) with cluster effects. Combining the two pairs of packets guarantees that there is at least some information on all of the interactions. However, the three-way interaction and the *IN-PERSON* × *PHONE* interaction will be somewhat underpowered, because each is informed by only half of the design. Adding more packets can help to balance out this partial aliasing, so that all the interactions will be closer to having optimum power. This is described further in the appendix.

This "foldover" method is probably the best way to choose packets, in the absence of strong assumptions that certain interactions between factors can be ignored. However, the method described earlier of choosing a pair of packets that alias only an interaction considered negligible, and then replicating only this pair, will be somewhat more powerful for tests of two-way interactions than the foldover approach. Therefore, it might be a better choice than the foldover approach if it were known that the three-way interaction could likely be ignored. Further methodological research would be helpful for comparing these approaches in order to provide clearer guidance to researchers planning a hybrid-PEC factorial.

The hybrid designs we discussed in this section are motivated by real-world constraints of implementing factorial designs in the field, namely, cases in which

it is impossible for a cluster to implement all experimental conditions. However, in some cases, it might be possible for a cluster to implement the conditions in cohorts; that is, some conditions would be implemented initially, and the rest would be implemented later. For example, consider a scenario in which each cluster can implement only half of the conditions in Year 1 and the remaining half in Year 2. In such cases, it is important that the investigators ensure that the implementation of conditions in cohorts entails minimal risk for contamination. Further, it is important to minimize the confounding effect of cohort implementation (e.g., the possibility that the effect of In-Person is due to implementing most of the conditions involving the on level of this factor in Year 2, in which implementation fidelity was higher because staff was more experienced).This can be done by stratifying the randomization based on the cohorts (ensuring that each condition is implemented equally in each of the cohorts), as well as including cohort effect and the interaction between cohort and the factors in the model.

## *2.4 Full-EIC Factorials*

We now move on from discussing factorial experiments with PECs to those in which the experimenter assigns individuals to clusters as part of the experimentation, namely, to factorial designs that involve EICs. We begin with the simplest case of EIC, in which every individual is assigned to a cluster (see Nahum-Shani, Dziak, et al. 2017 for a brief review of published examples with this kind of clustering).

**Example** Suppose again that an investigator wishes to develop an intervention for improving the coping skills of adolescents with mental illness; however, now the program will be delivered by external experts, rather than by local clinic staff. Further, although the outcome of interest is similar to the previous scenario, the scientific questions motivating the study are different. Specifically, for illustrative purposes alone, suppose that a specific form of group therapy that focuses on skill learning and social support (facilitated by a trained practitioner) has proven to be effective in several randomized controlled trials and is considered an effective gold standard treatment appropriate for use as a benchmark in this field. Still, in light of empirical evidence suggesting that this group therapy results in clinical improvement that is small to moderate in magnitude, the investigator is motivated to develop ways to further improve the short- and long-term effectiveness of this approach. Hence, the investigator seeks to address three scientific questions, namely, whether the targeted outcome would be improved by augmenting an intervention program that is based on weekly group sessions with each of the following three components: (1) weekly instructional and motivational videos that individuals will watch and discuss as a group (Video), (2) weekly phone-based individual coaching sessions (Phone), and (3) supportive text messages (Text). Hence, the factors of theoretical interest are *VIDEO* ($X_1$), *PHONE* ($X_2$), and *TEXT* ($X_3$). Although it is an

important part of the intervention, the group therapy itself is not one of the factors in the experimental design because it is offered to all individuals.

To address these questions, all individuals are assigned to therapy groups of about five individuals each. The therapy groups did not exist prior to the beginning of the study; they are generated as part of the study design. For the purposes of the experiment, individuals will be randomly assigned not only to an experimental condition but also to a therapy group. To avoid contamination of experimental factors and/or perceptions of inequalities within the group, individuals are first assigned randomly to groups, and then groups are randomly assigned to each of the eight experimental conditions. Thus, each subject is nested within a cluster, and each cluster will belong to one of the eight experimental conditions. In this scenario, the individual-level outcomes are independent at pretest but are no longer independent at posttest because (a) group members potentially influence each other and (b) group members will be influenced by the shared practitioner.

In this setting, the motivation for generating clustering as part of the experimentation is both scientific and therapeutic. Specifically, the intervention is designed to be delivered in group settings, in order to facilitate therapeutic group processes and to make therapeutic use of social reinforcers such as social support, sense of belonging, cohesiveness, and social accountability. In these cases, the outcome for treated individuals may be correlated due to common experiences, informal processes of socialization, and group dynamics. However, there are other practical reasons that often motivate the generation of clusters as part of a study. These often concern the availability of resources and/or the feasibility of intervention delivery. For example, intervention science experiments commonly include a staff of therapists, each of whom delivers the intervention to a subset of individuals (e.g., Cloitre, Koenen, Cohen, & Han, 2002). Hence, the outcomes of individuals may be correlated due to shared provider effects.

**Model** Similar to between-PEC factorials, the factors of interest in the full-EIC scenario described above cannot vary within a cluster. This is because clusters, rather than individuals, are assigned to the experimental conditions. However, the full-EIC scenario differs considerably from a between-PEC factorial in other ways as well. In a between-PEC factorial design, the clusters are pre-existing units, and hence the response is expected to have a positive ICC at both pretest and posttest. In a full-EIC design, the clusters are created by the investigator during the study by random assignment. Thus, the pretest ICC is expected to be zero because individuals have no shared experience prior to the intervention, whereas the posttest ICC is expected to be positive because individuals from the same group are likely to have shared experiences during the study. Despite these differences, Model (2) proposed above for between-PEC factorials can also be used to analyze data arising from factorial experiments with full EIC. This is because Model (2), which models the effects of cluster-level factors, does not specify whether the pretest response $P_{ij}$ is clustered or not. Note that for simplicity, we assume that it is not necessary to model a practitioner effect in addition to a therapy group (i.e., cluster) effect. In practice,

however, it might be desirable to model them both as nested random effects if a given practitioner must supervise multiple treatment groups.

**Power** Based on the work of Nahum-Shani, Dziak, et al. (2017), Table 3 presents a simplified formula for calculating the noncentrality parameter $\lambda$, as well as the denominator degrees of freedom $\upsilon$ when planning power for a full-EIC factorial. Here, in addition to the parameters influencing power in the case of a within-PEC factorial, power will be reduced to the extent that individuals within a cluster are similar in their posttest response (i.e., have a higher $\rho_{\text{post}}$). Further, as in a between-PEC factorial, the denominator degrees of freedom ($\upsilon$) are calculated based on the total number of clusters ($J$). Hence, the number of clusters in a full-EIC factorial has more influence on power than the number of individuals within a cluster. Both issues are illustrated in Fig. 3, which provides the expected power (based on the formula for Design 5 in Table 3) for various values of $\rho_{\text{post}}$, $J$, and $n$. This figure shows that to the extent that the posttest ICC is higher, more clusters will be needed to achieve adequate power. Further, although power improves as a function of both the number of clusters and the number of individuals within a cluster, the number of clusters plays a more important role than the number of individuals, especially if the posttest responses of cluster members are highly correlated.

**Design Considerations** As discussed earlier, similar to between-PEC factorials, power for factorial experiments with full EIC is more heavily influenced by the total number of clusters than by the number of individuals within a cluster. However, this feature might be more important to consider when designing factorials with EIC, rather than factorials with PEC. While investigators typically have limited control over the size of clusters that exist prior to experimentation, they are likely to have more influence on the size of clusters that are induced by experimentation. Hence, investigators designing factorial studies with EIC might consider dividing a given sample into more (smaller) clusters, rather than fewer (larger) clusters, in order to enhance power.

Of course, power is not the only aspect investigators should consider when selecting cluster size in factorial designs involving EIC. Practical limits might also arise concerning the cluster size. For example, in a group therapy setting, groups larger than, say, six individuals may be difficult for the therapist to manage. Another example concerns a setting where In-Person therapy is provided by several therapists. Here, each therapist might be able to treat only a limited number of individuals. Additionally, there may be theoretical reasons to expect a particular cluster size to have the most therapeutic effect. Hence, when planning factorial designs with EIC, careful consideration should be given to the number of individuals assigned to each cluster. In some situations, it would even be reasonable to empirically investigate the most effective cluster size. This can be done either by conducting a pilot study to investigate the feasibility and acceptability of various cluster sizes prior to conducting the factorial design or by including cluster size as one of the factors under investigation in a factorial study.

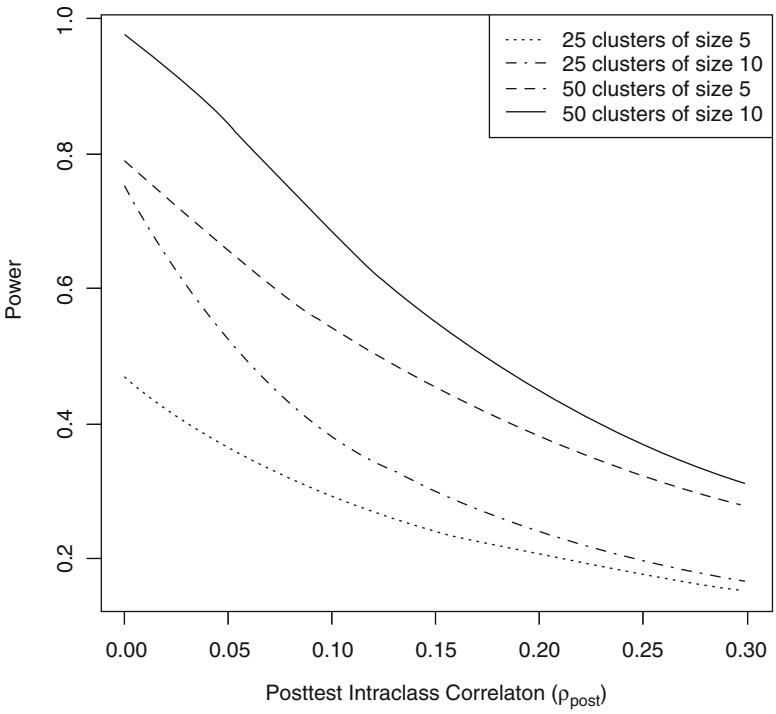


**Fig. 3** Expected power for full-EIC factorials. (Notes: *EIC* experimentally induced clusters. This figure shows the calculated power for a test of a main effect or interaction in Model (2) for a full-EIC multilevel factorial design. Model (2) assumes normally distributed responses and no cluster-by-treatment interaction. Here we further assume that the fitted model contains three dichotomous factors and all interactions (hence 9 regression coefficients including the pretest), although the number of parameters has only a modest effect on power for a given parameter. Balance is also assumed, namely, that each condition receives the same number of clusters and each cluster has the same number of individuals, although this may not be exactly possible in practice. To illustrate the link between the posttest ICC and power, it is assumed that the within-person correlation between pretest and posttest response ($\rho_{\text{pre, post}}$) is .65 and that the true value of the standardized regression coefficient is .15 (hence a Cohen's *d* of 0.3 if a main effect is being tested))

Planning the randomization scheme for full-EIC factorials should involve careful consideration of potential contamination. If individuals within a cluster receive different experimental conditions (i.e., different combinations of treatment components), then the potential for contamination could be high (see Dziak et al. 2012). If the potential for contamination is high, randomization should be done in a manner that ensures that everyone in a given cluster is also in the same condition. Specifically, investigators can either begin by assigning individuals to clusters and then randomly assigning *clusters* to experimental conditions or they can begin by assigning individuals to conditions and then assign individuals to clusters within each experimental condition. Either approach assures that all the members of a given cluster receive the same experimental condition.

## 2.5 Partial-EIC Factorials

Like full-EIC factorials, partial-EIC factorials involve generating clusters in the course of the study itself, as part of the experimentation. However, in partial-EIC, clusters are generated by experimentation for only a subset of the individuals (see Nahum-Shani, Dziak, et al. 2017 for published examples of this kind of clustering).

**Example** Suppose the same scenario as before, where an investigator wishes to develop an intervention for improving the coping skills of adolescents with mental illness. However, now assume that there is insufficient empirical evidence to determine whether the group-based therapy would improve symptom severity (the outcome of interest). Hence, unlike the previous example where the group therapy was assumed to be an effective gold standard treatment, in this scenario the investigator would like to address a scientific question concerning whether the targeted outcome would be improved by group-based weekly sessions (Group). As before, the investigator is also interested in addressing the scientific questions concerning weekly videos (Video) and supportive text messages (Texts). Hence, the factors of theoretical interest are *GROUP* ($X_1$), *VIDEO* ($X_2$), and *TEXT* ($X_3$).

To answer these questions, only individuals randomized to the on level of Group are assigned to groups of about five individuals each. As before, each individual in a given support group is given the same level of all of the assigned experimental factors (i.e., the same levels of Video and Text) as his/her fellow group members. In this scenario, the individuals in the on level of Group are clustered, whereas those in the off level remain independent.

**Model** In writing the model for a partial-EIC setting, it is not convenient for $Y$ to have a single subscript for some individuals and a double subscript for others. Based on the work of Roberts and Roberts (2005), Bauer, Sterba, and Hallfors (2008) and Nahum-Shani, Dziak, et al. (2017) recommended an approach where it is assumed that all individuals are clustered, yet the single individuals (i.e., those unclustered by design) are in "trivial" clusters of size 1. Specifically, for each cluster $j$, define a dummy-coded variable $C_j$, which is 1 if the cluster consists of multiple individuals assigned to be together and 0 if the cluster consists of a single individual. Thus, a cluster with $C_j = 1$ represents a "genuine" cluster, such as a therapy group, while a cluster with $C_j = 0$ is a trivial cluster of size one, such as a single individual not assigned to any therapy group. Assume there are $j = 1, \ldots, J$ clusters, both genuine and trivial, with $n_j$ individuals in the $j$th cluster (of course, $n_j = 1$ if $C_j = 0$), and denote the response for individual $i$ within cluster $j$ as $Y_{ij}$. Then, a model for a factorial design with partial EIC can take on the following form:

$$\begin{aligned} Y_{ij} = {} & \gamma_0 + \gamma_P P_{ij} + \\ & \gamma_1 X_{1j} + \gamma_2 X_{2j} + \gamma_3 X_{3j} + \\ & \gamma_4 X_{1j} X_{2j} + \gamma_5 X_{1j} X_{3j} + \gamma_6 X_{2j} X_{3j} + \\ & \gamma_7 X_{1j} X_{2j} X_{3j} + \\ & C_j u_j + e_{ij} \end{aligned} \tag{3}$$

where the $e_{ij}$ and the $u_j$ are independent $N(0, \sigma^2)$ and $N(0, \tau^2)$, respectively. Here, multiplying $u_j$ by $C_j$ assures that cluster-level variability is included in the model only for genuine and not for trivial clusters. This is an important feature of Model (3), because individual-level and cluster-level variability cannot meaningfully be distinguished from each other for trivial clusters. It may be desirable to allow the individual-level error variance $\sigma^2$ to differ between genuinely clustered and trivially clustered individuals.

Nahum-Shani, Dziak, et al. (2017) showed that even though the effect-coded cluster-generating factor $X_1$ and the dummy-coded clustering indicator $C_j$ contain the same information, they do not cause the model to be confounded in a deleterious way, because one is used only in the fixed-effects part of the model and the other is used only in the random-effects part. Here, the average effect of being in group therapy (versus not being in group therapy) would be expressed by $2\gamma_1$, while the deviation of the performance of a particular therapy group from the average would be expressed by $u_j$.

Model (3) differs from Model (2) only in that $u_j$ is multiplied by the clustering indicator $C_j$. In fact, Model (3) can be considered a generalization of models (2), if the values of $C_j$ are adjusted accordingly. Specifically, if $C_j = 1$ for all individuals, indicating that individuals are all clustered such that the factorial design is a full EIC, then Model (3) becomes Model (2).

**Power** Based on the work of Nahum-Shani, Dziak, et al. (2017), Table 3 (Design 6) presents a simplified formula for calculating the noncentrality parameter $\lambda$, as well as the denominator degrees of freedom $\upsilon$, when planning power for partial-EIC factorials. The formula for $\lambda$ is presented in two forms: one for a simplified scenario in which there is equal allocation of individuals to clustered and unclustered conditions and another for a more general case in which this allocation might be unequal.

Ordinarily, it is desirable to have balanced assignment on factors. However, individuals with $X_1 = +1$ will be subject to cluster-level variance $\tau^2$ in addition to their individual-level variance $\sigma^2$. Thus, if the individual-level variance $\sigma^2$ is equal for each cell, then cells with $X_1 = +1$ have a higher total error. Consistent with prior investigations of cluster allocation in randomized controlled trials with partial EIC (Baldwin et al. 2011), the formula for $\lambda$ provides a small increase in power when allocating more individuals to the clustered condition compared with equal allocation. This is illustrated in Fig. 4, which provides the expected power (based on the formula for Design 6 in Table 3) for various allocation proportions, as well as various cluster sizes and total sample size. This figure indicates that across various scenarios of cluster size and total sample size, the optimal allocation proportion to the clustered conditions is often between 0.6 and 0.7. The exact optimal allocation will depend heavily on the assumptions made about the variance components (see Nahum-Shani, Dziak, et al. 2017, for more information).

In the formula provided in Table 3, the denominator degrees of freedom $\upsilon$ is calculated conservatively based on the number of genuine clusters. However, as

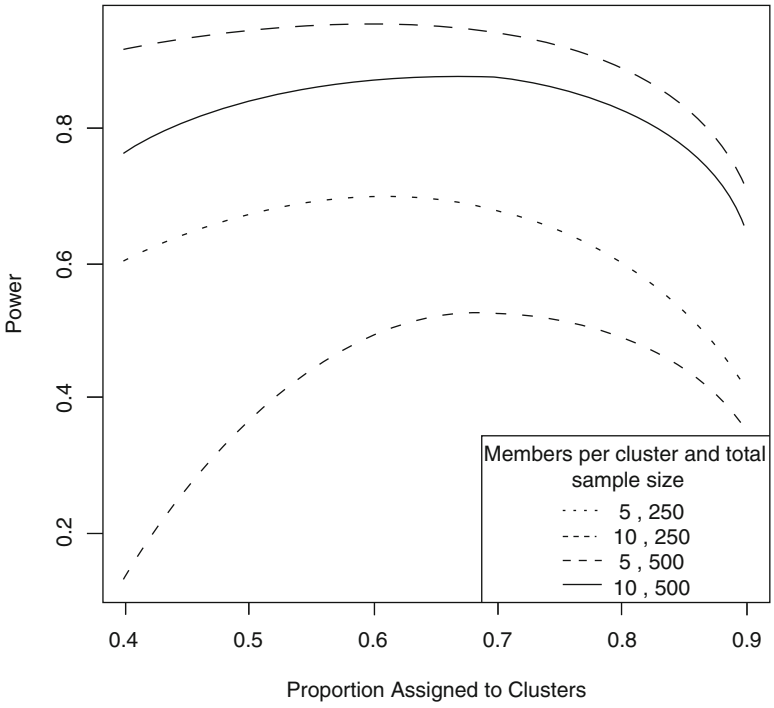


**Fig. 4** Expected power for partial-EIC factorials. (Notes: *EIC* experimentally induced clusters. This figure shows the calculated power for a test of a main effect or interaction in Model (3) for a partial-EIC multilevel factorial design. This means we assume a linear model with normally distributed responses, no cluster-by-treatment interaction, and a fitted model that contains three dichotomous factors and all interactions included (hence 9 regression coefficients including the pretest). Note that the number of parameters has only a modest effect on power for a given parameter. Balance is also assumed, namely, that each clustered condition receives the same number of clusters, each cluster has the same size, and each unclustered condition receives the same number of unclustered individuals, although this may not be exactly possible in practice. To illustrate the link between the allocation proportion (i.e., the proportion of individuals assigned to the clustered vs. the unclustered conditions) and power, it is assumed that the within-person correlation between pretest and posttest response ($\rho_{\text{pre, post}}$) is .65, the posttest ICC ($\rho_{\text{post}}$) is .10, and the true value of the standardized regression coefficient is .15 (hence a Cohen's $d$ of 0.3 if a main effect is being tested))

recommended by Nahum-Shani, Dziak, et al. (2017), when actually performing the hypothesis test, a Satterthwaite approximation can be used to provide slightly more power.

**Design Considerations** As with full-EIC factorials, careful consideration should be given to the number of individuals assigned to each group. The selection of the number of individuals within each group should be based on practical considerations, as well as empirical and theoretical evidence relating to the feasibility and the theoretical or clinical implications of employing various group sizes. Addi-

tionally, as with full-EIC factorials, the investigator should consider the potential for contamination when planning the randomization scheme for factorial designs with partial EIC. However, because in the partial-EIC setting certain experimental conditions are clustered while others are not, individuals cannot be assigned to clusters before they are assigned to conditions. A better option would be to first assign individuals to experimental conditions, and then randomly assign individuals to clusters within each clustered condition. Alternatively, the investigator can divide the randomization scheme into three steps. First, assign individuals to the two levels of the clustering factor $X_1$. Second, randomly assign individuals in the on level (i.e., the clustering level) of $X_1$ to clusters. Finally, randomly assign clusters in the on level of $X_1$, as well as single individuals in the off level of $X_1$, to the experimental conditions resulting from crossing the remaining (non-clustering) factors $X_2$ and $X_3$. Either approach will ensure that all members of a given cluster receive the same combination of treatment conditions.

## 3 Discussion

Multilevel factorial designs offer many opportunities for building the empirical evidence necessary for developing multicomponent behavioral, biobehavioral, and biomedical interventions. This chapter provides guiding principles for designing multilevel factorial designs and reviews five different classes of multilevel factorial designs that vary in terms of the nature of clustering and the randomization scheme employed. Our discussions of these classes highlight the importance of specifying the scientific questions of interest as a key to select the most appropriate class of multilevel factorial design and to plan procedures for randomizing individuals and/or clusters that would make the selected class as informative and efficient as possible. We also provide power planning resources, which overall demonstrate that multilevel factorial designs can offer adequate power for detecting the effects of scientific interest under reasonable scenarios in terms of number of clusters and individuals.

Nonetheless, our review highlights several important topics that require further research in order to improve the utility and accessibility of multilevel factorial designs for intervention scientists. These include the need for a systematic comparison of different approaches for selecting packets in hybrid-PEC factorials, as well as the importance of developing power planning resources for multilevel factorial studies that involve more than one level of clustering (e.g., factorial studies in which individuals are assigned to groups, and then multiple groups are led by each of a limited number of therapists). Below we elaborate on two additional important directions for future research. The first concerns accommodating cluster-by-treatment interactions, and the second concerns sequential randomizations.

### *3.1 Accommodating Cluster-by-Treatment Interactions*

For simplicity, cluster-by-treatment interactions are not included in any of the models reviewed in this chapter for multilevel factorial designs in which individuals (rather than clusters) are randomized to conditions. The underlying assumption here is that any cluster-by-treatment interaction is negligible. However, Raudenbush and Liu (2000) highlight the scientific benefit of estimating cluster-by-treatment interactions in settings where clustered individuals are randomized to experimental conditions. First, differences in cluster-based resources, such as in knowledge, skills, or environmental/social conditions, might lead to differences in the effects of certain intervention components, creating additional uncertainty concerning how and to what extent these components affect the targeted outcome. Second, cluster-by-treatment interactions can be viewed as an important test of the generalizability of treatment effects over the various settings in which intervention components may be implemented in the real world. From this point of view, each cluster contributes an independent study of intervention components, and the cluster-by-treatment interaction is conceptualized as a "meta-analysis" of treatment efficacy across clusters.

Even in situations where cluster-by-treatment interactions are not of theoretical interest, they may still affect the results of the analysis if they are of sufficient size. In general, with normally distributed response variables, the cluster-by-treatment interactions will not bias the estimates of the main effects or interactions of factors. However, they will add noise, which is not accounted for by the power formulas provided here. Power formulas for multilevel factorial designs with cluster-by-treatment interactions have not yet been derived, but this is an important topic for future research. Raudenbush and Liu (2000) discuss the power implications of cluster-by-treatment interactions in the case of a single-factor within-PEC experiment, namely, a multisite randomized controlled trial with one experimental condition and one control condition. They show that power becomes smaller to the extent that the variance component for the cluster-by-treatment interactions becomes larger, especially if there are a few large clusters instead of many small clusters. Hence, the formulas shown in Tables 3 and 4 offer the best-case scenario in terms of the predicted power, as the cluster-by-treatment variance is assumed to be zero. Generalizing the formula of Raudenbush and Liu (2000) to multiple factors may be a good starting point for future research aiming to develop power formulas for multilevel factorial designs with cluster-by-treatment interactions.

There are some special cases in which cluster-by-treatment interactions may cause estimation bias in addition to reduced power. One such case would be a hybrid-PEC design in which an effect of primary interest is aliased with a non-negligible cluster-by-treatment interaction. Another would be a non-normal response variable (e.g., binary) or, more generally, a situation in which a noticeable floor or ceiling effect occurs for the outcome variable. In these cases, effects or interactions that would classically be orthogonal may become related in more complicated ways (see Cox & Snell, 1988; Gail, Wieand, & Piantadosi, 1984).

Research is also needed on power and sample size planning for multilevel factorial experiments with binary or count responses (see Demidenko, 2007; VanderWeele, 2012; Wittes, 2002; Wolbers et al. 2011 for the non-multilevel case). In the absence of formulas for calculating power and planning sample size for such scenarios, computer simulation methods can offer a flexible alternative (Arnold, Hogan, Colford, & Hubbard, 2011).

## *3.2 Extensions to Sequential Randomizations*

In all the multilevel factorial designs reviewed in this chapter, it was assumed that all individuals are randomized to experimental conditions *once* in the course of the study. However, multilevel factorial designs can instead be sequential, multiple assignment, randomized trials (SMARTs; Murphy, 2005). A prototypical SMART is a special form of a factorial design in which some or all individuals can be randomized to experimental conditions more than once in the course of the experiment. Specifically, a SMART includes multiple, sequential stages of randomization, where each stage is designed to address a scientific question concerning the construction of an adaptive intervention. An adaptive intervention is a sequence of individualized intervention components that guide decision-making in clinical, educational, public health policy, and other settings (see Almirall, Nahum-Shani, Wang, & Kasari, 2018). The individualization in an adaptive intervention uses ongoing (dynamic) information about how well the individual is doing in the course of the intervention to modify the type, timing, and/or modality of treatment delivery (Collins, Murphy, & Bierman, 2004).

As in standard interventions, participants in an adaptive intervention can be members of PECs (e.g., schools or clinics) or can be assigned to clusters (e.g., therapy groups, online support networks) as part of the intervention. Hence, a SMART aiming to inform the development of *clustered adaptive interventions* (Kilbourne et al. 2013) might involve individuals that are clustered in pre-existing social or administrative units (PEC) and/or individuals that are assigned to clusters as part of the experimentation (EIC). Further, such multilevel SMARTs might employ various randomization schemes, randomizing individuals, clusters, or both to experimental conditions. However, in a SMART the nature of clustering as well as the randomization scheme might differ from one stage of randomization to another, depending on the motivating scientific questions and the type of intervention options compared at each stage.

To clarify this, consider a hypothetical SMART aiming to develop an adaptive intervention for improving the coping skills of adolescents with mental illness. As before, assume that the outcome of interest, $Y$, is a reverse-coded measure of symptom severity, so that higher values represent a more desirable outcome. Again, for illustrative purposes alone, suppose that a specific form of group therapy that focuses on skill learning and social support (facilitated by a trained practitioner) is considered an effective gold standard treatment appropriate for use as a benchmark

in this field. Still, assume that empirical evidence suggests that a sizable proportion of individuals do not benefit from this intervention; and in fact, it is possible to identify those individuals early in the course of the intervention (say, about 4 weeks after beginning the group therapy). Hence, the investigator is motivated to improve the effectiveness of the group therapy sessions by integrating other, potentially more engaging, intervention components in the sessions for all individuals, as well as by offering additional intervention components to individuals who by week 4 show early signs of non-response (i.e., early non-responders). Specifically, the investigator seeks to address two scientific questions, namely, whether the targeted outcome would be improved by (1) integrating weekly instructional and motivational videos in the group therapy sessions for all individuals (individuals will watch and discuss the videos in their therapy group; Video) and by (2) adding weekly phone-based individual coaching sessions to individuals who by week 4 show early signs of non-response (Phone). Note that based on the notion that "if it's not broken, don't fix it," the investigator decides that adolescents who show early signs of response (i.e., early responders) to group therapy sessions by week 4 should continue with that intervention approach.

Based on the hypothetical example above, the experimental design (see Fig. 5) should involve two factors: *VIDEO* ($X_1$) and *PHONE* ($X_2$). As before, each factor would have two levels, on and off. Notice that now the randomization to these factors should be sequential – individuals should be randomized to the two levels of *VIDEO* initially and then (at week 4) to the two levels of *PHONE*. Also notice that to answer the second scientific question, only early non-responders should get randomized to the two levels of *PHONE*. Specifically, a multilevel SMART design (see Almirall et al. 2018) aiming to answer these scientific questions would involve two stages of randomization. The first stage would involve randomly assigning all individuals to therapy groups (full EIC), as well as randomizing the therapy groups (clusters) to the two levels of *VIDEO*, namely, to either receive videos as part of the group therapy sessions (on level of $X_1$) or not (off level of $X_1$). The second stage of randomization would involve re-randomizing only those individuals who at week 4 show early signs of non-response to the two levels of *PHONE*, namely, to either add phone sessions to the initial intervention (on level of $X_2$) or not (off level of $X_2$). The exact definition of early non-response would be operationalized based on pre-specified evidence-based criteria.

Interestingly, in this scenario the intervention options at the second-stage randomization (add phone coaching vs. no phone coaching) are designed to be offered to individuals based on their individual-level non-response status. Hence, the second-stage randomization would involve assigning individuals, rather than therapy groups, to the two levels of $X_2$. However, these individuals are now clustered in social units that existed prior to the second-stage randomization; that is, although the therapy groups were experimentally induced during the first-stage randomization, they represent PECs for the second-stage randomization. Overall, in this hypothetical example, the first stage of the design is a full EIC, and the second stage is a within-PEC design. Besides the methodological issues involved, the investigator would also have to consider possible ethical concerns and risk of

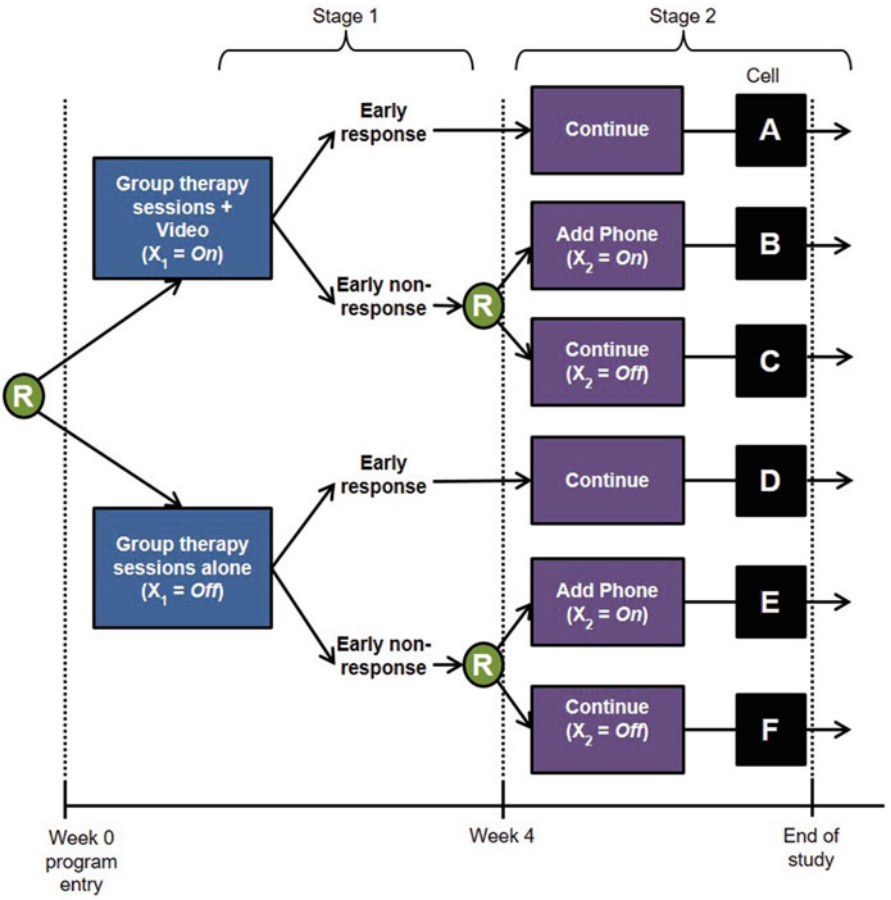


**Fig. 5** Hypothetical SMART study

contamination that might arise when offering different members of the same therapy group different treatments at the later stage.

Despite the increased interest in using SMARTs to build high-quality adaptive interventions, there has been little research on using SMARTs to develop an adaptive intervention in a setting where individuals are clustered. A comprehensive review of possible types of multilevel SMARTs and design principles to guide the design of such studies has the potential to advance the science of adaptive interventions. Further, existing data analytic methods for analyzing data arising from a SMART are not suitable for analyzing multilevel SMART data. These include methods for comparing adaptive interventions that are embedded by design in a SMART (Nahum-Shani et al. 2012a), as well as methods for exploring ways to further tailor (personalize) these embedded adaptive interventions (e.g., Q-learning; Laber, Linn, & Stefanski, 2014; Nahum-Shani et al. 2012b; Nahum-Shani, Ertefaie, et al. 2017).

Further research is needed to extend these methods to multilevel data arising from various types of multilevel SMARTs and provide power planning resources for multilevel SMARTs.

**Acknowledgment** The authors would like to thank Daniel Almirall for his advice and insights and Amanda Applegate for editorial assistance. Figures were drawn using the R software package (cran.r-project.org). This project was supported by Awards P50 DA010075, P50 DA039838, P01 CA180945, R01 DK097364, R01 AA022931, R01 DA039901, and K05 DA018206 from the National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or its institutes. Correspondence may be sent to Inbal Nahum-Shani, Survey Research Center, Institute for Social Research, University of Michigan, 426 Thompson Street, Suite 2204, Ann Arbor, MI 48104-2321, inbal@umich.edu.

## A.1 Appendix 1: Details Concerning Power Calculations

For all models provided in this chapter, the main effects and interactions of interest can be tested using significance tests for the corresponding regression coefficients. The power for such a significance test is estimated as the probability that a variable with a noncentral $F$ distribution with noncentrality parameter $\lambda$ exceeds the critical value for the test. To clarify this, the sampling distribution of the test statistic for the relevant coefficient is assumed to follow an $F$ distribution. Under the null hypothesis, this $F$ distribution has two parameters, the numerator ($\omega$) and denominator ($\upsilon$) degrees of freedom (determined based on the sample). This is called a central $F$ distribution, which is the $F$ distribution one can typically find tables for in statistics textbook and other resources. However, under the alternative hypothesis (i.e., when an effect exists in the population), the sampling distribution of the test statistic will have another parameter, in addition to $\omega$ and $\upsilon$. This parameter is known as the noncentrality parameter, which is often denoted by $\lambda$. This distribution is called a noncentral $F$ distribution. When the noncentral $F$ distribution has a noncentrality parameter of zero, it is identical to the standard (central) $F$ distribution. In other words, a noncentral $F$ distribution includes the $F$ distribution as a special case (see Koele, 1982; Raudenbush & Liu, 2000, for additional details).

To calculate the power to detect a particular alternative hypothesis, we first compute the critical value for the $F$ statistic under the null hypothesis using the selected type I error rate and the appropriate degrees of freedom. We then need to specify the alternative hypothesis in the form of its $\lambda$ value. As explained earlier, $\lambda$ equals $\gamma^2/\text{Var}(\widehat{\gamma})$, where $\gamma$ is the true value of the coefficient being tested (under the alternative hypothesis) and $\text{Var}(\widehat{\gamma})$ is the sampling variance (squared standard error) of its maximum likelihood estimate for samples of the size proposed. The power of the test to detect the given alternative hypothesis is then equal to the area under the noncentral $F$ distribution to the right of the critical value for the test. To the extent that $\lambda$ is larger, the power of the test increases. Hence, $\lambda$ represents the expected amount of evidence against the null hypothesis that will be available in

a sample with the proposed true parameters and the proposed sample size. Murphy and Myors (2004, p. 27) call it a measure of "how wrong" the null hypothesis is.

## B.1 Appendix 2: Sample R and SAS Code for Calculating Power

Here, we provide examples of how the noncentrality parameter can be used to calculate power using standard software such as R and SAS. For illustration, this sample code calculates power for detecting a standardized effect size of 0.15 (the true value of the coefficient being tested [`gamma`] divided by the square root of the variance of posttest response adjusting for treatment effect [`sigmaY`]) in a model with 9 regression coefficients (`p`), with 0.05 type I error rate, total sample size (`N`) of 250 individuals who are independent (not clustered), and 0.5 within-person correlation between pretest and posttest response (`rhoPrePost`). Here, "lambda" is the noncentrality parameter (calculated based on `N`, `gamma/sigmaY`, and `rhoPrePost`), and "critical" is the critical value for the hypothesis test (calculated based on the type I error rate and the appropriate degrees of freedom, which are 1 for the numerator, and `N-p` for the denominator).

Sample R code:

```
N <- 250;
rhoPrePost <- .5;
gamma <- .15;
sigmaY <- 1;
p <- 9;
df <- N - p;
lambda <- (N/(1-rhoPrePost))*((gamma/sigmaY)^2);
critical <- qf(p=.95,df1=1,df2=df,ncp=0,lower.tail=TRUE);
power <- 1-pf(q=critical,df1=1,df2=df,ncp=lambda,lower.
tail=TRUE);
print(power);
```

Sample SAS code:

```
DATA data1;
     N = 250;
     rhoPrePost = .5;
     gamma = .15;
     sigmaY = 1;
     p = 9;
     df = N - p;
     lambda = (N/(1-rhoPrePost))*((gamma/sigmaY)**2);
     critical = FINV(.95,1,df,0);
     power = 1-PROBF(critical,1,df,lambda);
RUN;
PROC PRINT DATA=data1; RUN;
```

## C.1 Appendix 3: SAS Code for Hybrid-PEC Designs

Sample SAS code to fit models for within-PEC and between-PEC factorial experiments is included in the online supplemental appendix of Dziak, Nahum-Shani, and Collins (2012). Likewise, sample SAS code for fitting models for full-EIC and partial-EIC factorial experiments is included in the appendix of Nahum-Shani, Dziak, et al. (2017). However, hybrid-PEC designs are not discussed in those papers. Therefore, sample SAS code for creating such a design and for analyzing the resulting data is included here.

We first present how to use the "foldover" method to choose the contents of packets. Our presentation here is based on Example 7.12 ("Replicated Blocked Design with Partial Confounding") on pages 679–683 of SAS Institute, Inc. (2011). In the example of hybrid EIC in the body of the chapter, there are three factors of interest, called *IN-PERSON*, *PHONE*, and *PARENT*, defining $2 \times 2 \times 2 = 8$ possible conditions. Also, we assume there are ten clusters, and that within each cluster, it is possible to implement only four of the eight conditions. Hence, we must repeatedly divide the eight conditions into two packets of four conditions each.

```
PROC FACTEX;
FACTORS InPerson Phone Parent;
BLOCKS NBLOCKS=2;
MODEL EST=(InPerson Phone Parent InPerson*Phone
           InPerson*Parent Phone*Parent);
EXAMINE CONFOUNDING ALIASING;
OUTPUT OUT=PacketPair1 BLOCKNAME=Packet NVALS=(1 2);
RUN;
```

In the FACTORS statement, we listed the names of the three factors in the example. Concerning the BLOCKS statement, it is important to note that here a "block" means a packet. We ask for two packets at a time because we decided that we wanted packets of four conditions each. If we ask for four packets, for example, the software will offer us four packets of two conditions each, which is not what we want here. In the MODEL statement, we specified the effects we would like to estimate. Here, seven effects might be of interest: three main effects, three two-way interactions, and one three-way interaction. Because the conditions will be divided into two packets and we are assuming that each packet will be given to only one cluster, the random effect of the cluster within this pair must be aliased with one of the effects. We choose to sacrifice the three-way interaction here; hence the MODEL statement does not include the three-way interaction. SAS will select the packets in order to provide us with interpretable information about the other six effects. Using the EXAMINE statement, we ask that the structure of the design will be outputted as well as the confounding rules (i.e., block with *IN-PERSON*PHONE*PARENT*).

The above code creates the first two packets (consistent with Strategy II in the book chapter) and writes them to a dataset (which we name here "PacketPair1"). We could just stop here if we had only two clusters, as in our first hybrid-PEC example. However, because we assume we have ten clusters, as in the case of our second hybrid-PEC example, we could either stop here and replicate the packets five times

or create more packets to which we can assign the ten clusters. With the former option, we would have some information to estimate the three-way interaction, albeit with very little power. With the latter option, we could enhance our ability to test the three-way interaction, by looking to add other combinations of conditions. Hence, we can ask PROC FACTEX to create more packets. Notice that we do not have to call PROC FACTEX a second time, because we used RUN but not QUIT to close this PROC. We can simply add the following code:

```
MODEL EST=(InPerson Phone Parent
           InPerson*Parent Phone*Parent InPerson*Phone*Parent);
OUTPUT OUT=PacketPair2 BLOCKNAME=Packet NVALS=(3 4);
RUN;
```

Following the "foldover" idea described in the SAS/QC(R) users' guide, we now ask for another two packets. We want these packets to provide us with information about the three-way interaction, which was assumed negligible earlier. However, in order to do this, we now need to confound these two packets with the two-way interaction between In-Person and Phone (*IN-PERSON*PHONE*). This means that using this pair of packets will not provide information about this two-way interaction. However, using the first pair of packets will provide information about this two-way interaction. Hence, when all four packets are combined, we will be able to estimate and test both the two-way interaction between In-Person and Phone and the three-way interactions. We could just stop here and have four packets, which are replicated for two or three clusters each (in our ten-cluster example). However, this would mean that the first two-way interaction and the three-way interaction will have somewhat lower power than the other effects, because they are each informed by only half of the packets in the study. If we want more balance in the amount of statistical power each test will possess, we must add more packets. This can be done by using the following code:

```
MODEL EST=(InPerson Phone Parent
           InPerson*Phone Phone*Parent InPerson*Phone*Parent);
OUTPUT OUT=PacketPair3 BLOCKNAME=Packet NVALS=(5 6);
RUN;
```

The packets created with the code above will provide information about everything except the two-way interaction between In-Person and Parent. Alternatively, the code below generates packets that provide information about everything except the interaction between Phone and Parent.

```
MODEL EST=(InPerson Phone Parent
           InPerson*Phone InPerson*Parent
           InPerson*Phone*Parent);
OUTPUT OUT=PacketPair4 BLOCKNAME=Packet NVALS=(7 8);
RUN;
```

The example in the SAS user's guide stops here and concludes that each main effect is informed by the whole sample, and each interaction is informed by 75% (= 6/8) of the sample. However, in our example in the body of the chapter, we assumed that we had ten clusters. Therefore, we can add more packets so that more

information will be used to estimate each interaction. We elect to add another pair of packets in which the three-way interaction is not aliased. This will provide 100% efficiency for the main effects, 80% efficiency for the two-way interactions, and 60% efficiency for the three-way interaction.

```
  MODEL EST=(InPerson Phone Parent
             InPerson*Phone InPerson*Parent
                InPerson*Phone*Parent);
  OUTPUT OUT=PacketPair5 BLOCKNAME=Packet NVALS=(9 10);
  RUN;
QUIT;
DATA Packets;
  SET PacketPair1 PacketPair2 PacketPair3 PacketPair4 PacketPair5;
RUN;
```

The dataset packets created with this code will contain 40 rows, corresponding to the 4 conditions in each of the 10 packets. Each of the ten packets must now be randomly assigned to one of ten clusters.

```
PROC PRINT DATA=Packets;
RUN;
```

We argue in this chapter that the data from the resulting experiment can then be analyzed in the usual way for a within-PEC factorial, without including packet as a level of nesting. For example, the following code fits a linear model with all interactions among factors:

```
PROC MIXED DATA=StudyData;
   MODEL Y = InPerson | Phone | Parent @ 3;
   RANDOM INTERCEPT / SUBJECT = clusterID;
RUN;
```

The following code also allows interactions between factors and clusters:

```
PROC MIXED DATA=StudyData;
   MODEL Y = InPerson | Phone | Parent @ 3 / DDFM=Satterthwaite;
   RANDOM INTERCEPT InPerson Phone Parent / SUBJECT = clusterID;
RUN;
```

However, as we discuss in the book chapter, the performance of factorial designs when there are interactions between factors and clusters has not yet been well studied in the social sciences, and further research is needed to better understand the role of possible interactions between treatments and clusters when analyzing a factorial design. Note that the Satterthwaite approximation is now used to calculate the denominator degrees of freedom because this code estimates the cluster-by-factors interactions (rather than just the effects of the factors); hence, the denominator degrees of freedom have to be calculated based not only on the number of individuals but also on the number of clusters.

## References


Almirall, D., Nahum-Shani, I., Wang, L., & Kasari, C. (2018). Experimental designs for research on adaptive interventions: Singly and sequentially randomized trials. In L. M. Collins & K. C. Kugler (Eds.), *Optimization of behavioral, biobehavioral, and biomedical interventions: Advanced topics*. New York, NY: Springer.

Arnold, B. F., Hogan, D. R., Colford, J. M., & Hubbard, A. E. (2011). Simulation methods to estimate design power: An overview for applied research. *BMC Medical Research Methodology, 11*(1), 1.

Baldwin, S. A., Bauer, D. J., Stice, E., & Rohde, P. (2011). Evaluating models for partially clustered designs. *Psychological Methods, 16*(2), 149–165.

Bauer, D. J., Sterba, S. K., & Hallfors, D. D. (2008). Evaluating group-based interventions when control participants are ungrouped. *Multivariate Behavioral Research, 43*, 210–236.

Cloitre, M., Koenen, K. C., Cohen, L. R., & Han, H. (2002). Skills training in affective and interpersonal regulation followed by exposure: A phase-based treatment for PTSD related to childhood abuse. *Journal of Consulting and Clinical Psychology, 70*(5), 1067.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Collins, L. M. (2018). *Optimization of behavioral, biobehavioral, and biomedical interventions: The multiphase optimization strategy (MOST)*. New York, NY: Springer.

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.

Collins, L. M., Murphy, S. A., & Bierman, K. L. (2004). A conceptual framework for adaptive preventive interventions. *Prevention Science, 5*(3), 185–196.

Cox, D. R., & Snell, E. J. (1988). *Analysis of binary data* (2nd ed.). London, UK: Chapman & Hall.

Demidenko, E. (2007). Sample size and optimal design for logistic regression with binary interaction. *Statistics in Medicine, 27*, 36–46.

Donner, A., & Klar, N. (2000). *Design and analysis of cluster randomization trials in health research*. London, UK: Arnold.

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.

Eldridge, S. M., Ashby, D., Feder, G. S., Rudnicka, A. R., & Ukoumunne, O. C. (2004). Lessons for cluster randomized trials in the twenty-first century: A systematic review of trials in primary care. *Clinical Trials, 1*(1), 80–90.

Gail, M. H., Wieand, S., & Piantadosi, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika, 71*, 431–444.

Kilbourne, A. M., Abraham, K. M., Goodrich, D. E., Bowersox, N. W., Almirall, D., Lai, Z., & Nord, K. M. (2013). Cluster randomized adaptive implementation trial comparing a standard versus enhanced implementation intervention to improve uptake of an effective re-engagement program for patients with serious mental illness. *Implementation Science, 8*(1), 1–14.

Koele, P. (1982). Calculating power in analysis of variance. *Psychological Bulletin, 92*(2), 513.

Kuehl, R. O. (2000). *Design of experiments: Statistical principles of research design and analysis*. Pacific Grove, CA: Duxbury Press.

Kugler, K. C., Dziak, J. J., & Trail, J. (2018). Coding and interpretation of effects in analysis of data from a factorial experiment. In L. M. Collins & K. C. Kugler (Eds.), *Optimization of behavioral, biobehavioral, and biomedical interventions: Advanced topics*. New York, NY: Springer.

Laber, E. B., Linn, K. A., & Stefanski, L. A. (2014). Interactive model building for Q-learning. *Biometrika, 101*, 831–847.

Mitchell, M. M., Bradshaw, C. P., & Leaf, P. J. (2010). Student and teacher perceptions of school climate: A multilevel exploration of patterns of discrepancy. *Journal of School Health, 80*(6), 271–279.

Moerbeek, M., & Teerenstra, S. (2015). *Power analysis of trials with multilevel data*. New York: Chapman and Hall/CRC.

Morgan, K. L., & Rubin, D. B. (2015). Rerandomization to balance tiers of covariates. *Journal of the American Statistical Association, 110*, 1412–1421.

Murphy, K. R., & Myors, B. (2004). *Statistical power analysis* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Murphy, S. A. (2005). An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine, 24*(10), 1455–1481.

Murray, D. M. (1998). *Design and analysis of group-randomized trials* (2nd ed.). New York, NY: Oxford.

Nahum-Shani, I., Dziak, J. J., & Collins, L. M. (2017). Multilevel factorial designs with experiment-induced clustering. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000128.

Nahum-Shani, I., Ertefaie, A., Lu, X. L., Lynch, K. G., McKay, J. R., Oslin, D. W., & Almirall, D. (2017). A SMART data analysis method for constructing adaptive treatment strategies for substance use disorders. *Addiction, 112*(5), 901–909.

Nahum-Shani, I., Qian, M., Almirall, D., Pelham, W. E., Gnagy, B., Fabiano, G. A., … Murphy, S. A. (2012a). Experimental design and primary data analysis methods for comparing adaptive interventions. *Psychological Methods, 17*(4), 457.

Nahum-Shani, I., Qian, M., Almirall, D., Pelham, W. E., Gnagy, B., Fabiano, G. A., … Murphy, S. A. (2012b). Q-learning: A data analysis method for constructing adaptive interventions. *Psychological Methods, 17*(4), 478.

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods, 5*(2), 199–213.

Roberts, C., & Roberts, S. A. (2005). Design and analysis of clinical trials with clustering effects due to treatment. *Clinical Trials, 2*, 153–162.

SAS Institute Inc. (2011). *SAS/QC® 9.3 users' guide.* SAS Institute, Inc.

Taljaard, M., Weijer, C., Grimshaw, J. M., Belle Brown, J., Binik, A., Boruch, R., … Saginur, R. (2009). Ethical and policy issues in cluster randomized trials: Rationale and design of a mixed methods research study. *Trials, 10*(1), 61.

VanderWeele, T. J. (2012). Sample size and power calculations for additive interactions. *Epidemiologic Methods, 1*(1): Article 8. https://doi.org/10.1515/2161-962X.1010.

Weijer, C., Grimshaw, J. M., Taljaard, M., Binik, A., Boruch, R., Brehaut, J. C., … Saginur, R. (2011). Ethical issues posed by cluster randomized trials in health research. *Trials, 12*(1), 100.

Wittes, J. (2002). Sample size calculations for randomized controlled trials. *Epidemiologic Reviews, 24*, 39–53.

Wolbers, M., Heemskerk, D., Chau, T. T., Yen, N. T., Caws, M., Farrar, J., & Day, J. (2011). Sample size requirements for separating out the effects of combination treatments: Randomized controlled trials of combination therapy vs. standard treatment compared to factorial designs for patients with tuberculous meningitis. *Trials, 12*, 26.

Wu, C. F. J., & Hamada, M. S. (2000). *Experiments: Planning, analysis, and parameter design optimization*. New York, NY: Wiley.

# Experimental Designs for Research on Adaptive Interventions: Singly and Sequentially Randomized Trials


**Daniel Almirall, Inbal Nahum-Shani, Lu Wang, and Connie Kasari**

**Abstract** In clinical or educational practice, it is often necessary to use an individually tailored, sequential approach to intervention in order to improve outcomes. Adaptive interventions (also known as dynamic treatment regimens) can be used to guide such sequential intervention decision-making. Adaptive interventions are multicomponent, multistage intervention packages. The multiphase optimization strategy (MOST) is a comprehensive research framework for development, optimization, and evaluation of multicomponent intervention packages, such as adaptive interventions. When working within the optimization phase of MOST, behavioral, biobehavioral, and educational intervention scientists often have important scientific questions about how best to optimize an adaptive intervention. This chapter discusses various types of experimental designs that can be used to optimize an adaptive intervention. In some of these, participants are randomized once over the course of the trial (i.e., singly randomized trials, or SRTs), and in others, participants are randomized at multiple stages (i.e., sequential, multiple assignment, randomized trials, or SMARTs). The choice between SRT and SMART ultimately is driven by the scientific questions that the intervention scientist seeks to answer. Motivated by the development of an adaptive intervention to improve social skills and academic

Funding was provided by the National Institutes of Health (P50 DA039838, R01 DA039901, R01 HD073975, R01 DK108678, R01 AA023187, R01 AA022113, and U54 EB020404) and the Institute for Education Sciences (R324U150001). This work is the sole responsibility of the authors and does not necessarily reflect the opinion of the funding agencies.

D. Almirall (✉)
Institute for Social Research and Department of Statistics, University of Michigan, Ann Arbor, Michigan
e-mail: dalmiral@umich.edu

I. Nahum-Shani · L. Wang
University of Michigan, Ann Arbor, Michigan

C. Kasari
University of California, Los Angeles, CA, USA



L. M. Collins, K. C. Kugler (eds.), *Optimization of Behavioral, Biobehavioral, and Biomedical Interventions*, Statistics for Social and Behavioral Sciences,
https://doi.org/10.1007/978-3-319-91776-4_4

engagement among children with autism, we illustrate these ideas by presenting four example of experimental designs: two examples of a SRT and two examples of a SMART. We present the rationale for each experimental design and the questions each is designed to answer. In doing so, this chapter provides an expanded set of tools that investigators aiming to develop an adaptive intervention can draw from within the MOST optimization phase toolbox.

## 1 Introduction

In many areas of clinical and educational practice, it is often necessary to use an individually tailored, sequential approach to intervention in order to improve outcomes in the long term. In a sequential approach, interventions (or intervention components) may be adapted and readapted over time based on changes in the individual, including changes that could occur as a result of prior intervention. Adaptive interventions are intervention designs that can be used to guide such sequential intervention decision-making. An adaptive intervention is a sequence of prespecified decision rules that can be used to guide whether, how, or when—and based on which measures—to alter an intervention or intervention component (e.g., treatment type, duration, frequency, or amount) at critical decision points during the course of care (Almirall, Nahum-Shani, Sherwood, & Murphy 2014; L. Collins, Murphy, & Bierman 2004; L. M. Collins, Nahum-Shani, & Almirall 2014; Nahum-Shani et al. 2012a).

An adaptive intervention is a type of multicomponent, behavioral, biobehavioral, or biomedical intervention (see companion book L. M. Collins 2018). Specifically, an adaptive intervention includes *decision points* or time points at which intervention decisions are made; *tailoring variables*, which can be used to make intervention decisions at each decision point; *intervention options* at each decision point; and a *decision rule*, which is used at each decision point to link the values of a (set of) tailoring variable(s) with a specific intervention option or set of intervention options. The intervention options at each decision point may focus on treatment, prevention, or both; they may include behavioral, pharmacological, or educational interventions.

As with all interventions, adaptive interventions are characterized by the unit(s) at which components are targeted (e.g., the individual, the parent, or the classroom environment), the entity or individual(s) providing each component (e.g., the clinician or the therapist), and the goals of the intervention—the short- and long-term outcomes the adaptive intervention is intended to impact. It is important to define the unit to which the intervention is delivered (e.g., the school, in a school-wide program) and also the unit at which the intervention should have an effect (e.g., individual students). For example, a classroom intervention may target the classroom environment, but its ultimate goal may be to impact the outcomes of the children within the classroom.

The multiphase optimization strategy (MOST) is a comprehensive framework for development, optimization, and evaluation of intervention packages, including adaptive interventions (L. M. Collins et al. 2011; L. M. Collins, Murphy, Nair, &

Strecher 2005; L. M. Collins, Murphy, & Strecher 2007). The objective of the optimization phase in MOST "is to make a multicomponent intervention more effective, efficient, economical, and scalable" (L. M. Collins 2018). The optimization phase is often based on factorial experimental (i.e., randomized) designs (Chakraborty, Collins, Strecher, & Murphy 2009; Nair et al. 2008). Many scientific questions arise when seeking to develop and optimize an adaptive intervention when organizing components into the stages that make up an adaptive intervention package. These questions include the following:

- Which intervention option is the best to start with?
- When, or under what conditions, is it best to transition from one intervention stage to the next?
- What subsequent intervention option is the best for individuals who are non-responders to initial intervention options?
- What is the best way to define response versus non-response?
- At each stage, based on what information/variables should we tailor the intervention options (i.e., treat one individual differently from another)?

Over the past 10–15 years, motivated in part by interest in research on "personalized" or "precision" medicine (F. S. Collins & Varmus 2015; National Research Council 2011), research has accelerated on developing and optimizing adaptive interventions. At the same time, there has been a surge of interest in experimental designs used to inform the construction of effective adaptive interventions.

The primary goal of this chapter is to review four types of experimental designs for optimizing adaptive interventions. Two of these designs are singly randomized trials (SRTs), in which participants are randomized only once in the course of the trial. The remaining two designs are sequential, multiple assignment, randomized trials (SMARTs), in which participants are randomized multiple times in the course of the trial (Chakraborty & Moodie 2013; Kosorok & Moodie 2015; P. W. Lavori & Dawson 2014; Lei, Nahum-Shani, Lynch, Oslin, & Murphy 2012; Murphy 2005). Ultimately, the choice of a particular design should be driven by the scientific questions the intervention scientist seeks to answer. Hence, we highlight the scientific questions that can be addressed with each design approach. Concrete examples, motivated by the development of an adaptive intervention to improve social skills and academic engagement among children with autism, are provided to ground the discussion.

By describing four different types of experimental designs that can be used to answer different sets of questions to develop an optimized adaptive intervention, this chapter also accomplishes the following. First, the chapter showcases diversity in the types of experimental designs that investigators aiming to optimize an adaptive intervention can employ within the MOST optimization phase toolbox. This addresses a common misconception that all research related to optimizing an adaptive intervention requires a SMART; in fact, two of the experimental designs described in this chapter employ a single randomization, rather than sequential randomizations. Second, the chapter demonstrates how data from an optimization trial can help refine the tailoring in an adaptive intervention.

The remainder of this chapter is outlined as follows: In Sect. 2, we define adaptive interventions and provide a concrete example of an adaptive intervention. In this section and throughout the chapter, our examples center around developing an adaptive intervention for improving social skills and academic outcomes among children with autism spectrum disorders (ASD). Motivated by the concrete example, in Sect. 3, we list various questions researchers may have concerning the development of an adaptive intervention. In Sect. 4, we describe the design of two SRTs and two SMARTs, the rationale for each, and the specific question(s) each is designed to answer.

## 2 Adaptive Interventions

An adaptive intervention aims to provide a replicable guide for the type of sequential intervention decision-making that is often typical (or even necessary) in practice. In an adaptive intervention, baseline information, such as demographics, biomarkers, or baseline severity or risk, may inform initial intervention decisions. Then, baseline information as well as post-initial intervention information, such as changes in severity, risk, context, intervention engagement, or adherence, may be used to make subsequent intervention decisions.

The information used in an adaptive intervention to make intervention decisions is referred to as *tailoring variables*. Their use in an adaptive intervention is operationalized via a *decision rule*, one at each *decision point*, which links values of the tailoring variable(s) with a recommendation for one or more subsequent *intervention options*. Thus, an adaptive intervention is a protocol comprising a sequence of individualized intervention options.

As stated earlier, an adaptive intervention is a type of behavioral, biobehavioral, or biomedical intervention (L. M. Collins 2018). Collins (2018) defines a component as "any part of an intervention that can be separated out for study" (p. 49). In an adaptive intervention, components may be items related to the tailoring variables, decision points, intervention options, or decision rules that make up an adaptive intervention, such as the following: (1) The schedule (e.g., daily versus weekly) used to measure an individual's progress to inform subsequent intervention is an example of a component related to the tailoring variables; (2) in an adaptive intervention that triggers an intervention if insufficient improvement is made, time span over which improvement is determined based on weekly measurements of progress (e.g., at week 6 vs. at week 12) is an example of a component related to the decision points; (3) the type of intervention to provide a child who is not improving sufficiently (e.g., a peer-mediated social skills intervention vs. a parent-mediated social skills intervention) is an example of a component related to the intervention options; and (4) the threshold or levels of a tailoring variable used to decide which (set of) intervention option(s) to provide is an example of a component related to the decision rules. Note that each of these components could, itself, be multicomponent. This has to do with the issue of the granularity of components (L. M. Collins 2018, see section 2.5.1). For example, the parent-mediated social skills intervention may

involve various components, such as the frequency with which the therapist meets with the parent, the meeting place, or components related to the content that is taught to the parent (e.g., how to organize play dates).

An adaptive intervention approach is often necessary in settings where there is wide treatment effect heterogeneity. This includes settings where there is between-unit heterogeneity in treatment effects (e.g., what works for one unit may not work for the other) or there is within-unit heterogeneity in treatment effects over time (e.g., what works now for one unit may not work in the future for the same unit or vice versa). Adaptive interventions may also be useful in settings in which effective interventions (or intervention components) cannot be made available to the entire population of interest or cannot be made available to the population of interest for the whole course of their disorder because of resource constraints (e.g., monetary cost, time cost, burden).

From the point of view of the entity (or entities) providing the intervention, an adaptive intervention is a sequence of decision rules guiding the individualized sequencing of intervention. From the point of view of the unit (or units) experiencing the intervention, an adaptive intervention may be experienced as a sequence of interventions.

Importantly, an adaptive intervention is an intervention design, not an experimental design. Thus, an adaptive intervention typically does not involve randomization for the purpose of scientifically investigating intervention components. Rather, as we discuss below, randomization can be used in experimental studies that seek to develop optimized adaptive interventions. (Randomization can also be used to evaluate an adaptive intervention.)

Adaptive interventions have been described and discussed most commonly in a treatment domain (e.g., to guide how to initially treat, adapt, and readapt treatment for children diagnosed with mood disorders in a psychiatric clinic setting (Dawson, Lavori, Luby, Ryan, & Geller 2007; Gunlicks-Stoessel, Mufson, Westervelt, Almirall, & Murphy 2015)). Here, the intervention target is typically the patient, and the entity providing the intervention is the clinician. However, adaptive interventions are not exclusive to the treatment domain or the medical setting. Adaptive interventions can also be used in prevention, recovery, maintenance, education, health policy, operations management, or combinations of these.

Adaptive interventions are also known as "adaptive treatment strategies" (Dawson & Lavori 2008; Murphy 2005; Murphy, Lynch, Oslin, McKay, & Tenhave 2007; Oetting, Levy, Weiss, & Murphy 2011), "treatment algorithms" (Trivedi, Fava, Marangell, Osser, & Shelton 2006), or, in the statistical literature, "dynamic treatment regimens" (Chakraborty & Moodie 2013; Ertefaie, Wu, Lynch, & Nahum-Shani 2016; Laber, Lizotte, Qian, Pelham, & Murphy 2014; P. W. Lavori & Dawson 2014; Murphy & Bingham 2009; Orellana, Rotnitzky, & Robins 2010; Wang, Rotnitzky, Lin, Millikan, & Thall 2012; Zhao, Zeng, Laber, & Kosorok 2015) or "individualized decision rules" (Chakraborty & Murphy 2014). A popular, special case of an adaptive intervention is the "stepped care intervention model" (Bower & Gilbody 2005; Sobell & Sobell 2000), where a less intensive/costly intervention is provided first and more intensive/costly interventions are provided to individuals

who are not responding sufficiently. Adaptive interventions are more general, in that they may also include strategies that "step down" intervention; for example, they may consider maintenance interventions once sufficient response has been achieved, or they may include switching from one type of intervention to another that is parallel in cost or intensity. Another special case of an adaptive intervention is when there is a single stage of intervention (one intervention decision point, e.g., one of two medications for a new patient, a single stage of intervention), and that decision is based on history of medications or likelihood of side effects based on individual characteristics of the patient.

Next, we provide an example of an adaptive intervention delivered in a school setting. Its primary goal is to improve social skills outcomes among students with ASD. Later, this example will be used to motivate a discussion about various types of experimental designs for developing adaptive interventions.

## *2.1 An Example Adaptive Intervention: Social Skills Intervention in a School Setting for Children with Autism Spectrum Disorder*

*Background:* Social impairment is one of the core deficits for children with ASD. These children often experience isolation, peer rejection, and lack of friends (Kasari, Locke, Gulsrud, & Rotheram-Fuller 2011). This worsens with age (Rotheram-Fuller 2005) and leads to poor academic outcomes (Steedly, Schwartz, Levin, & Luke 2008). Yet few interventions address social impairment in school-aged children with ASD in the "natural" school environment. Including children with ASD in schools (i.e., inclusive schools)—where they may interact with typically developing children—is a necessary, but likely insufficient (Ochs, Kremer-Sadlik, Solomon, & Sirota 2001), first step to address social impairment. Children with ASD may benefit more from a school-based intervention approach that provides evidence-based interventions to accelerate development of social and academic engagement.

*Need for an Adaptive Intervention Approach:* Various evidence-based interventions exist for improving social skills in children with ASD, including peer-, classroom-, parent-, and school-targeted interventions (these are reviewed below); however, not all children will benefit from such interventions. In general, heterogeneity in the characteristics of children with ASD and in response to treatment undermines the effectiveness of such interventions. Peers, parents, classrooms, and schools are all also expected to respond heterogeneously to interventions that are ultimately directed to their children with ASD. School leaders, teachers, and therapists also expect that not all children will respond equally well to a specific intervention but often have no guidelines on when and how to modify an approach (Kasari & Smith 2013). Further, due to cost and potential time burden for school employees and parents, not all evidence-based interventions can be provided in all schools, at all times, to all children with ASD. This suggests the need for an individualized, adaptive intervention approach.

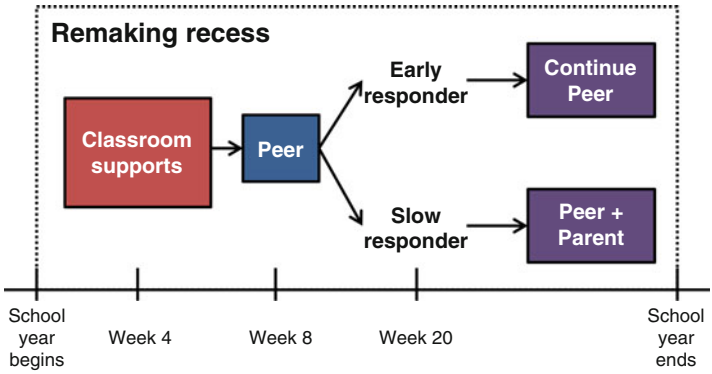


**Fig. 1** An example of adaptive intervention delivered in the school setting to improve academic outcomes among schoolchildren with autism spectrum disorder. Classroom supports intervention targets the classroom. Peer denotes peer-mediated social skills intervention, which targets the child's peers. Parent denotes parent-mediated social skills intervention, which targets the child's parents. The playground intervention, remaking recess, is provided to all children at the beginning of the school year. This is an intervention design, not an experimental trial design

Figure 1 shows a schematic of an example adaptive intervention for guiding how to individualize interventions in a school setting for children (ages 5–12) with ASD.

*Specific Intervention Goals:* The short-term goal of the adaptive intervention is to improve social engagement (i.e., reduce social impairment). The long-term goal of the adaptive intervention is to improve academic engagement, which, in school-aged children, can be affected by social impairment.

*Intervention Targets:* In order to improve the child's social engagement, there are four targets of intervention in this adaptive intervention: the child's peers, the child's classroom, the child's parent(s)/family, and the child's school playground.

*Intervention Options:* Remaking recess (RR; Kretzmann, Shih, & Kasari 2015), which targets the school playground, is a naturalistic, behavioral intervention that promotes social engagement during unstructured school times, such as recess and lunch breaks. RR includes a paraprofessional, employed by the school, who monitors and facilitates social interaction on the playground. The paraprofessional structures playground activities that seek to engage all children; importantly, this person is also responsible for monitoring the children for progress on the playground for purposes of deciding what intervention options to provide at later stages (see *Tailoring Variables* below). Classroom supports (CS; Lord & McGee 2001) targets the child's classroom. The main strategy comprising the CS intervention is to provide the teacher with skills related to developing visual supports with transition schedules in order to improve the behavioral regulation and classroom management of children. Peer-mediated intervention (Peer; Kasari & Patterson 2012) involves the use of a small group of typical peers from the target child's classroom (usually three peers). Peer targets the child's peers: the peers selected for each child with ASD are taught specific strategies for engaging that specific child on the playground.

Parent-assisted interventions (Parent) help parents support their children to develop friendships by teaching them how to facilitate social skills development in their children and teaching them how to host successful play dates (Frankel, Gorospe, Chang, & Sugar 2011). Parent, which targets the child's parent(s)/family, takes place in the home of the child with ASD. Most inclusive schools will have more than one child with ASD in the school and sometimes more than one child with ASD per classroom. Thus, RR and CS are considered "cluster-level" interventions, in the sense that an intact cluster of more than one child with ASD is potentially impacted by the intervention when the school or classroom, respectively, is targeted. On the other hand, in this adaptive intervention, Peer and Parent are considered "individual-level" interventions because only a single child is potentially impacted when the child's peers and parent(s)/family are targeted. In Peer, if there are multiple children with ASD in the same classroom, different peers are used for each target child. Concerning Parent, it is rare for a family to have more than one child with ASD at the same developmental age (i.e., in the same school). Further, play dates (one of the skills taught in Parent) are most often with typically developing peers.

*Tailoring Variables:* In this example, the intervention is tailored at the week 20 decision point. The tailoring variable—response status—used to make the week 20 intervention decision is collected by the paraprofessional who is part of the RR intervention. Using a "Clinical Global Impression" measure (CGI; Guy 1976), at week 20, the paraprofessional rates each child from 1 to 7 in terms of his/her improvement in peer engagement on the playground (1, *very much improved*; 2, *much improved*; 3, *minimally improved*; 4, *no change*; 5, *minimally worse*; 6, *much worse*; 7, *very much worse*). Based on prior literature and clinical expertise, children with CGI $< 3$ are identified as responders; children with CGI $\geq 3$ are identified as slower responders in need of a change in treatment.

*Intervention Points:* Intervention is delivered approximately over the course of a full school year. There are four intervention points in this example adaptive intervention: at the beginning of the school year and at 4, 8, and 20 weeks into the school year. These intervention points were selected, primarily, based on practical considerations. As stated above, in this example adaptive intervention, only one of these intervention points (the week 20 intervention point) involves a decision rule where treatment is tailored. Next, we discuss the intervention options provided at each intervention point. Intervention Point 1: At the first intervention point, which is at the beginning of the school year, all inclusive schools with children with ASD receive the playground intervention, RR. Intervention Point 2: At the second intervention point, which is at week 4, all classrooms with children with ASD receive CS. Teachers are often overburdened prior to and during the first 4 weeks of the school year; beginning at week 4 with CS provides sufficient time for teachers to settle into their new classes and be receptive to the CS intervention. Intervention Point 3: At the third intervention point, which is at week 8, all children with ASD receive Peer. Beginning this component at week 8 provides sufficient time for the paraprofessional to observe the typically developing children and the children with ASD on the playground and to select appropriate typical peers within the ASD child's classroom for the peer-mediated social skills intervention.

Beginning with Peer, an individual-level intervention, at week 8 also provides sufficient time to intervene prior to the holiday break.

The Decision Rule at Intervention Point 4: The fourth intervention point, which is at week 20 (after the holiday break), involves a decision rule. The intervention options provided at week 20 depend on whether the child is rated as a responder or slow responder (see above for definition). Children with ASD who are rated as responders stay the course with Peer, whereas those who are rated as slower responders are provided parent intervention (at home) in addition to Peer. The rationale for augmenting Peer with Parent for only children showing signs of slower response concerns (and not for all children) is that significant resources are needed to deliver Parent. By resources we mean the (time or monetary) cost to parents who must alter their schedules to be available for home visitations, as well as the cost to providers who would need to expend resources (e.g., clinicians) to intervene at both school and home.

As discussed above, only intervention point 4 involves a decision rule whereafter different intervention options are provided to different children, in this case depending on the child's week 20 response status. It is also possible to view intervention points 1, 2, and 3 as involving a special type of decision rule, in which the same intervention is assigned regardless of information known about the status of the child. Implicit in this are two key ideas: (1) that there exist (or could exist) other intervention options for schools at the beginning of the school year, for classrooms at week 4, and for children at week 8 and (2) that the adaptive intervention shown in Fig. 1 could be revised to use (or to collect and use) additional information about the school prior to the beginning of the school year, about the classroom prior to week 4, or about the child prior to week 8 to decide between these alternative intervention options.

Indeed, this chapter focuses on how adaptive interventions, such as the one shown in Fig. 1, could be optimized in such settings, including what kinds of questions behavioral intervention scientists might ask in such settings, and, therefore, what kinds of experimental studies behavioral intervention scientists could mount to address such questions about building an optimized adaptive intervention.

## 3 Unanswered Questions When Building an Optimized Adaptive Intervention

When developing an adaptive intervention, scientists often have important questions that cannot be answered based on existing (empirical, theoretical, practical) evidence. These unanswered questions typically concern the effectiveness of one or more components of an adaptive intervention, at one or more stages of the adaptive intervention. General examples include questions about how best to begin intervention for all individuals in a setting where it is has been decided (or it is known) how to adapt intervention in the future—whether to augment, intensify, or switch intervention for individuals who are non-responders (or non-adherers or non-engagers) to some initial intervention; the definition of non-response (e.g., which

cutoff should be used to determine non-response to the initial intervention); or the timing of the interventions (e.g., when or how frequently the decision points should be set).

Next, we describe three pairs of example questions motivated by the ASD adaptive intervention described above (Fig. 1). The scientific questions, which are not immediately evident from Fig. 1, center around alternatives to the intervention decisions at weeks 4, 8, and 20 in Fig. 1. Specifically, there is one pair of questions related to each of the week 4, week 8, and week 20 decision rules in the example adaptive intervention. In each pair the first question, denoted (i), is about a main effect (i.e., the average difference between two adaptive interventions); and the second question, denoted (ii), concerns candidate tailoring variables relevant for this effect (i.e., do certain factors moderate the effect of the adaptive intervention). Note that they are presented below in order of increasing complexity.

The following are questions related to the week 8 decision rule.

Question 1(i): Parent vs. Peer, on average? One question is whether, in the context of the adaptive intervention shown in Fig. 1, it is better to begin the individual-level intervention at week 8 with parent-mediated social skills intervention (Parent) rather than Peer. More specifically, "What is the effect on playground peer engagement of the adaptive intervention shown in Fig. 1 versus one that replaces Peer at week 8 with Parent?" This addresses the question of whether building social skills in the home via the parent is more important than building social skills at the school via the child's peers prior to deciding whether to provide combined Peer + Parent intervention. This is a critical question given the potential added cost or burden of intervening initially in the home (relative to the Peer intervention).

Question 1(ii): Parent vs. Peer, for whom? A second question concerns parent involvement or school attendance as candidate tailoring variables—variables that can inform the decision to offer Parent, rather than Peer, at week 8. Specifically, "Does the effect specified above (i.e., the difference between the adaptive intervention shown in Fig. 1 and an adaptive intervention that offers Parent instead of Peer at week 8) vary depending on the child's school attendance or parent involvement?" It may be that children with relatively lower parent involvement during the first 8 weeks of the school year benefit more from starting with Peer than starting with Parent at week 8 (because it would be more difficult to engage such parents in the parent-mediated social skills intervention). On the other hand, children with relatively lower school attendance during the first 8 weeks of the school year may benefit more from starting with Parent than starting with Peer at week 8 (because the child's peers have fewer opportunities to impact the child's social skills on the playground).

The following are questions related to the week 20 decision rule.

Question 2(i): Peer vs. Parent + Peer among slow responders? Another question is whether, among the slower responders to Peer, it is better to stay the course on Peer, rather than provide combined Parent + Peer. Specifically, "What is the effect

of the adaptive intervention shown in Fig. 1 versus one that stays the course on Peer regardless of the child's early response status?" It may be the case that, on average, children that are slower responders to Peer at week 20 simply need more time on Peer to see improved outcomes, indicating no need for Parent (which is costlier).

Question 2(ii): Peer vs. Parent+Peer, for different types of slow responders? A second question concerns social connections as a candidate tailoring variable—namely, as a variable that can inform the decision to add Parent, rather than to continue with Peer at week 20 for slow responders. Specifically, "Among slow responders, does the difference between adding Parent vs. continuing with Peer at week 20 vary depending on the extent to which a slow responding child developed additional social connections during the first 20 weeks of school?" It may be that children who are identified as slower responders and developed no additional, sustained social connections on the playground as a result of Peer are more likely to benefit from adding Parent versus those who developed any social connections.

The following are questions related to the week 4 decision rule.

Question 3(i): Classroom Supports, on average? Another question is whether, in the context of the adaptive intervention shown in Fig. 1, there is any evidence to support including the CS intervention. Specifically, "What is the effect of the adaptive intervention shown in Fig. 1 versus one without the CS intervention?" It may be the case that social skills and academic engagement outcomes are similar, on average, regardless of whether children receive CS. This may be important given the already busy schedules of teachers and the added cost to school administrators of implementing CS in every inclusive classroom.

Question 3(ii): Classroom Supports, for different types of classrooms? A second question concerns the classroom's inclusion environment as a candidate tailoring variable—a variable that can inform the decision of whether to intervene with CS at week 4. Specifically, "Does the effect of intervening in classrooms with CS versus not intervening in classrooms with CS vary depending on the extent to which the classroom is identified as being more or less inclusive during the first four weeks of the school year?" It may be that classrooms with a less inclusive environment benefit more from CS, whereas classrooms with a more inclusive environment do not benefit as much from CS.

## 4 Experimental Designs for Building Effective Adaptive Interventions

This section uses a case-study approach. We present four design examples, two employing a SRT and two employing a SMART. All four designs are motivated by one or more of the hypothetical scientific questions listed above concerning

the example adaptive intervention shown in Fig. 1 for improving social skills in schoolchildren with ASD. For each trial design, we present the following:

- Brief introduction related to this type of trial design;
- Which scientific questions the trial is intended to answer, from the list above;
- Schematic and description of the flow of the trial;
- Analytic comparisons associated with each question;
- Brief discussion about the trial design; and
- Review of the literature related to this type of trial design (if applicable).

In each of the four trial designs presented below, we presume all children are exposed to the remaking recess (RR) playground intervention at the beginning of the school year. In other words, this intervention component is not investigated in any of the trials below.

## *4.1 Example 1: A Two-Arm SRT*

*Introduction:* Singly randomized trials (SRTs) are trials where units are randomized only once. The simplest case is a two-arm SRT, such as the first example presented here (Fig. 2). In this example, each of the two arms is an adaptive intervention; the two adaptive interventions differ only in terms of a first-stage component (i.e., Peer vs. Parent at week 8). In SRTs that are more complex than the one described here, the single randomization may be among a subgroup of individuals/units (see Example 2), or the randomization may be to three or more arms (e.g., component factors with more than two levels). In the case where multiple components are being investigated (each with two or more levels), the single randomization may be a factorial randomization (L. M. Collins 2018, see Chapters 5 and 8).

*Scientific Questions Motivating the Design:* This trial addresses the first pair of questions in the previous section, concerning the week 8 decision rule. For convenience, we repeat the questions here.

Question 1(i): Parent vs. Peer, on average? "What is the effect on playground peer engagement of the adaptive intervention shown in Fig. 1 versus one that replaces Peer at week 8 with Parent?"

Question 1(ii): Parent vs. Peer, for whom? "Does the effect specified above (i.e., the difference between the adaptive intervention shown in Fig. 1 and an adaptive intervention that offers Parent instead of Peer at week 8) vary depending on the child's school attendance or parent involvement?"

*Schematic and Flow:* See Fig. 2. Beginning at week 4, all children are exposed to the CS intervention. At week 8, children are randomized with equal probability to peer- versus parent-mediated social skills intervention (Peer vs. Parent). At week 20, all children are assessed for early versus slow response status. Children identified as early responders stay the course on their initially assigned intervention (Peer or Parent). Children identified as slower responders are provided both Peer and Parent (Peer+Parent).

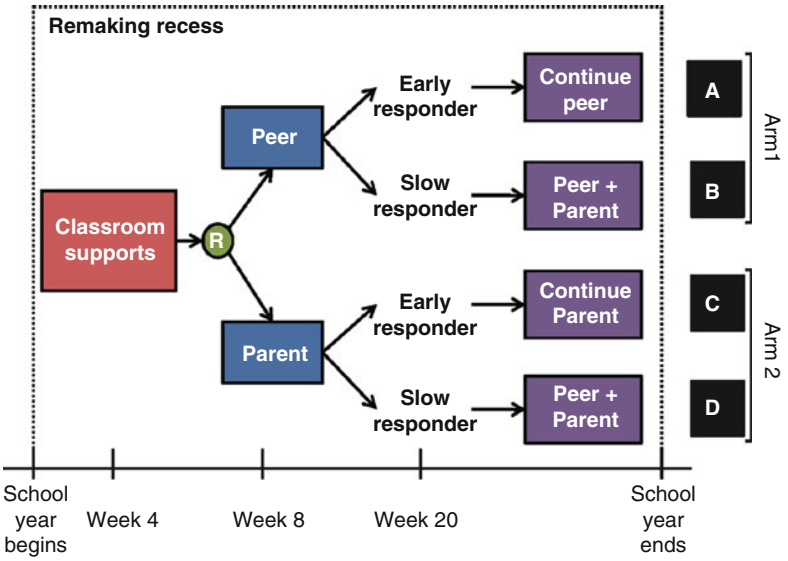


**Fig. 2** A two-arm singly randomized trial for developing an adaptive intervention to improve academic outcomes among schoolchildren with autism. Classroom supports intervention targets the classroom. Peer denotes peer-mediated social skills intervention, which targets the child's peers. Parent denotes parent-mediated social skills intervention, which targets the child's parents. The playground intervention, remaking recess, is provided to all children at the beginning of the school year. A circled "R" denotes randomization. The adaptive interventions given by cells A+B (Arm 1 in this trial) comprise the adaptive intervention shown in Fig. 1

*Comparisons:* To describe the analytic comparisons associated with each of the scientific questions this design is intended to answer, we refer to cells A–D in Fig. 2, which denote the *intervention-response pathways* a child may experience. Question 1(i) compares the two adaptive interventions embedded in this trial. Children assigned to A or B are consistent with an adaptive intervention that offers Peer at week 8, whereas children assigned to C or D are consistent with an adaptive intervention that offers Parent at week 8. Both adaptive interventions offer CS to all children at week 4, examine slow response at week 20, assign children who are early responders to stay the course, and provide Peer + Parent to slower responders. To address question 1(i), outcomes at the end of the study (or change in outcomes from week 8 onward) are compared between children assigned to cell A or B and children assigned to cell C or D. To address question 1(ii), outcomes are compared between cells A+B and C+D for children with different values of school attendance and parent involvement. These comparisons correspond to a standard moderation analysis.

*Discussion:* The trial design shown in Fig. 2 is a two-arm randomized trial, where each arm is an adaptive intervention; cells A+B represent one adaptive intervention, and cells C+D represent a second adaptive intervention. Thus, standard data analytic methods for multi-arm randomized trials can be used to analyze the above-listed

questions using data arising from such a trial. However, note that in this example, the two-arm SRT is not a "standard" randomized controlled trial (RCT), in the sense that there is not a "business as usual" control group (i.e., neither of the two adaptive interventions represents the current standard used in schools to improve social and academic outcomes in children with ASD).

## 4.2 Example 2: An Enhanced, Non-responder SRT

*Introduction:* The second example trial design (Fig. 3) is an *enhanced non-responder SRT*. It is an SRT because there is a single randomization (at week 20). It is a non-responder trial because the single randomization is among a subset of study participants characterized as non-responders (referred to as slow responders in our examples). In standard non-responder randomized trials, often only non-responders to previous intervention are recruited to participate in the study, consented, and randomized to the subsequent intervention options—responders often do not participate in the study. See Prasad (2007) for a brief discussion of standard non-responder trials of pharmacological agents in cardiovascular disease. (Standard responder trials are similar, except the focus is on recruiting, consenting, and randomizing responders.) By contrast, in an enhanced non-responder trial, such as the one shown in Fig. 3, individuals are recruited and consented at the start of intervention (e.g., at the beginning of the school year), and both responders and

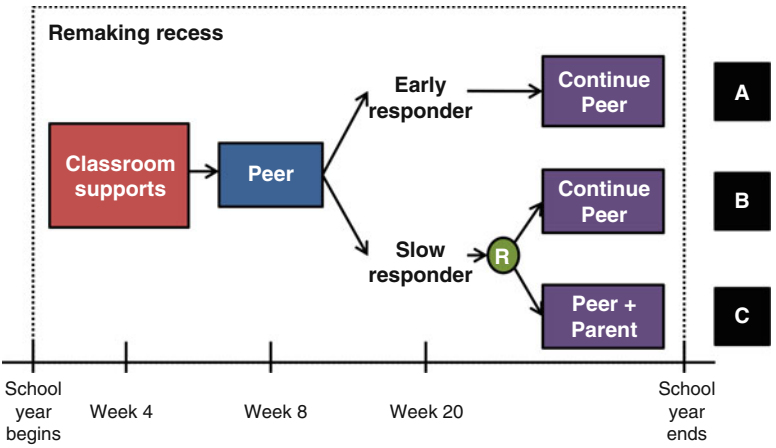


**Fig. 3** An enhanced non-responder trial for developing an adaptive intervention to improve academic outcomes among schoolchildren with autism spectrum disorders. Classroom supports intervention targets the classroom. Peer denotes peer-mediated social skills intervention, which targets the child's peers. Parent denotes parent-mediated social skills intervention, which targets the child's parents. The playground intervention, remaking recess, is provided to all children at the beginning of the school year. A circled "R" denotes randomization

non-responders are followed over all stages of the intervention. Non-responder trials are common in the behavioral intervention sciences; below we provide a review of standard non-responder trials (and analogous "responder trials").

*Scientific Questions Motivating the Design:* This trial addresses the second pair of questions in the previous section, concerning the week 20 decision rule. For convenience, we repeat the questions here:

Question 2(i): Peer vs. Parent+Peer, among slow responders? "What is the effect of the adaptive intervention shown in Fig. 1 versus one that stays the course on Peer regardless of the child's early response status?"

Question 2(ii): Peer vs. Parent+Peer, for different types of slow responders? "Among slow responders, does the difference between adding Parent vs. continuing with Peer at week 20 vary depending on the extent to which a slow responding child developed additional social connections during the first 20 weeks of school?"

*Schematic and Flow:* See Fig. 3. Beginning at week 4, all classrooms are provided the CS intervention. At week 8, all children are provided Peer. At week 20, all children are assessed for response status. All children identified as early responders stay the course on Peer (the initially assigned intervention). Children identified as slower responders to Peer are randomized with equal probability to stay the course versus Peer+Parent intervention.

*Comparisons:* Question 2(i) compares the two interventions embedded in this trial. Children assigned to A or B receive a fixed intervention that offers Peer at week 8 and remains on Peer regardless of early or slower response, whereas children assigned to A or C are assigned to an adaptive intervention that offers Peer at week 8 and then, at week 20, stays the course on Peer for early responders and provides Peer+Parent for slower responders. To address question 2(i), outcomes at the end of the study (or change in outcomes from week 20 onward) are compared between children assigned to A or B versus A or C. The analysis associated with this comparison is an easy-to-implement weighted analysis because children assigned to A are shared between the two interventions being compared and they are also overrepresented (slower responders in cell B or C have 1/2 the chance of being represented in each of the two interventions). To accommodate these design considerations, the data for slower responders are upweighted relative to those for early responders in the corresponding regression analysis. The weights here are known by design (specifically, slower responders are given a weight of 2, whereas early responders are given a weight of 1); see Nahum-Shani et al. (2012a) for details.

To address question 2(ii), outcomes are compared between cells B and C for children with different values of social connectedness on the playground during the first 20 weeks of the school year. A straightforward data analysis associated with this comparison is a standard moderation analysis where the data are restricted to slow responders (i.e., only slow responders are included in this analysis).

*Discussion:* As mentioned above, children in cell A or B receive a fixed intervention (L. M. Collins 2018, Section 1.7), that is, an intervention that is not adaptive. Children in this intervention are offered Peer at week 8 and remain on

Peer regardless of their response status. On the other hand, children in cell A or C receive an adaptive intervention. Hence, the comparison associated with question 2(i) is a contrast of a fixed versus an adaptive intervention. As such, 2(i) evaluates all of the components making up the second decision rule as a package: the choice of week 20 as a second decision point, the use of response status at week 20 as a tailoring variable, and the choice of interventions offered for early responders and slow responders.

Recall that in a standard non-responder SRT, often only non-responders to previous intervention are recruited to participate in the study, consented, and randomized to the subsequent intervention options. Both the standard and enhanced non-responder SRTs may yield data that can inform the effect of staying the course on Peer versus adding Parent among slower responders. However, the enhanced non-responder trial design has at least five potential advantages over the standard non-responder trial design. First, there is a potential for selection bias with the standard non-responder trial because in many applications participants are recruited/consented *after* they have become a non-responder (by contrast, in an enhanced non-responder trial, participants are consented to be part of the study prior to any intervention). Thus, participants in an enhanced non-responder trial should be a better representation of the non-responders that will be part of an intervention in the field.

Second, generalizability of the results of a non-responder trial may be compromised in some applications of the standard non-responder trial because other aspects of the adaptive interventions prior to or after becoming a non-responder are not well operationalized. For example, in studies where the earlier stage treatments are observed rather than provided, unclear criteria might be used for what constitutes earlier stages of intervention, or, in standard non-responder trials where recruitment/consent takes place after non-response, the timing of the transition between the two stages of intervention may be unclear or not well operationalized.

Third, standard non-responder trials may have limited ability to investigate certain important time-varying moderators (type 2(ii) questions). For example, information about social connectedness on the playground during the first 20 weeks of the school year or adherence or engagement to interventions prior to becoming a slower responder at week 20—which may be of interest in investigating who may benefit more or less from staying the course with Peer versus Peer+Parent—may be unavailable or imprecise. Because enhanced non-responder trials collect data over the entire course of the intervention, these data will usually be available for analysis.

Fourth, the enhanced design provides an opportunity to estimate the difference in mean outcome under each of the two adaptive interventions, whereas (when there is no selection bias) the standard design only ensures an opportunity to estimate the mean outcome under the subsequent intervention options for non- (or slow) responders. This is because, for some of the reasons provided above, the rate of non-response *at week 20 following RR, CS, and Peer interventions* may be unavailable or imprecisely estimated in standard non-responder trials, and this rate is necessary for estimating the difference in mean outcome between the two adaptive interventions (see Appendix for more details). If the focus is on testing whether the difference is

zero, the non-response rate is not necessary, but the assumption of no selection bias remains critical.

Fifth, the enhanced design would allow for secondary (noncausal) analyses about the time course of the responders; these analyses may be used to design future studies on adaptive interventions. For example, in the design in Fig. 3, investigators might observe how well early responders maintain their response while receiving Peer. If at some point there is a reduction in social connectedness on the playground for some children, such observations may help investigators develop new intervention strategies for further improving or maintaining the social connections on the playground. Such observations are not possible in a standard non-responder trial.

The primary limitation of the enhanced non-responder trial (relative to the standard non-responder trial) is the additional complexity and cost of the trial (e.g., the added cost of the longer study duration and of measuring outcomes for the individuals in cell A).

*Review of (Non-)responder Trials in Behavioral Intervention Science:* A number of studies have utilized standard non-responder trials to study how to improve long-term outcomes among youth with mental health disorders who do not respond sufficiently to a first-stage course of intervention. One example is the adolescent depression antidepressant and psychotherapy trial (Goodyer et al. 2007). This was a pragmatic, non-responder clinical trial that compared selective serotonin reuptake inhibitor (SSRI) medication versus the combination of SSRI plus cognitive behavioral therapy (CBT) among depressed youth attending routine child and adolescent mental health service centers in the UK who had not responded to an initial course of brief psychosocial intervention. A second example is the treatment of SSRI-resistant depression in adolescents (TORDIA) trial, in which adolescents with depression who did not respond to an adequate course of SSRI were randomized to switch to another SSRI or to venlafaxine medication, with or without CBT (Brent et al. 2008). A third example is the pediatric obsessive-compulsive disorder (OCD) treatment study II (POTS II) study (Freeman et al. 2009), which compared two forms of second-stage OCD-specific CBT among youth with OCD who were partial responders to first-stage treatment with selective serotonin reuptake inhibitor (SSRI) medication. In a fourth example, a group of Scandinavian investigators compared continued CBT versus a switch to sertraline medication among youth with OCD who were non-responders to 14 weeks of first-stage CBT (Skarphedinsson et al. 2015).

A number of other studies have utilized standard responder trials to study how to maintain early improvements over the longer term (or prevent subsequent relapse) among individuals who respond to an initial course of intervention. In one example, Emslie et al. (2004) examined the impact of discontinuing medication among children and adolescents with major depression who had an adequate response after 12 weeks of fluoxetine medication. In a second example, Kennard et al. (2008) examined the feasibility and acceptability of continued antidepressant medication versus medication plus CBT among youth with major depressive disorder who had responded to first-line medication. A third example is a study

that examined the effect of discontinuing versus continuing risperidone medication (an antipsychotic) among children with ASD with severe disruptive behaviors who had already responded to an initial 8-week course of medication (McCracken et al. 2002). A fourth example is a weight loss maintenance study (Svetkey et al. 2008) that compared a personal-contact intervention; an interactive, technology-based intervention; and a self-directed intervention for sustaining weight loss among obese or overweight individuals who successfully lost a significant amount of weight during a first-stage standard behavioral weight loss intervention.

## 4.3 *Example 3: A Prototypical SMART Design*

*Introduction:* The third example trial design is a SMART. SMART designs were developed explicitly for the purpose of answering multiple open questions (at

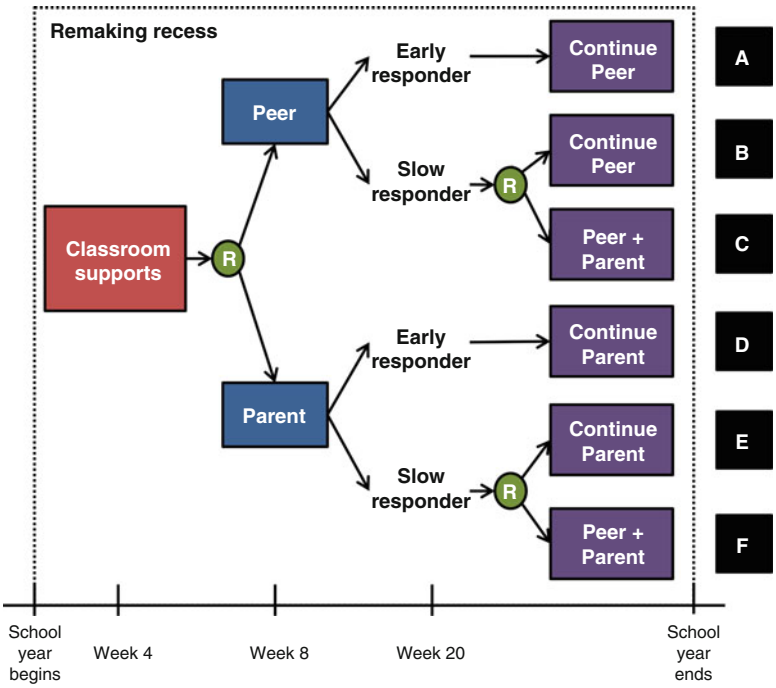


**Fig. 4** A sequential, multiple assignment, randomized trial (SMART) for developing an adaptive intervention to improve academic outcomes among schoolchildren with autism. Classroom supports intervention targets the classroom. Peer denotes peer-mediated social skills intervention, which targets the child's peers. Parent denotes parent-mediated social skills intervention, which targets the child's parents. The playground intervention, remaking recess, is provided to all children at the beginning of the school year. A circled "R" denotes randomization

multiple intervention stages) when building an optimized adaptive intervention (P. Lavori & Dawson 2000; Murphy 2005). The SMART design presented in Fig. 4 is considered a "prototypical SMART," in that it has become one of the most popular types of SMART designs: it employs two stages of randomization, where all individuals are randomized initially to two stage 1 intervention options and non-responding individuals are randomized subsequently to two stage 2 intervention options. However, SMART designs can have various forms, including SMARTs that do not restrict subsequent randomizations based on response to previous intervention (Example 4 is like this), SMARTs that re-randomize only responders, and SMARTs that randomized both responders and non-responders.

*Scientific Questions Motivating the Design:* This trial addresses both the first and second pairs of questions in the previous section. For convenience, we repeat the questions here:

Question 1(i): Parent vs. Peer, on average? "What is the effect on playground peer engagement of the adaptive intervention shown in Fig. 1 versus one that replaces Peer at week 8 with Parent?"

Question 1(ii): Parent vs. Peer, for whom? "Does the effect specified above (i.e., the difference between the adaptive intervention shown in Fig. 1 and an adaptive intervention that offers Parent instead of Peer at week 8) vary depending on the child's school attendance or parent involvement?"

Question 2(i): Peer vs. Parent+Peer, among slow responders? "What is the effect of the adaptive intervention shown in Fig. 1 versus one that stays the course on Peer regardless of the child's early response status?"

Question 2(ii): Peer vs. Parent+Peer, for different types of slow responders? "Among slow responders, does the difference between adding Parent vs. continuing with Peer at week 20 vary depending on the extent to which a slow responding child developed additional social connections during the first 20 weeks of school?"

This trial also addresses additional questions concerning two fixed interventions. One of them provides Parent at week 8 and stays the course regardless of response at week 20, and the second one provides Peer at week 8 and stays the course regardless of early or slower response at week 20.

*Schematic and Flow:* See Fig. 4. Participants in this trial may be randomized twice, at week 8 and at week 20. Beginning at week 4, all children are given the CS intervention. At week 8, children are randomized with equal probability to Peer or Parent. At week 20, all children are assessed for response. Children identified as early responders stay the course on their initially assigned intervention (Peer or Parent). Children identified as slower responders are re-randomized with equal probability to stay the course versus Peer+Parent.

The SMART in Fig. 4 has four sequential intervention strategies embedded within it. Two of these interventions are adaptive, and two are fixed. These are shown in Table 1. Note that the two fixed interventions provide the same intervention at week 8 and at week 20—they do not adjust the intervention based on the response status at week 20. For example, the "Always Peer" intervention provides CS at week 4 and provides Peer from week 8 to the end of the school year, regardless of response status at week 20. Similarly, the "Always Parent" intervention provides CS at week 4

**Table 1** The four sequences of intervention embedded in the example prototypical SMART in Fig. 4. The two intervention sequences labeled with a dagger (†) are fixed interventions, whereas the other two are adaptive interventions. The playground intervention, remaking recess—which is provided to all children at the beginning of the school year—is not shown in the Table

| Intervention sequence | Intervention decision at week 4 | Intervention decision at week 8 | Intervention decision at week 20 | Cells (Fig. 4) |
|---|---|---|---|---|
| "Always Peer" † | CS | Peer | Peer | A+B |
| "Always Parent" † | CS | Parent | Parent | D+E |
| "Peer First" | CS | Peer | Early responders get Peer<br>Slower responders get Peer+Parent | A+C |
| "Parent First" | CS | Parent | Early responders get Parent<br>Slower responders get Peer+Parent | D+F |

and provides Parent from week 8 to the end of the school year regardless of response status at week 20.

*Comparisons:* Question 1(i) is a comparison of children assigned to A or C versus those assigned to D or F. Question 2(i) is a comparison of children assigned to A or C versus those assigned to A or B. Because this trial has four (rather than two) embedded sequences of intervention, additional questions are possible. For example, with this SMART, it is possible to compare adaptive interventions D and F to D and E. (Intervention sequences D and F were not available in any of the other designs.) In addition, it is possible to examine how the week 8 and week 20 intervention decisions interact with each other—how Peer vs. Parent at week 8 interacts with staying the course versus Peer+Parent at week 20 among slower responders.

*Discussion:* As in the enhanced non-responder trial, the analysis associated with the comparison of the four embedded adaptive interventions requires an easy-to-implement weighted regression. Here, children assigned to A or D are shared between multiple adaptive interventions; these children are also overrepresented relative to slower responders in cells B, C, E, and F. To accommodate these design considerations, slower responders are given twice the weight of early responders in the corresponding regression analysis. Again, see Nahum-Shani et al. (2012a) for details.

The SMART can also be used to investigate candidate tailoring variables for both week 8 and week 20 intervention options in the context of the four embedded adaptive interventions. This includes the ability to investigate questions 1(ii) and 2(ii) and others involving the fixed interventions "Always Peer" and "Always Parent."

*Review of SMART Designs in Behavioral Intervention Science:* SMARTs have become popular over the past 10 years. They have been (and are currently being) used to address important questions in the development of adaptive interventions across a wide spectrum of the behavioral, biobehavioral, biomedical, and

educational sciences. In oncology, SMART designs have been used to develop adaptive interventions for prostate cancer (Kidwell 2014; Thall, Wooten, Logothetis, Millikan, & Tannir 2007; Wang et al. 2012). Lei et al. (2012) provided an excellent review of four example SMART studies to develop behavioral adaptive interventions, one each in ASD and ADHD and two in adult substance use; this article includes a description of each SMART and the types of scientific questions they were designed to answer. The Clinical Antipsychotics Trial of Intervention Effectiveness (Lieberman et al. 2005; Shortreed & Moodie 2012) and the Sequenced Treatment Alternatives to Relieve Depression (P. W. Lavori et al. 2001; Rush et al. 2004) studies are examples of early precursors to the SMART in adult mental health research. Recently, there has been increased interest in adaptive interventions and SMART designs in child and adolescent mental health (Almirall & Chronis-Tuscano 2016; Kasari et al. 2014).

The Methodology Center at Penn State University hosts a web page with an updated list of SMART designs across a wide range of the behavioral, biobehavioral, and educational sciences (Methodology Center 2016), including studies in weight loss (Naar-King et al. 2015; Sherwood et al. 2016; Spring & Nahum-Shani 2016) and smoking cessation (Joseph 2016).

## *4.4 Example 4: A Clustered, Non-restricted SMART Design*

*Introduction:* The fourth example trial design (Fig. 5) is also a SMART because it has two randomizations, one at week 4 and another at week 8. This example demonstrates how a SMART can be designed to optimize sequences of intervention options at multiple levels of intervention, here, at both a cluster level (the classroom) and an individual level. Accordingly, unlike the SMART design in the third example (Fig. 4), which randomizes only at the individual level, the SMART in this fourth example (Fig. 5) randomizes to intervention options *both* at the classroom level and at the level of the individual child with ASD within the classroom.

*Scientific Questions Motivating the Design:* This trial addresses the first and third pairs of questions in the previous section. For convenience, we repeat the questions here:

Question 1(i): Parent vs. Peer, on average? "What is the effect on playground peer engagement of the adaptive intervention shown in Fig. 1 versus one that replaces Peer at week 8 with Parent?"

Question 1(ii): Parent vs. Peer, for whom? "Does the effect specified above (i.e., the difference between the adaptive intervention shown in Fig. 1 and an adaptive intervention which offers Parent instead of Peer at week 8) vary depending on the child's school attendance or parent involvement?"

Question 3(i): Classroom Supports, on average? "What is the effect of the adaptive intervention shown in Fig. 1 versus one without the CS intervention?"

Question 3(ii): Classroom Supports, for different types of classrooms? "Does effect of intervening in classrooms with CS versus not intervening in classrooms with

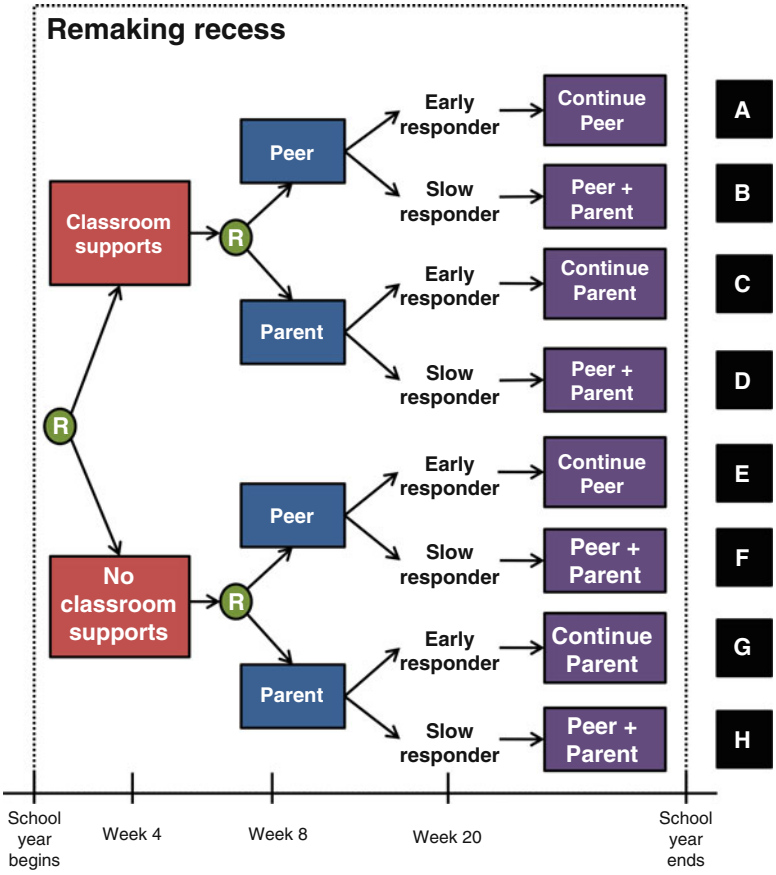


**Fig. 5** A cluster-randomized, sequential, multiple assignment, randomized trial (SMART) for developing an adaptive intervention to improve academic outcomes among schoolchildren with ASD. Classroom supports intervention targets the classroom. Peer denotes peer-mediated social skills intervention, which targets the child's peers. Parent denotes parent-mediated social skills intervention, which targets the child's parents. The playground intervention, remaking recess, is provided to all children at the beginning of the school year. A circled "R" denotes randomization

CS vary depending on the extent to which the classroom is identified as being more or less inclusive during the first four weeks of the school year?"

This trial also addresses additional questions concerning two adaptive interventions that, at week 8, provide Parent instead of Peer.

*Schematic and Flow:* See Fig. 5. Beginning at week 4, classrooms are randomized with equal probability to receive or not receive the CS intervention. At week 8, all children in all classrooms (regardless of whether CS is provided or not) are randomized with equal probability to Peer versus Parent. At week 20, all children are assessed for early versus slow response status. All children identified as early

**Table 2** The four adaptive interventions embedded in the example cluster-randomized SMART in Fig. 5. The playground intervention, remaking recess—which is provided to all children at the beginning of the school year—is not shown in the Table

| Adaptive intervention | Intervention decision at week 4 | Intervention decision at week 8 | Intervention decision at week 20 | Cells (Fig. 5) |
|---|---|---|---|---|
| "CS, Peer" | CS | Peer | Early responder get Peer<br>Slower responder get Peer+Parent | A+B |
| "No CS, Peer" | No CS | Peer | Early responder get Peer<br>Slower responder get Peer+Parent | C+D |
| "CS, Parent" | CS | Parent | Early responder get Parent<br>Slower responder get Peer+Parent | E+F |
| "No CS, Parent" | No CS | Parent | Early responder get Parent<br>Slower responder get Peer+Parent | G+H |

responders stay the course on their initially assigned intervention (Peer or Parent). Children identified as slower responders are offered Peer+Parent intervention.

The SMART in Fig. 5 has four adaptive interventions embedded within it. These are shown in Table 2. In all of them, children who are early responders at week 20 stay the course, whereas children who are slower responders at week 20 are provided Peer+Parent. The four adaptive interventions differ in terms of whether or not classrooms are provided with CS at week 4 and their initial individual-level intervention (Peer vs. Parent) at week 8.

*Comparisons:* To address question 1(i), outcomes at the end of the study (or change in outcomes from week 8 onward) are compared between children in cell A or B and children in cell C or D. To address question 1(ii), outcomes are compared between children in cell A or B and those in C or D for children with different values of school attendance and parent involvement.

To address question 3(i), outcomes at the end of the study (or change in outcomes from week 4 onward) are compared between children in cell A or B and children in cell E or F. To address question 3(ii), outcomes are compared between children in cell A or B and those in cell E or F for children with different values on the extent to which the classroom is more or less inclusive during the first 4 weeks of the school year.

Questions 1(i)–(ii) concern the effect of Peer vs. Parent at week 8 in a setting where CS is provided to all classrooms (A and B vs. C and D), whereas questions 3(i)–(ii) concern the effect of CS vs. no CS in a setting where Peer is provided to all children at week 8. However, in this SMART, because both factors are varied (CS vs. no CS at week 4, as well as Peer vs. Parent at week 8), additional questions are possible. For example, it is possible to examine how the week 4 and week 8 intervention decisions interact with each other: how Parent vs. Peer at week 8 interacts with previous exposure to CS vs. not at week 4.

*Discussion:* In addition to being different from the SMART in Fig. 4 because this trial includes cluster randomization, this example SMART also differs in another respect: in this cluster-randomized SMART, none of the randomizations are "restricted," whereas in the SMART design in the third example, the second randomization was restricted to only slower responders. (In other words, in the trial in Example 3, the second randomization is conditional on response status.) The decision to "restrict" randomizations in a SMART is based on ethical, scientific, or practical/feasible considerations having to do with whether particular intervention options at particular stages ought to be or ought not to be considered for certain subgroups. For example, in the third example (Fig. 4), due to its cost, Peer+Parent was deemed not to be a feasible option for early responders at week 20; for this reason, the SMART in Fig. 4 considers randomizing only slower responders at week 20 to Peer+Parent. By contrast, in the SMART in Fig. 5, it was feasible at week 4 for all classrooms to be provided or not provided CS (and this was an interesting scientific question), and therefore all classrooms were randomized. Similarly, in both of the SMART designs presented (Figs. 4 and 5), at week 8, it was feasible for all children with ASD within all classrooms to be provided Peer or Parent (and this was an interesting question), and therefore in these two designs, all children were randomized at week 8.

*Review of Cluster-Randomized SMART Designs in Behavioral Intervention Science:* Research on optimizing multilevel (or cluster-level) adaptive interventions—and, especially, the use of cluster-randomized SMARTs for their optimization—is not as common as research on optimizing individual-level adaptive interventions.

Here, we review two studies in the field of implementation science, which (for the most part) concerns the development and evaluation of multilevel interventions to promote the integration of evidence-based individual-level interventions into authentic clinical practice settings, such as hospitals or clinics. Such multilevel interventions, which are known as implementation interventions, are designed to address barriers that impede the adoption of evidence-based practices in clinical practice settings, with the ultimate goal of improving individual-level (patient) health outcomes. The first study (Kilbourne et al. 2013; Kilbourne, Almirall, Goodrich, et al. 2014) is a SRT to develop an optimized adaptive implementation intervention to improve the adoption of an outreach program for patients with a diagnosis of serious mental illness in VA facilities nationwide. The study design was an enhanced non-responder trial, much like the study shown in Fig. 3, but with the randomization at the level of the VA facilities.

The second study (Kilbourne, Almirall, Eisenberg, et al. 2014) is a cluster-randomized SMART, still in the field collecting data as of this writing, to develop an optimized adaptive implementation intervention designed to improve the uptake/adoption of a brief, evidence-based intervention for mood disorders in community mental health settings across Colorado and Michigan. This SMART seeks to understand when and for which sites to utilize a relatively costly internal facilitator component among sites having difficulty adopting the evidence-based intervention.

## 5 Discussion

Motivated by the development of an adaptive intervention in an authentic school setting for improving social skills and academic engagement among children with ASD, this chapter provides a review of adaptive interventions and the types of questions investigators often have in the process of building these interventions. The four examples of experimental designs provided here demonstrate how various trial design tools can be used in the optimization phase of MOST to address different types of open scientific questions concerning the construction of an adaptive intervention. As intervention scientists increasingly acknowledge the critical role the optimization phase plays in the process of developing high-quality multicomponent interventions, we expect the use of these (and other) experimental designs to grow accordingly.

The four hypothetical trial examples provided in this chapter highlight the various ways experiments can be designed to optimize an adaptive intervention. Experimental designs for optimization of adaptive interventions may be singly randomized or sequentially randomized; randomizations may be restricted among a subgroup of participants (including on the basis of response to prior intervention) or unrestricted, and randomizations may be at the individual level or a cluster level.

Ultimately, the design of the experiment should be motivated by the scientific questions the investigator wishes to answer. For example, the design in Fig. 2 focuses on which individual-level component (Peer or Parent) to offer at week 8 in the context of an intervention that subsequently adapts the individual-level intervention at week 20. This trial design does not address questions about the effectiveness of Peer + Parent among slower responders at week 20 or about the classroom-level intervention component (CS) at week 4. This study also does not address questions about the school-level intervention component (RR). For a scientist utilizing the design in Fig. 2, such questions are not of scientific interest, there is previous evidence (e.g., clinical, theoretical, or practical) suggesting that these questions need not be addressed via randomization, or the questions are of scientific interest but are to be examined in a future study. As an example of a practical consideration, suppose that the rate of slow response at week 20 is very small; in such a setting, it may not be practical to randomize slower responders to two intervention options at week 20. The SMART design is a special case of the factorial experimental trial design (Almirall et al. 2014; L. M. Collins 2018, see Chapter 3). This idea is most easily appreciated by observing how the SMART in Example 3 (Fig. 4) effectively crosses the single $2^1$ randomization in Example 1 (Fig. 2) with the single $2^1$ randomization in Example 2 (Fig. 3), leading to the $2^2 = 4$ embedded adaptive interventions described in Table 1. In a similar way, the SMART in Example 4 (Fig. 5) crosses the single $2^1$ randomization that investigates whether or not to employ the CS component with the single $2^1$ randomization in Example 1 (Fig. 2), leading to the four embedded adaptive interventions described in Table 2.

Note that, for simplicity, in all of the example trials presented in this chapter (including the SMART designs), each randomization involves a single factor with two levels (i.e., in all example trials, each randomization was a $2^1$-way randomization). Some investigators, however, may be interested in investigating a single factor with three levels at a particular stage of intervention, necessitating a $3^1$-way randomization at that stage of intervention. Other investigators may be interested in investigating multiple factors (e.g., screening multiple intervention components)—say, $p > 1$ factors—each with two levels at a particular stage of intervention, necessitating a $2^p$-way randomization at that stage of intervention.

In addition, for simplicity, the two SMART designs presented in this chapter only considered randomization at two stages. However, some investigators may be interested in addressing important questions at more than two stages of an adaptive intervention. Indeed, if justified by the science, the two SMARTs in Figs. 4 and 5 could be combined into one, three-stage sequentially randomized trial. At this point in research on adaptive interventions using sequentially randomized trials, randomizations at more than two stages remain rare.

The focus of this chapter has been on experimental designs for the optimization phase of MOST, specifically for optimization of adaptive interventions, and not experimental designs for the preparation or evaluation phases (L. M. Collins 2018). For preparation, intervention scientists may conduct pilot studies that focus on feasibility or acceptability considerations in the development of an adaptive intervention (Almirall, Compton, Gunlicks-Stoessel, Duan, & Murphy 2012; Kim, Ionides, & Almirall 2016) or analyses using data from existing observational or experimental studies that provide the rationale for exploring new questions in the development of an optimized adaptive intervention (e.g., analyses suggesting new targets for intervention or new decision points at which changes in intervention are critical). Such an approach is consistent with MOST (L. M. Collins 2018). For evaluation, intervention scientists typically would use a standard two- or multi-arm RCT design. For example, following any one of the example trial designs presented in this chapter, a behavioral intervention scientist may choose to conduct a follow-up two- or multi-arm RCT of the optimized adaptive intervention versus a suitable control, or a behavioral intervention scientist may choose to conduct another optimization trial to answer different questions.

In addition, because the focus in this chapter was on the design of the experiments—and especially the types of questions motivating the use of different types of experiments for optimizing adaptive interventions—data analytic methods or sample size/power resources were not discussed. This is an active and ever-growing area of methodological research. Two books describe design and (primarily) analytic methods for SMARTs in greater detail (Chakraborty & Moodie 2013; Kosorok & Moodie 2015). Analysis methods for the comparison of embedded adaptive interventions (Nahum-Shani et al. 2012a) have been extended for survival (Li & Murphy 2011) and longitudinal continuous (Lu et al. 2016) outcomes. A number of manuscripts focus on sample size calculators for different type of SMART designs (Almirall et al. 2012; Kim et al. 2016; Li & Murphy 2011; Oetting

et al. 2011). A large and growing literature now exists on analytic methods for examining candidate baseline and time-varying tailoring variables using data from a SMART (see Nahum-Shani et al. 2012b and Zhao et al. 2015).

There are various directions for future work. In terms of design, there is currently a dearth of methodological work on experimental design considerations for multilevel SMART designs. The SRT and SMART designs by Kilbourne and colleagues (Kilbourne et al. 2013; Kilbourne, Almirall, Eisenberg, et al. 2014; Kilbourne, Almirall, Goodrich, et al. 2014) and the design shown in Fig. 4 are exciting first steps in this direction. Such designs could have great appeal to educational intervention scientists who, by definition, work in clustered settings (e.g., repeated outcome measures for children nested within classrooms and nested within schools); implementation scientists (e.g., repeated outcome measures for patients, nested within clinics/hospitals); and prevention scientists, who often work on interventions at multiple levels (e.g., universal community-level interventions followed by selective individual-level interventions).

There is also a need for new methodological research on primary aim data analytic methods that behavioral intervention scientists have become accustomed to. These include the development of random effects or mixed models for comparing the embedded adaptive interventions in a restricted SMART design and methods for handling different types of outcomes (e.g., over-dispersed or zero-inflated outcomes in substance use research).

## Appendix

Let $z \in (-1, 1)$ be an indicator for the two interventions embedded in the enhanced non-responder trial shown in Fig. 3. Specifically, let $z = -1$ indicate the (non) adaptive intervention that provides RR at week 0, CS at week 4, Peer at week 8, and a continuation of Peer at week 20 regardless of response status. In Fig. 3, individuals in cells A and B are consistent with this intervention. Let $z = 1$ indicate the adaptive intervention that provides RR at week 0, CS at week 4, Peer at week 8, continued Peer at week 20 for slower responders, and Peer+Parent at week 20 for slower responders. In Fig. 3, individuals in cells A and C are consistent with this intervention.

Let $Y(z)$ denote an end-of-study outcome under the adaptive intervention indexed by $z$. Let $R$ denote binary response status (= 1 for early responder; = 0 for slower responder) at week 20 under the previous sequence of interventions: RR followed by CS followed by Peer. Response status is not indexed by $z$ because $z$ is unknown at the time the response status is assessed. Let $\pi = Pr(R = 1)$ be the response rate at the end of week 20 under RR followed by CS followed by Peer.

For a fixed embedded adaptive intervention $z$, the marginal mean outcome had the entire population followed adaptive intervention $z$, denoted $\mu(z) = E(Y(z))$, can be written as a weighted average of the mean outcome given week 20 response status:

$$\mu(z) = E(Y(z)) = E(Y(z) \mid R = 1) \cdot \pi + E(Y(z) \mid R = 0) \cdot (1 - \pi).$$

In question 2(i), the goal is to estimate and test $\Delta = \mu(1) - \mu(-1)$; that is, the difference in mean outcomes had the entire population followed adaptive intervention $z = 1$ versus had the entire population followed adaptive intervention $z = -1$.

Plugging in the above expression for $\mu(z)$ and utilizing the fact that in the enhanced non-responder trial

$$E(Y(-1) \mid R = 1) = E(Y(1) \mid R = 1),$$

that is, both adaptive interventions have the same mean outcome among responders, then we have that the causal effect is equal to

$$\Delta = E\big(Y(1) - Y(-1) \mid R = 0\big) \times (1 - \pi).$$

The above derivation shows that (in the population) the difference in outcome between the two adaptive interventions is a product of the non-response rate $1 - \pi$ and the difference in mean outcomes among non-responders (i.e., $E(Y(1) - Y(-1) \mid R = 0)$).

These derivations may suggest that data arising from a standard non-responder trial could be used to estimate $\Delta$. However, there are two caveats to this. First, the standard non-responder trial must have recruited from the same population of non-responders to RR, CS, and Peer (no selection bias). Second, an appropriate estimate of the non-response rate under RR, CS, and Peer must be available. If interest is solely in testing whether $\Delta = 0$ (i.e., there is no interest in estimating $\Delta$), then only an estimate of the difference in mean outcomes among non-responders is necessary, since $1 - \pi$ is expected to be nonzero (otherwise, it would be difficult to justify conducting either a standard or an enhanced non-responder trial). However, the requirement that there is no selection bias in the sample of non-responders still applies.

## References

Almirall, D., & Chronis-Tuscano, A. (2016). Adaptive interventions in child and adolescent mental health. *Journal of Clinical Child & Adolescent Psychology, 45*(4), 383–395.

Almirall, D., Compton, S. N., Gunlicks-Stoessel, M., Duan, N., & Murphy, S. A. (2012). Designing a pilot sequential multiple assignment randomized trial for developing an adaptive treatment strategy. *Statistics in Medicine, 31*(17), 1887–1902.

Almirall, D., Nahum-Shani, I., Sherwood, N., & Murphy, S. (2014). Introduction to SMART designs for the development of adaptive interventions: With application to weight loss research. *Translational Behavioral Medicine, 4*, 260–274.

Bower, P., & Gilbody, S. (2005). Stepped care in psychological therapies: Access, effectiveness and efficiency. *The British Journal of Psychiatry, 186*(1), 11–17.

Brent, D., Emslie, G., Clarke, G., Wagner, K., Asarnow, J., Keller, M., … others (2008). The treatment of adolescents with SSRI-resistant depression (TORDIA): A comparison of switch to venlafaxine or to another SSRI, with or without additional cognitive behavioral therapy. *JAMA, 299*(8), 901–913.

Chakraborty, B., Collins, L. M., Strecher, V. J., & Murphy, S. A. (2009). Developing multicomponent interventions using fractional factorial designs. *Statistics in Medicine, 28*(21), 2687–2708.

Chakraborty, B., & Moodie, E. (2013). *Statistical methods for dynamic treatment regimes*. New York: Springer.

Chakraborty, B., & Murphy, S. A. (2014). Dynamic treatment regimes. *Annual Review of Statistics and Its Application, 1*, 447.

Collins, F. S., & Varmus, H. (2015). A new initiative on precision medicine. *New England Journal of Medicine, 372*(9), 793–795.

Collins, L., Murphy, S., & Bierman, K. (2004). A conceptual framework for adaptive preventive interventions. *Prevention Science, 5*, 185–196.

Collins, L. M. (2018). *Optimization of behavioral and biobehavioral interventions: The multiphase optimization strategy (MOST)*. Cham: Springer.

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 effective tobacco use interventions. *Annals of Behavioral Medicine, 41*(2), 208–226.

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., Murphy, S. A., & Strecher, V. (2007). The multiphase optimization strategy (most) and the sequential multiple assignment randomized trial (SMART): New methods for more potent ehealth interventions. *American Journal of Preventive Medicine, 32*(5), S112–S118.

Collins, L. M., Nahum-Shani, I., & Almirall, D. (2014). Optimization of behavioral dynamic treatment regimens based on the sequential multiple assignment randomized trial (SMART). *Clinical Trials, 11*, 426–434.

Dawson, R., & Lavori, P. W. (2008). Sequential causal inference: Application to randomized trials of adaptive treatment strategies. *Statistics in Medicine, 27*(10), 1626–1645.

Dawson, R., Lavori, P. W., Luby, J. L., Ryan, N. D., & Geller, B. (2007). Adaptive strategies for treating childhood mania. *Biological Psychiatry, 61*(6), 758–764.

Emslie, G. J., Heiligenstein, J. H., Hoog, S. L., Wagner, K. D., Findling, R. L., McCRACKEN, J. T., … Jacobson, J. G. (2004). Fluoxetine treatment for prevention of relapse of depression in children and adolescents: A double-blind, placebo-controlled study. *Journal of the American Academy of Child & Adolescent Psychiatry, 43*(11), 1397–1405.

Ertefaie, A., Wu, T., Lynch, K. G., & Nahum-Shani, I. (2016). Identifying a set that contains the best dynamic treatment regimes. *Biostatistics, kxv025, 17*(1), 135–148.

Frankel, F. D., Gorospe, C. M., Chang, Y.-C., & Sugar, C. A. (2011). Mothers' reports of play dates and observation of school playground behavior of children having high-functioning autism spectrum disorders. *Journal of Child Psychology and Psychiatry, 52*(5), 571–579.

Freeman, J. B., Choate-Summers, M. L., Garcia, A. M., Moore, P. S., Sapyta, J. J., Khanna, M. S., … Franklin, M. E. (2009). The pediatric obsessive-compulsive disorder treatment study II: Rationale, design and methods. *Child and Adolescent Psychiatry and Mental Health, 3*(1), 1.

Goodyer, I., Dubicka, B., Wilkinson, P., Kelvin, R., Roberts, C., Byford, S., … others (2007). Adolescent depression antidepressant and psychotherapy trial (ADAPT). A randomised controlled trial of SSRIs with and without cognitive behaviour therapy in adolescents with major depression. *British Medical Journal, 335*, 142–146.

Gunlicks-Stoessel, M., Mufson, L., Westervelt, A., Almirall, D., & Murphy, S. (2015). A pilot SMART for developing an adaptive treatment strategy for adolescent depression. *Journal of Clinical Child & Adolescent Psychology, 45*(4), 480–494.

Guy, W. (1976). *ECDEU assessment manual for psychopharmacology: 1976*. National Institute of Mental Health us Department of Health and welfare, Rockville, MD.

Joseph, A. (2016). *SMART for smoking cessation in lung cancer screening.* Retrieved from https://clinicaltrials.gov/ct2/show/NCT02597491

Kasari, C., Kaiser, A., Goods, K., Nietfeld, J., Mathy, P., Landa, R., … Almirall, D. (2014). Communication interventions for minimally verbal children with autism: Sequential multiple assignment randomized trial. *Journal of the American Academy of Child and Adolescent Psychiatry, 53*, 635–646.

Kasari, C., Locke, J., Gulsrud, A., & Rotheram-Fuller, E. (2011). Social networks and friendships at school: Comparing children with and without ASD. *Journal of Autism and Developmental Disorders, 41*(5), 533–544.

Kasari, C., & Patterson, S. (2012). Interventions addressing social impairment in autism. *Current Psychiatry Reports, 14*(6), 713–725.

Kasari, C., & Smith, T. (2013). Interventions in schools for children with autism spectrum disorder: Methods and recommendations. *Autism, 17*(3), 254–267.

Kennard, B. D., Emslie, G. J., Mayes, T. L., Nightingale-Teresi, J., Nakonezny, P. A., Hughes, J. L., … Jarrett, R. B. (2008). Cognitive-behavioral therapy to prevent relapse in pediatric responders to pharmacotherapy for major depressive disorder. *Journal of the American Academy of Child & Adolescent Psychiatry, 47*(12), 1395–1404.

Kidwell, K. (2014, in press). SMART designs in cancer research: Past, present and future. *Clinical Trials, 11*, 445–456

Kilbourne, A. M., Abraham, K. M., Goodrich, D. E., Bowersox, N. W., Almirall, D., Lai, Z., & Nord, K. M. (2013). Cluster randomized adaptive implementation trial comparing a standard versus enhanced implementation intervention to improve uptake of an effective re-engagement program for patients with serious mental illness. *Implementation Science, 8*(1), 1.

Kilbourne, A. M., Almirall, D., Eisenberg, D., Waxmonsky, J., Goodrich, D. E., Fortney, J. C., … others (2014). Protocol: Adaptive implementation of effective programs trial (adept): Cluster randomized smart trial comparing a standard versus enhanced implementation strategy to improve outcomes of a mood disorders program. *Implement Science, 9*, 132.

Kilbourne, A. M., Almirall, D., Goodrich, D. E., Lai, Z., Abraham, K. M., Nord, K. M., & Bowersox, N. W. (2014). Enhancing outreach for persons with serious mental illness: 12-month results from a cluster randomized trial of an adaptive implementation strategy. *Implementation Science, 9*(1), 1.

Kim, H., Ionides, E., & Almirall, D. (2016, June). A sample size calculator for SMART pilot studies. *SIAM Undergraduate Research Online*. http://dx.doi.org/10.1137/15S014058

Kosorok, M. R., & Moodie, E. E. (2015). *Adaptive treatment strategies in practice: Planning trials and analyzing data for personalized medicine* (Vol. 21). Philadelphia: SIAM.

Kretzmann, M., Shih, W., & Kasari, C. (2015). Improving peer engagement of children with autism on the school playground: A randomized controlled trial. *Behavior Therapy, 46*(1), 20–28.

Laber, E. B., Lizotte, D. J., Qian, M., Pelham, W. E., & Murphy, S. A. (2014). Dynamic treatment regimes: Technical challenges and applications. *Electronic Journal of Statistics, 8*(1), 1225.

Lavori, P., & Dawson, D. (2000). A design for testing clinical strategies: Biased individually tailored within-subject randomization. *Journal of the Royal Statistical Society, Series A, 163*, 29–38.

Lavori, P. W., & Dawson, R. (2014). Introduction to dynamic treatment strategies and sequential multiple assignment randomization. *Clinical Trials, 11*(4), 393–399.

Lavori, P. W., Rush, A. J., Wisniewski, S. R., Alpert, J., Fava, M., Kupfer, D. J., … others (2001). Strengthening clinical effectiveness trials: Equipoise-stratified randomization. *Biological Psychiatry, 50*(10), 792–801.

Lei, H., Nahum-Shani, I., Lynch, K., Oslin, D., & Murphy, S. (2012). A SMART design for building individualized treatment sequences. *Annual Review of Clinical Psychology, 8*, 21–48.

Li, Z., & Murphy, S. A. (2011). Sample size formulae for two-stage randomized trials with survival outcomes. *Biometrika, 98*(3), 503–518.

Lieberman, J., Stroup, T., McEvoy, J., Swartz, M., Rosenheck, R., Perkins, D., … Clinical Antipsychotic Trials of Intervention Effectiveness Investigators (2005). Effectiveness of antipsychotic drugs in patients with chronic schizophrenia. *New England Journal of Medicine, 353*(12), 1209–1223.

Lord, C., & McGee, J. P. (2001). *Educating children with autism: Committee on educational interventions for children with autism.* National Academy Press, Washington, DC.

Lu, X., Nahum-Shani, I., Kasari, C., Lynch, K. G., Oslin, D. W., Pelham, W. E., … Almirall, D. (2016). Comparing dynamic treatment regimes using repeated-measures outcomes: Modeling considerations in smart studies. *Statistics in Medicine, 35*(10), 1595–1615.

McCracken, J. T., McGough, J., Shah, B., Cronin, P., Hong, D., Aman, M. G., … others (2002). Risperidone in children with autism and serious behavioral problems. *New England Journal of Medicine, 347*(5), 314–321.

Methodology Center. (2016, May). *Example SMART studies.* Retrieved from https://methodology.psu.edu/ra/adap-inter/projects

Murphy, S. (2005). An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine, 24*, 1455–1481.

Murphy, S., & Bingham, D. (2009). Screening experiments for developing dynamic treatment regimes. *Journal of the American Statistical Association, 104*, 391–408.

Murphy, S., Lynch, K., Oslin, D., McKay, J., & Tenhave, T. (2007). Developing adaptive treatment strategies in substance abuse research. *Drug and Alcohol Dependence, 88*, S24–S30.

Naar-King, S., Ellis, D. A., Idalski Carcone, A., Templin, T., Jacques-Tiura, A. J., Brogan Hartlieb, K., … Jen, K.-L. C. (2015). Sequential multiple assignment randomized trial (SMART) to construct weight loss interventions for African American adolescents. *Journal of Clinical Child & Adolescent Psychology, 45*(4), 428–441.

Nahum-Shani, I., Qian, M., Almirall, D., Pelham, W., Gnagy, B., Fabiano, G., … Murphy, S. (2012a). Experimental design and primary data analysis methods for comparing adaptive interventions. *Psychological Methods, 17*, 457–477.

Nahum-Shani, I., Qian, M., Almirall, D., Pelham, W., Gnagy, B., Fabiano, G., … Murphy, S. (2012b). Q-learning: A data analysis method for constructing adaptive interventions. *Psychological Methods, 17*, 478–494.

Nair, V., Strecher, V., Fagerlin, A., Ubel, P., Resnicow, K., Murphy, S., … Zhang, A. (2008). Screening experiments and the use of fractional factorial designs in behavioral intervention research. *American Journal of Public Health, 98*(8), 1354–1359.

National Research Council. (2011). *Toward precision medicine: Building a knowledge network for biomedical research and a new taxonomy of disease*. Washington, DC: National Academies Press (US).

Ochs, E., Kremer-Sadlik, T., Solomon, O., & Sirota, K. G. (2001). Inclusion as social practice: Views of children with autism. *Social Development, 10*(3), 399–419.

Oetting, A., Levy, J., Weiss, R., & Murphy, S. (2011). Statistical methodology for a smart design in the development of adaptive treatment strategies. In P. Shrout, K. Keyes, & K. Ornstein (Eds.), *Causality and psychopathology: Finding the determinants of disorders and their cures* (pp. 179–205). Arlington, VA: Oxford University Press.

Orellana, L., Rotnitzky, A., & Robins, J. (2010). Dynamic regime marginal structural mean models for estimating optimal dynamic treatment regimes, part I: Main content. *International Journal of Biostatistics, 6*(2), Article 8.

Prasad, K. (2007). Cardiovascular disease. In D. Machin, S. Day, & S. Green (Eds.), *Textbook of clinical trials* (pp. 215–239). New York: Wiley.

Rotheram-Fuller, E. J. (2005). Age-related changes in the social inclusion of children with autism in general education classrooms. Unpublished thesis, University of California.

Rush, A. J., Fava, M., Wisniewski, S. R., Lavori, P. W., Trivedi, M. H., Sackeim, H. A., … others (2004). Sequenced treatment alternatives to relieve depression (star* d): rationale and design. *Controlled Clinical Trials, 25*(1), 119–142.

Sherwood, N. E., Butryn, M. L., Forman, E. M., Almirall, D., Seburg, E. M., Crain, A. L., … Jeffery, R. W. (2016). The bestfit trial: A smart approach to developing individualized weight loss treatments. *Contemporary Clinical Trials, 47*, 209–216.

Shortreed, S. M., & Moodie, E. E. (2012). Estimating the optimal dynamic antipsychotic treatment regime: Evidence from the sequential multiple-assignment randomized clinical antipsychotic trials of intervention and effectiveness schizophrenia study. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 61*(4), 577–599.

Skarphedinsson, G., Weidle, B., Thomsen, P. H., Dahl, K., Torp, N. C., Nissen, J. B., … others (2015). Continued cognitive-behavior therapy versus sertraline for children and adolescents with obsessive–compulsive disorder that were non-responders to cognitive-behavior therapy: a randomized controlled trial. *European Child & Adolescent Psychiatry, 24*(5), 591–602.

Sobell, M. B., & Sobell, L. C. (2000). Stepped care as a heuristic approach to the treatment of alcohol problems. *Journal of Consulting and Clinical Psychology, 68*(4), 573.

Spring, B., & Nahum-Shani, I. (2016, September). *SMART weight loss management.* Retrieved from https://projectreporter.nih.gov/project_info_description.cfm?aid=9126830icde=31096021

Steedly, K. M., Schwartz, A., Levin, M., & Luke, S. D. (2008). Social skills and academic achievement. *Evidence for Education, 3*(2), 1–8.

Svetkey, L. P., Stevens, V. J., Brantley, P. J., Appel, L. J., Hollis, J. F., Loria, C. M., … others (2008). Comparison of strategies for sustaining weight loss: The weight loss maintenance randomized controlled trial. *JAMA, 299*(10), 1139–1148.

Thall, P. F., Wooten, L. H., Logothetis, C. J., Millikan, R. E., & Tannir, N. M. (2007). Bayesian and frequentist two-stage treatment strategies based on sequential failure times subject to interval censoring. *Statistics in Medicine, 26*(26), 4687–4702.

Trivedi, M. H., Fava, M., Marangell, L. B., Osser, D. N., & Shelton, R. C. (2006). Use of treatment algorithms for depression. *Prim Care Companion J Clin Psychiatry, 8*(5), 258.

Wang, L., Rotnitzky, A., Lin, X., Millikan, R., & Thall, P. (2012). Evaluation of viable dynamic treatment regimes in a sequentially randomized trial of advanced prostate cancer. *Journal of the American Statistical Association, 107*(498), 493–508.

Zhao, Y.-Q., Zeng, D., Laber, E. B., & Kosorok, M. R. (2015). New statistical learning methods for estimating optimal dynamic treatment regimes. *Journal of the American Statistical Association, 110*(510), 583–598.

# Intensively Adaptive Interventions Using Control Systems Engineering: Two Illustrative Examples


**Daniel E. Rivera, Eric B. Hekler, Jennifer S. Savage, and Danielle Symons Downs**

**Abstract** Control systems engineering is a diverse field that examines how system variables can be adjusted over time to improve targeted outcomes. In recent years, control engineering approaches have shown significant appeal in the optimization phase of the multiphase optimization strategy (MOST). Control engineering can provide the basis for modeling intensive longitudinal data and using these models to optimize personalized, time-varying interventions. This chapter describes how control systems engineering principles, particularly system identification and model predictive control, can be applied to serve as dynamic modeling methods and optimal decision frameworks, respectively, for two intensively adaptive interventions: Just Walk, an intervention to promote walking in sedentary middle-age adults, and Healthy Mom Zone, an intervention to manage gestational weight gain in obese and overweight pregnant women. The integrated system identification-model predictive control strategy described in this chapter, as well as the pivotal role that behavioral theory plays in developing dynamical models, is illustrated with examples taken from these two interventions.

D. E. Rivera (✉)
Control Systems Engineering Laboratory, School for Engineering of Matter, Transport, and Energy, Arizona State University, Tempe, AZ, USA
e-mail: daniel.rivera@asu.edu

E. B. Hekler
Center for Wireless Health and Population Systems, Department of Family Medicine and Public Health, University of California, San Diego, La Jolla, CA, USA
e-mail: ehekler@ucsd.edu

J. S. Savage
Center for Childhood Obesity Research and the Department of Nutritional Sciences, Penn State University, State College, PA, USA
e-mail: jfs195@psu.edu

D. S. Downs
Exercise Psychology Laboratory, Department of Kinesiology, College of Health and Human Development, Department of Obstetrics and Gynecology, Penn State College of Medicine, Penn State University, State College, PA, USA
e-mail: dsd11@psu.edu



L. M. Collins, K. C. Kugler (eds.), *Optimization of Behavioral, Biobehavioral, and Biomedical Interventions*, Statistics for Social and Behavioral Sciences,
https://doi.org/10.1007/978-3-319-91776-4_5

## 1 Introduction

In the development of effective behavioral, biobehavioral, and biomedical interventions, it is becoming increasingly important to seek alternatives to standard intervention practices, such as the multiphase optimization strategy (MOST; see the companion volume, Collins, 2018). Conventional fixed interventions are designed for a standard response that may not recognize individual participant characteristics or be readily amenable to optimization procedures. One promising approach lies in *time-varying adaptive interventions*; these adjust treatment dosages over the course of the intervention based on the values of *tailoring variables*; these are often measures of participant response or adherence (Collins, Murphy, & Bierman, 2004) but can also include external factors that influence behavior (e.g., weather). In this chapter, we describe how control systems engineering (Aström & Murray, 2010) offers a potentially useful framework for optimizing the effectiveness of broad classes of adaptive behavioral interventions, particularly those that are referred to as *intensively adaptive interventions*, which feature frequent (e.g., weekly, daily, or even more frequent) adaptations. Specifically, the development of decision policies from control systems engineering, coupled with technological enhancements in information and computer technology, will result in novel forms of time-varying adaptive interventions that increase compliance, enhance overall intervention potency, and reduce waste of time and other resources (Bekiroglu, Lagoa, Murphy, & Lanza, 2017; Chakraborty & Murphy, 2014; Riley et al. 2011; Rivera, 2012; Rivera, Pew, & Collins, 2007; Zafra-Cabeza, Rivera, Collins, Ridao, & Camacho, 2011).

Control systems are widely used in industrial practice to achieve desired behavior of processes by systematically adjusting manipulated variables based on measured information (Aström & Murray, 2010; Ogata, 2010; Skogestad & Postlethwaite, 1996). Prior work (Rivera et al. 2007) has established that time-varying adaptive interventions featuring repeated assessments, intensive data collection, and frequent decision-making can be conceptualized as engineering control systems. In a control engineering approach to time-varying adaptive interventions, the controller assigns dosages to each participant as dictated by the solution of a formal optimization problem that fully incorporates the parameters or predictions arising from a dynamical systems model (in control systems this corresponds to a mathematically specified conceptual model; the conceptual model is discussed in detail in Chapter 2 of the companion volume Collins, 2018).

In this chapter we illustrate how control engineering is applied for two behavioral interventions: Just Walk, an intervention focused on increasing walking activity in sedentary adults, and Healthy Mom Zone, an intervention to manage gestational weight gain (GWG) in obese and overweight pregnant women. Just Walk and Healthy Mom Zone are two ongoing intervention studies, for which we present some experimental results to date and discuss the envisioned full outcome of the intervention in simulation. For both interventions, we discuss how a data-based modeling technique from engineering known as *system identification* (Ljung, 1999) can be used to estimate relevant dynamical system models that relate treatment

components of an adaptive intervention and external factors to outcomes of interest. System identification is an $N = 1$ or idiographic approach (Molenaar & Campbell, 2009; Velicer, 2010). The estimated dynamical systems models then serve as the basis for applying a control technology known as *model predictive control* (MPC) as a decision algorithm for automatic dosage selection. In the adaptive intervention literature, decision algorithms are often called decision rules. A multiple-degree-of-freedom formulation for MPC is presented that enables the interventionist to adjust the speed at which a desired target outcome should be reached, with the capability to independently adjust to both anticipated and unanticipated changes in symptoms (as well as possible side effects). Simulation results are presented to illustrate the performance of the proposed decision scheme, which incorporates individual participant response, intervention constraints, modeling errors, and variability typically present in a real-life application.

This chapter is organized as follows: Sect. 2 defines and provides a general overview of "open-loop" dynamical systems modeling and how the presence of a controller achieves "closed-loop" control (defined in Table 1). Section 3 discusses the use of system identification to arrive at dynamical system models for intensively adaptive interventions; this can be done with both black-box and semi-physical approaches, the latter of which can be related directly to behavioral theories. Section 4 presents the use of MPC as an effective means to accomplish closed-loop control in an intervention setting, while Sects. 5 and 6 present the application of these ideas (either already implemented or envisioned) to the Just Walk and Healthy Mom Zone interventions, respectively. The chapter ends with a summary and conclusions in Sect. 7, which includes a discussion on future directions in this area.

## 2 A Control Systems Engineering Approach for Modeling Interventions

Control systems engineering is a diverse field that examines how a system variable (i.e., an intervention component) can be adjusted over time so that its dynamical response (i.e., its behavior over time) is transformed from undesirable to desirable. Cruise and climate control in automobiles, the "sensor reheat" feature in a microwave oven, the home thermostat, and the artificial pancreas are all examples of control systems engineering at work; these are just a few of many applications and success stories (the interested reader can learn about these and more in http://ieeecss.org/general/impact-control-technology). The conventional approach to control systems engineering is to first accomplish some form of dynamical systems modeling to understand the response of the system in open loop (i.e., describe system behavior in the absence of a formal set of decision rules; Table 1 and Sect. 2.1) and then to use this model to design a controller and then implement automatic or closed-loop control (Table 1 and Sect. 2.2). These concepts are further developed in the subsections below.

**Table 1** Some relevant control engineering terminology

| Term | Definition |
|---|---|
| Block diagram | Graphical representation of the signals and systems that comprise a closed-loop control system |
| Closed loop | Mode of operation of the control system, once a controller/set of decision rules is implemented |
| Controller | Mathematical set of relationships that translate error (i.e., deviation from a goal or set point) into settings for a manipulated variable (which defines an intervention dosage). Also referred to as a set of decision rules in the context of this chapter. Sometimes referred to in the scientific literature as a decision policy |
| Control engineering | Field that considers how to manipulate system variables in order to transform dynamic behavior to desirable from undesirable |
| Control error ($e = r - y$) | Difference between the controlled variable and the set point; in a walking intervention, if an individual walks 6000 steps in a day and her daily set point target is 10,000 steps, the control error is 4000 steps. The ultimate goal of a control system is to have the controlled variable perfectly track the set point, in which case the control error would correspond to zero |
| Control loop | Closed-loop system |
| Control structure | Refers primarily to whether feedback or feedforward strategies (or their combination) are applied in a closed-loop system |
| Controlled variables ($y$) | System variables that we wish to keep at a reference value or set point ($r$) |
| Disturbance inputs ($d$) | System variable that influences the controlled variable response but cannot be manipulated by the controller; disturbance changes occur external to the system (hence sometimes referred to as exogenous variables). Disturbance inputs can be measured or unmeasured |
| Disturbance rejection | Ability of the control system to manipulate system variables such that the controlled variable is kept as close as possible to the set point, in spite of significant changes in the disturbance variables |
| Error projection | Signal created by the model predictive control algorithm that indicates the discrepancy between predicted outputs and their reference trajectories (set points) over a future horizon. The model predictive control algorithm relies on the error projection and the dynamical model to optimize future control actions |
| Feedback control | Control strategy in which a controlled variable ($y$) is measured and compared to a reference value or set point ($r$). The controller issues actions (decisions on the values of a manipulated variable ($u$)) on the basis of the discrepancy between $y$ and $r$ (which is referred to as the control error $e$) |
| Feedforward control | Control strategy in which changes in a disturbance variable ($d$) are monitored and the manipulated variable ($u$) is chosen to counteract anticipated changes in $y$ as a result of $d$ |
| Fluid analogy | A representation of a control engineering problem as an interconnection of tanks in which the amount of liquid in a tank (its corresponding "inventory") changes as a consequence of the actions of inputs to the system (represented as inflows and outflows) |
| Manipulated inputs ($u$) | System variable whose adjustment influences the response of the controlled variable $y$; the magnitude of $u$ is determined by the controller |

(continued)

**Table 1** (continued)

| Term | Definition |
|---|---|
| Model error | Discrepancy between the parameter values and structure of the dynamical model that is used to design the control system, versus that of the model that actually describes the plant. Control systems are ideally designed to be robust to specified levels of model error. Not to be confused with the control error $e$. Also referred to as plant-model mismatch |
| Model predictive control (MPC) | Control engineering algorithm that optimizes, using a moving horizon philosophy, an explicit objective function under constraints. Model predictive control presents an optimization strategy that accomplishes feedback and feedforward control in this problem space |
| Offset | Sustained discrepancy between the controlled variable response and the set point in a closed-loop system; it is reflected in a nonzero control error $e$ during all time in the operation of the control system |
| Open loop | Dynamical system behavior without a controller (i.e., without a set of decision rules) |
| Output variables | Dependent variables in the system; typically these reflect outcomes of interest. Controlled variables $y$ correspond to an output variable, which has a desired reference set point $r$ |
| Process | The dynamical system under study, for which a closed-loop controller or decision rule will be applied |
| Set point tracking | The ability of the control system to manipulate system variables such that the controlled variable follows a reference (set point) trajectory as closely as possible |

## *2.1 Open-Loop Dynamical Systems Modeling*

A foundational starting point in MOST is the conceptual model. As defined in Chapter 2 of the companion volume, "the conceptual model expresses all of what is known or hypothesized about how the intervention under development is to intervene on the behavioral, biobehavioral, or biomedical process" (Collins, 2018, p. 36). Similarly, control systems engineering requires clear specification of what is known, hypothesized, or conjectured; this takes the form of a dynamical systems model, along with any information regarding constraints that are in place (or need to be enforced) during the intervention. Control systems engineering is particularly interested in supporting decision-making in systems whose behavior varies over time; dynamical systems modeling considers how to characterize the change over time or *transient response* resulting from changes in manipulated inputs (e.g., intervention components whose dosage can be adjusted by the intervention scientist, denoted by $u$) and disturbance inputs (e.g., external influences which are not manipulated directly in the intervention, denoted by $d$) on outputs (e.g., proximal or distal outcomes; could also represent mediators, denoted by $y$) measured in an intensive longitudinal setting. In a typical intervention, the input ($u$) will represent the dosage of a primary intervention component such as activity goals, medication, or counseling, while a disturbance ($d$) corresponds to behavioral constructs associated

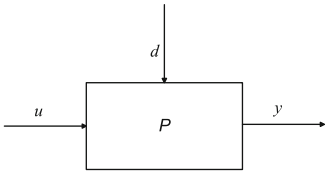


**Fig. 1** An input-output "block" diagram representation of the effect of an input ($u$) and disturbance ($d$) on the output ($y$) via a dynamical system ($P$). This is an open-loop configuration, as $u$ is exogenous to the system and not dictated by a set of decision rules. Changes in the input $u$ (e.g., daily step goals, meal demonstrations) produce, based on the system dynamics $P$, a resulting change in an output $y$ (e.g., daily steps walked, weight gain) while in the presence of external disturbances $d$ (e.g., busyness, stress, bad weather)

with the intervention that influence outcomes but are independent of treatment, such as reported levels of busyness, anxiety, or stress. The output ($y$) can be an outcome of interest that the intervention aims to modify, such as physical activity, steps walked, or weight gain. In Fig. 1, we show that *u, d,* and $y$ can be related to each other by a model $P$, which symbolically represents a dynamical system. $P$ can be expressed mathematically using a system of ordinary differential equations (in continuous time $t$) or a system of difference equations (in sampled or discrete time $k$), among various other representations. A dynamical systems approach allows for an efficient mapping of the causal relationship between variables by capturing the concepts of change and effect in interventions.

The problem of an individual driving a car is an everyday illustration of an open-loop dynamical system. The steering wheel, gas pedal, and brake pedal are examples of manipulated variables that are adjusted by an operator (the driver). The velocity and direction of the vehicle are proximal outcomes of interest (which we refer to as outputs); being able to arrive at a particular location in a given amount of time could be considered a distal outcome. The ability of the vehicle to achieve its objectives (e.g., go in a particular direction or take an individual or group of individuals from point A to B) is influenced by external factors such as road conditions, the presence of other vehicles, and weather; these are examples of variables that we would consider as disturbance inputs. For example, in behavioral interventions, a disturbance input might be weather conditions that inhibit (or promote) physical activity or stresses, such as managing childcare, that may prevent a pregnant woman from pursuing healthy eating. While some disturbances will be inherently unmeasured, it is most beneficial to measure and model how disturbance effects influence the outputs (i.e., outcomes of interest).

## *2.2 Controller Design and Closed-Loop Control*

The ultimate goal in the use of control systems engineering is to develop a closed-loop intervention in which an algorithmic control system, relying on a model and repeated measurements of important system variables, assigns appropriate

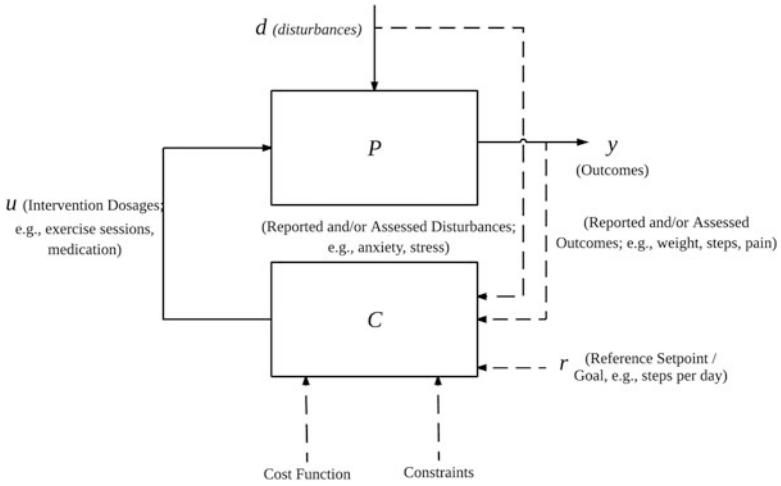


**Fig. 2** Conceptual block diagram representation of a control system showing a closed-loop strategy using a desired reference for a designated cost function and clinical constraints. This figure extends the ideas introduced in Fig. 1 by closing the loop. In a closed-loop system, dosages $u$ are assigned by the controller $C$ based on the solution of an optimization problem that minimizes a cost function (subject to intervention constraints) to take outcomes to a specified goal (the reference set point)

dosage magnitudes automatically over time (Table 1). Because the controller now determines the response of the manipulated variable $u(t)$ (Fig. 2), the control loop is considered closed. Many controller forms (and control design strategies) are available, but our emphasis in this chapter will be on MPC approaches that solve an optimization problem in real time, taking into account problem constraints and relying on measurements of both outcomes of interest (controlled variables) and external factors (disturbances) provided by or assessed from the participant. In the adaptive intervention language of Collins et al. (2004) and Chapter 8 of the companion volume, the controller represents a set of decision rules, and any measured variables $y$ or $d$ that influence the decisions made by the controller correspond to tailoring variables.

To illustrate a control engineering strategy, we examine the block diagram shown in Fig. 2. $P$ represents the dynamics expected of the treatment intervention on outcomes of interest; this is the same block as in Fig. 1. The controller, represented by the symbol $C$, is given a set of operating intervention constraints and a cost function providing a performance metric for the optimization problem. Intervention constraints include limits on the dosages of intervention components, while the cost function represents a performance metric such as the average deviation from a goal over time. These constraints are the control systems engineering equivalent of MOST optimization constraints. Indeed, efficiency, economy, and scalability (see Chapter 2 of the companion volume) are domains that can feasibly indicate constraints to be accounted for by a controller. As control systems engineering

is particularly interested in responses over time, added constraints (e.g., plausible dosages, amount of change in dosage from 1 day to the next, anticipated rate of response to an intervention that is deemed healthy vs. unhealthy) must also be specified and understood when establishing constraints for the controller; these ideas will be illustrated in detail later in the chapter. Based on the discrepancy between the measured outcomes $y$ (e.g., steps walked, weight gained) and their desired reference values (which we refer to as the *control error* $e(t) = y(t) - r(t)$), the controller assigns dosages $u(t)$ to enable a desired closed-loop response for each participant (Ogunnaike & Ray, 1994; Rivera et al. 2007). The assessed values of current and possibly future disturbance signals (e.g., stress, busyness, day of week) can influence the model and ultimately have a role in decision-making; when these measures are available, they can be part of the closed-loop control system. The closed-loop system aims to achieve control by accomplishing the following three primary functional tasks.

1. *Reaching a desired goal (set point tracking).* Intervention dosages are assigned that will ultimately change an outcome of interest, to reach a desired goal. For example, when designing an intervention to increase physical activity, the intervention scientist may decide on a goal of a sustained increase of 3000 steps walked per day within 6 weeks of the start of the intervention.
2. *Accounting for measured external factors (rejection of measured disturbances).* The controller can decide on component dosages to mitigate the effect from reported or assessed external influences (e.g., anxiety) using disturbance models. For instance, if some external event that leads to increased stress or anxiety is known a priori, then dosages can be adjusted to compensate for this disturbance. The mode of operation that accomplishes this task is referred to as *feedforward* control. The controller algorithm possesses the functionality to allow this feature to be turned on or off or otherwise "tuned" (as discussed later in the chapter) for a desired level of performance.
3. *Accounting for unmeasured (and/or unmodelled) external factors (rejection of unmeasured disturbances).* The controller can adjust component dosages to mitigate the effect of unknown or unmodelled external influences. For example, there could be a change in physical activity or weight gain that may not be directly explained by changes in measured or reported conditions. In such cases, the controller, through the process of *feedback* control, is able to adjust dosages to mitigate the effects of this unmeasured disturbance.

It is valuable to examine a representative controller equation that reflects the previous discussion. Consider an intervention where the dosage of an intervention component $u(t)$ is determined by a controller equation incorporating both feedback and feedforward functionality. One such equation would be

$$u(k) = u\,(k-1) + K_1\,e(k) + K_2\,e\,(k-1) + K_3\,d(k). \tag{1}$$

Equation (1) indicates that the dosage at the current review interval $k$ ($u(k)$) is determined from the dosage from the prior review interval $u(k-1)$ with some scaled

corrections that arise from current and previous control errors ($e(k) = r(k) - y(k)$ and $e(k-1) = r(k-1) - y(k-1)$, respectively) and the current measured value for a disturbance $d(k)$. $K_1$, $K_2$, and $K_3$ are controller coefficients that are dependent on the model and other considerations; they represent controller tuning. Note the contrast between the controller determination of dosage in (1) and classical *if-then* decision rules used to assign dosage in an adaptive intervention (which may assign $u$ based on the current measurement of $y$, using a threshold and without articulating a control error or taking advantage of a measured disturbance (Hekler et al. 2018)).

The three functional tasks of the control system have to be fulfilled under a number of practical requirements; hence this functionality has to be integrated into controller design and implementation. In conventional practice, intervention dosage limits are often set to avoid adverse iatrogenic effects (e.g., drug toxicity, inadequate GWG). In addition, intervention dosages are generally designed at categorical (i.e., discrete) levels. For example, counseling sessions can either be weekly, biweekly, or monthly. Similarly, a commercially produced drug may be available only in certain fixed dosages; these limits need to be recognized by the control algorithm. Furthermore, dosage changes occurring day-to-day should not be very abrupt, due to potential negative consequences that the participant may experience, such as confusion related to frequent changes in counseling session frequency or withdrawal symptoms when medication dosages are changed. Hence, the controller should be "tuned" in such a way that dosing can be adjusted from more aggressive settings (where intervention dosages change rapidly over a relatively short time frame) to more conservative settings where dosage changes relatively slowly over time. These decisions should ultimately be made on the basis of theoretical, conceptual, and clinical insights.

## 3 System Identification to Obtain Dynamical System Models

In traditional control engineering problems, the dynamics of a system can be modeled on the basis of physical laws ("first principles"), by applying conservation and accounting concepts on extensive variables of interest, such as total mass and total energy. However, the underlying complexity and often poorly understood mechanisms of many systems of practical interest present challenges to first-principles modeling, leading to the use of modeling methods based on experimental data, that is, system identification methods (Ljung, 1999). In general terms, the problem of system identification focuses on modeling dynamical systems from intensive data, using systems engineering concepts (such as optimization and signal processing) and statistical principles. The presence of disturbances, particularly unmeasured ones, presents a challenge that makes system identification a nontrivial task in many situations of industrial and practical importance. System identification is traditionally broken down into four steps.

1. *Experimental design*. This is often the most important step and most critical to success of the overall system identification procedure. Part of this step is choosing the parameters that define the input into the system; these inputs will be experimentally manipulated. This chapter will illustrate the use of a system identification-based experimental design in the description of Just Walk (Sect. 5).
2. *Definition and selection of a model structure*. The structure can be either a general, "ready-made" black-box structure (such as autoregressive with external input (ARX) models) or other members of the family of prediction-error models (Ljung, 1999). Other alternatives include semi-physical structures based on first principles such as fluid analogies. Black-box and semi-physical structures are both illustrated in this chapter.
3. *Parameter estimation*. The parameter estimation step involves using a numerical procedure to obtain estimates of the model parameters. The type of objective function (e.g., squared sum of the prediction error), the model structure, and the nature of any prefiltering operations on the data are among a myriad of design variables that must be specified in parameter estimation. Other factors that must be considered are potential numerical problems that might impact the quality of the parameter estimates.
4. *Model validation*. Having estimated a model, its adequacy must be assessed. Among the issues that must be considered is whether the output predicted by the model compares favorably with the measured output, whether the "step response" of the estimated model agrees with expert intuition, and to what extent unmodelled dynamics and model uncertainty will impact the ability to design a well-performing control system.

In practice, system identification is an iterative procedure (Fig. 3). The lack of a priori information regarding a system will require that some of the steps initially be examined in a cursory manner. After each stage, the user must discern whether any previous steps were accomplished incorrectly; the procedure is then repeated until a suitable model is obtained. This is in line with the continual optimization principle used in MOST, with two different types of system identification experiments available that roughly correspond to that which would be used in the preparation stage versus the optimization phase. In the preparation phase, an open-loop system identification experiment can be conducted as the control systems equivalent of a pilot study. As described in Chapter 2 of the companion volume (Collins, 2018), the purpose of pilot studies is both to support initial examination of the feasibility to conduct the research study in general (e.g., ability to recruit, run the trial, gather data, analyze the results) and to test the feasibility of the approach. As control systems engineering places such high priority on a mathematical specification via a dynamical model, there is a requirement for a pilot study that can provide initial insights, beyond just simulation studies, for creating dynamical models. The open-loop system identification experiment is an appropriate methodology for this type of pilot testing because it can support the other targets of pilot testing in the preparation phase of MOST (i.e., pilot the methods, examine feasibility of the intervention for real-world use) while also supporting dynamical systems modeling. This will be illustrated with the Just Walk case study in Sect. 5. In contrast, a closed-loop

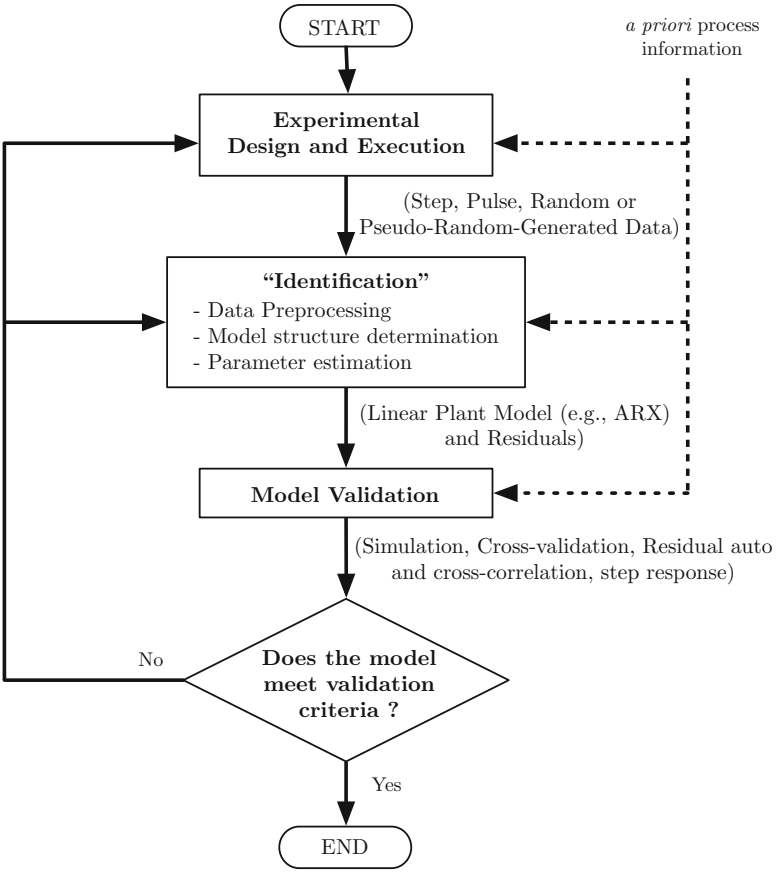


**Fig. 3** System identification problem – iterative loop schematic

experiment that includes the controller/set of decision rules is the design often more appropriate for the optimization phase. In particular, a closed-loop experiment enables examining the quality of the decision rules, intervention constraints, and other procedures used to create an automated intensively adaptive intervention.

In this section we provide some additional details and illustrations of dynamic modeling via system identification in two forms: black-box/fully empirical methods and semi-physical approaches.

## *3.1 Black-Box Approaches*

Black-box approaches are best suited in scenarios in which little is known about the mechanisms of change between inputs and outputs. These methods can be

valuable in settings where only limited involvement can be expected from personnel analyzing the intervention data and building models. The estimated model can serve myriad purposes, such as providing predictions that are used by a controller to assign dosages based on measured participant responses. The modeling procedure undertaken in this study is summarized in three subparts, as follows.

1. *Data preprocessing*. Data preprocessing operations include interpolation or imputation for missing entries, mean subtraction or removal of trends, and filtering operations to provide the appropriate level of description in the data. For instance, in Deshpande, Nandola, Rivera, and Younger (2011), a 3-day moving average filter is applied to attenuate the high frequency (i.e., rapid day-to-day) variations in the data.
2. *Discrete-time modeling using multi-input ARX models*. Preprocessed data are fitted to an ARX parametric model (ARX $[n_a\ n_{b_1}\ n_{b_2}\ldots n_{b_{n_u}}\ n_{k_1}\ n_{k_2}\ldots n_{k_{n_u}}]$) that is defined by a difference equation involving current and prior values of $y$ and $u$:

$$\begin{aligned} &y(k)+a_1y\left(k-1\right)+\cdots+a_{n_a}y\left(k-n_a\right)=b_{11}u_1\left(k-n_{k_1}\right)+\ldots\\ &+b_{1n_{b_1}}u_1\left(k-n_{k_1}-n_{b_1}+1\right)+b_{21}u_2\left(k-n_{k_2}\right)+\ldots+b_{2n_{b_2}}u_2\left(k-n_{k_2}-n_{b_2}+1\right)\\ &\qquad\qquad\qquad\vdots\\ &+b_{n_u1}u_{n_u}\left(k-n_k\right)+\cdots+b_{n_un_{b_{n_u}}}u_{n_u}\left(k-n_{k_{n_u}}-n_{b_{n_u}}+1\right)\\ &+e(k). \end{aligned} \tag{2}$$

$n_u$ represents the number of inputs $[u_1\ u_2\ldots u_{n_u}]$; $n_a$, $n_{b_1}, n_{b_2}, \ldots, n_{b_{n_u}}$ are the model orders; and $n_{k_1}, n_{k_2}, \ldots, n_{k_{n_u}}$ are the input delays. $e(k)$ is the prediction error, while $k$ is a discrete time index (e.g., day). The model parameter vector

$$\left[a_1,\ldots,a_{n_a},b_{11},\ldots,b_{1n_{b_1}},b_{21},\ldots,b_{2n_{b_2}},\ldots,b_{n_u1},\ldots,b_{n_un_{b_{n_u}}}\right] \tag{3}$$

is estimated using regression. ARX parameter estimation constitutes a linear least-squares regression problem (Ljung, 1999) and has attractive statistical properties, such as consistency.

3. *Model validation*. Model validation in black-box modeling can take many forms. Simulation of the estimated output from the model to measured data is typical and is best when done with a "fresh" dataset (i.e., one that was not used for estimation); this dataset is referred to as a validation dataset, and the validation procedure is referred to as cross-validation. In the absence of a cross-validation dataset, good fit does not necessarily imply a predictive model. Model fit can be quantified based on the value of the objective function (from parameter estimation); however, a popular metric in system identification to determine the

percentage of output variance explained by the model is the normalized root-mean-square error (NRMSE) fit index:

$$\text{Model fit } (\%) = 100 \times \left(1 - \frac{\left\| y(k) - \hat{y}(k) \right\|_2}{\| y(k) - \overline{y} \|_2}\right), \tag{4}$$

where $y(k)$ is the measured output, $\widehat{y}(k)$ is the simulated output, $\overline{y}$ is the mean of all measured $y(k)$ values, and $\|\cdot\|_2$ indicates a vector 2-norm ($\| x \|_2 \stackrel{\text{def}}{=} \sqrt{x^T x}$). Various statistical tests can be performed as part of model validation, including testing the residual time series from model fitting to determine whether it can be classified as white noise. If a residual time series consists of white noise, this indicates that all the important correlation and dynamic behavior have been captured by the estimated model, so what remains unmodelled in the data is completely random. There are many benefits to ARX estimation, and it is possible to synergistically integrate the various steps involved in identification to develop effective, practical procedures. Substantial literature exists that can provide guidance for this process. For instance, Ljung (1994) suggests starting with fourth-order ($[n_a = 4\ n_b = 4\ n_k = 1]$) ARX models; if this is not satisfactory, the recommendation is to increase the model order to eighth order, include additional inputs, or move to a nonlinear model structure. Because ARX parameter estimation consists of linear regression and is computationally inexpensive, it is possible to exhaustively examine ranges of ARX model orders and then, using cross-validation and goodness-of-fit measures, systematically identify model structure(s) and parameters that adequately describe the system. Computational tools such as the system identification toolbox in MATLAB support this functionality; however, reproducing these methods in standard statistical packages is not insurmountable.

## *3.2 Dynamic Modeling Beyond Black-Box Approaches: The Use of Behavioral Theory*

The previous section described an empirical "black-box" modeling approach for determining the system dynamics where the choice of model structure is primarily driven by aspects of goodness of fit on a validation dataset. However, it is possible, and in many instances desirable, to incorporate theories from behavioral science in a dynamical systems framework relevant to interventions (Hekler, Klasnja, Froehlich, & Buman, 2013; Martín et al. 2014; Navarro-Barrientos, Rivera, & Collins, 2011). In the following subsections, we discuss the use of the theory of planned behavior (Ajzen, 1991, 2005) and social cognitive theory (Bandura, 1986) to inform the development of dynamical systems models for behavior change.

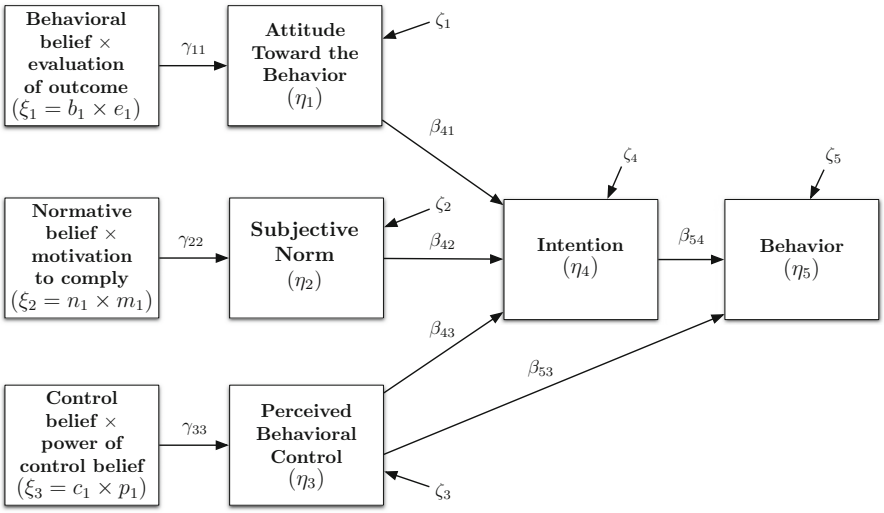


**Fig. 4** Path diagram for the theory of planned behavior

### 3.2.1 Theory of Planned Behavior and Its Fluid Analogy

The theory of planned behavior (Ajzen, 1991, 2005) is used broadly to describe the relationship between behaviors, intentions, and other psychosocial constructs. Figure 4 shows a path diagram for the theory based on structural equation modeling (Bollen, 1983). It depicts the steady-state relationships between variables. $\eta_i$ represents endogenous variables, $\xi_i$ represents exogenous variables, $\beta_{ij}$ and $\gamma_{ij}$ are regression weights, and $\zeta_i$ are disturbance variables. The theory of planned behavior says that an individual's behavior $\eta_5$ is influenced by perceived behavioral control (perceived ease or difficulty to adopt the behavior) $\eta_3$ and intention (level of motivation) $\eta_4$. Intention, in turn, is determined by attitude toward the behavior (positive and negative evaluation of the behavior) $\eta_1$, subjective norms (influence of significant others) $\eta_2$, and perceived behavioral control $\eta_3$. The exogenous variables are expressed as follows:

$$\xi_1 = b_1 \times e_1 \quad \xi_2 = n_1 \times m_1 \quad \xi_3 = c_1 \times p_1, \tag{5}$$

where $b_1$ represents the strength of beliefs about the outcome, $e_1$ is the evaluation of the outcome, $n_1$ is the strength of normative beliefs, $m_1$ is the strength of the motivation to comply to the different normative beliefs, $c_1$ is the strength of the control belief, and $p_1$ is the perceived power of the control factor.

The path diagram representation in Fig. 4 is typically used to describe phenomena cross-sectionally, and it does not take time into account. It is possible, however, to rely on information provided by the path diagram to develop a dynamical model corresponding to the theory of planned behavior, in which variables are expressed as

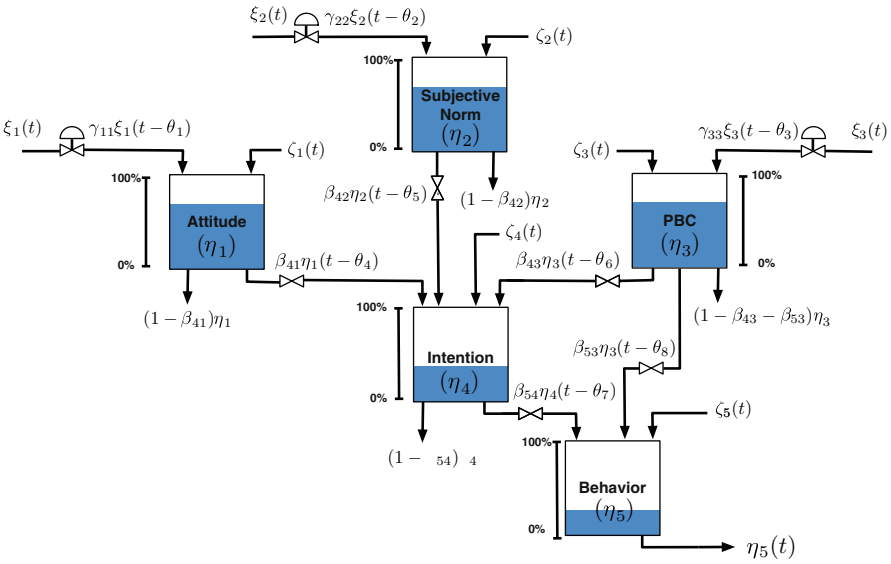


**Fig. 5** Fluid analogy for the theory of planned behavior, consistent with the path diagram in Fig. 3

functions of time, as demonstrated using the concept of fluid analogies in Navarro-Barrientos et al. (2011). Each of the endogenous variables in the model can be postulated as liquid (e.g., water) in a tank; the value of the endogenous variable defines the fluid inventory, as depicted in Fig. 5. The exogenous variables $\xi_1$, $\xi_2$, and $\xi_3$ are inflows into the inventories (i.e., dosage of each intervention component influencing the target behavior). A dynamical system description can be generated by applying the principle of conservation of mass (Ogunnaike & Ray, 1994) to each inventory, leading to a system of ordinary differential equations:

$$\tau_1 \frac{d\eta_1}{dt} = \gamma_{11}\, \xi_1\,(t-\theta_1) - \eta_1(t) + \zeta_1(t) \tag{6}$$

$$\tau_2 \frac{d\eta_2}{dt} = \gamma_{22}\, \xi_2\,(t-\theta_2) - \eta_2(t) + \zeta_2(t) \tag{7}$$

$$\tau_3 \frac{d\eta_3}{dt} = \gamma_{33}\, \xi_3\,(t-\theta_3) - \eta_3(t) + \zeta_3(t) \tag{8}$$

$$\tau_4 \frac{d\eta_4}{dt} = \beta_{41}\, \eta_1\,(t-\theta_4) + \beta_{42}\, \eta_2\,(t-\theta_5) + \beta_{43}\, \eta_3\,(t-\theta_6) - \eta_4(t) + \zeta_4(t) \tag{9}$$

$$\tau_5 \frac{d\eta_5}{dt} = \beta_{54}\, \eta_4\,(t-\theta_7) + \beta_{53}\, \eta_3\,(t-\theta_8) - \eta_5(t) + \zeta_5(t) \tag{10}$$

where $\tau_i$ are time constants, $\theta_i$ represent time delays, and $\zeta_i$ represent disturbances. In this dynamical representation, the regression weights $\beta_{ij}$ and $\gamma_{ij}$ from the structural equation model correspond to gains of the dynamical system. The dynamical model can be used for modeling idiographic behavior; these parameters can be used to determine the shape and the speed of the response of the individual's behavior changes. While (6), (7), (8), (9), and (10) are written using first-order derivatives, higher-order derivatives with additional parameters can be used to augment the model equations to reflect more complex dynamical system behavior such as inverse response, underdamped (e.g., oscillatory) response, and the like. The use of higher-order derivatives is discussed in Martín et al. (2014) and Navarro-Barrientos et al. (2011).

This section describes one means for arriving at a dynamical system model for the theory of planned behavior; additional examination in a control engineering context has been developed by Vanderwater and Davison (2011).

### 3.2.2 Social Cognitive Theory Dynamical Model

Social cognitive theory describes a human agency model in which individuals proactively self-reflect, self-regulate, and self-organize (Bandura, 1989). It estimates an individual's ability to engage in a targeted behavior, based on internal and external parameters and their interrelationships, with some self-perceived and others externally measured. Social cognitive theory components occur as a consequence of variation in external or internal stimuli. From an engineering perspective, a number of social cognitive theory components can be treated as output variables ($y$), such as the following.

- *Self-efficacy*, which is the perceived confidence in one's ability to perform a target behavior. It is an essential factor that influences behavior and that is influenced by behavior and the environment (e.g., one's belief that one is able to go for a walk every day).
- *Outcome expectancies* are the perceived likelihood that performing a target behavior will result in specific, anticipated outcome (e.g., one's belief that going for a walk every day will result in weight loss).
- *Behavior* is the action of interest (e.g., walking).
- *Behavioral outcomes*, which are outcomes obtained as a result of the behavior. These are directly related to outcome expectancies and the future behavior (e.g., increasing daily steps by 3000). In the case of physical activity, for example, a behavioral outcome could be weight loss (positive) or pain resulting from exercise (negative).

According to the theory, there are variables that act as stimuli to promote or inhibit behavior and to attenuate or enhance the effect of the components. These can be considered inputs to the system, and they can be external or internal to the individual. The following are some examples of these types of variables.

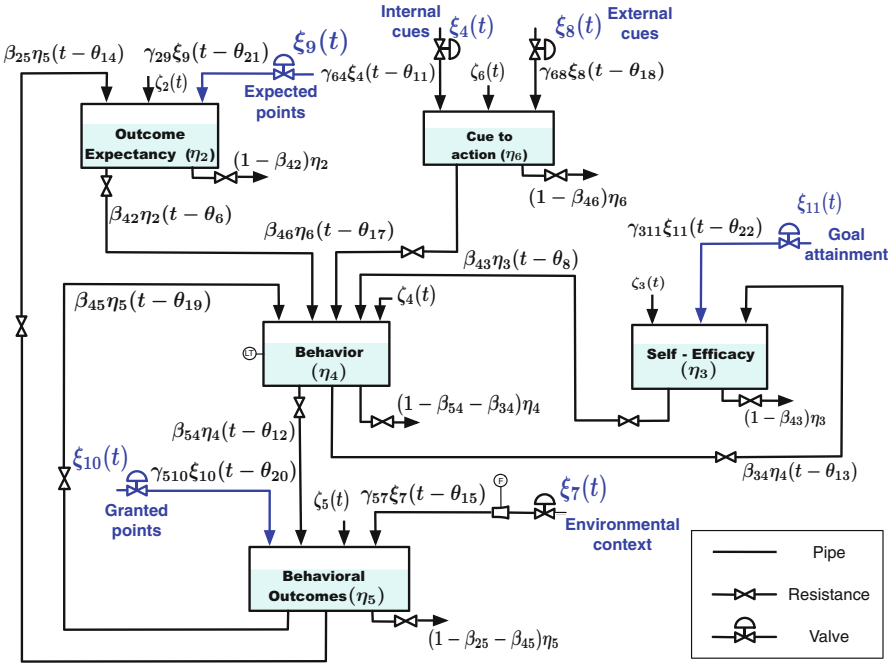


**Fig. 6** Fluid analogy for a version of the dynamic social cognitive theory model meaningful to a physical activity intervention. Inputs are represented as inflows and outputs as inventory levels

- *Environmental context* in which the behavior occurs and which directly influences the resultant behavioral outcomes. For physical activity, this could include factors such as weather or whether a day is a weekday or weekend day.
- *Internal and external cues to action* that directly influence behavior. In social cognitive theory, beliefs (e.g., self-efficacy) are conceptualized as predispositions for engaging in a behavior that is then triggered by a *cue to action*.

A fluid analogy of the social cognitive theory has been developed (Martín et al. 2014; Riley et al. 2016) that depicts how the various components relate with one another over time, particularly to understand behavior. Figure 6 is a simplified version of the model that represents how a behavioral scientist might articulate the determinants of behavior (Ferster, 1970). Main constructs are treated as inventories; other components and properties are depicted as inflows and/or outflows.

In Fig. 6, behavior ($\eta_4$) is represented as a fluid inventory that increases and decreases in frequency and/or duration over time. Self-efficacy ($\eta_3$) is represented as an inventory of varying levels that differs between different behaviors, different individuals, and even within an individual. Prior experience engaging in the behavior ($\beta_{34}$) is a critical learning feedback loop that adds or depletes self-efficacy to subsequently engage in the behavior.

Martín et al. (2014) and Riley et al. (2016) describe a model that is broadly applicable to many health behaviors. Consider the Just Walk intervention (explained further in Sect. 5), where the main goal is to promote physical activity among sedentary individuals. Such an intervention relies on the systematic delivery of the following intervention components, based on individual performance.

- *Daily goals* ($u_8$), to quantify the desired behavior (e.g., 10,000 steps per day).
- *Expected points* ($u_9$), the announced daily reward points that will be granted to individuals if they achieve the daily goal.
- *Granted points* ($u_{10}$), given every day if individuals reach the set goal. Points can later be exchanged for tangible rewards (e.g., gift cards).
- As depicted in Fig. 6, additional inputs include the following.
- *Outcome expectancy (OE) for reinforcement:* expected daily reward points ($\xi_9$).
- *Reinforcement:* received daily reward points resulting from behavior ($\xi_{10}$).
- *Goal attainment:* ($\xi_{11}$) computed as the difference between the daily goal and the actual performed behavior, which influences self-efficacy. This signal is particularly useful in developing a social cognitive theory model that properly predicts "ambitious but doable" goals, since individuals might react negatively to a goal that seems unattainable.

To obtain a mathematical model, it is necessary to describe how the inventories and their respective inflows and outflows fit within a dynamical system. The procedure is as described by Navarro-Barrientos et al. in a dynamic model for the theory of planned behavior Navarro-Barrientos et al. (2011). Five inventories represented by the variables $\eta_2, \ldots, \eta_6$ are considered in the fluid analogy in Fig. 6. The exogenous inputs are represented by $\xi_4$, $\xi_7$, $\xi_8$, $\xi_9$, $\xi_{10}$, and $\xi_{11}$. From each inventory there are a number of inflow "resistances" (represented by the coefficients $\gamma_{25}, \ldots, \gamma_{68}$) and outflow "resistances" (represented by $\beta_{25}, \ldots, \beta_{46}$). There are additional parameters that represent the physical characteristics of each inventory and flow; these have an important effect on the dynamic behavior of the system. There are the time constants $\tau_2, \ldots, \tau_6$ that denote capacity and allow for exponential decay (or growth) of the inventory, and time delay parameters ($\theta_2, \ldots, \theta_{22}$) can be used for each flow signal to capture dead time. Unmeasured disturbances (which can reflect unmodelled dynamics) are also considered as $\zeta_2$, $\ldots, \zeta_6$.

In the fluid analogy, the principle of conservation of mass is used, such that for each inventory, an accumulation term is defined as the sum of all the inflows minus the sum of all the outflows. The accumulation term is denoted by the time constant $\tau$ times the rate of change (derivative) in the inventory. The following equations define the system for each tank.

$$\tau_2 \frac{d\eta_2}{d\mathrm{t}} = \gamma_{29}\xi_9\,(t-\theta_{21}) + \beta_{25}\eta_5\,(t-\theta_{14}) - \eta_2(t) + \zeta_2(t) \tag{11}$$

$$\tau_3 \frac{d\eta_3}{d\mathrm{t}} = \gamma_{311}\xi_{11}\,(t-\theta_{22}) + \beta_{34}\eta_4\,(t-\theta_{13}) - \eta_3(t) + \zeta_3(t) \tag{12}$$

$$\tau_4 \frac{d\eta_4}{d\mathrm{t}} = \beta_{42}\eta_2\left(t-\theta_6\right) + \beta_{43}\eta_3\left(t-\theta_8\right) + \beta_{46}\eta_6\left(t-\theta_{17}\right) + \beta_{45}\eta_5\left(t-\theta_{19}\right) - \eta_4(t) + \zeta_4(t) \tag{13}$$

$$\tau_5 \frac{d\eta_5}{d\mathrm{t}} = \gamma_{57}\xi_7\left(t-\theta_{15}\right) + \gamma_{510}\xi_{10}\left(t-\theta_{20}\right) + \beta_{54}\eta_4\left(t-\theta_{12}\right) - \eta_5(t) + \zeta_5(t) \tag{14}$$

$$\tau_6 \frac{d\eta_6}{d\mathrm{t}} = \gamma_{64}\xi_4\left(t-\theta_{11}\right) + \gamma_{68}\xi_8\left(t-\theta_{18}\right) - \eta_6(t) + \zeta_6(t) \tag{15}$$

As in the case of the theory of planned behavior, the system represented by (11), (12), (13), (14), and (15) is written using first-order differential equations, but in order to describe more elaborate transient responses (that may include overshoot and oscillatory behavior), a higher-order model relying on second-order derivatives could be used (Navarro-Barrientos et al. 2011). More detailed descriptions of the constructs, model considerations, simulation scenarios, and additional features that can be incorporated to the model are provided in Martín et al. (2014) and Riley et al. (2016).

### 3.2.3 Other Theories: Self-Regulation and Mediation

Self-regulation theory in psychology has been largely influenced by the work of Carver and Scheier (1998). As discussed in Dong et al. (2012) and Carver and Scheier (1998), many theories of self-regulation are conceptualized on the basis of feedback control systems. How self-regulation can be incorporated within the scope of the theory of planned behavior is presented in the work of Dong et al. (2012). Likewise, Timms and co-workers (Timms, 2014; Timms, Rivera, Collins, & Piper, 2014) have developed dynamical systems models from path models for mediation, using the ideas and concepts described in this chapter. The interested reader can consult these references for details.

## 4 Model Predictive Control

This section presents MPC as the algorithmic framework for making systematic dosage assignments in closed-loop, intensively adaptive interventions. MPC has widespread application in diverse industries, ranging from petrochemicals to aerospace (Qin & Badgwell, 2003). This technology has also been useful in designing treatment regimens for diverse medical applications, from diabetes mellitus control to HIV/AIDS treatment (Deshpande, 2011; Nandola & Rivera, 2013; Wang,

Dassau, & Doyle, 2010; Zurakowski & Teel, 2006). As noted previously, control engineering, such as MPC, is one approach to optimizing an adaptive intervention, and a closed-loop experiment for the purpose of developing a controller is one type of optimization trial (Collins, 2018). In this section we further describe the logic behind MPC.

Figure 7 depicts the "receding horizon" strategy that forms the conceptual basis for MPC. The predicted change in outputs in the future (calculated using the estimated model from system identification) augmented with predictions from measured disturbance variables (when available) can be used to generate an *error projection* signal that represents expected deviations between the outputs and the goal, based on current and past control actions. Because past actions of the control system have already taken place and cannot be reversed, the only option left to the controller is to determine future values of the manipulated variables that will bring the outputs closer to desired reference trajectories. On the basis of the error projection signal and the dynamical model, the optimization step in MPC calculates a sequence of future control actions that will minimize the projected error with respect to a desired goal. The control actions are determined by an online optimization algorithm as follows: the constrained optimization problem shown in Eqs. (16), (17), (18), and (19) is numerically solved at each decision point (e.g., daily) based on model-based predictions over a period of time (defined by the prediction horizon $p$) and optimal manipulated variable changes over a period of time (defined by the move horizon $m$). Instead of implementing all of the calculated control actions, only the one corresponding to the immediate future decision point is applied, with the procedure repeated at the next assessment interval (e.g., daily) and then continuously until the end of the intervention. This strategy allows the controller to respond to unforeseen or unpredictable changes (i.e., unmeasured disturbances) and account for plant-model mismatch (i.e., model error) as it systematically relies on up-to-date participant response information to make its calculations. In Eq. (16), the optimization objective is related to the error between the predicted values ($y(k+1), \ldots, y(k+p)$) and the reference set point $y_r$, with $Q_y$ as a user-defined weight. It is solved under constraints on the allowable minimum and maximum outcome values, input treatment dosages, and their rate of change (represented by $\Delta u(k) = u(k) - u(k-1)$). Variables $p$ and $m$ correspond to the prediction horizon and the control horizon, respectively.

$$\min_{\left\{[u(k+i)]_{i=0}^{m-1}\right\}} J \triangleq \sum_{i=1}^{p} \left(y\left(k+i\right) - y_r\right)^T Q_y \left(y\left(k+i\right) - y_r\right) \tag{16}$$

$$y_{\min} \leq y\left(k+i\right) \leq y_{\max}, 1 \leq i \leq p \tag{17}$$

$$u_{\min} \leq u\left(k+i\right) \leq u_{\max}, 0 \leq i \leq m-1 \tag{18}$$

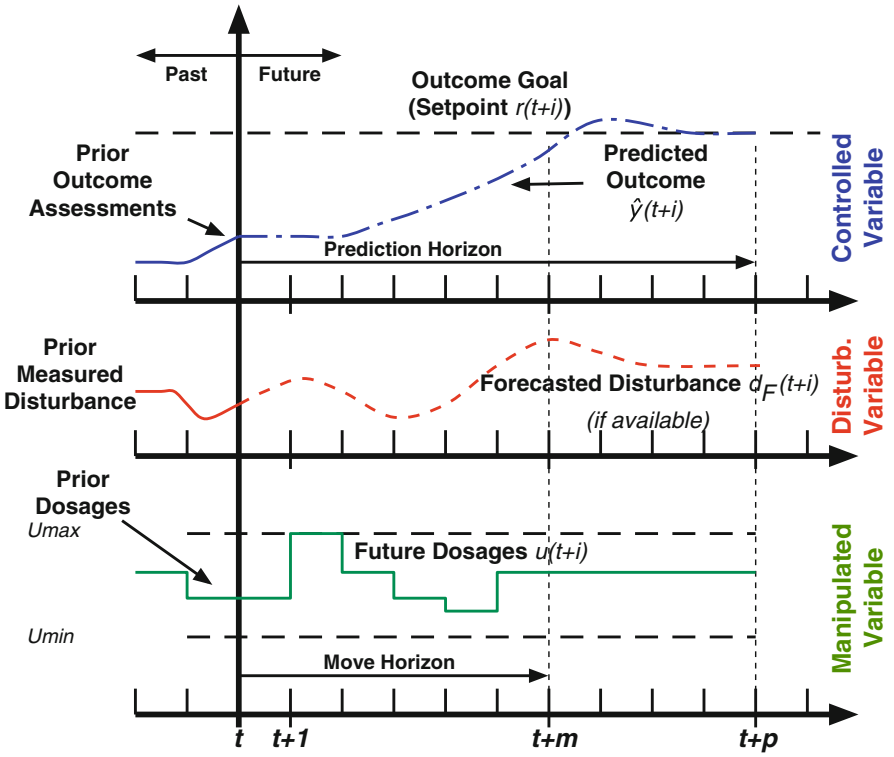


**Fig. 7** Conceptual representation of the receding horizon strategy used by the MPC algorithm, depicted for an intervention featuring scalar controlled ($y$), manipulated ($u$), and measured disturbance ($d$) variables. In an intervention to promote increased physical activity (such as the Just Walk intervention described in Sect. 5), the controlled variable represents daily steps walked by a participant, which should ultimately reach a desired outcome goal (denoted by the set point signal $r$). The controller algorithm determines daily step goals (the manipulated variable) over a future horizon of decision intervals (denoted by $m$) to achieve this goal with minimum deviation (as determined by an objective function). Measurements and forecasts of a disturbance variable (such as predicted busyness) are used by the MPC algorithm to improve the prediction of future outcomes (over a prediction horizon $p$) and establish a closed-loop error projection signal that is then mitigated by the controller

$$\Delta u_{\min} \leq \Delta u\,(k+i) \leq \Delta u_{\max}, 0 \leq i \leq m-1 \tag{19}$$

The optimization problem in Eq. (16) is solved through established numerical procedures from operations research. For linear dynamical models with categorical inputs, the optimization problem in Eq. (16) is solved using a mixed-integer quadratic program (MIQP). Details of the algorithm and its solution are provided in Nandola and Rivera (2013).

There are many possible ways to formulate an MPC algorithm, all featuring the receding horizon approach described previously. The MPC algorithm evaluated in this chapter relies on what is referred to as a "three degree-of-freedom" (3DoF) tuning approach to achieve desired levels of performance (Lee & Yu, 1994; Wang & Rivera, 2008). The 3DoF tuning method enables the performance requirements associated with reaching a desired goal and accounting for measured and unmeasured disturbances to be adjusted independently by varying three "knobs" represented by the parameters $\alpha_r$, $\alpha_d$, and $f_a$, respectively (Deshpande, 2011; Deshpande et al. 2011; Nandola & Rivera, 2013). This tuning approach provides the intervention scientist a flexible and intuitive method to adjust the controller so that the outputs achieve a desired speed and shape of response. In the ensuing sections, we demonstrate how the 3DoF formulation enables obtaining desirable participant response profiles over time while accommodating clinical constraints.

An important aspect of the 3DoF approach is its ability to address the presence of plant-model mismatch (i.e., model error) and provide *robustness* to the control system. In industrial practice, robustness to modeling errors and other forms of uncertainty is an important practical consideration (Aström & Murray, 2010; Skogestad & Postlethwaite, 1996). In the context of behavioral interventions, robustness issues arise from the statistical errors resulting from parameter estimation during system identification, the variability that may exist between participants, unmodelled dynamics that are not captured during system identification, or changes in the model that may occur within the participant over time. Space limitations prevent describing these concepts in more detail, but control system robustness is evaluated in a series of simulated scenarios relevant to the adaptive intervention that is described in Deshpande (2011, 2014).

## 5 Control Systems Engineering for a Physical Activity Intervention: Just Walk

### *5.1 Description of the Just Walk Intervention*

Just Walk (Freigoun et al. 2017; Hekler et al. 2018; Korinek et al. 2018; Phatak et al. 2018) was developed as an mHealth adaptive, walking intervention for sedentary, overweight adults, designed primarily as a tool to generate individualized computational models for understanding behavior via system identification. The pilot intervention system included a front-end Android app (Just Walk), a backend server, and wearable devices (Fitbit Zip) to objectively measure physical activity and automatically sync with the smartphone application. Participants were recruited nationally to partake in a walking intervention in which they would be assigned daily step goals via the Just Walk app, and points were awarded if the goals were achieved. Points were converted into Amazon gift cards after a certain threshold was reached. Participants were also required to complete a series of daily morning

and evening ecological momentary assessment (EMA; Shiffman, Stone, & Hufford, 2008) measures (e.g., confidence in achieving goal, predicted busyness for that day, previous night's sleep quality) for the entire duration of the study.

The total study duration was 14 weeks, including an initial 2-week baseline period in which no step goals were delivered. Each participant's step goals were based on his/her median daily step value, calculated from the 14-day baseline period, and were designed to establish a mechanism for individualizing the definition of an ambitious but achievable step count. The methods used to determine each daily goal are described below. All physical activity data were collected from the Fitbit Zip (provided to participants as a part of the study) and stored both locally and in Fitabase (Small Steps Labs, San Diego, CA, USA; a secure research platform that stores information provided to it by the Fitbit API in an SQL server database).

Participants were generally healthy, inactive, 40–65 years old, with a body mass index (BMI) of 25–45 kg/m$^2$ who currently owned an Android phone capable of connecting to a Fitbit Zip via Bluetooth 4.0 and were willing to engage with the mHealth intervention for 14 weeks. Participants were considered inactive and eligible for the study if they engaged in less than 1000 metabolic equivalent of task (MET)-minutes/week, as measured by the International Physical Activity Questionnaire (IPAQ) (Craig et al. 2003). Individuals were excluded if they did not speak English, were pregnant, had a BMI over 45 kg/m$^2$, indicated medical problems that precluded unsupervised physical activity based on the Physical Activity Readiness Questionnaire (PAR-Q), or were currently participating in a commercial or research-related diet or exercise program. Participants were recruited nationally through community advertising techniques (e.g., emails to listservs, word-of-mouth, social media advertisements) and were provided a Fitbit Zip upon enrolling in the study.

## *5.2 System Identification Design of the Open-Loop Intervention*

The Just Walk intervention features an innovative idiographic (single-subject) experimental design based on system identification, resulting in a unique dataset that is the basis of this study. The goal of this research activity is to perform a number of analyses leading to more predictive models that will ultimately inform intervention development, including development of the set of decision rules that will define a personalized and perpetually adapting intervention (i.e., one that can take into account daily changes in variables such as personal stress levels, sleep, and social context). Figure 8 shows screenshots of the mobile application that was specifically developed for this study. This mobile application automatically retrieves intensive measurements (e.g., physical activity, location, weather, day of the week) and also records scaled states from daily participant feedback (e.g., sleep, stress, mood).

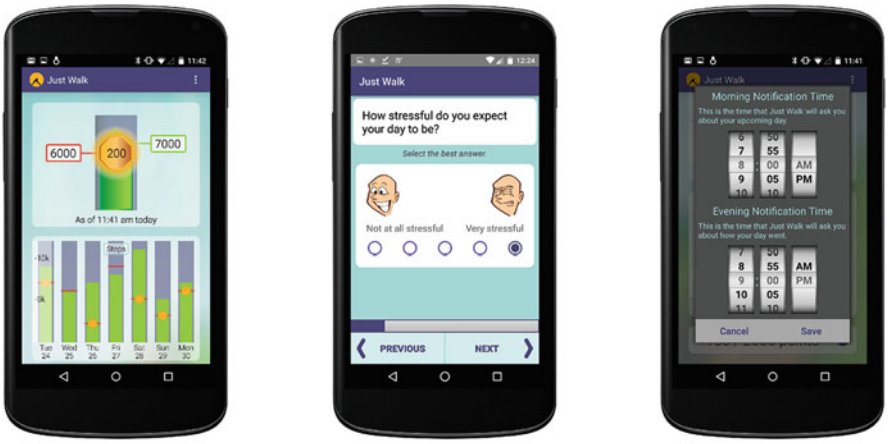


**Fig. 8** Selected screenshots from the Just Walk app. (**a**) displays the opened app displaying progress towards the current goal (red box), daily steps achieved (as determined from the latest sync with Fitbit; green box) and expected points that the participant will receive if the goal is achieved (gold medallion). The lower portion of the screenshot displays the step and goal achievement history. (**b**) displays an example of a daily morning question. (**c**) shows how the user can specify morning and evening notification times

### 5.2.1 Input Signal Design

The input signal design procedure utilized in the Just Walk study relies on using deterministic yet "pseudo-random" signals that are orthogonal (i.e., independent) in both time and frequency domains. The procedure is described in detail in the work of Martín, Rivera, and Hekler (2015a). In Just Walk, two manipulated variables are adjusted experimentally: *goals* and *expected points*. *Goals* quantify the desired behavior at a daily level, while *expected points* are the daily available points announced each morning that are granted upon achievement of the daily goal. These two manipulated input signals $u_n, n = 1, 2$ are generated from a sum of sinusoid (i.e., multisine) signals, defined as

$$u_n(k) = \lambda_n \sum_{j=1}^{n_s} \sqrt{2\alpha_{[n,j]}} \cos\left(\omega_j k T_s + \phi_{[n,j]}\right), \quad \omega_j = \frac{2\pi j}{N_s T_s}, \quad k = 1, \ldots, N_s. \tag{20}$$

As manipulated variables, the parameters and properties of the multisines in (20) are specified by the intervention scientist. $N_s$ corresponds to the number of samples per period, while $T_s$ is the sampling time ($T_s = 1$ day for Just Walk). $\omega_j$ is the frequency for the $j$th sinusoid; these are linearly spaced over an interval. The number of sinusoids $n_s (\leq N_s/2)$ and their corresponding amplitudes (specified through $\lambda_n$

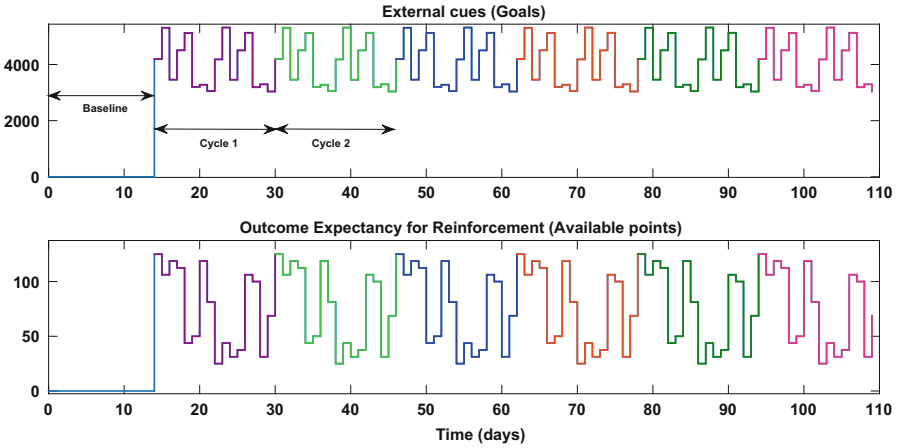


**Fig. 9** Illustration of the time series corresponding to multisine excitation for manipulated variables (goals and available points) in Just Walk. Signal magnitudes are personalized

and $\alpha_{[n,j]}$) are chosen to meet identification requirements for achieving sufficient variation in the signal (formally referred to as persistency of excitation); persistent excitation is dictated by the order of the models to be identified (Ljung, 1999). In this experiment, $n_s = 6$ is used. To obtain signals that are statistically independent, $\alpha_{[n,j]}$ is chosen so that the signals are orthogonal in frequency, leading to a so-called "zippered" power spectral density, as described in Rivera, Lee, Mittelmann, and Braun (2009). A signal length corresponding to $N_s = 16$ days was chosen to avoid any connection of signal periodicity with the number of days in a week, and the phases $\phi_{[n,j]}$ were selected to minimize the crest factor (i.e., time-domain distribution) of the signal using the algorithm developed by Guillaume, Schoukens, Pintelon, and Kollar (1991).

Figure 9 illustrates the time series corresponding to manipulated inputs *goals* and *expected points* as two orthogonal multisine signals following a baseline (Martín et al. 2015a; Phatak et al. 2018). Magnitudes of the input signals were chosen relying on experiences from previous studies (Adams et al. 2013; King et al. 2013) designed to obtain an expected profile of physical activity. The maximum number of steps to be set as a goal is calculated as a factor of the initial baseline level of physical activity. For most cases in our experimental design, this factor is equal to 2; however, it was varied if the actual baseline step level of individuals was too high or low. Specifically, if a participant's baseline median steps were below 3000, then the range for the goals was between 1 and 2.5 of their baseline median steps, to increase the likelihood of "ambitious" goals. If baseline median steps were greater than 7500 steps, then the range was set between 1 and 1.75 (to reduce the likelihood of overly ambitious goals, such as 15,000 steps in 1 day). The cycles are repeated for the duration of the study; each additional cycle that is implemented increases the length of the overall dataset, offering the opportunity to collect validation data and helping to reduce variance errors in the subsequent model estimates.

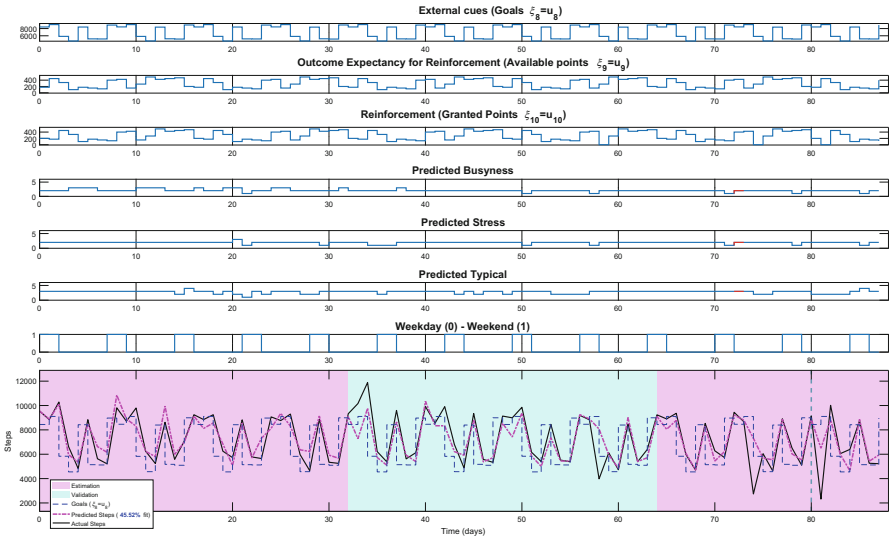


**Fig. 10** Time series plot showing seven selected input sequences (manipulated inputs and measured disturbances), predicted behavior (from an ARX black-box model), actual behavior, model overall fit, and estimation and validation cycles (1st, 2nd, and 5th for estimation; 3rd and 4th for validation) for a selected Just Walk participant

In addition to the two manipulated inputs, a large set of disturbances were also measured using mHealth technologies (smartphone and wearable devices). Overall experimental duration beyond the baseline varied between five and six cycles for each participant. A time series plot for a representative participant that depicts the behavior and seven inputs is shown in Fig. 10.

### 5.2.2 Data Preprocessing and Model Structure Considerations

Having executed the experiment and collected data, the next step is data preprocessing (see Fig. 3). Data preprocessing tasks include interpolation (to impute missing values) and mean subtraction.

Model structure selection decisions consist of determining the input signals to be included for each participant and the corresponding ARX model orders for the output and each input, in accordance with Eq. (2). Taking advantage of the computational simplicity of ARX modeling, the approach taken here is to exhaustively examine a range of model orders and use model validation procedures to determine the most suitable structure. For this problem, ARX model order ranges for $n_a$ and $n_b$ from 1 to 3 (i.e., $\max(n_a) = 3$ and $\max\left(n_{b_j}\right) = 3\ \forall j = 1, \ldots n_u$) seem reasonable. A priori knowledge of the social cognitive theory fluid analogy model developed in Martín et al. (2014) implies that very high-order models should not be necessary to characterize these behavior change dynamics. From

inspecting the intervention data, it is reasonable to assume a basic unit input delay (i.e., $n_{k_j} = 1 \; \forall j$). Finally, the absence of drifts in the data leads us to assume stationary (though potentially time-varying) noise characteristics over the course of the intervention period.

In determining the inputs to be considered, our approach is to start with a basic three-input model consisting of *goals* ($u_8$), *expected points* ($u_9$), and *granted points* ($u_{10}$); from this basic model then additional inputs (e.g., *predicted busyness*, *predicted stress*) are added, with all possible combinations of these inputs estimated. Model validation following estimation ultimately determines which of these inputs are most important in describing individual behavior; nonetheless, in the preprocessing stage, correlation analysis can be used to determine inputs that may be significantly cross-correlated with each other or to identify inputs that appear to have no significant effect on the output. In both scenarios, the number of inputs that need to be considered in the parameter estimation procedure can be reduced, ultimately leading to parsimonious models that can be generated with less effort.

### 5.2.3 Model Parameter Estimation and Validation

Model estimation and concomitant validation with the Just Walk intervention data are now considered. As noted previously, initially a core three-input model consisting of *goals* ($u_8$), *expected points* ($u_9$), and *granted points* ($u_{10}$) is evaluated, then additional inputs (e.g., *predicted busyness*, *predicted stress*) are added, with all possible combinations of the additional inputs being estimated. At an individual level, the full dataset is segmented into informative 16-day cycles for model estimation/validation. The cycle length is defined by the multisine input signal described in Sect. 5.2.1.

Cross-validation (the process of evaluating model fit over data not used for estimation) represents one of the most valuable activities in system identification (Ljung, 1994). The conventional approach in system identification is to assign a certain percentage of data for estimation, followed by validation (e.g., 50% estimation, 50% validation). Such an approach assumes that the noise characteristics of the problem remain unchanged during the course of the intervention. However, it is reasonable to expect that noise and disturbance characteristics will vary over long-duration interventions such as Just Walk. In our analysis, each data cycle is assigned to either estimation or validation; all combinations of data cycles involving at least two cycles for validation are generated and evaluated.

Table 2 summarizes results of this procedure for a four-input model (*goals*, *expected points*, *granted points*, and *predicted busyness*) of a selected participant. The NRMSE fit index from Eq. (4) is calculated for each cycle and averaged for estimation and validation data, respectively. All data cycle combinations that feature at least two cycles for validation or estimation (20 candidate ARX models) are evaluated. For each of these combinations of estimation and validation cycles (corresponding to a specific row in Table 2), ARX orders are determined from an exhaustive search routine that selects a stable ARX model with the highest

**Table 2** Intermediate results for a 4-input ARX model of a selected participant from Just Walk. The model highlighted in yellow corresponds to the best model (balancing estimation and validation data fits) for four inputs, for this participant

| E* | V* | NRMSE Fit (%) | | | | | Average Estimation NRMSE Fit (%) | Average NRMSE Validation Fit (%) | Overall NRMSE Fit (%) | ARX Order (4-input) [$na_1$,$nb_1$,$nb_2$,$nb_3$,$nb_4$] |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Exp. #1 | Exp. #2 | Exp. #3 | Exp. #4 | Exp. #5 | | | | |
| [1,2] | [3,4,5] | 77.40% | 85.44% | 79.27% | 27.68% | 13.70% | 81.42% | 40.22% | 40.11% | [1,1,1,1,3] |
| [1,3] | [2,4,5] | 77.39% | 82.25% | 81.30% | 26.88% | 15.36% | 79.35% | 41.50% | 40.60% | [1,2,1,1,3] |
| [1,4] | [2,3,5] | 64.82% | 71.25% | 67.27% | 45.89% | 21.04% | 55.36% | 53.19% | 42.29% | [1,3,1,1,1] |
| [1,5] | [2,3,4] | 61.36% | 59.51% | 60.96% | 40.14% | 24.47% | 42.92% | 53.54% | 37.40% | [1,1,1,3,1] |
| [2,3] | [1,4,5] | 70.46% | 90.25% | 84.15% | 25.00% | 11.19% | 87.20% | 35.55% | 37.70% | [3,3,1,2,3] |
| [2,4] | [1,3,5] | 49.06% | 71.94% | 67.25% | 52.39% | 22.98% | 62.17% | 46.43% | 40.56% | [3,1,2,1,3] |
| [2,5] | [1,3,4] | 54.89% | 61.75% | 60.36% | 47.35% | 23.68% | 42.72% | 54.20% | 39.33% | [3,1,1,1,1] |
| [3,4] | [1,2,5] | 45.97% | 61.27% | 69.24% | 51.46% | 24.02% | 60.35% | 43.75% | 41.15% | [1,3,3,1,3] |
| [3,5] | [1,2,4] | 63.11% | 66.96% | 52.29% | 41.52% | 19.47% | 35.88% | 57.20% | 41.12% | [1,1,1,1,1] |
| [4,5] | [1,2,3] | 36.37% | 52.47% | 50.06% | 49.24% | 25.88% | 37.56% | 46.30% | 32.75% | [1,1,1,3,2] |
| [3,4,5] | [1,2] | 53.63% | 64.61% | 49.26% | 46.59% | 19.93% | 38.59% | 59.12% | 40.12% | [1,1,1,1,1] |
| [2,4,5] | [1,3] | 50.12% | 59.76% | 59.36% | 49.92% | 23.64% | 44.44% | 54.74% | 38.71% | [3,1,1,1,1] |
| [2,3,5] | [1,4] | 58.63% | 66.76% | 64.91% | 49.62% | 27.28% | 52.98% | 54.13% | 40.59% | [3,1,3,2,1] |
| [2,3,4] | [1,5] | 59.43% | 76.99% | 70.11% | 41.51% | 22.32% | 62.87% | 40.88% | 41.61% | [2,3,3,2,3] |
| [1,4,5] | [2,3] | 57.91% | 61.11% | 60.18% | 45.69% | 24.92% | 42.84% | 60.65% | 38.81% | [1,1,1,3,1] |
| [1,3,5] | [2,4] | 66.34% | 66.02% | 67.24% | 42.13% | 22.57% | 52.05% | 54.08% | 41.31% | [1,3,1,1,1] |
| [1,3,4] | [2,5] | 68.39% | 77.75% | 73.46% | 41.86% | 18.78% | 61.24% | 48.27% | 42.26% | [1,3,2,1,1] |
| [1,2,5] | [3,4] | 61.85% | 56.05% | 68.43% | 44.82% | 35.02% | 50.97% | 56.63% | 46.03% | [2,3,1,2,3] |
| [1,2,4] | [3,5] | 71.99% | 73.18% | 72.36% | 43.28% | 20.40% | 62.82% | 46.38% | 43.61% | [1,2,1,1,3] |
| [1,2,3] | [4,5] | 75.95% | 87.02% | 80.67% | 26.39% | 13.36% | 81.21% | 19.88% | 39.87% | [1,1,1,1,3] |

$E \equiv$ Estimation Cycles (magenta), $V \equiv$ Validation Cycles (cyan)

predictive ability (based on the maximum average validation fit). This step provides a safeguard against overparameterization.

The final chosen model should reflect, in addition to a good fit to validation data, a good fit for the entire dataset (consisting of both estimation and validation cycles). This suggests that the final model choice should correspond to the model that yields highest overall fit (the "overall NRMSE fit" column in Table 2). Incorporating the overall fit criterion with the fit to cross-validation data balances good prediction with model accuracy over the entire dataset. Note that using this analysis, the best results for the specific participant occur in the model resulting from row 18 (cycles 1, 2, and 5 for estimation; 3 and 4 for validation) with an overall NRMSE index of 46.03% for a model with structure $n_a = 2$, $n_{b_1} = 3$, $n_{b_2} = 1$, $n_{b_3} = 2$, and $n_{b_4} = 3$. This model performs close to the model with best fit over the validation data (56.63% for row 18 vs. 60.65% in row 15); however, the model with the best fit to validation data does not exhibit the best fit to data overall (38.81% in lieu of 46.03%).

### 5.2.4 Overall Fit Analysis and Assessment of Individual Participant Characteristics

Analyses similar to those in Table 2 can be performed with additional inputs for all possible combinations. For example, for a total of seven inputs, 16 different input models can be generated for each participant (since *goals* ($u_8$), *expected*

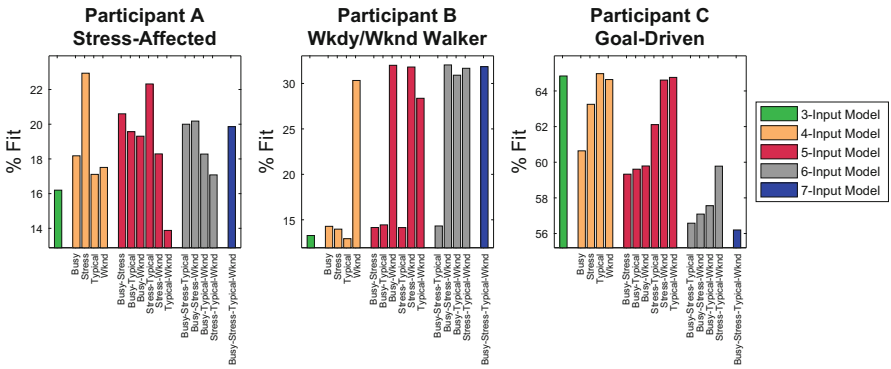


**Fig. 11** Average validation fit percentages of individualized ARX models from black-box system identification for three individual participants from Just Walk. The base model (in green) contains the inputs goals, expected points, and granted Points. Additional inputs consist of *Busy* (i.e., predicted busyness), *stress* (i.e., predicted stress), *typical* (i.e., predicted typicality of the participant's day), and *wknd* (i.e., weekday vs. weekend)

*points* ($u_9$), and *granted points* ($u_{10}$) are always grouped). Evaluating these 16 input combinations enables making some important conclusions on participant characteristics resulting from the intervention.

Figure 11 depicts model validation fit percentages from three different participants from Just Walk. The Y-axis indicates the fit percentage of the 3, 4, 5, 6, and 7-input models, and the X-axis corresponds to the psychosocial measures (busyness, stress, weekday, typicality (i.e., how typical the day was)) measured daily. Here, we can see that participant A's walking behavior is largely driven by stress (highest fit percentage seen for the stress bar in the four-input model), participant B's behavior is driven by day of the week, and participant C has the highest fit percentage for the three-input model, indicating that the daily step goal had the greatest impact on behavior. Step responses from the individual ARX models can be used to reveal more precise direction and magnitude information. This strategy has significant implications for personalized and adaptive behavior change interventions; if one can determine the inputs that are most meaningful for a given individual in a given context, it is possible then to optimize the target behavior over a specified time (hours, days, weeks, months).

## *5.3 Conceptual Design of a Hybrid MPC-Based Closed-Loop Intervention for Just Walk*

As noted in the Introduction, Just Walk represents an ongoing effort. One goal of future research is closed-loop implementation and validation in a real-world setting. Substantial efforts have been made to formulate and evaluate closed-loop

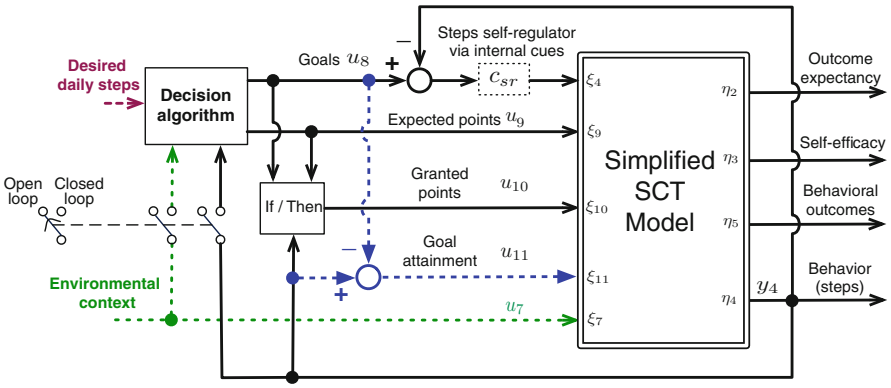


**Fig. 12** Conceptual diagram for the proposed intervention based on the simplified social cognitive theory model

intervention strategies via simulation (Martín, 2016; Martín, Rivera, & Hekler, 2016a). The proposed closed-loop intervention is depicted in Fig. 12. It relies on the simplified version of the dynamic social cognitive theory model described in Sect. 3.2.2 and includes a self-regulator via internalized cues as described in Martín (2016). The intervention considers measurements of the actual steps (*behavior* $y_4$) and environmental context (e.g., weather), which are used by the decision algorithm, now implemented as a hybrid MPC controller. While actual steps and measured environmental context constitute controlled and measured disturbance variables, respectively, from the standpoint of control engineering, they are both considered tailoring variables of the adaptive intervention, as changes in these signals have an influence on how the MPC decision algorithm decides on intervention dosages.

The goal of the Just Walk adaptive intervention is to have participants achieve a desired sustained level of daily steps, while considering some important physical and operational constraints, such as the following.

- Maximum and minimum values for goals and points ($u_8$, $u_9$, and $u_{10}$) depending on physical conditions (e.g., maximum and minimum daily step goals for an individual). Financial limitations lead to bounds on the expected reward points, since these have a direct conversion into monetary value.
- Goals and reward points must be drawn from discrete sets of integer values that may represent meaningful effects on the intervention. As prior physical activity experiments have shown (Adams et al. 2013; King et al. 2013), having a fixed set of goals and points could be important to analyze specific aspects of interest on the intervention.
- The intervention may be configured in different stages, where some of the inputs may be deactivated or partially activated. For instance, when the behavior has reached the desired level and is successfully sustained, a gradual decrease in rewards may be activated.

The control strategy for intervention design must incorporate the defined requirements and constraints for the physical activity behavioral intervention. Hybrid MPC (HMPC) (Martín et al. 2016a; Nandola & Rivera, 2013), summarized in Sect. 4, is applied to this problem since it incorporates hybrid dynamics (Bemporad & Morari, 1999). Hybrid dynamical systems consider discrete and continuous events simultaneously; they can be represented by differential (or difference) equations and logical conditions describing their categorical or binary response. The aim of control design is directed to the following tasks:

- *Set point tracking*: Goals and expected reward points are assigned to obtain the desired amount of daily steps following continuous and discrete constraints.
- *Rejection of measured disturbances*: The controller manipulates goals and expected points to mitigate the effect from measured external disturbances (e.g., environmental context) relying on the subsystem of the social cognitive theory model that is related to those signals. For instance, if an environmental event (e.g., bad weather) is known a priori, then goals or expected rewards can be adjusted to compensate for that disturbance.
- *Rejection of unmeasured disturbances*: Inputs are manipulated to mitigate the effect of unknown and possibly unmodelled external influences. For example, any unexpected situation that may impact the individual's likeliness to engage in physical activity (e.g., sickness of a family member, sudden party invitation) can be mitigated by adjustments on goals or points by the controller.

For the envisioned Just Walk closed-loop intervention, the input $u$ and output $y$ are

$$u = \left[ u_8 \; u_9 \; u_{10} \right]^T, \; n_u = 3 \tag{21}$$

$$y = \left[ y_2 \; y_3 \; y_4 \; y_5 \right]^T, \; n_y = 4 \, . \tag{22}$$

### 5.3.1 Maintenance Training Stage

In Just Walk, once the desired goal has been reached and sustained for a predetermined number of days, a maintenance training stage of the intervention is initiated, and the number of points is reduced. Here the HMPC algorithm must be reconfigured to maintain the daily performed steps in spite of a reduction in the number of points, and, if needed, reactivate the use of points if a significant relapse of sedentary behavior occurs. To adapt the HMPC performance to these new considerations, the objective function in (16) is modified to include targets on the manipulated variables:

$$\begin{aligned} J \; \triangleq & \sum_{i=1}^{p} \left( y\left(k+i\right) - y_r \right)^T Q_y \left( y\left(k+i\right) - y_r \right) \\ & + \sum_{i=1}^{m-1} \left( u\left(k+i\right) - u_r \right)^T Q_u \left( u\left(k+i\right) - u_r \right) . \end{aligned} \tag{23}$$

Beyond $y_r$ and $Q_y$ that existed previously in (16), $u_r$ and $Q_u$ represent reference values and penalty weights for $u$, respectively, which can be adjusted during the course of the intervention.

During the initiation phase, the main goal is to achieve the required daily steps. The reference output set point is $y_r = \begin{bmatrix} y_{r_2} & y_{r_3} & y_{r_4} & y_{r_5} \end{bmatrix}^T$, where $y_{r_4}$ is the desired amount of daily steps (e.g., 10,000). Considering vectors $u$ and $y$ defined in (21) and (22), the weight matrices $Q_u$ and $Q_y$ are specified in the objective function (23) such that set point tracking is applied only to the variable $y_4$ (daily steps). The maintenance stage is enabled when the goal has been achieved and sustained at least $n_s - 2$ times during the last $n_s$ days. The goal is considered achieved when the difference between the actual steps and the reference is within a predefined tolerance $\text{Tol}_4$. A logical constraint is added to the HMPC optimization problem to enforce this requirement. During this phase it is necessary to reconfigure the controller to target a low use of points ($u_9$). With target inputs $u_r = \begin{bmatrix} u_{r_8} & u_{r_9} & u_{r_{10}} \end{bmatrix}^T$, an appropriate value for $u_{r_9}$ must be selected (e.g., $u_{r_9} = 0$ points) with the weight matrix $Q_u$ changed such that this requirement is now part of the objective function of the control system. The value of $w_{u_9}$ depends on the expected performance of the set point tracking versus the input targeting. The matrix $Q_y$ remains unchanged. If at any time $k$ the logical condition is not satisfied (e.g., a relapse), the initiation phase is reactivated.

## 5.4 *Simulation Scenario for the Closed-Loop Adaptive Intervention*

The simulation results presented in this section assume a hypothetical individual with a sedentary lifestyle, performing an average (i.e., baseline) of 5000 steps per day with an intervention starting at day zero. This simulation scenario considers the same model parameters used in the open-loop intervention. Delays ($\theta_i$) and internal disturbance parameters ($\zeta_i$) are not considered. The sampling time is $T = 1$ day; controller horizons are $p = 7$ and $m = 5$ days, while maximum and minimum bounds on $u$, $\Delta u$, and $y$ are

- $u_{\min} = \begin{bmatrix} 5000 & 0 & 0 \end{bmatrix}^T$, $u_{\max} = \begin{bmatrix} 10000 & 500 & 500 \end{bmatrix}^T$
- $\Delta u_{\min} = \begin{bmatrix} -1000 & -500 & -500 \end{bmatrix}^T$, $\Delta u_{\max} = \begin{bmatrix} 1000 & 500 & 500 \end{bmatrix}^T$
- $y_{\min} = \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix}^T$, $y_{\max} = \begin{bmatrix} 10000 & 10000 & 12000 & 10000 \end{bmatrix}^T$

The categorical values of the intervention components are defined by the sets

- $U_8 = \{5000, 6000, 7000, 8000, 9000, 10000\}$
- $U_9 = \{100, 200, 300, 400, 500\}$

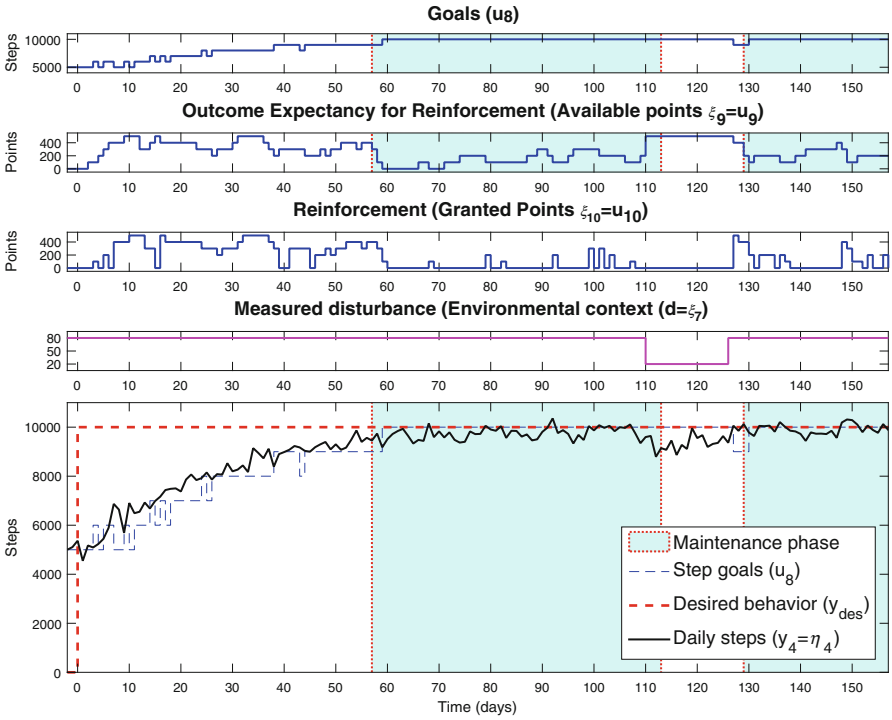


**Fig. 13** Simulation results for the HMPC-based adaptive intervention for a participant with low physical activity

The unmeasured disturbance is assumed Gaussian with $d'(k) \sim \mathcal{N}(0, 40000)$. To allow for a progressive increase on the performed steps and a fast disturbance rejection, the tuning parameters are

$$\alpha_r = \begin{bmatrix} 0 & 0 & 0.96 & 0 \end{bmatrix}^T \; \alpha_d = 0.1 \; f_a = \begin{bmatrix} 0 & 0 & 0.3 & 0 \end{bmatrix}^T .$$

Simulation results are shown in Fig. 13, where goals ($u_8$) and available points ($u_9$) are generated by the HMPC algorithm. The value for granted points ($u_{10}$) is taken from the available points only when the previous day goal is achieved, as enforced by model constraints. The maintenance stage is illustrated via a shaded region; this phase starts when the goal has been achieved at least 4 times during the last $n_s = 6$ days with a tolerance of $\text{Tol}_4 = 600$ steps. During this stage a reduction in the number of available and granted points results from the actions of the control system. The impact of measured disturbances (e.g., environmental context) is tested via a downward pulse starting at day 110 and lasting for 15 days; as a result, the individual tends to reduce his or her steps, so the controller reacts by discontinuing the maintenance phase and relying again on points to compensate for any deviations.

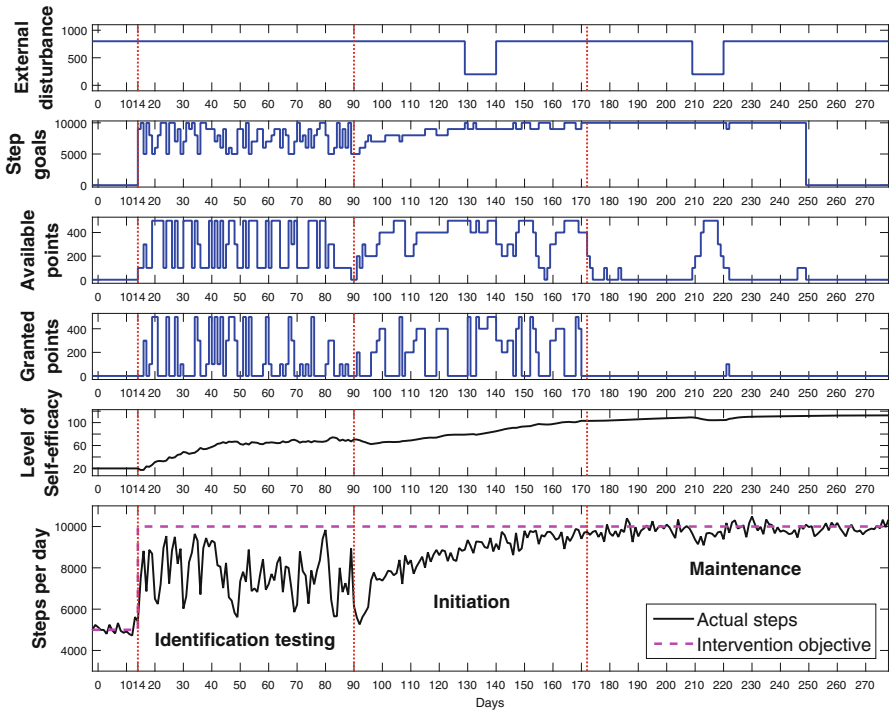


**Fig. 14** Comprehensive illustration of a representative time series resulting from the Just Walk intervention that features four phases: measurement-only, identification testing, initiation, and maintenance

Figure 14 summarizes the overall life cycle of the Just Walk intervention, as a closed-loop experiment with four phases. Phase 1 is an initial measurement-only period, which provides a baseline measurement of a person's current activity. Phase 2 is an open-loop experiment, which is similar to our pilot study. This phase enables estimating/validating our dynamical model and individualized tailoring variable selection as per our pilot study. In Phase 3, we conduct a closed-loop experiment focused on behavioral initiation. Specifically, this phase incorporates MPC to make daily decisions, and the system will refine these decisions based on data gathered each day. During Phase 3, the MPC controller will strive for appropriate targets for our at-risk group (i.e., 10,000 steps/day on average or, if a participant does not seem capable of meeting that goal, 3000 steps/day above the participant's baseline median steps). Finally, Phase 4 of the study is a closed-loop experiment, meant to optimize the controller for maintenance. Specifically, the goal is to optimize our approach for providing the minimum support at which the participant maintains set-point targets.

# 6 Control Systems Engineering for a Gestational Weight Gain Intervention: Healthy Mom Zone

## *6.1 The Healthy Mom Zone Intervention*

### 6.1.1 Description of the Problem and Need for Intervention Approach

The majority of women in the childbearing years are overweight, which elevates the risk for myriad health problems (Flegal, Kruszon-Moran, Carroll, Fryar, & Ogden, 2016) particularly with the transition to motherhood. Maternal obesity and subsequent high GWG are strongly related to and independently predict adverse obstetric outcomes (e.g., preterm delivery, gestational diabetes, hypertension, preeclampsia) and elevate risks for macrosomia and early onset of obesity in the offspring (Institute of Medicine & National Research Council Committee to Reexamine IOM Pregnancy Weight Guidelines, 2009). GWG is a modifiable factor that can be targeted to reduce risks, and managing it can impact the offspring's likelihood of being obese.

In 2009, the Institute of Medicine (IOM) report reexamining the GWG guidelines called for effective interventions to manage weight gain, especially in overweight and obese women, who often gain more weight in pregnancy than is recommended (see Table 3). However, there is currently no "gold standard" intervention to prevent high GWG in overweight/obese pregnant women. Past randomized interventions have shown that GWG can be effectively managed by "mirroring" effective programs used in non-pregnant adults (e.g., frequent contact, weight/dietary intake monitoring, engaging in exercise); however, the effects have largely been limited to women who are not overweight or obese (Olson, Strawderman, & Reed, 2004; Phelan et al. 2011; Polley, Wing, & Sims, 2002). Overweight/obese pregnant women may require a more individualized approach, such as a program that helps each overweight or obese pregnant woman to control her GWG on a weekly basis and adapts to her unique needs over pregnancy. In other words, the intervention strategy would involve varying the component dosages in response to an individual's needs, much like clinical practice (Kumar, Nilsen, Pavel, & Srivastava, 2013). We have developed such an intervention, described in this chapter as Healthy Mom Zone, that uses control systems engineering to construct a dynamical model of energy balance (Dong, 2014; Dong et al. 2012, 2013) and considers how GWG responds to changes

**Table 3** Institute of Medicine (2009) GWG guidelines

| Category | Prepreg BMI ($kg/m^2$) | GWG range (pounds) | Rates of GWG 2nd–3rd TRI (M range in pounds/week) |
|---|---|---|---|
| Underweight | < 19.8 | 28–40 | 1 (1–1.3) |
| Normal | 19.9–24.9 | 25–35 | 1 (0.8–1) |
| Over weight | 25.0–29.9 | 15–25 | 0.6 (0.5–0.7) |
| Obese | ≥ 30.0 | 11–20 | 0.5 (0.4–0.6) |

Institute of Medicine and National Research Council Committee to Reexamine IOM Pregnancy Weight Guidelines (2009)

in energy intake, exercise, and planned/self-regulatory behaviors for a customized program for each woman. This novel, ongoing intervention has the potential to shift the focus of weight management from a "one size fits all" method to an individually tailored and adaptive approach to effectively manage GWG and promote maternal and infant health.

### 6.1.2 Goals of the Healthy Mom Zone Intervention

The Healthy Mom Zone intervention is an individually tailored behavioral intervention designed specifically to manage GWG among overweight/obese pregnant women throughout their pregnancy. The conceptual framework of the intervention (see Fig. 15) is based on a dynamical model of energy balance that describes how a behavioral intervention can influence GWG (Dong, 2014; Dong et al. 2012, 2013). It relies on integrating mechanistic energy balance and dynamical models of planned/self-regulatory behaviors describing how internal psychological processes can reinforce positive program outcomes. This model includes (a) a two-compartment energy balance model predicting changes in body mass as a result of energy intake and physical activity, (b) two theory of planned behavior (Ajzen,

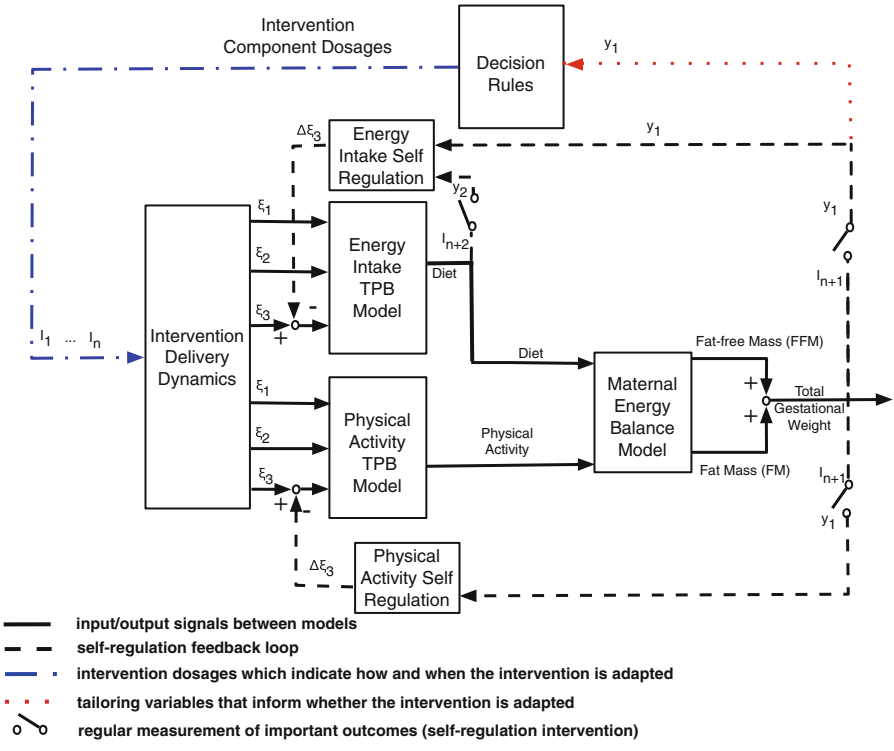


**Fig. 15** Conceptual framework of the Healthy Mom Zone intervention

2005) models describing how energy intake and exercise are affected by behavioral variables (including women's attitude, social influences, perceived control, and motivation), (c) a program delivery module relating magnitude and duration of components to inflows of the theory of planned behavior models, and (d) two self-regulation units modeling how success expectancies in the intervention influence one's goal achievement motivation. The final module of Healthy Mom Zone is comprised of decision rules that rely on assessed values of the tailoring variable (GWG) to make decisions on intervention components. Phase 2 of Healthy Mom Zone (currently ongoing) is relying on *if-then* decision rules, but the overall goal of the research is to determine the feasibility and usefulness of hybrid MPC. This is illustrated in a simulation later in this chapter (Sect. 6.2).

A fluid analogy for this dynamical system model is shown in Fig. 16. A complete description and the mathematics behind the analogy are found in Dong (2014) but are briefly explained here. The energy balance model corresponds to two inventories for fat mass and fat-free mass, which are built by energy intake but depleted through basic metabolic function and physical activity. Theory of planned behavior models for energy intake and physical activity corresponds to the fluid analogy described in Sect. 3.2.1 and Fig. 5. Self-regulation loops and intervention delivery inventories (which account for the accumulation or depletion of the inflows to the models) form part of the fluid analogy as well. A set of decision rules were developed based on the IOM (2009) GWG guidelines, our own research (e.g., Dong et al. 2012, 2013; Symons Downs, 2016; Symons Downs, Savage, & Rauff, 2014; Thomas et

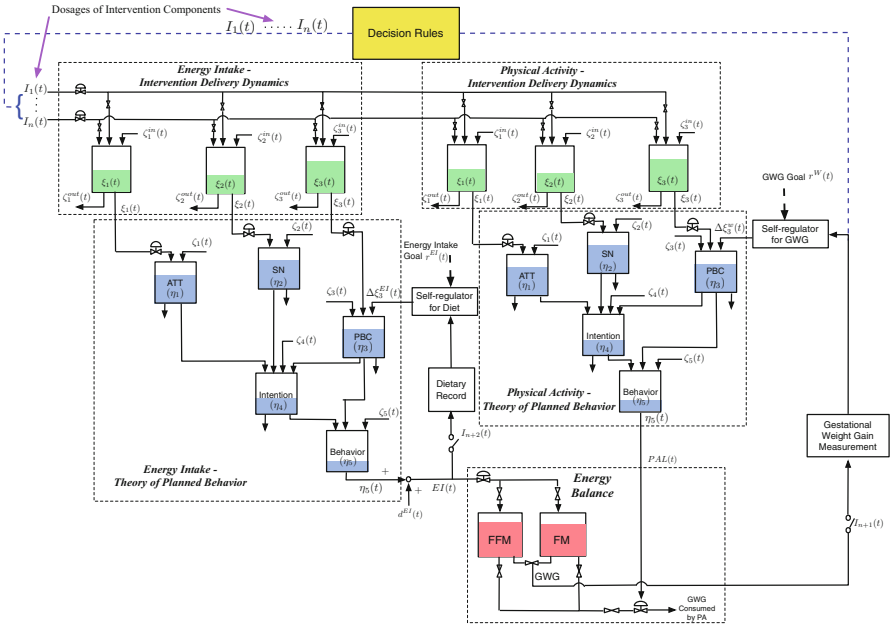


**Fig. 16** Fluid analogy for the closed-loop Healthy Mom Zone intervention

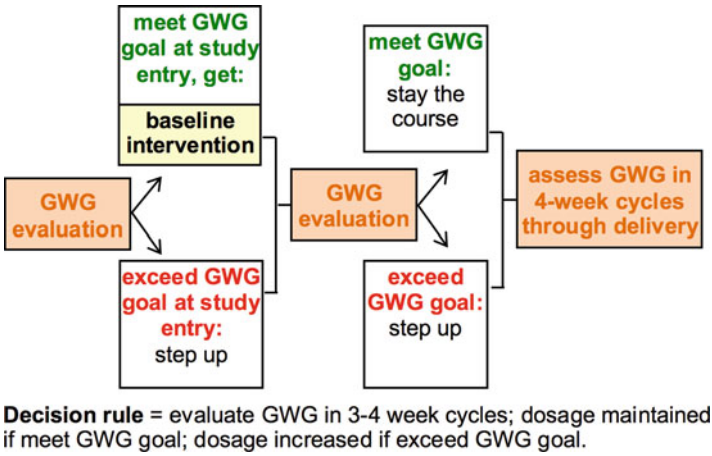


**Fig. 17** If-then decision rules for evaluating GWG and adapting the Healthy Mom Zone intervention

al. 2012), and clinical insight that inform when and how to adapt the components. Decision rules define changes to the intervention and correspond with altering the dosage (Collins et al. 2004). The dosage level is based on tailoring variables that are expected to impact the effect of the component (e.g., effect of exercise on GWG), and the level of intervention required to address the needs of individuals varies according to the tailoring variable (GWG). In *Healthy Mom Zone*, GWG is evaluated weekly, and the collective weight gain is assessed over a 3–4 week period. If a woman is within her GWG goal, she continues to receive the same dosage of the intervention. If she does not meet her goal, her intervention dosage is adapted or "stepped up" (Fig. 17). If she is under her goal, we use clinical guidance to decide whether and how the dosage change should be made.

### 6.1.3 Components of the Healthy Mom Zone Intervention

The Healthy Mom Zone intervention components were informed by past research and our pilot data (Diabetes Prevention Program Research Group, 2002; Dong et al. 2012, 2013; Symons Downs, DiNallo, & Kirner, 2007; Symons Downs & Hausenblas, 2004; The Look AHEAD Research Group, 2006) showing that when people are taught how to set appropriate goals, self-monitor, and effectively manage their time, they are more likely to achieve their goals and see positive behavioral outcomes (e.g., eating healthy, engaging in exercise, managing weight; see Fig. 18). All women start in Healthy Mom Zone with the baseline intervention, which includes standard prenatal care and education on GWG, healthy eating, exercise, and self-monitoring. The intervention adapts or "steps up" based on the GWG evaluation and decision rule criteria described above and includes different variations of hands-on active learning strategies that are added to the baseline intervention

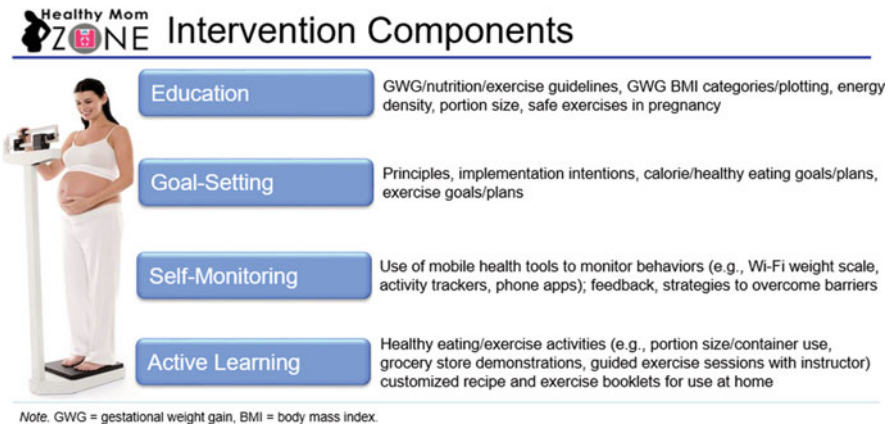


**Fig. 18** Healthy Mom Zone intervention components

sequentially (e.g., step-up 1 = baseline intervention + active learning healthy eating demonstrations + exercise session; step-up 2 = baseline intervention + step-up 1 + additional active learning components (e.g., additional on-site exercise session, portion size, and containers)). Self-monitoring of GWG, healthy eating, and exercise behaviors includes the use of mHealth tools (e.g., Wi-Fi scale, dietary intake smartphone app, activity monitors) to facilitate self-regulation, motivation, and behavior change.

### 6.1.4 Study Assessments

Intensive longitudinal data are used in Healthy Mom Zone to assess the primary study outcome of GWG and several biobehavioral/psychological secondary outcomes (Table 4). Pre- and post-intervention assessments are conducted at the Pennsylvania State University Clinical Research Center, and the secure data capture (RedCAP) system is used to collect electronic survey data. Women use a Wi-Fi scale and an activity monitor (daily) and a smartphone app (weekly) at home to measure their weight and kcal activity expenditure and intake.

### 6.1.5 Flow of Participants

The target for Healthy Mom Zone is 30 overweight and obese pregnant women (BMI 25; 40 with physician approval). Eligible participants (e.g., singleton pregnancies, ages 18–40 years, able to read and understand English, no obstetric/medical complications limiting participation) are screened, enrolled, and consented. They complete pre-intervention assessments both on-site at the research center (e.g., body composition, bloodwork) and at home (e.g., electronic surveys) and are then randomized to either the control condition (standard of care) or treatment condition

**Table 4** Healthy Mom Zone assessments

| Variable | Assessments |
|---|---|
| Weight and height | High precision stand-on adult scale; stadiometer for height<br>Wi-Fi smart scale |
| Metabolism | Mobile metabolism device |
| Biomarkers | Blood, urine |
| Adiposity | Body composition |
| Healthy eating behaviors | Smartphone dietary intake app (kcal intake)<br>Back-calculation method to estimate energy intake (kcal intake)<br>Eating inventory |
| Exercise behaviors | Activity monitors and survey (expenditure)<br>Exercise log |
| Motivational determinants | Attitude, subjective norm, perceived behavioral control, intention surveys |
| Self-regulation | Self-regulation index and questionnaire |
| Sociodemographic | Health and history questionnaire |

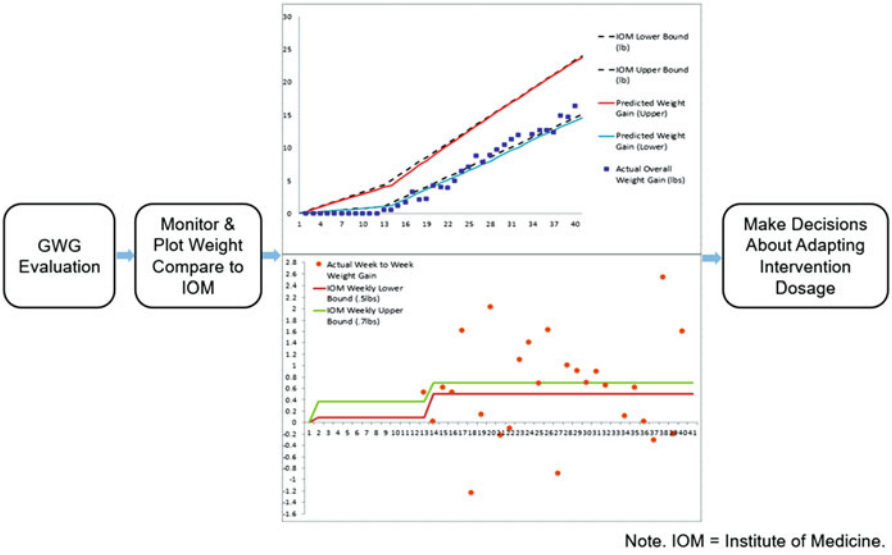


**Fig. 19** Process for evaluating maternal GWG in Healthy Mom Zone: monitor and plot weekly weight, compare to IOM guidelines, and evaluate every 3–4 weeks to make decisions about adapting intervention dosages

(Healthy Mom Zone intervention) from early (e.g., 6–12 weeks gestation) through late pregnancy (e.g., 37 weeks gestation). GWG is evaluated in 3–4 week cycles, and the intervention dosage is adapted as necessary to help women stay within their GWG goals. An illustration of how participant weights are plotted and evaluated in Healthy Mom Zone is shown in Fig. 19.

### 6.1.6 Illustration from Phase 1 Participant Data

As noted previously, Healthy Mom Zone is an ongoing study. Phase 1 of the intervention has focused on feasibility of dosages and measurements, in order to better understand user acceptability of the intervention. Phase 2 is a proof-of-concept study that is implementing a fully adaptive intervention using *if-then* decision rules and is intended to characterize the effects of energy balance and planned and self-regulatory behaviors on GWG. In this section, we illustrate with selected Phase 1 participant data some aspects of the conceptual framework described in Sect. 6.1.2. The first is the energy balance model. As noted in Dong (2014), Guo, Rivera, Downs, and Savage (2016), and Thomas et al. (2012), under conditions of constant resting metabolic rate, the effect of changes in energy intake ($\Delta EI$) and physical activity level ($\Delta PAL$) on change in gestational weight ($\Delta W$) is reasonably approximated by the differential equation:

$$\frac{\Delta W(t)}{\mathrm{d}t} = K_{\mathrm{EI}} \Delta EI(t) + K_{\mathrm{PA}} \Delta PAL(t) \tag{24}$$

$K_{\mathrm{EI}}$ and $K_{\mathrm{PA}}$ correspond to the gain parameters of integrating systems (Ogunnaike & Ray, 1994) corresponding to changes in energy intake and physical activity, respectively. Figure 20 illustrates how this simple dynamic energy balance model adequately captures the data for a Phase 1 participant. It should be noted that proper use of the energy balance model involves addressing issues of missing data, energy intake underreporting, and the possibility of changing resting metabolic rate over time; these are problems that are being investigated at this time, with a series of approaches presented in Guo et al. (2016) and Guo, Rivera, Savage, and Downs (2017).

Phase 1 data have also been used to perform some initial parameter estimation of dynamic theory of planned behavior models. Figure 21 shows show a dynamical model in accordance with Sect. 3.2.1 and Eqs. (6), (7), (8), (9), and (10) for a Phase 1 participant in a reduced structure involving subjective norm, perceived behavioral control, intention (INT), and physical activity as constructs. The model was estimated using semi-physical identification routines in MATLAB, with the fit calculated using the NRMSE criteria from Eq. (4).

## 6.2 *Simulated Comparison of Decision Rules Versus an MPC-Based Closed-Loop Intervention*

In this section, we consider a simulated comparison of "if-then" decision rules (patterned after the general structure in Fig. 17) and a hybrid MPC-based controller as described in Sect. 4 based on a hypothetical participant, a 25-year-old female with pregravid body mass 75 kg, 160 cm in height, which classifies her as overweight (BMI = 29.30). The open-loop model for GWG interventions is as described in Dong (2014) and Dong et al. (2012, 2013) and is conceptually depicted in Fig. 15.

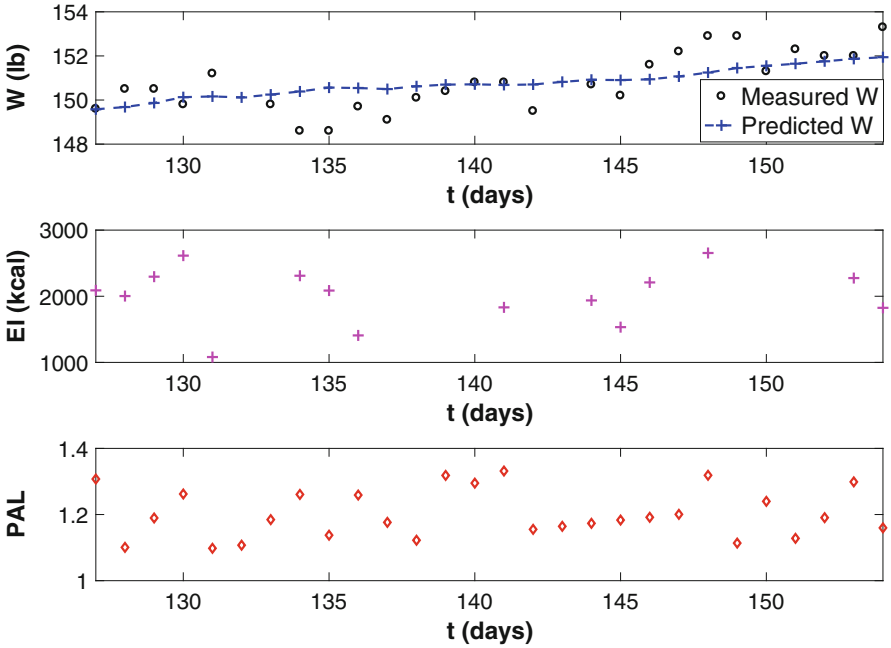


**Fig. 20** Energy balance model prediction using (24) for a Phase 1 Healthy Mom Zone participant. Weight (top) is recorded on a WIFI scale, energy intake (middle) is reported on a smartphone app, and physical activity level (bottom) is derived from an accelerometer

For the sake of simplicity, we focus only on the effects that intervention components and self-regulation play on the perceived behavioral control inflow in the theory of planned behavior models. The 2009 IOM guidelines for GWG in Table 3 are used as reference trajectory for both the HMPC controller and "if-then" decision rules.

In this illustration, we will consider three ($n_c = 3$) intervention components ($u_1$, $u_2$, $u_3$), which are augmented or reduced to prespecified dosage levels during the GWG intervention. These are healthy eating active learning (as $u_1(k) \in \{0,1,2,3\}$), physical activity active learning (as $u_2(k) \in \{0,1,2,3,4\}$), and goal setting (as $u_3(k) \in \{0,1\}$). In the baseline program, all three components are introduced simultaneously to define the base dosage as $u_j(k) = 1, j \in \{1,2,3\}$.

The availability of multiple intervention components in this problem forces a decision regarding which component should be augmented or reduced first at each decision point (biweekly in this illustration) when the dosage can be updated per the individual's measured outcomes and performance. These decisions (which apply to both the *if-then* decision rules and the HMPC controller) are summarized in Table 5. We consider that physical activity active learning ($u_2$) will be augmented from the baseline only when healthy eating active learning ($u_1$) reaches its maximum dosage and healthy eating active learning ($u_1$) will not be reduced from full dosage until physical activity active learning ($u_2$) returns back to the base dosage (augmentation

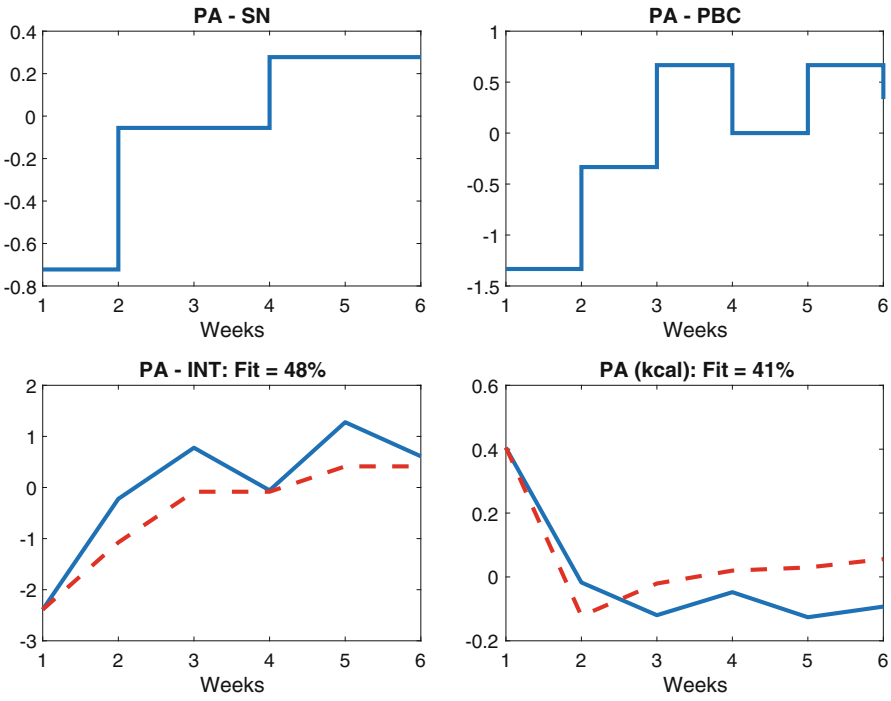


**Fig. 21** Theory of planned behavior dynamic model for Phase 1 Healthy Mom Zone participant for a reduced structure involving subjective norm, perceived behavioral control, and intention (INT) to describe physical activity. Solid line is measured data, while dashed line corresponds to model prediction. Fit corresponds to the NRMSE index shown in Eq. (4)

and reduction sequence above baseline). When it is necessary to decrease the dosage from the baseline, $u_2$ is reduced first, followed by $u_1$ and $u_3$; the augmentation sequence from zero dosage to baseline will be in the opposite order, with $u_3$ increased to base dosage first, followed by $u_1$ and $u_2$ (augmentation and reduction sequence below baseline). At each decision point, there will be only one intervention component augmented or reduced, which necessitates the selection of only one input change (if, $\Delta u_i(k) \neq 0$ then $\Delta u_j(k) = 0$ for $j \neq i$; $i, j \in \{1, 2, \ldots, n_c\}$). The logic in the set of decision rules is sequential and restricts how the future dosage can be specified, based on the current dosage. We note that in Phase 2 of Healthy Mom Zone, we used an exclusively step-up approach and did not reduce dosage (Fig. 17); this was a conscious decision because most women need support managing their weight, so it was important to not take the support away, even if this implies greater consumption of intervention resources. However, for sake of illustration in this simulation, we consider decision rules in which the dosage can be adapted up or down.

In our simulation we assume that with no intervention, this participant will have a ramp increase in her energy intake from day 14 to day 91 and her energy intake

**Table 5** Summary of dosage augmentations and reductions followed by the *if-then* decision rules and the HMPC controller for the simulations of a GWG adaptive intervention in Sect. 6.2

| Options | Description |
|---|---|
| Step down 3 | Reduction of goal setting ($u_3$) |
| Step down 2 | Reduction of healthy eating active learning ($u_1$) |
| Step down 1 | Reduction of physical activity active learning ($u_2$) |
| Baseline | Baseline dose for all components |
| Step up 1 | First augmentation of healthy eating active learning ($u_1$) |
| Step up 2 | Second augmentation of healthy eating active learning ($u_1$) |
| Step up 3 | First augmentation of physical activity active learning ($u_2$) |
| Step up 4 | Second augmentation of physical activity active learning ($u_2$) |
| Step up 5 | Third augmentation of physical activity active learning ($u_2$) |

will keep constant throughout the remainder of the pregnancy. The participant has maintained a sedentary lifestyle up to the point of entry in the intervention and, in the absence of intervention, would potentially engage in less physical activity from the second to third trimester as she gains weight. The intervention can help improve her physical activity level in the second trimester, but she will still decrease her physical activity in the third trimester. These are two physical activity disturbances for the intervention and no intervention cases, which will reduce her energy expenditure in the energy balance model.

The hypothetical intervention scenarios assume the participant enters the intervention with baseline program at gestation week 14 and starts engaging in self-regulatory behaviors (e.g., weighing herself to monitor GWG, using dietary records to monitor energy intake and pedometer to monitor physical activity). The dosage of the intervention components is adapted every 2 weeks until week 36. In the simulation study, we assume the anticipated disturbance occurs late in the participant's second trimester and continues until her delivery, when there is no intervention. The intervention will help the participant attenuate this energy intake increase.

Table 6 summarizes the model parameters in this simulation study, including the behavioral parameters, time constants $\tau_i$, time delays $\theta_i$, gains assumed for the participant, and filter parameters. The definition of these variables can be found in Dong (2014) and Dong et al. (2012, 2013). All values are hypothetical but can be selected such that the simulated responses mimic those of an actual participant. The parameters for the HMPC are as follows: $p = 30$ and $m = 28$, $Q_y = 10$, $\alpha_r = 0.9$, $\alpha_d = 0.3$, $f_a = 1.0$, the sampling time for the participant to measure her GWG and monitor dietary record is $T_s = 1$ day, and the decision interval is $T_{\text{sw}} = 14$ days. The move size constraints for the manipulated variables at the decision intervals for the HMPC controller are $\Delta u(k)_{\max} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^T$ and $\Delta u(k)_{\min} = \begin{bmatrix} -1 & -1 & -1 \end{bmatrix}^T$, to match those of the "if-then" decision rules (where no more than one augmentation or reduction is allowed per decision point). However, HMPC faces fewer restrictions in this regard, as can be seen in simulations presented in Dong (2014), in which the HMPC algorithm is not limited to move sizes of $\pm 1$.

**Table 6** Model parameters for the simulation study in Sect. 6.2. Time constants ($\tau_i$), delays ($\theta_i$), and self-regulation adjustable parameters ($\lambda_r$, $\lambda_d$) are in units of days

| Parameter | EI-TPB | PA-TPB | Parameter | EI-TPB | PA-TPB |
|---|---|---|---|---|---|
| $b_1$ | 3 | 1 | $e_1$ | 6 | 4 |
| $n_1$ | 2 | 7 | $m_1$ | 3 | 8 |
| $p_1$ | 1 | 4 | $c_1$ | 2 | 2 |
| $\tau_1$ | 1 | 30 | $\gamma_{11}$ | 1 | 0.7 |
| $\tau_2$ | 1 | 30 | $\gamma_{22}$ | 1 | 0.5 |
| $\tau_3$ | 1 | 10 | $\gamma_{33}$ | 1 | 0.7 |
| $\tau_4$ | 1 | 20 | $\beta_{41}$ | 1 | 0.34 |
| $\tau_5$ | 1 | 30 | $\beta_{42}$ | 1 | 0.27 |
| $\theta_1, \ldots, \theta_3$ | 0 | 0 | $\beta_{43}$ | 1 | 0.13 |
| $\theta_4, \ldots, \theta_6$ | 0 | 0 | $\beta_{53}$ | 0 | 0.08 |
| $\theta_7, \theta_8$ | 0 | 0 | $\beta_{54}$ | 1 | 0.42 |
| $k_{u_1}$ | 0.004 | 0 | $\theta_{u_1}$ | 0 | 0 |
| $k_{u_2}$ | 0 | 0.01 | $\theta_{u_2}$ | 0 | 0 |
| $k_{u_3}$ | 0.002 | 0.006 | $\theta_{u_3}$ | 0 | 0 |
| $\lambda_r$ | 70 | 100 | $\lambda_d$ | 90 | 155 |

Dong (2014) and Dong et al. (2012, 2013)
*TPB* theory of planned behavior, *EI* energy intake

The simulation responses for maternal body mass, energy intake, energy expenditure, the intervention component dosages, and the perceived behavioral control inflows to theory of planned behavior models under different scenarios (HMPC-based intervention, adaptive intervention using decision rules, and no intervention) are shown in Fig. 22. The scenario in the absence of the intervention is presented in order to illustrate how the interventions play an important role in the participant's improvement in behavior, specifically how significant disturbances in energy intake and physical activity influence weight gain. From looking at the energy intake profile, we can see that in the absence of intervention, the participant will increase her energy intake not only in the first trimester before the intervention due to her awareness of the pregnancy but also late in the second trimester after the intervention starts. At the beginning of the intervention, the participant's weight is within the IOM guidelines, although her energy intake (3302 kcal/day) is already 100 kcal higher than the energy intake reference values (3202 kcal/day) assumed for the third trimester. Therefore, for both the HMPC-based intervention and the "if-then" decision rules, the initial dosage this participant receives is the baseline program.

In the HMPC-based intervention, this participant's weight is always within the IOM guidelines after the intervention starts. In order not to have this participant reduce too much weight, which might even be below the lower bound of IOM guidelines, the HMPC controller first reduces the intervention for this participant by setting the dosage to zero for component $u_2(k)$ at week 16. At this reduced dosage, the other two components at their baseline doses can still maintain this participant within the IOM guidelines during the week 16–20 when there is no additional energy

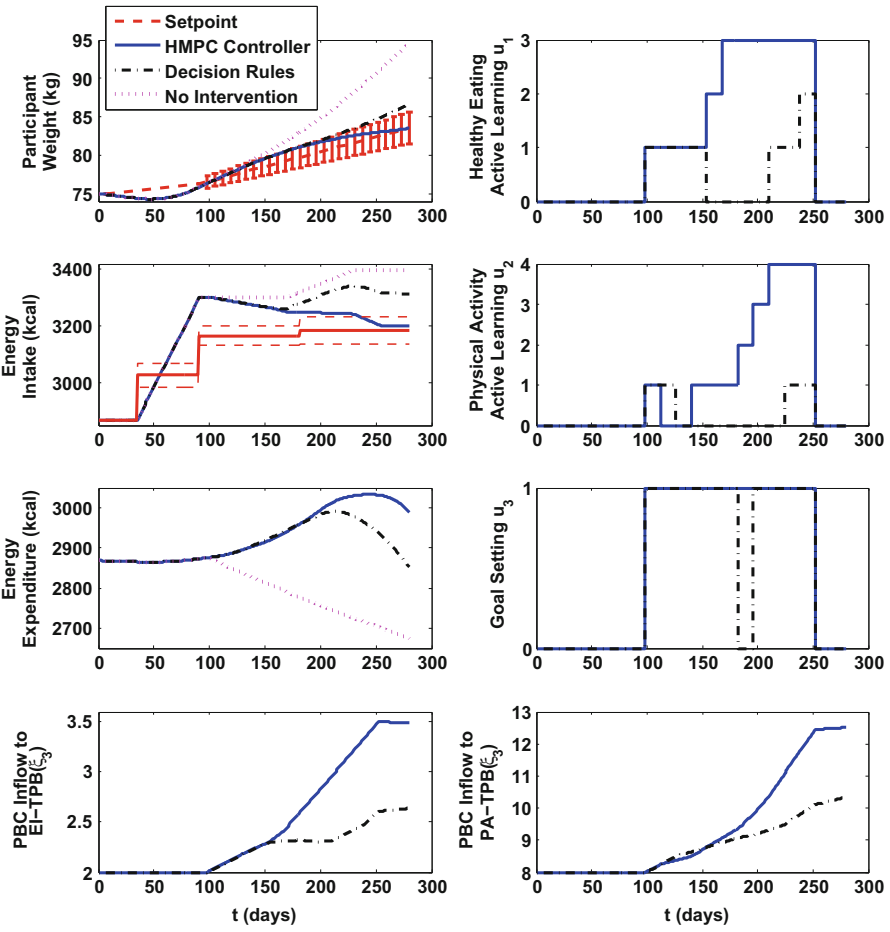


**Fig. 22** Simulation results in Sect. 6.2 for a hypothetical Healthy Mom Zone participant, showing responses for maternal weight, energy intake, energy expenditure, intervention component dosages, and perceived behavioral control inflows to the theory of planned behavior models for an HMPC-based intervention, an adaptive intervention using *if-then* decision rules, and no intervention cases. The red curves (with high/low intervals) shown in the upper left participant weight and energy intake plots correspond to the 2009 IOM Guidelines ((Institute of Medicine and National Research Council Committee to Reexamine IOM Pregnancy Weight Guidelines, 2009), Table 3) applied on a daily basis; the nominal (center) lines represent set points for the HMPC weight controller and the energy intake self-regulator, respectively. The blue solid lines are the case with an HMPC-based intervention, the black dashed-dotted lines represent the case with adaptive intervention using "if-then" decision rules, and the magenta dotted lines are the case in the absence of an intervention. Both the "if-then" decision rules and the HMPC controller rely on the dosage sequence described in Table 5

intake increase (disturbance). At week 20, the HMPC controller augments the intervention by adding this component $u_2(k)$, which restores the intervention to the

baseline program. The energy intake disturbance occurs at day 170, and therefore the HMPC controller continues augmenting the intervention at week 22 (day 154) by increasing the dose for $u_1(k)$, considering the prediction horizon $m = 30$. The component $u_2(k)$ does not get augmented until 2 weeks later, on week 26, which is 2 weeks after $u_1(k) = 3$. The move size constraint on $u_2(k)$ in this scenario is also $\pm 1$ to match the case using *if-then* decision rules, to allow for a proper comparison. The intervention reaches the maximum dosages for all the components at week 30 and remains unchanged until the end of the intervention. This hypothetical participant with the HMPC-based intervention is able to keep her weight within the IOM guidelines throughout her whole pregnancy, even in the presence of an energy intake increase as a disturbance. This participant also manages to control her energy intake within the reference values in the mid-third trimester. The participant's perceived behavioral control inflows to the theory of planned behavior models increase faster when the intervention gets augmented and more slowly when the intervention gets reduced.

Examining the adaptive intervention using the "if-then" decision rules, the participant will receive the same baseline program at entry to the intervention due to her weight being within the IOM guidelines. Because she is able to meet her GWG goal under the baseline program for two intervention decisions (4 weeks), her intervention gets reduced by setting the dosage of component $u_2(k)$ to zero at week 18. With this reduced intervention, for the next 4 weeks, this participant's weight gain continues to remain within the IOM guidelines, and, therefore, the dosage of component $u_1(k)$ gets reduced to zero at week 22, at which time only the component $u_3(k)$ with base dosage is assigned to this participant. This participant still succeeds in maintaining her GWG within the IOM guidelines for the next 4 weeks, and, hence, at week 26, the dosage of $u_3(k)$ is also decreased to zero. The energy intake increase occurs at day 170 (late second trimester). With this energy intake disturbance, her high energy intake increase in the first trimester, and her low-intensity (or even no) intervention, and her weight gain moves outside the IOM guidelines around day 195, at which time the intervention is resumed with component $u_3(k)$ added first at week 28, $u_1(k)$ and stepped up at week 30, and $u_2(k)$ added at week 32. At week 34, her intervention is augmented above the baseline by increasing $u_1(k)$ to step 2, and the intervention stops at week 36. Because the *if-then* decision rules do not incorporate a dynamical model, all these augmented actions do not take place in the most timely manner. As a result, this participant's GWG falls outside the IOM guidelines late in her pregnancy, and her energy intake is above the reference values throughout her pregnancy. The "if-then" decision rules are better than having no intervention, but the improvements in perceived behavioral control to the theory of planned behavior models are only around half of those using HMPC-based intervention.

This hypothetical case study illustrates how aspects of the control engineering formulation provide potential benefits of HMPC over "if-then" rules that rely only on current values of tailoring variables. Specifically, the feedforward control action in the HMPC controller is very useful in addressing anticipated disturbances that may be known a priori or can be measured/predicted in the course of

the intervention, while the feedback control action will respond to unmeasured disturbances. The HMPC controller will assign the future dosages based not only on the participant's past and current dosages, but also on the predicted measured outcomes over the prediction horizon, while the "if-then" decision rules described in the chapter adapt the intervention only based on the participant's current responses. Hence, when the anticipated disturbance is available, the HMPC-based intervention can better predict the future responses of the participant and make the dosage adjustment earlier than the intervention using decision rules that may or may not provide the augmentation in time. The "if-then" decision rules do offer the benefit of simplicity, so it is up to the intervention scientist to decide whether the additional implementation effort required by the HMPC controller is justifiable.

## 7 Summary and Future Directions

In this chapter, our goal has been to show how behavioral interventions can benefit from a control systems engineering perspective. This has been accomplished through descriptions of a comprehensive control engineering methodology that is illustrated in two behavioral application settings: Just Walk and Healthy Mom Zone. The control engineering approach consists of two major steps: system identification to estimate (black-box or semi-physical) dynamical models and control algorithms that rely on these estimated dynamical models to optimize decision-making in the intervention. Behavioral theory can influence the task of dynamic model building and system identification; this was demonstrated for the theory of planned behavior and social cognitive theory. In all cases, the common algorithmic framework for achieving closed-loop control is hybrid MPC (HMPC), which consists of a constrained optimization problem that is solved in real time via a receding horizon approach. Through the choice of horizon lengths, weight values, and filter parameters, HMPC can be tuned to achieve desired levels of performance and robustness and thus represents a flexible, extensible framework for decision-making in intensively adaptive mHealth interventions.

Space limitations keep us from describing some interesting extensions of this work currently underway. These include modeling efforts (as part of Healthy Mom Zone research) to enhance the adaptive intervention to include fetal weight and infant birth outcomes (Savage, Downs, Dong, & Rivera, 2014) and ways to apply identification and state estimation methods with energy balance models to estimate energy intake (Guo et al. 2016). Identification test monitoring (Martín, Rivera, & Hekler, 2015b, 2016b) is being studied as part of Just Walk to systematically determine the optimal number of multisine input cycles and thereby establish a minimal duration for the intervention.

The intensively adaptive interventions presented in this chapter involve a daily timescale for decisions. Augmenting an intensively adaptive intervention with a "just-in-time" adaptive intervention (JITAI; Nahum-Shani, Hekler, & Spruijt-Metz, 2015) that can provide support when needed, multiple times within a day, represents

an important future direction for this work. Developing an effective JITAI requires recognizing just-in-time states when a participant has the opportunity to engage in a behavior and is receptive to support. State estimation techniques, such as model on demand (Stenman, 1999) or machine learning, can be used to infer the existence of just-in-time states on the basis of available measurements and models. This research activity calls for alternative approaches to system identification experiments (e.g., micro-randomization (Klasnja et al. 2015)) in order to generate informative databases.

**Acknowledgment** The authors wish to acknowledge the participation of multiple students, postdocs, staff, and collaborators who have participated in the development and analysis of the interventions described in this chapter. These include doctoral student Sayali Phatak, Dr. Elizabeth Korinek, and Kevin Hollingshead (staff) in the Designing Health Laboratory, Arizona State University; former and current doctoral students Yuwen (Shirley) Dong, Penghong Guo, Mohammad Freigoun, and César Martin Moreno in the Control Systems Engineering Laboratory, Arizona State University; Penn State Exercise Psychology Laboratory graduate students (Krista Leonard and Abigail Pauley), undergraduate research assistants, and Dr. Theodore Hovick (Mount Nittany Physician Group, State College PA); and Dr. Emily Hohman and Katie McNitt (Registered Dietitian) with the Center for Childhood Obesity Research at Penn State University.

**Funding** Support from the US National Institutes of Health (NIH; grants R21 DA024266, K25 DA021173, R01 HL119245 and R56 HL126799) and the National Science Foundation (NSF; grant IIS-1449751) is gratefully acknowledged. Additional support has been received from the Piper Health Solutions Consortium at Arizona State University. The opinions expressed in this article are the authors' own and do not necessarily reflect the views of NIH, NSF, or the Virginia G. Piper Charitable Trust.

## References

Adams, M. A., Sallis, J. F., Norman, G. J., Hovell, M. F., Hekler, E. B., & Perata, E. (2013). An adaptive physical activity intervention for overweight adults: A randomized controlled trial. *PLoS One, 8*(12), e82901.

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes, 50*(2), 179–211.

Ajzen, I. (2005). *Attitudes, personality, and behavior* (2nd ed.). New York, NY: Open University Press.

Aström, K. J., & Murray, R. M. (2010). *Feedback systems: An introduction for scientists and engineers*. Princeton, NJ: Princeton University Press.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice-Hall Inc.

Bandura, A. (1989). Human agency in social cognitive theory. *American Psychologist, 44*(9), 1175.

Bekiroglu, K., Lagoa, C., Murphy, S. A., & Lanza, S. T. (2017). Control engineering methods for the design of robust behavioral treatments. *IEEE Transactions on Control Systems Technology, 25*(3), 979–990.

Bemporad, A., & Morari, M. (1999). Control of systems integrating logic, dynamics, and constraints. *Automatica, 35*(3), 407–427.

Bollen, K. (1983). *Structural equations with latent variables*. New York, NY: Wiley. https://doi.org/10.1002/9781118619179

Carver, C. S., & Scheier, M. F. (1998). *On the self-regulation of behavior*. New York, NY: Cambridge University Press.

Chakraborty, B., & Murphy, S. A. (2014). Dynamic treatment regimes. *Annual Review of Statistics and Its Application, 1*, 447–464.

Collins, L. M. (2018). *Optimization of behavioral and biobehavioral interventions: The multiphase optimization strategy*. Cham, Switzerland: Springer.

Collins, L. M., Murphy, S. A., & Bierman, K. L. (2004). A conceptual framework for adaptive preventive interventions. *Prevention Science, 5*(3), 185–196.

Craig, C. L., Marshall, A. L., Sjöström, M., Bauman, A. E., Booth, M. L., Ainsworth, B. E., … Sallis, J. F. (2003). International physical activity questionnaire: 12-country reliability and validity. *Medicine & Science in Sports & Exercise, 35*(8), 1381–1395.

Deshpande, S. (2011). *A control engineering approach for designing an optimized treatment plan for fibromyalgia* (M.S. thesis). Arizona State University.

Deshpande, S. (2014). *Optimal input signal design for data-centric identification and control with applications to behavioral health and medicine* (Ph.D. dissertation). Arizona State University.

Deshpande, S., Nandola, N. N., Rivera, D. E., & Younger, J. (2011). A control engineering approach for designing an optimized treatment plan for fibromyalgia. *Proceedings of the American Control Conference*, 4798–4803.

Diabetes Prevention Program Research Group. (2002). The diabetes prevention program (DPP). *Diabetes Care, 25*(12), 2165–2171.

Dong, Y. (2014). *A novel control engineering approach to designing and optimizing adaptive sequential behavioral interventions* (Ph.D. dissertation). Arizona State University.

Dong, Y., Rivera, D. E., Downs, D. S., Savage, J. S., Thomas, D. M., & Collins, L. M. (2013). Hybrid model predictive control for optimizing gestational weight gain behavioral interventions. *Proceedings of the American Control Conference*, 1970–1975.

Dong, Y., Rivera, D. E., Thomas, D. M., Navarro-Barrientos, J. E., Downs, D. S., Savage, J. S., & Collins, L. M. (2012). A dynamical systems model for improving gestational weight gain behavioral interventions. *Proceedings of the American Control Conference (ACC)*, 4059–4064.

Ferster, C. B. (1970). Schedules of reinforcement with Skinner. In P. B. Dews (Ed.), *Festschrift for B. F. Skinner, Century psychology series* (pp. 37–46). Appleton-Century-Crofts: New York, NY.

Flegal, K. M., Kruszon-Moran, D., Carroll, M. D., Fryar, C. D., & Ogden, C. L. (2016). Trends in obesity among adults in the United States, 2005 to 2014. *JAMA, 315*(21), 2284–2291.

Freigoun, M. T., Martín, C. A., Magann, A. B., Rivera, D. E., Phatak, S. S., Korinek, E. V., & Hekler, E. B. (2017). System identification of Just Walk: A behavioral mHealth intervention for promoting physical activity. *Proceedings of the American Control Conference*, 116–121.

Guillaume, P., Schoukens, J., Pintelon, R., & Kollar, I. (1991). Crest-factor minimization using nonlinear Chebyshev approximation methods. *IEEE Transactions on Instrumentation and Measurement, 40*(6), 982–989.

Guo, P., Rivera, D. E., Downs, D. S., & Savage, J. S. (2016). Semi-physical identification and state estimation of energy intake for interventions to manage gestational weight gain. *Proceedings of the American Control Conference*, 1271–1276.

Guo, P., Rivera, D. E., Savage, J. S., & Downs, D. S. (2017). State estimation under correlated partial measurement losses: Implications for weight control interventions. *IFAC-PapersOnLine, 50*(1), 13532–13537.

Hekler, E. B., Klasnja, P., Froehlich, J. E., & Buman, M. P. (2013). Mind the theoretical gap: Interpreting, using, and developing behavioral theory in HCI research. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 3307–3316.

Hekler E.B., Rivera D.E., Martin C.A., Phatak S.S., Freigoun M.T., Korinek E., Klasnja P., Adams M.A., Buman M.P. (2018). Tutorial for Using Control Systems Engineering to Optimize Adaptive Mobile Health Interventions *J Med Internet Res*, 20(6):e214, doi: https://doi.org/10.2196/jmir.8622

Institute of Medicine, & National Research Council Committee to Reexamine IOM Pregnancy Weight Guidelines. (2009). *Weight gain during pregnancy: Reexamining the guidelines*. Washington, DC: National Academies Press.

King, A. C., Hekler, E. B., Grieco, L. A., Winter, S. J., Sheats, J. L., Buman, M. P., ... Cirimele, J. (2013). Harnessing different motivational frames via mobile phones to promote daily physical activity and reduce sedentary behavior in aging adults. *PLoS One, 8*(4), e62613.

Klasnja, P., Hekler, E. B., Shiffman, S., Boruvka, A., Almirall, D., Tewari, A., & Murphy, S. A. (2015). Microrandomized trials: An experimental design for developing just-in-time adaptive interventions. *Health Psychology, 34*(S), 1220.

Korinek, E. V., Phatak, S. S., Martín, C. A., Freigoun, M. T., Rivera, D. E., Adams, M. A., ... Hekler, E. B. (2018). Adaptive step goals and rewards: A longitudinal growth model of daily steps for a smartphone-based walking intervention. *Journal of Behavioral Medicine, 41*(1), 74–86.

Kumar, S., Nilsen, W., Pavel, M., & Srivastava, M. (2013). Mobile health: Revolutionizing healthcare through transdisciplinary research. *Computer, 46*(1), 28–35.

Lee, J., & Yu, Z. (1994). Tuning of model predictive controllers for robust performance. *Computers & Chemical Engineering, 18*(1), 15–37.

Ljung, L. (1994). From data to model: A guided tour. *International Conference on Control, 1*, 422–430.

Ljung, L. (1999). *System identification: Theory for the user* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Martín, C. A. (2016). *A system identification and control engineering approach for optimizing mHealth behavioral interventions based on social cognitive theory* (Ph.D. dissertation). Arizona State University.

Martín, C. A., Rivera, D. E., & Hekler, E. B. (2015a). Design of informative identification experiments for behavioral interventions. *IFAC-PapersOnLine, 48*(28), 1325–1330.

Martín, C. A., Rivera, D. E., & Hekler, E. B. (2015b). An identification test monitoring procedure for MIMO systems based on statistical uncertainty estimation. *Proceedings of the IEEE Conference on Decision and Control*, 2719–2724.

Martín, C. A., Rivera, D. E., & Hekler, E. B. (2016a). A decision framework for an adaptive behavioral intervention for physical activity using hybrid model predictive control. *Proceedings of the American Control Conference*, 3576–3581.

Martín, C. A., Rivera, D. E., & Hekler, E. B. (2016b). An enhanced identification test monitoring procedure for MIMO systems relying on uncertainty estimates. *Proceedings of the IEEE Conference on Decision and Control*, 2091–2096.

Martín, C. A., Rivera, D. E., Riley, W. T., Hekler, E. B., Buman, M. P., Adams, M. A., & King, A. C. (2014). A dynamical systems model of social cognitive theory. *Proceedings of the American Control Conference*, 2407–2412.

Molenaar, P. C., & Campbell, C. G. (2009). The new person-specific paradigm in psychology. *Current Directions in Psychological Science, 18*(2), 112–117.

Nahum-Shani, I., Hekler, E. B., & Spruijt-Metz, D. (2015). Building health behavior models to guide the development of just-in-time adaptive interventions: A pragmatic framework. *Health Psychology, 34*(S), 1209.

Nandola, N. N., & Rivera, D. E. (2013). An improved formulation of hybrid model predictive control with application to production-inventory systems. *IEEE Transactions on Control Systems Technology, 21*(1), 121–135.

Navarro-Barrientos, J.-E., Rivera, D. E., & Collins, L. M. (2011). A dynamical model for describing behavioural interventions for weight loss and body composition change. *Mathematical and Computer Modelling of Dynamical Systems, 17*(2), 183–203.

Ogata, K. (2010). *Modern control engineering*. Pearson, Upper Saddle River, New Jersey.

Ogunnaike, B. A., & Ray, H. (1994). *Process dynamics, modeling, and control*. Oxford University Press, New York.

Olson, C. M., Strawderman, M. S., & Reed, R. G. (2004). Efficacy of an intervention to prevent excessive gestational weight gain. *American Journal of Obstetrics and Gynecology, 191*(2), 530–536.

Phatak, S. S., Freigoun, M., Martín, C. A., Rivera, D. E., Korinek, E. V., Adams, M. A., … Hekler, E. B. (2018). Modeling individual differences: A case study of the application of system identification for personalizing a physical activity intervention. *Journal of Biomedical Informatics, 79*, 82–97. https://doi.org/10.1016/j.jbi.2018.01.010

Phelan, S., Phipps, M. G., Abrams, B., Darroch, F., Schaffner, A., & Wing, R. R. (2011). Randomized trial of a behavioral intervention to prevent excessive gestational weight gain: The Fit for Delivery Study. *The American Journal of Clinical Nutrition, 93*(4), 772–779.

Polley, B. A., Wing, R., & Sims, C. (2002). Randomized controlled trial to prevent excessive weight gain in pregnant women. *International Journal of Obesity & Related Metabolic Disorders, 26*(11), 1494–1502.

Qin, S. J., & Badgwell, T. A. (2003). A survey of industrial model predictive control technology. *Control Engineering Practice, 11*(7), 733–764.

Riley, W. T., Martín, C. A., Rivera, D. E., Hekler, E. B., Adams, M. A., Buman, M. P., … King, A. C. (2016). Development of a dynamic computational model of social cognitive theory. *Translational Behavioral Medicine, 6*(4), 483–495.

Riley, W. T., Rivera, D. E., Atienza, A. A., Nilsen, W., Allison, S. M., & Mermelstein, R. (2011). Health behavior models in the age of mobile interventions: Are our theories up to the task? *Translational Behavioral Medicine, 1*(1), 53–71.

Rivera, D. E. (2012). Optimized behavioral interventions: What does system identification and control engineering have to offer? *IFAC Proceedings Volumes, 45*(16), 882–893.

Rivera, D. E., Lee, H., Mittelmann, H. D., & Braun, M. W. (2009). Constrained multisine input signals for plant-friendly identification of chemical process systems. *Journal of Process Control, 19*(4), 623–635.

Rivera, D. E., Pew, M. D., & Collins, L. M. (2007). Using engineering control principles to inform the design of adaptive interventions: A conceptual introduction. *Drug and Alcohol Dependence, 88*, S31–S40.

Savage, J. S., Downs, D. S., Dong, Y., & Rivera, D. E. (2014). Control systems engineering for optimizing a prenatal weight gain intervention to regulate infant birth weight. *American Journal of Public Health, 104*(7), 1247–1254.

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology, 4*, 1–32.

Skogestad, S., & Postlethwaite, I. (1996). *Multivariable feedback control: Analysis and design* (Vol. 1). Chichester, UK: Wiley.

Stenman, A. (1999). *Model on demand: Algorithms, analysis and applications* (Ph.D. dissertation). Department of Electrical Engineering, Linköping University.

Symons Downs, D. (2016). Obesity in special populations: Pregnancy. *Primary Care, 43*(1), 109.

Symons Downs, D., DiNallo, J. M., & Kirner, T. L. (2007). Pregnant women's exercise motivation and behavior: The moderating influence of parental status. *Journal of Sport & Exercise Psychology, 29*, S160.

Symons Downs, D., Savage, J. S., & Rauff, E. L. (2014). Falling short of guidelines? Lacking knowledge to achieve gestational weight gain, diet, and physical activity recommendations in pregnancy. *Journal of Women's Health Care, 3*(184), 1–6. https://doi.org/10.4172/2167-0420.1000184

Symons Downs, D., & Hausenblas, H. (2004). Exercising during pregnancy and postpartum: An elicitation study using the framework of the theory of planned behavior. *Journal of Midwifery & Women's Health, 49*, 138–144.

The Look AHEAD Research Group. (2006). Baseline characteristics of the randomised cohort from the Look AHEAD (Action for Health in Diabetes) study. *Diabetes and Vascular Disease Research, 3*(3), 202–215.

Thomas, D. M., Navarro-Barrientos, J. E., Rivera, D. E., Heymsfield, S. B., Bredlau, C., Redman, L. M., ... Butte, N. F. (2012). Dynamic energy-balance model predicting gestational weight gain. *The American Journal of Clinical Nutrition, 95*(1), 115–122.

Timms, K. P. (2014). *A novel engineering approach to modeling and optimizing smoking cessation interventions* (Ph.D. dissertation). Arizona State University.

Timms, K. P., Rivera, D. E., Collins, L. M., & Piper, M. E. (2014). Continuous-time system identification of a smoking cessation intervention. *International Journal of Control, 87*(7), 1423–1437.

Vanderwater, R., & Davison, D. (2011). Using rewards to change a person's behavior: A double-integrator output-feedback dynamic control approach. *Proceedings of the American Control Conference*, 1861–1866.

Velicer, W. (2010). *Applying idiographic research methods: Two examples.* Paper presented at the 8th International Conference on Teaching Statistics, Ljubljana, Slovenia.

Wang, W., & Rivera, D. E. (2008). Model predictive control for tactical decision-making in semiconductor manufacturing supply chain management. *IEEE Transactions on Control Systems Technology, 16*(5), 841–855.

Wang, Y., Dassau, E., & Doyle, F. J. (2010). Closed-loop control of artificial pancreatic β-Cell in type 1 diabetes mellitus using model predictive iterative learning control. *IEEE Transactions on Biomedical Engineering, 57*(2), 211–219.

Zafra-Cabeza, A., Rivera, D. E., Collins, L. M., Ridao, M. A., & Camacho, E. F. (2011). A risk-based model predictive control approach to adaptive interventions in behavioral health. *IEEE Transactions on Control Systems Technology, 19*(4), 891–901.

Zurakowski, R., & Teel, A. R. (2006). A model predictive control based scheduling method for HIV therapy. *Journal of Theoretical Biology, 238*(2), 368–382.

# Coding and Interpretation of Effects in Analysis of Data from a Factorial Experiment


**Kari C. Kugler, John J. Dziak, and Jessica Trail**

**Abstract** This chapter is intended to describe the differences between effect coding and dummy coding when the multiple regression approach is used to perform analysis of variance (ANOVA) with balanced (i.e., an equal number of subjects in each experimental condition) factorial designs. Using a hypothetical example of a $2^3$ factorial experiment, we present these two coding schemes for categorical independent variables and explain how the effects estimated can have different interpretations depending on the coding scheme used. Particular attention is paid to highlighting how and why differences exist and when a researcher might want to use one coding scheme over another. We demonstrate that effect coding is usually preferred for analyzing data from factorial experiments in the optimization phase of the multiphase optimization strategy.

## 1 Introduction

Factorial experiments are increasingly being used in the behavioral sciences, in part due to the rising interest in optimization of multicomponent behavioral, biobehavioral, and biomedical interventions. In the companion to this volume, Collins (2018) describes an approach for optimization of interventions called the multiphase optimization strategy (MOST). As part of this approach, the investi-

K. C. Kugler (✉)
The Pennsylvania State University, The Methodology Center, University Park, PA, USA
e-mail: kck18@psu.edu

J. J. Dziak
The Methodology Center, The Pennsylvania State University, University Park, PA, USA

J. Trail
Argonne National Laboratory, Chicago, IL, USA



L. M. Collins, K. C. Kugler (eds.), *Optimization of Behavioral, Biobehavioral, and Biomedical Interventions*, Statistics for Social and Behavioral Sciences,
https://doi.org/10.1007/978-3-319-91776-4_6

gator conducts an experiment to gather the information needed to optimize the intervention; the approach to experimentation varies depending on the resources available and the objectives of the optimization. Frequently the optimization of the intervention requires assessing the individual and combined effects of a set of intervention components, so that the investigator can determine which components and component levels meet the optimization criterion that has been selected. When this kind of information is needed, often factorial designs emerge as the most efficient and economical way to gather the information needed for optimization.

Many behavioral scientists have had some training in factorial experiments, mostly the $2 \times 2$ factorial, and are familiar with the textbook definitions of the terms *main effect* and *interaction*. They are also familiar with the facts that these experiments are often analyzed using factorial analysis of variance (ANOVA) and that ANOVA analyses can be reexpressed as multiple regression models by using certain numerical coding systems for the categorical factors. However, researchers may not be aware that how factor levels are coded in the regression factorial analysis of variance (ANOVA) determines how the resulting effect estimates should be interpreted. In fact, when dummy coding (0s and 1s) is used, significance tests of the resulting effect estimates generally do not represent tests of the statistical significance of main effects and interactions according to the ANOVA textbook definitions, but rather are more complicated linear combinations of these effects.

The purpose of this chapter is to demonstrate how the choice between dummy coding and effect coding can change the interpretation of the effects estimated based on the results of a factorial experiment. As will be shown, the two coding schemes always yield the same omnibus $F$ statistic used for testing the overall predictive ability of the fitted linear model. However, they can yield different estimates, test statistics, and $p$-values for the main and interaction effects. Neither dummy coding nor effect coding is right or wrong per se, but because the resulting estimates have different meanings, one or the other of the coding schemes will be more appropriate for any given set of research questions. Therefore, it is important for investigators to have a clear understanding of the differences. Hardy (1993) provides an excellent discussion of this topic in the context of nonexperimental studies; however, in this chapter we focus on the differences between effect and dummy coding for factorial experiments. For simplicity, the discussion is limited to factorial experiments in which all factors have two levels, often referred to as $2^k$ experiments, where $k$ denotes the number of factors. This type of factorial design is typical in screening experiments (see the companion volume; Collins, 2018).

The chapter is organized as follows: First we provide a brief introduction to the factorial design. We follow this with a description of the classical definitions of main effects and interactions. We then present the two different coding schemes for categorical factors. We provide a discussion of how to interpret regression coefficients when effect coding is used and when dummy coding is used. Next we provide a numerical example comparing effect and dummy coding. We then discuss how to analyze data generated from a factorial experiment using two of the most widely used statistical software packages (SAS and SPSS). We conclude with a discussion about how behavioral scientists might integrate this information into their research.

## 2 Brief Introduction to the Factorial Design

Suppose an investigator is interested in examining the effects of three components of an obesity preventive intervention designed to increase adherence to protocols related to weight loss (e.g., Pellegrini et al., 2014, 2015): coaching calls, text messages, and a report provided to the individual's primary care physician (PCP). The investigator wants to optimize this intervention using MOST (Collins, Dziak, & Li, 2009; Collins, Kugler, & Gwadz, 2016) and decides to conduct a factorial experiment. The following three factors are identified: (1) *COACH* (yes/no), (2) *TEXT* (yes/no), and (3) *PCP* (yes/no). This is a $2 \times 2 \times 2$, or $2^3$, factorial experiment. The design of this experiment is depicted in Table 1. In the experimental condition column of Table 1, a "+" represents yes (i.e., this component is provided) and a "−" represents no (i.e., this component is not provided). Instead of yes and no, the levels could also be high/low or any other reasonable two-level comparison. As Table 1 shows, there are $2^3 = 8$ experimental conditions. Each experimental condition represents a unique combination of the levels of the three factors, and all unique combinations are included in the design.

### *2.1 Classical Definition of Effects*

Let us review the classical definitions of effects of factors with two levels.

#### 2.1.1 Main Effects

The classical definition of the main effect in a $2^k$ ANOVA is the difference between the mean response at one level of a particular factor and the mean response at the other level, collapsing over the levels of all remaining factors (Montgomery, 2009). For a $2^3$ factorial experiment with factors $A$, $B$, and $C$, the main effect of factor $A$ is represented by

**Table 1** Experimental conditions for the hypothetical $2^3$ factorial experiment examining weight loss adherence intervention components

| Experimental condition number | Experimental condition | Factor 1 *COACH* | Factor 2 *TEXT* | Factor 3 *PCP* |
|---|---|---|---|---|
| 1 | − − − | No | No | No |
| 2 | − − + | No | No | Yes |
| 3 | − + − | No | Yes | No |
| 4 | − + + | No | Yes | Yes |
| 5 | + − − | Yes | No | No |
| 6 | + − + | Yes | No | Yes |
| 7 | + + − | Yes | Yes | No |
| 8 | + + + | Yes | Yes | Yes |

$$ME_A = \overline{\mu}_{(+\cdot\cdot)} - \overline{\mu}_{(-\cdot\cdot)},$$

where $\overline{\mu}_{(+\cdot\cdot)}$ and $\overline{\mu}_{(-\cdot\cdot)}$ represents the mean response for the "+" and "−" levels, respectively, of factor $A$, collapsing over the levels of factors $B$ and $C$ (the dot subscript means "summed over"). Here the notation $\overline{\mu}$ denotes a mean of means, that is, an average of the mean responses across several experimental conditions. In the experiment depicted in Table 1, $\overline{\mu}_{(+\cdot\cdot)}$ is the mean response of experimental conditions 5–8, and $\overline{\mu}_{(-\cdot\cdot)}$ is the mean response of experimental conditions 1–4. Thus, in the factorial experiment described above, the main effect of *COACH* is the difference between mean adherence for *COACH* at the yes and no levels, collapsing over levels of *TEXT* and *PCP*.

After some algebra, using facts such as $\overline{\mu}_{(+\cdot\cdot)} = \left(\overline{\mu}_{(+++)} + \overline{\mu}_{(++-)} + \overline{\mu}_{(+-+)} + \overline{\mu}_{(+--)}\right)/4$, the main effect equals

$$\begin{aligned} ME_A = &\frac{1}{4}\overline{\mu}_{(+++)} + \frac{1}{4}\overline{\mu}_{(++-)} + \frac{1}{4}\overline{\mu}_{(+-+)} + \frac{1}{4}\overline{\mu}_{(+--)} \\ &- \frac{1}{4}\overline{\mu}_{(-++)} - \frac{1}{4}\overline{\mu}_{(-+-)} - \frac{1}{4}\overline{\mu}_{(--+)} - \frac{1}{4}\overline{\mu}_{(---)}. \end{aligned}$$

The main effects of $B$ and $C$ are defined similarly:

$$ME_B = \overline{\mu}_{(\cdot+\cdot)} - \overline{\mu}_{(\cdot-\cdot)}$$

and

$$ME_c = \overline{\mu}_{(\cdot\cdot+)} - \overline{\mu}_{(\cdot\cdot-)}.$$

### 2.1.2 Two-Way Interaction Effects

There is an $A \times B$ interaction if the effect of factor $A$ is different depending on the level of factor $B$ (or, equivalently, if the effect of factor $B$ is different depending on the level of factor A). In a $2^k$ ANOVA, a two-way interaction is the average of the difference in the effect of a particular factor across the levels of a second factor, collapsing over all other factors (Montgomery, 2009). For example, consider a classical $2^3$ ANOVA. There is an interaction between factors $A$ and $B$ if the average effect of $A$ differs over levels of $B$:

$$\left(\overline{\mu}_{(++\cdot)} - \overline{\mu}_{(-+\cdot)}\right) \neq \left(\overline{\mu}_{(\pm\cdot)} - \overline{\mu}_{(--\cdot)}\right).$$

Therefore, one straightforward definition of the $A \times B$ interaction could be

$$\left(\overline{\mu}_{(++\cdot)} - \overline{\mu}_{(-+\cdot)}\right) - \left(\overline{\mu}_{(+-\cdot)} - \overline{\mu}_{(--\cdot)}\right),$$

which is zero if the effect of A does not depend on that of B. After some algebra, using facts such as $\overline{\mu}_{(++\cdot)} = \frac{1}{2}(\mu_{+++} + \mu_{++-})$, this definition of the interaction equals

$$\frac{1}{2}\mu_{+++} + \frac{1}{2}\mu_{++-} - \frac{1}{2}\mu_{-++} - \frac{1}{2}\mu_{-+-} - \frac{1}{2}\mu_{+-+} - \frac{1}{2}\mu_{+--} + \frac{1}{2}\mu_{--+} + \frac{1}{2}\mu_{---}.$$

However, there is a problem with this definition because it essentially inflates the scale of measurement for the interaction relative to the main effect; notice the $\frac{1}{2}$ coefficients above relative to the $\frac{1}{4}$ coefficients in the similar decomposition of the main effect. Therefore, a more consistent definition of the interaction would be *half* the difference in differences, and we will use that definition here. This allows the interaction to be twice its effect-coded regression coefficient, just as the main effect is twice its own effect-coded regression coefficient. Specifically, we define the $A \times B$ interaction as

$$INT_{A\times B} = \frac{1}{2}\left[\left(\overline{\mu}_{(++\cdot)} - \overline{\mu}_{(-+\cdot)}\right) - \left(\overline{\mu}_{(+-\cdot)} - \overline{\mu}_{(--\cdot)}\right)\right].$$

The rescaling does not matter for testing purposes, because rescaling the mean also rescales the standard error. However, when reporting effect sizes, it is important to be clear about what definition is being used (see Dziak, Nahum-Shani, & Collins, 2012).

In the example experiment, the *COACH* × *TEXT* interaction is the average of the difference in the effect of *COACH* on adherence when *TEXT* = yes and the effect of *COACH* when *TEXT* = no, collapsed over levels of *PCP*. If the effect of *COACH* on adherence is the same when *TEXT* = yes and when *TEXT* = no, there is no *COACH* × *TEXT* interaction, and this difference is zero. For a more detailed discussion of interactions, see Chap. 4 in the companion volume.

### 2.1.3 Three-Way Interaction Effects

There is an $A \times B \times C$ interaction when the $A \times B$ interaction effect is different depending on the level of factor $C$. (This can alternatively be thought of as any of the two-way interaction effects differs depending on the level of the third factor.) In a classical $2^k$ ANOVA, a three-way interaction is (half) the average of the difference in the two-way interaction effects at differing levels of a third factor, collapsing over any remaining factors. For example, in a $2^3$ ANOVA, this is represented by the following:

$$INT_{A\times B\times C} = \frac{1}{2}\left[\frac{1}{2}\left[\left(\mu_{(+++)} - \mu_{(-++)}\right) - \left(\mu_{(+-+)} - \mu_{(--+)}\right)\right] - \frac{1}{2}\left[\left(\mu_{(++-)} - \mu_{(-+-)}\right) - \left(\mu_{(+--)} - \mu_{(---)}\right)\right]\right].$$

Notice the additional $\frac{1}{2}$ multiplier, similar to the multiplier for $INT_{A \times B}$. In the companion volume (Collins, 2018), the 2 or 4 in the denominator of $INT_{A \times B}$ or $INT_{A \times B \times C}$ is called a coefficient correction.

The $COACH \times TEXT \times PCP$ interaction is the average of the difference between the $COACH \times TEXT$ interaction when $PCP =$ yes and the $COACH \times TEXT$ interaction when $PCP =$ no. If there were additional factors in the experiment, this difference would be collapsed over those factors as well.

## 2.2 Coding Schemes for Categorical Variables

In the section that follows, we present the details of dummy coding and effect coding of experimental factors. In later sections we demonstrate how these two coding schemes differ, how the effects they estimate are different, and how one could come to a different conclusion depending on the type of coding scheme used. In particular, we will demonstrate that significance tests of the coefficients of effect-coded factors and their products are equivalent to tests of ANOVA main effects and interactions, but significance tests of the coefficients of dummy-coded factors are not equivalent to tests of ANOVA effects unless there are no interactions in the model. When there is not a product term in the model, it does not matter what type of coding scheme is used: The test for significance will be the same for each of the effects. However, there will be differences in the scale of the regression coefficients. For the dummy-coded additive model, the regression coefficient represents the classically defined main effects, whereas with effect coding, the regression coefficient needs to be multiplied by a scaling factor of 2 in order to be equivalent to the classically defined main effect. The rescaling by two occurs because the high and low levels of effect codes are placed further apart $(+1 - (-1) = 2)$ than those of dummy codes $(1 - 0 = 1)$; it could be avoided by using $+1/2$ and $-1/2$ as codes instead of the usual $+1$ and $-1$ (as in, e.g., Raudenbush & Liu, 2000), but we do not address this form of notation here because it would make other notational issues related to factorial experiments become much more complicated.

We begin with dummy coding, which is the type of coding perhaps most familiar to behavioral scientists.

### 2.2.1 Dummy Coding

In a dummy coding system, 1 represents membership in one level/category of a factor, and 0 represents membership in the other. For participants who were provided *COACH*, the dummy variable has a value of 1, and for those who were not provided *COACH*, the dummy variable has a value of 0. The same approach would be used for *TEXT* and *PCP*. (Readers interested in learning more about dummy coding for factors with more than two levels are referred to Hardy (1993, pp. 8–9).)

Table 2 shows the dummy coding scheme for the factorial design in Table 1. The rows correspond to the experimental conditions, and the columns correspond to the dummy variable vector for each effect. We are avoiding the use of the terms *main effect* and *interaction* here because, as will be demonstrated below, the effects estimated when dummy codes are used are generally not main effects and interactions according to the classical definitions above. Instead, we will use the terms *first-order effect* (FOE) instead of main effect, *second-order effect* (SOE) instead of two-way interaction, *third-order effect* (TOE) instead of three-way interaction, and so on.

The column labeled $Z_{COACH}$ represents the dummy variable corresponding to the FOE of the factor *COACH* (the letter $Z$ was arbitrarily chosen and should not be confused with a $Z$ score). All those at the no level are given a 0 (Rows 1–4), and all those at the yes level are given a 1 (Rows 5–8). The column labeled $Z_{TEXT}$ represents the dummy variable corresponding to the FOE of the factor *TEXT*. Again, a 0 is listed for all those at the no level (Rows 1,2,5,6); a 1 is listed for all those at the yes level (Rows 3,4,7,8). The column $Z_{PCP}$ represents the dummy variable corresponding to the FOE of the factor *PCP*. Here, Rows 1,3,5, and 7 are given a 0 for the no level, and Rows 2,4,6, and 8 are given a 1 for the yes level.

The next three columns contain the dummy codes for the SOEs. These are constructed by multiplying the elements in the FOE vectors for the individual factors involved in the SOE. For example, $Z_{COACH \times TEXT}$ represents the $COACH \times TEXT$ SOE. It is created by multiplying the elements in the vector for $Z_{COACH}$ by the elements in the vector for $Z_{TEXT}$. Finally, the last column represents the TOE, and it is obtained by multiplying the three FOE vectors.

The role these vectors play in conducting a factorial ANOVA will be discussed in Sect. 2.3.5.

### 2.2.2 Effect Coding

Effect coding is similar to dummy coding, in that a 1 represents membership in one of the levels/categories. However, in effect coding for two-level variables, the other level is coded $-1$ rather than 0 (see Hardy, 1993, pp. 64–71, for a description of effect coding for variables with more than two levels). The variables representing the interactions are constructed by multiplying the vectors corresponding to the factors involved in the interactions, in a manner analogous to what was shown above for dummy coding. Table 3 presents the effect coding scheme for the hypothetical factorial experiment in Table 1.

**Table 2** Dummy coding scheme for the eight experimental conditions

| Experimental condition number | Experimental condition | Dummy codes | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | $Z_{COACH}$ | $Z_{TEXT}$ | $Z_{PCP}$ | $Z_{COACH \times TEXT}$ | $Z_{COACH \times PCP}$ | $Z_{TEXT \times PCP}$ | $Z_{COACH \times TEXT \times PCP}$ |
| 1 | − − − | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | − − + | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 3 | − + − | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 4 | − + + | 0 | 1 | 1 | 0 | 0 | 1 | 0 |
| 5 | + − − | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | + − + | 1 | 0 | 1 | 0 | 1 | 0 | 0 |
| 7 | + + − | 1 | 1 | 0 | 1 | 0 | 0 | 0 |
| 8 | + + + | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

**Table 3** Effect coding scheme for the eight experimental conditions

| Experimental condition number | Experimental condition | Effect codes | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | $X_{COACH}$ | $X_{TEXT}$ | $X_{PCP}$ | $X_{COACH \times TEXT}$ | $X_{COACH \times PCP}$ | $X_{TEXT \times PCP}$ | $X_{COACH \times TEXT \times PCP}$ |
| 1 | − − − | −1 | −1 | −1 | 1 | 1 | 1 | −1 |
| 2 | − − + | −1 | −1 | 1 | 1 | −1 | −1 | 1 |
| 3 | − + − | −1 | 1 | −1 | −1 | 1 | −1 | 1 |
| 4 | − + + | −1 | 1 | 1 | −1 | −1 | 1 | −1 |
| 5 | + − − | 1 | −1 | −1 | −1 | −1 | 1 | 1 |
| 6 | + − + | 1 | −1 | 1 | −1 | 1 | −1 | −1 |
| 7 | + + − | 1 | 1 | −1 | 1 | −1 | −1 | −1 |
| 8 | + + + | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

## 2.3 Use of Codes in Regression Models and Interpretation of Coefficients

In this chapter we compare the algebraic implications of dummy codes and effect codes on the interpretation of regression coefficients. Although we introduced dummy codes first above because they are more familiar to most researchers, it is convenient in this section to discuss effect codes first and then show how dummy codes differ from them.

### 2.3.1 Population Means for Each Experimental Condition When Effect Coding Is Used

Below is the hypothetical experiment with three factors, each with two levels, expressed as an effect-coded regression equation:

$$\mu = \beta_0 + \beta_1 X_{COACH} + \beta_2 X_{TEXT} + \beta_3 X_{PCP} + \beta_{12} X_{COACH \times GUM} + \beta_{13} X_{COACH \times PCP} + \beta_{23} X_{TEXT \times PCP} + \beta_{123} X_{COACH \times TEXT \times PCP}. \quad (1)$$

In Eq. (1) above, $\mu$ represents the population mean of the outcome of interest, that is, adherence to weight loss protocol; and each $X$ corresponds to a column from Table 3. Each regression coefficient ($\beta_1, \beta_2, \beta_3$, etc.) expresses the expected change in $\mu$ given a one-unit change in the associated $X$, holding all else constant. For example, for an individual in experimental condition (1) in Table 3, the expression would be

$$\mu_{---} = \beta_0 + \beta_1 (-1) + \beta_2 (-1) + \beta_3 (-1) + \beta_{12}(1) + \beta_{13}(1) + \beta_{23}(1) + \beta_{123} (-1) .$$

Table 4 shows the population means for each experimental condition, expressed in terms of Eq. (1) using effect coding.

When effect coding is used, the intercept, $\beta_0$, is the grand mean. This can be seen by examining Table 4 carefully. Assuming a balanced design (see Chap. 3 in the companion volume), in which each experimental condition has the same sample size $n$, the grand mean is the average of all the experimental condition means. Expressing these means in terms of regression coefficients, as is shown in Table 4, and taking the average shows that the result is $\beta_0$.

**Table 4** Population means for each experimental condition when effect coding is used

| Experimental condition number | Experimental condition | Population means |
|---|---|---|
| 1 | − − − | $\mu_{(---)} = \beta_0 - \beta_1 - \beta_2 - \beta_3 + \beta_{12} + \beta_{13} + \beta_{23} - \beta_{123}$ |
| 2 | − − + | $\mu_{(--+)} = \beta_0 - \beta_1 - \beta_2 + \beta_3 + \beta_{12} - \beta_{13} - \beta_{23} + \beta_{123}$ |
| 3 | − + − | $\mu_{(-+-)} = \beta_0 - \beta_1 + \beta_2 - \beta_3 - \beta_{12} + \beta_{13} - \beta_{23} + \beta_{123}$ |
| 4 | − + + | $\mu_{(-++)} = \beta_0 - \beta_1 + \beta_2 + \beta_3 - \beta_{12} - \beta_{13} + \beta_{23} - \beta_{123}$ |
| 5 | + − − | $\mu_{(+--)} = \beta_0 + \beta_1 - \beta_2 - \beta_3 - \beta_{12} - \beta_{13} + \beta_{23} + \beta_{123}$ |
| 6 | + − + | $\mu_{(+-+)} = \beta_0 + \beta_1 - \beta_2 + \beta_3 - \beta_{12} + \beta_{13} - \beta_{23} - \beta_{123}$ |
| 7 | + + − | $\mu_{(++-)} = \beta_0 + \beta_1 + \beta_2 - \beta_3 + \beta_{12} - \beta_{13} - \beta_{23} - \beta_{123}$ |
| 8 | + + + | $\mu_{(+++)} = \beta_0 + \beta_1 + \beta_2 + \beta_3 + \beta_{12} + \beta_{13} + \beta_{23} + \beta_{123}$ |

### 2.3.2 Interpretation of Effects in Effect Coding

Main Effects

In (1), $\beta_1$ represents the main effect of *COACH*. This main effect is computed by taking the difference between the average of Rows 5–8 (*COACH* = yes) and average of Rows 1–4 (*COACH* = no).

$$\begin{aligned}\overline{\mu}_{(+\cdot\cdot)} = \frac{1}{4}[&(\beta_0 + \beta_1 - \beta_2 - \beta_3 - \beta_{12} - \beta_{13} + \beta_{23} + \beta_{123}) \\ &+ (\beta_0 + \beta_1 - \beta_2 + \beta_3 - \beta_{12} + \beta_{13} - \beta_{23} - \beta_{123}) \\ &+ (\beta_0 + \beta_1 + \beta_2 - \beta_3 + \beta_{12} - \beta_{13} - \beta_{23} - \beta_{123}) \\ &+ (\beta_0 + \beta_1 + \beta_2 + \beta_3 + \beta_{12} + \beta_{13} + \beta_{23} + \beta_{123})] \\ = \frac{1}{4}[4\beta_0 + 4\beta_1] &= \beta_0 + \beta_1 \overline{\mu}_{(-\cdot\cdot)} = \frac{1}{4}[4\beta_0 - 4\beta_1] = \beta_0 - \beta_1\end{aligned}$$

Using the population means for each experimental condition in Table 4, the main effect is expressed as follows:

$$ME_{COACH} = \overline{\mu}_{(+\cdot\cdot)} - \overline{\mu}_{(-\cdot\cdot)} = (\beta_0 + \beta_1) - (\beta_0 - \beta_1) = 2\beta_1,$$

and the regression coefficient is expressed as follows:

$$\beta_1 = \frac{1}{2} * ME_{COACH}.$$

As shown in this equation, with effect coding the $\beta$s *correspond to the main effects according to the classical definition*. In order to compute the actual main effect, a simple linear transformation is required, that is, multiplying the regression coefficient by a scaling constant of 2. This scaling constant has no effect on the hypothesis test, because it is also applied to the standard errors.

Two-Way Interaction Effects

Using the same approach, the effect of the *COACH* × *TEXT* interaction is calculated as follows:

$$\overline{\mu}_{(++\cdot)} = \frac{1}{2}\left[2\beta_0 + 2\beta_1 + 2\beta_2 + 2\beta_{12}\right]$$

$$\overline{\mu}_{(-+\cdot)} = \frac{1}{2}\left[2\beta_0 - 2\beta_1 + 2\beta_2 - 2\beta_{12}\right]$$

$$\overline{\mu}_{(+-\cdot)} = \frac{1}{2}\left[2\beta_0 + 2\beta_1 - 2\beta_2 - 2\beta_{12}\right]$$

$$\overline{\mu}_{(--\cdot)} = \frac{1}{2}\left[2\beta_0 - 2\beta_1 - 2\beta_2 + 2\beta_{12}\right]$$

$$INT_{COACH\times TEXT} = \frac{1}{2}\left[(2\beta_1 + 2\beta_{12}) - (2\beta_1 - 2\beta_{12})\right] = 2\beta_{12}$$

$$\beta_{12} = \frac{1}{2} INT_{COACH\times TEXT}$$

Thus, with effect coding, $\beta_{12}$ represents one-half the classical definition of a two-way interaction. In other words, the effect of the *COACH* × *TEXT* interaction is simply $2\beta_{12}$.

Three-Way Interaction Effects

Again, using the same approach, the following represents the *COACH* × *TEXT* × *PCP* interaction effect:

$$\begin{aligned} &INT_{COACH\times TEXT\times PCP} \\ &\quad = \frac{1}{2}\left[\frac{1}{2}\left[(2\beta_1 + 2\beta_{12} + 2\beta_{13} + 2\beta_{123}) - (2\beta_1 - 2\beta_{12} + 2\beta_{13} - 2\beta_{123})\right]\right. \\ &\qquad \left. -\frac{1}{2}\left[(2\beta_1 + 2\beta_{12} - 2\beta_{13} - 2\beta_{123}) - (2\beta_1 - 2\beta_{12} - 2\beta_{13} + 2\beta_{123})\right]\right] \\ &\quad = 2\beta_{123} \end{aligned}$$

and so,

$$\beta_{123} = \frac{1}{2} INT_{COACH \times TEXT \times PCP}.$$

### 2.3.3 Summary: Effect-Coded Effects

When effect coding is used, the regression coefficients in a regression model are equivalent to the classically defined main effects and interactions, except for a scaling constant of 2 that does not affect hypothesis tests. Simply multiplying the regression coefficient by this scaling constant produces the estimated main and interaction effects for factors with two levels.

### 2.3.4 Population Means for Each Experimental Condition When Dummy Coding Is Used

Below is the hypothetical experiment with three factors, each with two levels, now expressed as a regression equation using dummy coding:

$$\begin{aligned} \mu = {} & \alpha_0 + \alpha_1 Z_{COACH} + \alpha_2 Z_{TEXT} + \alpha_3 Z_{PCP} + \alpha_{12} Z_{COACH \times TEXT} \\ & + \alpha_{13} Z_{COACH \times PCP} + \alpha_{23} Z_{TEXT \times PCP} + \alpha_{123} Z_{COACH \times TEXT \times PCP}. \end{aligned} \quad (2)$$

In Eq. (2), $\mu$ again represents the outcome of interest, that is, adherence to weight loss protocol. Each $Z$ corresponds to a column from Table 2. We use $\alpha$ to represent regression coefficients to distinguish these from the regression coefficients in Eq. (1).

Table 5 shows the population means for each of the experimental conditions expressed in terms of Eq. (2) when dummy coding is used. For example, the expected mean for experimental condition 1, in which each of the three factors is set to the no (−) level, is

$$\mu_{(---)} = \alpha_0 + \alpha_1(0) + \alpha_2(0) + \alpha_3(0) + \alpha_{12}(0) + \alpha_{13}(0) + \alpha_{23}(0) + \alpha_{123}(0) = \alpha_0,$$

as shown in Row 1 of Table 5. This shows that the intercept, $\alpha_0$, equals $\mu_{---}$, the mean when all of the independent variables are set to zero—in this case, the no level of each factor. Each of the remaining coefficients represents the expected change in $\mu$ given a one-unit change in the associated $Z$, holding all else constant. As another example, the population mean for experimental condition 8 is

$$\begin{aligned} \mu_{(+++)} &= \alpha_0 + \alpha_1(1) + \alpha_2(1) + \alpha_3(1) + \alpha_{12}(1) + \alpha_{13}(1) + \alpha_{23}(1) + \alpha_{123} \\ &= \alpha_0 + \alpha_1 + \alpha_2 + \alpha_3 + \alpha_{12} + \alpha_{13} + \alpha_{23} + \alpha_{123}, \end{aligned}$$

as shown in Row 8 of Table 5.

**Table 5** Population means for each experimental condition when dummy coding is used

| Experimental condition number | Experimental condition | Population means |
|---|---|---|
| 1 | $---$ | $\mu_{(---)} = \alpha_0$ |
| 2 | $--+$ | $\mu_{(--+)} = \alpha_0 + \alpha_3$ |
| 3 | $-+-$ | $\mu_{(-+-)} = \alpha_0 + \alpha_2$ |
| 4 | $-++$ | $\mu_{(-++)} = \alpha_0 + \alpha_2 + \alpha_3 + \alpha_{23}$ |
| 5 | $+--$ | $\mu_{(+--)} = \alpha_0 + \alpha_1$ |
| 6 | $+-+$ | $\mu_{(+-+)} = \alpha_0 + \alpha_1 + \alpha_3 + \alpha_{13}$ |
| 7 | $++-$ | $\mu_{(++-)} = \alpha_0 + \alpha_1 + \alpha_2 + \alpha_{12}$ |
| 8 | $+++$ | $\mu_{(+++)} = \alpha_0 + \alpha_1 + \alpha_2 + \alpha_3 + \alpha_{12} + \alpha_{13} + \alpha_{23} + \alpha_{123}$ |

Comparing Tables 4 and 5 shows that the type of coding scheme used produces dramatically different expressions for the population means for each experimental condition. It is also worth noting that the intercept has a different meaning. When effect coding is used, the intercept, $\beta_0$, is the grand mean, but the intercept in a dummy-coded regression model, $\alpha_0$,is the mean response when each of the three factors is set to the no $(-)$ level.

### 2.3.5 Expressing Dummy-Coded Effects in Terms of Classical Effects

*When dummy coding is used and product terms are included in the model, the regression coefficients do not correspond to the main effects and interactions according to the classical definition* (Hardy, 1993). This is not inherently bad or a mistake; however, it is important to be aware of the interpretational differences to ensure that the coding scheme selected is appropriate for the research questions at hand. In this section we directly compare the two coding schemes by expressing dummy-coded effects in terms of classical effect-coded effects.

FOEs

As discussed above, the main effect of *COACH* is computed by taking the difference between the average of Rows 5–8 (*COACH* = yes)

$$\begin{aligned}\overline{\mu}_{(+\cdot\cdot)} &= \tfrac{1}{4}\left[(\alpha_0 + \alpha_1) + (\alpha_0 + \alpha_1 + \alpha_3 + \alpha_{13}) + (\alpha_0 + \alpha_1 + \alpha_2 + \alpha_{12})\right. \\ &\quad \left. + (\alpha_0 + \alpha_1 + \alpha_2 + \alpha_3 + \alpha_{12} + \alpha_{13} + \alpha_{23} + \alpha_{123})\right] \\ &= \tfrac{1}{4}\left[4\alpha_0 + 4\alpha_1 + 2\alpha_2 + 2\alpha_3 + 2\alpha_{12} + 2\alpha_{13} + \alpha_{23} + \alpha_{123}\right]\end{aligned}$$

and Rows 1–4 ($COACH$ = no)

$$\begin{aligned}
\overline{\mu}_{(-\cdot\cdot)} &= \tfrac{1}{4}\left[(\alpha_0) + (\alpha_0+\alpha_3) + (\alpha_0+\alpha_2) + (\alpha_0+\alpha_2+\alpha_3+\alpha_{23})\right] \\
&= \tfrac{1}{4}\left[4\alpha_0 + 2\alpha_2 + 2\alpha_3 + \alpha_{23}\right] \\
ME_{COACH} &= \overline{\mu}_{(+\cdot\cdot)} - \overline{\mu}_{(-\cdot\cdot)} \\
&= \tfrac{1}{4}\left[4\alpha_0 + 4\alpha_1 + 2\alpha_2 + 2\alpha_3 + 2\alpha_{12} + 2\alpha_{13} + \alpha_{23} + \alpha_{123}\right] \\
&-\tfrac{1}{4}\left[4\alpha_0 + 2\alpha_2 + 2\alpha_3 + \alpha_{23}\right] \\
&= \tfrac{1}{4}\left[4\alpha_1 + 2\alpha_{12} + 2\alpha_{13} + \alpha_{123}\right] \\
&= \alpha_1 + \tfrac{1}{2}\alpha_{12} + \tfrac{1}{2}\alpha_{13} + \tfrac{1}{4}\alpha_{123}.
\end{aligned}$$

Based on this, it is now possible to solve for $\alpha_1$:

$$\alpha_1 = FOE_{COACH} = ME_{COACH} - \frac{1}{2}\alpha_{12} - \frac{1}{2}\alpha_{13} - \frac{1}{4}\alpha_{123}.$$

As shown in this equation, when dummy coding is used, $\alpha_1$ does not represent the main effect of *COACH* according to the classical definition. Instead, $\alpha_1$ is a linear combination of the classically defined main effect of *COACH* and the higher-order effects that involve *COACH*, in this case two SOEs and the TOE. Thus, a test of the null hypothesis that $\alpha_1$ equals zero is not a test of the null hypothesis that the main effect of *COACH* equals zero.

This raises the question of how $\alpha_1$ should be interpreted. $\alpha_1$ is the effect of *COACH* when all the other factors in the model are set to zero, also known as a simple effect. An explanation of this is provided in Appendix 1.

SOEs

Using the same approach as above, the *COACH* × *TEXT* interaction is calculated as follows: First, notice that the means conditional on the first two factors can be written as

$$\begin{aligned}
\overline{\mu}_{(++\cdot)} &= \frac{1}{2}\left[2\alpha_0 + 2\alpha_1 + 2\alpha_2 + \alpha_3 + 2\alpha_{12} + \alpha_{13} + \alpha_{23} + \alpha_{123}\right], \\
\overline{\mu}_{(-+\cdot)} &= \frac{1}{2}\left[2\alpha_0 + 2\alpha_2 + \alpha_3 + \alpha_{23}\right], \\
\overline{\mu}_{(+-\cdot)} &= \frac{1}{2}\left[2\alpha_0 + 2\alpha_1 + \alpha_3 + \alpha_{13}\right],
\end{aligned}$$

and

$$\overline{\mu}_{(--\cdot)} = \frac{1}{2}\left[2\alpha_0 + \alpha_3\right].$$

Substituting these into the definition of the interaction term, we have

$$\begin{aligned} INT_{COACH\times TEXT} &= \frac{1}{2}\left[\frac{1}{2}\left[(2\alpha_1 + 2\alpha_{12} + \alpha_{13} + \alpha_{123}) - (2\alpha_1 + \alpha_{13})\right]\right] \\ &= \frac{1}{2}\alpha_{12} + \frac{1}{4}\alpha_{123}. \end{aligned}$$

That is, the second-order coefficient $\alpha_{12}$ equals

$$\alpha_{12} = SOE_{COACH\times TEXT} = 2INT_{COACH\times TEXT} - \frac{1}{2}\alpha_{123}$$

and hence does not represent the interaction itself, but a combination of the two-way and three-way interactions.

Thus the test of the null hypothesis that $\alpha_{12}$ equals zero is not a test of the null hypothesis that the *COACH* × *TEXT* interaction equals 0 (Chakraborty, Collins, Strecher, & Murphy, 2009). Instead, $\alpha_{12}$ is a linear combination of the classically defined *COACH* × *TEXT* interaction and the TOE. Appendix 1 shows that the hypothesis test should be interpreted as a test of the null hypothesis that the *COACH* × *TEXT* interaction equals 0 when *PCP* is set to zero. The same holds for the other SOEs in the model. That is, interpreting the *COACH* × *PCP* (e.g., $\alpha_{13}$) coefficient as an interaction requires setting *TEXT* to zero, and interpreting the *TEXT* × *PCP* (e.g., $\alpha_{23}$) coefficient as an interaction requires setting *COACH* to zero.

TOEs

Using the same approach, the following is the *COACH* × *TEXT* × *PCP* interaction effect:

$$\begin{aligned} &INT_{COACH\times TEXT\times PCP} \\ &= \frac{1}{4}\left[\left[(\alpha_1 + \alpha_{12} + \alpha_{13} + \alpha_{123}) - (\alpha_1 + \alpha_{13})\right] - \left[(\alpha_1 + \alpha_{12}) - (\alpha_1)\right]\right] \\ &= \frac{1}{4}\alpha_{123}. \end{aligned}$$

Therefore,

$$\alpha_{123} = TOE_{COACH\times TEXT\times PCP} = 4\, INT_{COACH\times TEXT\times PCP}.$$

This TOE is the only effect in this model that corresponds to the classically defined effect, in this case a three-way interaction. *In general, the test of significance for the highest-order effect is equivalent in dummy coding and effect coding.*

### 2.3.6 Summary: Dummy Coding Effects

When dummy coding is used and product terms are included in the model, the regression coefficients no longer estimate effects corresponding to the classically defined main effects or interactions; rather, they estimate simple effects—the effect of a factor when the other factors are set to zero. The one exception is the highest-order effect. Although the magnitude of the highest-order effect estimates is typically different, the hypothesis tests are identical. Table 6 provides a summary of the effects with effect coding and dummy coding.

## *2.4 A Numerical Example*

### 2.4.1 Experimental Design

To illustrate the difference between effect coding and dummy coding, we present a hypothetical numerical example of the aforementioned factorial experiment and compare analyses of the data using effect coding and dummy coding.

For this experiment, 32 adults are randomly assigned to one of the eight experimental conditions. Below are the results of the hypothetical experiment, presented first for effect-coded independent variables and then for dummy-coded independent variables. We assume balanced data (i.e., there are an equal number of subjects in each condition at the end of our study). Please note that the numerical example is based on artificial data.

### 2.4.2 Effect Coding Results and Interpretation

Table 7 shows the output for an effect-coded multiple regression model that includes the three main effects, three two-way interaction effects, and one three-way interaction effect using PROC REG procedure in SAS/STAT software (SAS version 9.4, SAS Institute Inc.).

For researchers interested in making decisions about which levels of a factor to include in an optimized intervention, focus is drawn to the parameter estimates corresponding to the main effects (for a more detailed discussion of decision-making, see Chap. 7 in the companion volume). Beginning with the regression coefficient for $X_{COACH}$, the parameter estimate of 6.52 is significant ($p < 0.01$), meaning that, on average, the receipt of *COACH* increases adherence by 13.04 units (6.52*2) compared to not getting *COACH*, collapsing over levels of *TEXT* and *PCP*. The parameter estimates for the *TEXT* and *PCP* regression coefficients are not significant, $p = 0.53$ and $p = 0.27$, respectively, meaning that there is not enough evidence to suggest that these factors have an effect on mean adherence in these artificial data.

**Table 6** Summary of effects with effect coding and dummy coding

| Effect coding | | | Dummy coding | | |
|---|---|---|---|---|---|
| Regression coefficient | Regression coefficient expressed in terms of experimental condition means | Definition of regression coefficient | Regression coefficient | Regression coefficient expressed in terms of experimental condition means | Definition of regression coefficient |
| $\beta_0$ | $\overline{\mu}_{\cdots}$ | Grand mean | $\alpha_0$ | $\mu_{---}$ | Mean response when all factor levels are 0 |
| $\beta_1$ | $\frac{1}{2}\left[\overline{\mu}_{(+\cdot\cdot)} - \overline{\mu}_{(-\cdot\cdot)}\right]$ | ½ the main effect of *COACH* | $\alpha_1$ | $\mu_{(+--)} - \mu_{(---)}$ | The effect of *COACH*, when *TEXT* and *PCP* are 0 |
| $\beta_2$ | $\frac{1}{2}\left[\overline{\mu}_{(\cdot+\cdot)} - \overline{\mu}_{(\cdot-\cdot)}\right]$ | ½ the main effect of *TEXT* | $\alpha_2$ | $\mu_{(-+-)} - \mu_{(---)}$ | The effect of *TEXT*, when *COACH* and *PCP* are 0 |
| $\beta_3$ | $\frac{1}{2}\left[\overline{\mu}_{(\cdot\cdot+)} - \overline{\mu}_{(\cdot\cdot-)}\right]$ | ½ the main effect of *PCP* | $\alpha_3$ | $\mu_{(--+)} - \mu_{(---)}$ | The effect of *PCP*, when *COACH* and *TEXT* are 0 |
| $\beta_{12}$ | $\frac{1}{4}\left[\left(\overline{\mu}_{(++\cdot)} - \overline{\mu}_{(-+\cdot)}\right) - \left(\overline{\mu}_{(+-\cdot)} - \overline{\mu}_{(--\cdot)}\right)\right]$ | ½ the interaction effect of *COACH* × *TEXT* | $\alpha_{12}$ | $(\mu_{(++-)} - \mu_{(-+-)}) - (\mu_{(+--)} - \mu_{(---)})$ | The effect of *COACH* × *TEXT* when *PCP* is 0 |
| $\beta_{13}$ | $\frac{1}{4}\left[\left(\overline{\mu}_{(+\cdot+)} - \overline{\mu}_{(-\cdot+)}\right) - \left(\overline{\mu}_{(+\cdot-)} - \overline{\mu}_{(-\cdot-)}\right)\right]$ | ½ the interaction effect of *COACH* × *PCP* | $\alpha_{13}$ | $(\mu_{(+-+)} - \mu_{(--+)}) - (\mu_{(+--)} - \mu_{(---)})$ | The effect of *COACH* × *PCP* when *TEXT* is 0 |
| $\beta_{23}$ | $\frac{1}{4}\left[\left(\overline{\mu}_{(\cdot++)} - \overline{\mu}_{(\cdot-+)}\right) - \left(\overline{\mu}_{(\cdot+-)} - \overline{\mu}_{(\cdot--)}\right)\right]$ | ½ the interaction effect of *TEXT* × *PCP* | $\alpha_{23}$ | $(\mu_{(-++)} - \mu_{(--+)}) - (\mu_{(-+-)} - \mu_{(---)})$ | The effect of *TEXT* × *PCP* when *COACH* is 0 |
| $\beta_{123}$ | $\frac{1}{8}\left[\left[\left(\mu_{(+++)} - \mu_{(-++)}\right) - \left(\mu_{(+-+)} - \mu_{(--+)}\right)\right] - \left[\left(\mu_{(++-)} - \mu_{(-+-)}\right) - \left(\mu_{(+--)} - \mu_{(---)}\right)\right]\right]$ | ½ the interaction effect of *COACH* × *TEXT* × *PCP* | $\alpha_{123}$ | $\left[\left(\mu_{(+++)} - \mu_{(-++)}\right) - \left(\mu_{(+-+)} - \mu_{(--+)}\right)\right] - \left[\left(\mu_{(++-)} - \mu_{(-+-)}\right) - \left(\mu_{(+--)} - \mu_{(---)}\right)\right]$ | 4 × the interaction effect of *COACH* × *TEXT* × *PCP* |

**Table 7** Annotated regression output for effect-coded variables

The REG procedure

Model: MODEL1
Dependent variable: adherence

Number of observations read 32
Number of observations used 32

Analysis of variance

| Source | DF | Sum of squares | Mean square | F value | Pr > F |
|---|---|---|---|---|---|
| Model | 7 | 1686.53173 | 240.93310 | 217.32 | <.0001 |
| Error | 24 | 26.60724 | 1.10863 | | |
| Corrected total | 31 | 1713.13897 | | | |

| | |
|---|---|
| Root MSE | 1.05292 |
| R-square | 0.9845 |
| Dependent mean | −0.17256 |
| Adj R-Sq | 0.9799 |
| Coeff Var | −610.15872 |

| Variable | DF | Parameter estimate | Standard error | t value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | −0.17256 | 0.18613 | −0.93 | 0.3631 |
| $X_{COACH}$ | 1 | 6.51605 | 0.18613 | 35.01 | <.0001 |
| $X_{TEXT}$ | 1 | 0.11797 | 0.18613 | 0.63 | 0.5322 |
| $X_{PCP}$ | 1 | −0.21210 | 0.18613 | −1.14 | 0.2657 |
| $X_{COACH \times TEXT}$ | 1 | 3.13873 | 0.18613 | 16.86 | <.0001 |
| $X_{COACH \times PCP}$ | 1 | 0.40815 | 0.18613 | 2.19 | 0.0383 |
| $X_{TEXT \times PCP}$ | 1 | −0.31576 | 0.18613 | −1.70 | 0.1027 |
| $X_{COACH \times TEXT \times PCP}$ | 1 | 0.26153 | 0.18613 | 1.41 | 0.1728 |

There are two significant two-way interactions: the *COACH* × *TEXT* interaction ($p < 0.01$) and *COACH* × *PCP* interaction ($p = 0.04$). We can interpret the regression coefficient for the *COACH* × *TEXT* interaction as follows: The expected difference between the effect of *COACH* on adherence when *TEXT* is used and the effect of *COACH* on adherence when *TEXT* is not used, collapsing over levels of *PCP*, is 6.28 units (3.14*2). In other words, the effect of *COACH* on mean adherence increases by 6.28 units when *TEXT* is used compared to when *TEXT* is not used, collapsing over levels of *PCP*.

An interesting feature about the results of a multiple regression analyses when effect coding is used is that all of the standard errors are the same and the effect estimates are uncorrelated. This is demonstrated in Table 8; all of the variances are 0.03, and all of the covariances are 0.

**Table 8** Covariance matrix of effect-coded parameter estimates

| **Variable** | Intercept | $X_{COACH}$ | $X_{TEXT}$ | $X_{PCP}$ | $X_{COACH \times TEXT}$ | $X_{COACH \times PCP}$ | $X_{TEXT \times PCP}$ | $X_{COACH \times TEXT \times PCP}$ |
|---|---|---|---|---|---|---|---|---|
| Intercept | **0.03** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| $X_{COACH}$ | 0.00 | **0.03** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| $X_{TEXT}$ | 0.00 | 0.00 | **0.03** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| $X_{PCP}$ | 0.00 | 0.00 | 0.00 | **0.03** | 0.00 | 0.00 | 0.00 | 0.00 |
| $X_{COACH \times TEXT}$ | 0.00 | 0.00 | 0.00 | 0.00 | **0.03** | 0.00 | 0.00 | 0.00 |
| $X_{COACH \times PCP}$ | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.03** | 0.00 | 0.00 |
| $X_{TEXT \times PCP}$ | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.03** | 0.00 |
| $X_{COACH \times TEXT \times PCP}$ | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.03** |

**Table 9** Annotated regression output for dummy-coded variables

The REG procedure

Model: MODEL1
Dependent variable: adherence

Number of observations read 32

Number of observations used 32

Analysis of variance

| Source | DF | Sum of squares | Mean square | F value | Pr > F |
|---|---|---|---|---|---|
| Model | 7 | 1686.53173 | 240.93310 | 217.32 | <.0001 |
| Error | 24 | 26.60724 | 1.10863 | | |
| Corrected total | 31 | 1713.13897 | | | |

| | |
|---|---|
| Root MSE | 1.05292 |
| R-square | 0.9845 |
| Dependent mean | −0.17256 |
| Adj R-Sq | 0.9799 |
| Coeff Var | −610.15872 |

| Variable | DF | Parameter estimate | Standard error | t value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | −3.62490 | 0.52646 | −6.89 | <.0001 |
| $Z_{COACH}$ | 1 | 6.46140 | 0.74452 | 8.68 | <.0001 |
| $Z_{TEXT}$ | 1 | −4.88693 | 0.74452 | −6.56 | <.0001 |
| $Z_{PCP}$ | 1 | −0.08592 | 0.74452 | −0.12 | 0.9091 |
| $Z_{COACH \times TEXT}$ | 1 | 11.50880 | 1.05292 | 10.93 | <.0001 |
| $Z_{COACH \times PCP}$ | 1 | 0.58648 | 1.05292 | 0.56 | 0.5827 |
| $Z_{TEXT \times PCP}$ | 1 | −2.30917 | 1.05292 | −2.19 | 0.0382 |
| $Z_{COACH \times TEXT \times PCP}$ | 1 | 2.09224 | 1.48905 | 1.41 | 0.1728 |

### 2.4.3 Dummy Coding Results and Interpretation

Table 9 shows the output for a multiple regression model with the same data as in Table 7 but this time using dummy coding, so that three FOEs, three SOEs, and one TOE are estimated.

The overall $F$ statistic is exactly the same ($F = 217.32$, $p < 0.01$) as that obtained in the effect-coded model shown in Table 7. In addition, the test for significance for the highest-order effect, $Z_{COACH \times TEXT \times PCP}$, is the same ($p = 0.17$). However, this is where the similarities end. All of the other regression weights, standard errors, and hypothesis tests are different and must be interpreted differently. For example, the regression coefficient for $Z_{COACH}$ is significant ($t = 8.68$, $p < 0.01$). The dummy-coded analysis indicates that, on average, *COACH* increases adherence by 6.46 units compared to no *COACH*, when *TEXT* and *PCP* are set to zero. Stated another way,

the $t$ test for this regression coefficient tests the null hypothesis that the effect of *COACH* on mean adherence among those who did not receive *TEXT* or *PCP* is zero. This is very different from the classical definition of a main effect. Note that the regression coefficient $\alpha_{TEXT}$ is statistically significant, but its counterpart in effect coding, $\beta_{TEXT}$, is not significant. This illustrates how the results of hypothesis tests of individual effects can differ between effect and dummy coding. There are also noticeable differences in the variance-covariance matrix, as shown in Table 10. First, the variances of the dummy-coded effects vary depending on the order of the effect. Second, the covariances are non-zero and substantial, and the covariance matrix displays a complicated pattern of positive and negative signs.

## *2.5 Regression Software Packages*

### 2.5.1 SAS (v 9.4)

The analyses reported in the numerical example above were obtained using SAS PROC REG. We recommend this procedure when using SAS because in PROC REG the user is responsible for coding the independent variables. We find with this approach the interpretation of effects is more straightforward, because it is always clear what coding scheme has been used.

Another procedure, PROC GLM, can be used to analyze data from a factorial experiment. In the GLM procedure, the user has the option to use or not use the CLASS statement. The CLASS statement can be used to indicate that one or more variables represent discrete levels or categories, such as the levels of experimental factors. If the CLASS statement is not used, the independent variables are considered quantitative (as they are in PROC REG), that is, they are handled the same way as they are in PROC REG, so the way in which they are coded by the user matters. In other words, different results will be obtained depending on whether effect coding or dummy coding is used. By contrast, if the CLASS statement is used, which seems like an intuitively appealing approach when analyzing data from a factorial experiment, the output can be confusing. The ANOVA table for the Type I sums of squares (described below) corresponds to the effects using effect coding; however, for the regression parameter estimates, *the SAS default is to report results for dummy-coded variables.* Thus if the CLASS statement is used, the ANOVA table, which is based on effect coding, reports different hypothesis tests than those in the regression section of the output.

Particular ANOVA-related tests reflect different partitions of the sums of squares associated with particular sources of variability. In this context, these correspond to tests of coefficients in different coding schemes. In PROC REG, Type I (hierarchical) and Type II (partial) can be requested as additional output. In PROC GLM, Type I (hierarchical) and Type III (partial) are the default with the request of a solution. Table 11 explains how the coding scheme and use of a CLASS statement can change the meaning of the effects for the PROC GLM procedure. In general,

**Table 10** Covariance matrix of dummy-coded effects

| **Variable** | Intercept | $Z_{COACH}$ | $Z_{TEXT}$ | $Z_{PCP}$ | $Z_{COACH \times TEXT}$ | $Z_{COACH \times PCP}$ | $Z_{TEXT \times PCP}$ | $Z_{COACH \times TEXT \times PCP}$ |
|---|---|---|---|---|---|---|---|---|
| Intercept | **0.28** | −0.28 | −0.28 | −0.28 | 0.28 | 0.28 | 0.28 | −0.28 |
| $Z_{COACH}$ | −0.28 | **0.55** | 0.28 | 0.28 | −0.55 | −0.55 | −0.28 | 0.55 |
| $Z_{TEXT}$ | −0.28 | 0.28 | **0.55** | 0.28 | −0.55 | −0.28 | −0.55 | 0.55 |
| $Z_{PCP}$ | −0.28 | 0.28 | 0.28 | **0.55** | −0.28 | −0.55 | −0.55 | 0.55 |
| $Z_{COACH \times TEXT}$ | 0.28 | −0.55 | −0.55 | −0.28 | **1.11** | 0.55 | 0.55 | −1.11 |
| $Z_{COACH \times PCP}$ | 0.28 | −0.55 | −0.28 | −0.55 | 0.55 | **1.11** | 0.55 | −1.11 |
| $Z_{TEXT \times PCP}$ | 0.28 | −0.28 | −0.55 | −0.55 | 0.55 | 0.55 | **1.11** | −1.11 |
| $Z_{COACH \times TEXT \times PCP}$ | −0.28 | 0.55 | 0.55 | 0.55 | −1.11 | −1.11 | −1.11 | **2.22** |

**Table 11** Interpretations of results from different approaches to conducting ANOVA using PROC GLM

| Coding scheme | Interpretation of effect estimates | |
|---|---|---|
| Effect coding | Without CLASS statement | With CLASS statement |
| Type I sums of squares | Classical[a] | Classical |
| Type III sums of squares | Classical | Classical |
| *b* weights and hypothesis tests | Classical | Dummy |
| Dummy coding | | |
| Type I sums of squares | Classical | Classical |
| Type III sums of squares | Dummy[b] | Classical |
| *b* weights and hypothesis tests | Dummy | Dummy |

[a]Interpreted as classical effects
[b]Interpreted as first-order, second-order, third-order, etc., effects as produced by dummy-coded independent variables

if PROC GLM is used, we recommend *not* using the CLASS statement and coding the independent variables yourself to avoid confusion.

### 2.5.2 SPSS (PASW Statistics 23)

As in SAS, there are two options available to SPSS users to analyze factorial experiments. Users can invoke either the "LINEAR" procedure under the "Regression" pull-down tab or the "UNIVARIATE" procedure under the "General Linear Model" pull-down tab. Similar to SAS PROC REG, the LINEAR procedure assumes the independent variables are quantitative; therefore, how the variables are coded determines whether or not the parameter estimates and corresponding tests reflect the classical definitions. Within the UNIVARIATE procedure, the user can enter the variables as covariates or factors. If the variables are entered as covariates, then the way in which the variables are coded matters; it is only with effect coding that the results will be consistent with the classical definitions of effects (see Table 10). However, if the variables are entered as factors, then SPSS will dummy code the variables. The partitioning of the sums of squares (SS) is also included in the output for the UNIVARIATE procedure. Type III SS (partial) is the default, although Type I and Type II can be requested. Table 12 demonstrates how the coding scheme and the choice of how the variables are characterized changes the meaning of the effects. To avoid confusion, we recommend that if the UNIVARIATE procedure is used, the independent variables should be entered as covariates, not as factors.

**Table 12** Interpretations of results from different approaches to conducting ANOVA using SPSS UNIVARIATE procedure

| Coding scheme | Interpretation of effect estimates | |
|---|---|---|
| Effect coding | Variables as covariates | Variables as factors |
| Type III sums of squares | Classical[a] | Classical |
| *b* weights and hypothesis tests | Classical | Dummy |
| Dummy coding | | |
| Type III sums of squares | Dummy[b] | Classical |
| *b* weights and hypothesis tests | Dummy | Dummy |

[a]Interpreted as classical effects
[b]Interpreted as first-order, second-order, third-order, etc., effects as produced by dummy-coded independent variables

## *2.6 Discussion*

In this chapter we have presented a hypothetical example of a $2^3$ factorial experiment to demonstrate that how a categorical variable is coded determines the interpretation of its effects in a regression analysis. We showed that different coding schemes for categorical variables estimate different effects and, therefore, can result in different results and different conclusions. If dummy coding is used and product terms are included in an analysis, the significance tests for the regression coefficients do not readily correspond to the classical ANOVA tests of main effects and interaction effects. Rather, the effects in a regression with dummy-coded variables should be interpreted as the effect of the variable, when all other variables in the analysis are set to zero (i.e., a simple effect); the only exception is the highest-order effect. This is contrasted to effect coding, where if product terms are modeled in an analysis, significance tests of the regression coefficients *do* correspond to the classical ANOVA tests of main effects and interaction effects. Effect-coded coefficients actually represent exactly half the corresponding ANOVA effect, but this cancels out when doing a significance test because they also have exactly half the standard error. We also discussed two of the many statistical software options available to researchers to analyze data from factorial experiments using multiple regression. Researchers should be aware of the default options and how use of these defaults may affect their results.

Researchers trained in using dummy coding may be surprised at the emphasis given in this chapter to the differences between dummy coding and effect coding and the advantages of the latter when testing interactions. Indeed, readers may have experience in testing interactions with dummy-coded covariates, and we are not suggesting that their results were wrong. In many cases outside an optimization trial, the main focus of a regression model that includes interactions is on testing the highest-order interaction. As was shown above, the test of the highest-order interaction is equivalent regardless of whether factors are dummy-coded or effect-coded. However, in the context of an optimization trial, a researcher is interested in using information from multiple main effects and interactions considered together

to make decisions; thus, it is important to be more mindful of the coding of factors and the interpretation of their coefficients.

In the beginning we noted that this chapter is based on experiments with balanced data. This was done to avoid making the discussion overly complex; however, the general principles discussed here hold even when the number of subjects is not exactly the same across experimental conditions. In addition, the hypothetical example did not include other covariates in the regression analyses. Similar to analyses of other experimental data, the benefits of adding covariates to the regression models (e.g., reduction in variance) are the same for analyses of data from a factorial experiment. This extends to including all the variables corresponding to the effects from the factorial model; in models where all the variables of the effects are included, the standard errors are smaller than models where only the variables corresponding to main effects are included, and this changes the interpretation of the hypothesis tests.

This chapter has only considered the coding of effects in linear models, but factorial experiments are also performed in settings with binary (e.g., Cook et al., 2007; Ledolter & Swersey, 2006), count, or even survival time (e.g., Day et al., 2002; Wolbers et al., 2011) outcomes. The implications of the coding of factors for these outcomes are beyond the scope of this chapter. We briefly consider the coding of factors with binary outcomes in order to illustrate the issues involved. In a two-factor binary logistic regression with an interaction, one could write

$$\text{log odds } (Y = 1) = \beta_0 + \beta_1 X_A + \beta_2 X_B + \beta_{12} X_{A\times \mathrm{B}}.$$

If there is no interaction ($\beta_{12} = 0$), then one can test the effects of factors $A$ and $B$ by testing the significance of $\beta_1$ and $\beta_2$, respectively, regardless of whether the $X$ factors are dummy-coded or effect-coded. However, if there is an interaction, the interpretation of the main effect becomes complicated, again regardless of whether the $X$ factors are dummy-coded or effect-coded.

Thus, although one could correctly write

$$\text{log odds } (Y = 1 \mid X_B = +1) = \beta_0 + \beta_1 X_A + \beta_2 + \beta_{12} X_A$$

and could also correctly write

$$\text{log odds } (Y = 1 \mid X_B = -1) = \beta_0 + \beta_1 X_A - \beta_2 - \beta_{12} X_A,$$

it is not possible to average these together and conclude that

$$\text{log odds } (Y = 1) = \beta_0 + \beta_1 X_A,$$

in the same way that one could previously collapse across rows or columns in a balanced linear model to get an overall mean for each column or row. This is because the average of two log-odd estimates does not give the overall log odds in the way that an average of two means gives an overall mean; the log-odd function is not a

linear function like the mean. This is not pointed out in order to discourage the use of non-normally distributed outcome variables, but only to say that some additional care may be required. Cox and Snell (1988) and Collet (1991) describe the analysis of binary data, including factorial experiments with binary outcomes.

This chapter has also not considered the issue of *unbalanced* effect coding (e.g., te Grotenhuis et al., 2016). This is a modified form of effect coding that is mostly applicable to observed rather than randomized factors, and is therefore not likely to be very helpful in the context of a randomized factorial experiment.

In conclusion, the goals of this chapter were to highlight how the coding scheme of categorical variables determines the interpretation of its effects in a regression analysis. This becomes apparent in the analysis of data generated from a factorial experiment as part of the optimization phase of MOST. While either coding scheme can be used, the different coding schemes estimate different effects, particularly when product terms are included in the model; it is important for the researcher to be clear of what question they are interested in answering to guide the selection of a coding scheme.

**Acknowledgments** The preparation of this report was supported by National Institute on Drug Abuse grants P50 DA039838 and P50 DA010075. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute, the National Institute on Drug Abuse, or the National Institutes of Health.

## Appendices

### *Appendix 1: Basis for Interpretation of Regression Coefficients When Dummy Coding Is Used*

The equations below show the correspondence between the population means and the regression coefficients for dummy-coded variables.

1. Interpretation of $\alpha_1$
   Notice that $\mu_{(---)} = \alpha_0$ is the mean response when factor $A =$ no, $B =$ no, and $C =$ no.
   Also, $\mu_{(+--)} = \alpha_0 + \alpha_1$ is the mean response when $A =$ yes, $B =$ no, and $C =$ no.
   Therefore, $\alpha_1 = \mu_{(+--)} - \mu_{(---)}$, that is, the effect of $A =$ yes vs. $A =$ no, when $B =$ no and $C =$ no.
   Note the similarities and differences between this and the classical definition of a main effect of $A$: $\overline{\mu}_{(+\cdot\cdot)} - \overline{\mu}_{(-\cdot\cdot)}$.

2. Interpretation of $\alpha_{12}$
   $\mu_{(---)} = \alpha_0$ is the mean response when $A =$ no, $B =$ no, and $C =$ no.
   $\mu_{(+--)} = \alpha_0 + \alpha_1$ is the mean response when $A =$ yes, $B =$ no, and $C =$ no.
   $\mu_{(-+-)} = \alpha_0 + \alpha_2$ is the meanresponse when $A =$ no, $B =$ yes, and $C =$ no.

$\mu_{(+--)} = \alpha_0 + \alpha_1 + \alpha_2 + \alpha_{12}$ is the mean response when $A =$ yes, $B =$ yes, and $C =$ no.

Therefore, $\alpha_{12} = \mu_{(++-)} - \mu_{(+--)} - \mu_{(-+-)} + \mu_{(---)} = (\mu_{(++-)} - \mu_{(-+-)}) - (\mu_{(+--)} - \mu_{(---)})$, that is, the difference in the effect of $A =$ yes vs. $A =$ no when $B =$ yes, $C =$ no and $B =$ no, $C =$ no.

Note the similarities and differences between this and the classical definition of a two-way interaction effect of $A$ and $B$:

$$A \times B = \frac{1}{2}\left[\left(\overline{\mu}_{(++\cdot)} - \overline{\mu}_{(-+\cdot)}\right) - \left(\overline{\mu}_{(+-\cdot)} - \overline{\mu}_{(--\cdot)}\right)\right].$$

3. Interpretation of $\alpha_{123}$

The eight condition means are as follows:

$\mu_{(---)} = \alpha_0$: Mean response when $A =$ no, $B =$ no, and $C =$ no.

$\mu_{(+--)} = \alpha_0 + \alpha_1$: Mean response when $A =$ yes, $B =$ no, and $C =$ no.

$\mu_{(-+-)} = \alpha_0 + \alpha_2$: Mean response when $A =$ no, $B =$ yes, and $C =$ no.

$\mu_{(++-)} = \alpha_0 + \alpha_1 + \alpha_2 + \alpha_{12}$: Mean response when $A =$ yes, $B =$ yes, and $C =$ no.

$\mu_{(--+)} = \alpha_0 + \alpha_3$: Mean response when $A =$ no, $B =$ no, and $C =$ yes.

$\mu_{(+-+)} = \alpha_0 + \alpha_1 + \alpha_3 + \alpha_{13}$: Mean response when $A =$ yes, $B =$ no, and $C =$ yes.

$\mu_{(-++)} = \alpha_0 + \alpha_2 + \alpha_3 + \alpha_{23}$: Mean response when $A =$ no, $B =$ yes, and $C =$ yes.

$\mu_{(+++)} = \alpha_0 + \alpha_1 + \alpha_2 + \alpha_3 + \alpha_{12} + \alpha_{13} + \alpha_{23} + \alpha_{123}$: Mean response when $A =$ yes, $B =$ yes, and $C =$ yes.

Therefore,

$$\begin{aligned}\alpha_{123} &= \mu_{(+++)} - \mu_{(-++)} - \mu_{(+-+)} - \mu_{(++-)} + \mu_{(+--)} \\ &\quad + \mu_{(-+-)} + \mu_{(--+)} - \mu_{(---)} \\ &= \left[\left(\mu_{(+++)} - \mu_{(-++)}\right) - \left(\mu_{(+-+)} - \mu_{(--+)}\right)\right] \\ &\quad - \left[\left(\mu_{(++-)} - \mu_{(-+-)}\right) - \left(\mu_{(+--)} - \mu_{(---)}\right)\right].\end{aligned}$$

That is, the difference in the interaction effect of $A$ and $B$ when $C =$ yes and $C =$ no.

Note that this is essentially the same (except for rescaling) as three-way interaction effect of $A$, $B$, and $C$:

$$\begin{aligned}A \times B \times C = \frac{1}{4}\Big[&\left[\left(\mu_{(+++)} - \mu_{(-++)}\right) - \left(\mu_{(+-+)} - \mu_{(--+)}\right)\right] \\ &- \left[\left(\mu_{(++-)} - \mu_{(-+-)}\right) - \left(\mu_{(+--)} - \mu_{(---)}\right)\right]\Big].\end{aligned}$$

4. Another way to look at $\alpha_1$ is as follows: As shown on Page 14,

$$\alpha_1 = ME_A - \left(\frac{1}{2}\alpha_{12} + \frac{1}{2}\alpha_{13} + \frac{1}{4}\alpha_{123}\right) \ (*)$$

$$ME_A = \overline{\mu}_{(+\cdot\cdot)} - \overline{\mu}_{(-\cdot\cdot)}$$

$$\alpha_1 = \mu_{(+--)} - \mu_{(---)} \ (**)$$

$$\begin{aligned}\alpha_{12} &= \mu_{(++-)} - \mu_{(-+-)} - \mu_{(+--)} + \mu_{(---)} \\ &= \left(\mu_{(++-)} - \mu_{(-+-)}\right) - \left(\mu_{(+--)} - \mu_{(---)}\right)\end{aligned}$$

$$\begin{aligned}\alpha_{12} &= \mu_{(+-+)} - \mu_{(--+)} - \mu_{(+--)} + \mu_{(---)} \\ &= \left(\mu_{(+-+)} - \mu_{(--+)}\right) - \left(\mu_{(+--)} - \mu_{(---)}\right)\end{aligned}$$

$$\begin{aligned}\alpha_{123} &= \mu_{(+++)} - \mu_{(-++)} - \mu_{(+-+)} - \mu_{(++-)} + \mu_{(+--)} + \mu_{(-+-)} \\ &+ \mu_{(--+)} - \mu_{(---)} \\ &= \left[\left(\mu_{(+++)} - \mu_{(-++)}\right) - \left(\mu_{(+-+)} - \mu_{(--+)}\right)\right] \\ &- \left[\left(\mu_{(++-)} - \mu_{(-+-)}\right) - \left(\mu_{(+--)} - \mu_{(---)}\right)\right]\end{aligned}$$

So, $(_*)$ becomes

$$\begin{aligned}\alpha_1 &= \tfrac{1}{4}\left\{\left[\mu_{(+++)} + \mu_{(++-)} + \mu_{(+-+)} + \mu_{(+--)}\right]\right. \\ &\quad \left. - \left[\mu_{(-++)} + \mu_{(-+-)} + \mu_{(--+)} + \mu_{(---)}\right]\right\} \\ &\quad - \tfrac{1}{2}\left\{\left[\mu_{(++-)} - \mu_{(-+-)}\right] - \left[\mu_{(+--)} - \mu_{(---)}\right]\right\} \\ &\quad - \tfrac{1}{2}\left\{\left[\mu_{(+-+)} - \mu_{(--+)}\right] - \left[\mu_{(+--)} - \mu_{(---)}\right]\right\} \\ &\quad - \tfrac{1}{4}\left\{\left[\left(\mu_{(+++)} - \mu_{(-++)}\right) - \left(\mu_{(+-+)} - \mu_{(--+)}\right)\right]\right. \\ &\quad \left. - \left[\left(\mu_{(++-)} - \mu_{(-+-)}\right) - \left(\mu_{(+--)} - \mu_{(---)}\right)\right]\right\}\end{aligned}$$

$$\begin{aligned}&= \tfrac{1}{4}\mu_{(+++)} + \tfrac{1}{4}\mu_{(++-)} + \tfrac{1}{4}\mu_{(+-+)} + \tfrac{1}{4}\mu_{(+--)} - \tfrac{1}{4}\mu_{(-++)} - \tfrac{1}{4}\mu_{(-+-)} \\ &- \tfrac{1}{4}\mu_{(--+)} - \tfrac{1}{4}\mu_{(---)} - \tfrac{1}{2}\mu_{(++-)} + \tfrac{1}{2}\mu_{(-+-)} + \tfrac{1}{2}\mu_{(+--)} - \tfrac{1}{2}\mu_{(---)} \\ &- \tfrac{1}{2}\mu_{(+-+)} + \tfrac{1}{2}\mu_{(--+)} + \tfrac{1}{2}\mu_{(+--)} - \tfrac{1}{2}\mu_{(---)} - \tfrac{1}{4}\mu_{(+++)} + \tfrac{1}{4}\mu_{(-++)} \\ &+ \tfrac{1}{4}\mu_{(+-+)} - \tfrac{1}{4}\mu_{(--+)} + \tfrac{1}{4}\mu_{(++-)} - \tfrac{1}{4}\mu_{(-+-)} - \tfrac{1}{4}\mu_{(+--)} + \tfrac{1}{4}\mu_{(---)}\end{aligned}$$

$$= \mu_{(+--)} - \mu_{(---)},$$

which is exactly the same equation in $(_{**})$.

## *Appendix 2: Example Dataset and Code*

```
options ls=100 ps=150 nodate formdlim='-';

data anova;
      input A$ B$ C$ Y;
      datalines;
no    no    no    -4.535654
yes   no    no    2.263351
no    yes   no    -8.841193
yes   yes   no    8.046294
no    no    yes   -3.736171
yes   no    yes   2.307247
no    yes   yes   -12.173537
yes   yes   yes   8.624994
no    no    no    -3.778953
yes   no    no    3.883212
no    yes   no    -9.232583
yes   yes   no    9.034533
no    no    yes   -3.738127
yes   no    yes   4.092454
no    yes   yes   -10.897736
yes   yes   yes   9.106033
no    no    no    -3.995389
yes   no    no    2.009244
no    yes   no    -8.046562
yes   yes   no    9.562785
no    no    yes   -3.498379
yes   no    yes   1.601328
no    yes   yes   -10.072611
yes   yes   yes   10.104139
no    no    no    -2.189597
yes   no    no    3.190188
no    yes   no    -7.926976
yes   yes   no    11.189867
no    no    yes   -3.870586
yes   no    yes   5.347232
no    yes   yes   -10.483766
yes   yes   yes   11.132855
;
run;

data dummy;
      set ANOVA;
      dummyA=(A="yes");
      dummyB=(B="yes");
      dummyC=(C="yes");
      dummyAB=dummyA*dummyB;
      dummyAC=dummyA*dummyC;
      dummyBC=dummyB*dummyC;
      dummyABC=dummyA*dummyB*dummyC;
run;

data effect;
      set ANOVA;
      if A="yes" then effectA=1;
            else effectA=-1;
      if B="yes" then effectB=1;
            else effectB=-1;
      if C="yes" then effectC=1;
            else effectC=-1;
      effectAB=effectA*effectB;
      effectAC=effectA*effectC;
run;

proc glm data=anova;
      class a b c;
      model y = a|b|c / solution;
      lsmeans A B C A*B A*C B*C A*B*C;
run;

proc reg data=dummy;
      model Y=dummyA dummyB dummyAB dummyC dummyAC dummyBC dummyABC;
run;

proc reg data=effect;
      model Y=effectA effectB effectAB effectC effectAC effectBC effectABC;
run;

proc glm data=dummy;
      class dummyA dummyB dummyC dummyAB dummyAC dummyBC dummyABC;
      model y = dummyA dummyB dummyC dummyAB dummyAC dummyBC dummyABC / solution;
      lsmeans dummyA dummyB dummyC dummyAB dummyAC dummyBC dummyABC;
run;
proc glm data=effect;
      class effectA effectB effectC effectAB effectAC effectBC effectABC;
      model y = effectA effectB effectC effectAB effectAC effectBC effectABC/ solution;
      lsmeans effectA effectB effectC effectAB effectAC effectBC effectABC;
run;

quit;
```

## References

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

Collet, D. (1991). *Modelling binary data*. London, UK: Chapman & Hall.

Collins, L. M. (2018). *Optimization of behavioral, biobehavioral, and biomedical interventions: The multiphase optimization strategy (MOST)*. New York, NY: Springer.

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

Collins, L. M., Kugler, K. C., & Gwadz, M. V. (2016). Optimization of multicomponent behavioral and biobehavioral interventions for the prevention and treatment of HIV/AIDS. *AIDS and Behavior, 20*(1), 197–214.

Cook, N. R., Albert, C. M., Gaziano, J. M., Zaharris, E., MacFadyen, J., Danielson, E., … Manson, J. E. (2007). A randomized factorial trial of vitamins C and E and beta carotene in the secondary prevention of cardiovascular events in women: Results from the Women's Antioxidant Cardiovascular Study. *Archives of Internal Medicine, 167*, 1610–1618.

Cox, D. R., & Snell, E. J. (1988). *Analysis of binary data* (2nd ed.). London, UK: Chapman & Hall.

Day, L., Fildes, B., Gordon, I., Fitzharris, M., Flamer, H., & Lord, S. (2002). Randomized factorial trial of falls prevention among older people living in their own homes. *British Medical Journal, 325*, 128.

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*(2), 153–175.

Hardy, M. A. (1993). *Regression with dummy variables*. Newbury Park, CA: Sage.

Ledolter, J., & Swersey, A. J. (2006). Using a fractional factorial design to increase direct mail response at Mother Jones Magazine. *Quality Engineering, 18*, 469–475.

Montgomery, D. C. (2009). *Design and analysis of experiments*. Hoboken, NJ: Wiley.

Pellegrini, C. A., Hoffman, S. A., Collins, L. M., & Spring, B. (2014). Optimization of remotely delivered intensive lifestyle treatment for obesity using the Multiphase Optimization Strategy: Opt-IN study protocol. *Contemporary Clinical Trials, 38*, 251–259.

Pellegrini, C. A., Hoffman, S. A., Collins, L. M., & Spring, B. (2015). Corrigendum to "Optimization of remotely delivered intensive lifestyle treatment for obesity using the Multiphase Optimization Strategy: Opt-IN study protocol" [Contemp. Clin. Trials 38 (2014) 251-259]. *Contemp Clin Trials, 45*(Pt B), 468-469. doi:https://doi.org/10.1016/j.cct.2015.09.001.

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods, 5*(2), 199.

te Grotenhuis, M., Pelzer, B., Schmidt-Catran, A., Nieuwenhuis, R., Konig, R., & Eisinga, R. (2016). When size matters: advantages of weighted effect coding in observational studies. *International Journal of Public Health, 62*, 163–167.

Wolbers, M., Heemskerk, D., Chau, T. T. H., Yen, N. T. B., Caws, M., Farrar, J., & Day, J. (2011). Sample size requirements for separating out the effects of combination treatments: Randomised controlled trials of combination therapy vs. standard treatment compared to factorial designs for patients with tuberculous meningitis. *Trials, 12*, 26.

# Optimizing the Cost-Effectiveness of a Multicomponent Intervention Using Data from a Factorial Experiment: Considerations, Open Questions, and Tradeoffs Among Multiple Outcomes


**John J. Dziak**

**Abstract** Cost-effectiveness—increasing the benefit obtained for a given expenditure of time or money—is an important idea in many applied research fields. It is one important quality that a researcher interested in the multiphase optimization strategy (MOST) may wish to optimize. However, further research is needed about how to best incorporate cost information into the analysis of factorial experiments typically used during the optimization phase of MOST. This chapter will review the issues involved in making cost-effectiveness judgments using the results of factorial experiments and explore some possibilities for further methodological research on how best to estimate and compare cost-effectiveness using the results of factorial experiments.

## 1 Introduction

Because of great needs and limited resources, cost-effectiveness is a very important goal in much research both in the social and behavioral sciences and in the medical and life sciences. The multiphase optimization strategy (MOST) can be used to make a multicomponent behavioral, biobehavioral, or biomedical intervention more efficient by selecting high-performing components or omitting low-performing ones. This is closely aligned with the goal of cost-effectiveness, which is to provide as much improvement on an outcome as possible per unit cost. Demonstrating and improving the cost-effectiveness of interventions may lead to increased adoption of these interventions and to great public health benefits (Guyll, Spoth, & Crowley, 2011; see also Crowley & Jones, 2017; Embry & Biglan, 2008; Van Ryzin, Roseth, Fosco, Lee, & Chen, 2016). MOST offers the possibility of analyzing

J. J. Dziak (✉)
The Methodology Center, The Pennsylvania State University, University Park, PA, USA
e-mail: Jjd264@psu.edu



L. M. Collins, K. C. Kugler (eds.), *Optimization of Behavioral, Biobehavioral, and Biomedical Interventions*, Statistics for Social and Behavioral Sciences,
https://doi.org/10.1007/978-3-319-91776-4_7

effectiveness and cost-effectiveness for individual intervention components rather than only packages as a whole, thereby potentially developing more efficient interventions (Collins, Dziak, & Li, 2009; Collins, 2018, companion volume). 3However, the question of how to directly compare effectiveness and cost has not yet been adequately considered in the MOST literature. This chapter will explore the relationship between cost-effectiveness and factorial experiments within the context of MOST and how data from an optimization or evaluation trial can be used to improve cost-effectiveness.

Consider the choice of whether to include or omit a component in an intervention. If a component were known to be iatrogenic or ineffective, then omitting it would be a straightforward decision. However, it may happen that a component shows at least some evidence of effectiveness, but is perhaps not effective enough to justify its cost of inclusion. This suggests a need to balance effectiveness and cost in an analysis. Such analyses can be described as cost-effectiveness analyses, a kind of economic decision-making method. This requires a perspective that differs somewhat from classic practice in many fields of research in the social sciences, which often focus on effectiveness as defined by statistical significance. Cost-effectiveness analysis is focused on the practical significance of the effect of an intervention or a component of an intervention, especially in relation to its cost, and it requires that both effectiveness and cost be operationally defined and measured. There is an extensive literature on cost-effectiveness analyses of interventions as a whole, as in a two-condition randomized controlled trial (RCT) comparing a new intervention to a standard one, but much less literature exists on cost-effectiveness analysis based on factorial experiments.

## *1.1 Cost-Effectiveness as an Optimization Criterion in MOST*

Researchers using MOST should specify an optimization criterion: roughly speaking, a performance measure they wish to improve in an intervention. If cost-effectiveness is the criterion of interest, then the researchers will have to do some kind of cost-effectiveness analysis, both in the optimization phase (choosing a proposed intervention) and in the evaluation phase (confirming that the proposed intervention performs satisfactorily). The data for the evaluation phase will often be based on an RCT. However, the optimization phase of MOST often involves a factorial or fractional factorial experiment (see Collins et al., 2009; Collins, Kugler, & Gwadz, 2016; Wu & Hamada, 2009, and the Collins, 2018, companion volume). The literature on making cost-effectiveness decisions from factorial experiments is in its infancy, and many questions await further explanation. Therefore, there is much room for further methodological research on how to approach such an analysis. Because of this the current chapter cannot serve as a complete instruction manual. Instead, it is intended as an introduction to the issues and questions involved, in order to stimulate further thinking and research.

This chapter will focus mainly on cost-effectiveness analysis, rather than cost-benefit analysis (see Crowley, Hill, Kuklinski, & Jones, 2014; Gift, Haddix, & Corso, 2003; Messonier & Meltzer, 2003; National Academies, 2016, for a discussion of the distinction). Cost-effectiveness analysis is focused on determining what interventions obtain the best health outcomes per unit of cost. Cost-benefit analysis goes further and assigns monetary values to different health outcomes, in order to try to decide how much money is worth spending on various interventions. Cost-benefit decisions are more difficult and controversial (Appelbaum, 2011; Carroll, 2014; Kolata, 1992) and are beyond the scope of this chapter, although they are relevant and closely related to some of the ideas herein (particularly the weighting of multiple outcomes, described later in this chapter). Of course, decisions about how much money is to be spent must ultimately still be made by someone—so focusing on cost-effectiveness rather than cost-benefit analysis does not remove the question, but rather passes it on to a decision-maker (Gift et al., 2003).

### *1.2 Outline of Chapter*

This chapter will begin by considering the basic ideas of cost-effectiveness in simple contexts (i.e., two-arm RCT) and then explore how they might be extended to more complex ones (i.e., factorial experiments). Specifically, cost-effectiveness methods will first be introduced for analysis of an RCT comparing two conditions and then for an RCT comparing more than two conditions. This will be done by first introducing the cost-effectiveness plane (a graphical aid for making decisions) and then considering ways of combining cost and effectiveness information, such as by specifying a willingness-to-pay parameter. Finally, cost-effectiveness ideas will be applied to the more complex situation of a factorial design, with special emphasis on factorial designs as they are used for component-screening experiments in the context of the optimization phase of MOST. After this, the issue of how to balance multiple dimensions of effectiveness or cost will be considered. Finally, other special issues which may merit further research are described.

## 2 Economic Evaluation in a Two-Condition Randomized Controlled Trial

As mentioned above, the simplest case of interest here is the evaluation of an RCT comparing a new intervention to an alternative choice, such as an existing standard of care. This alternative choice could be nothing at all (e.g., a waitlist control) or some placebo-like intervention, or a standard regimen of care already known to be reasonably effective. Suppose that a scientist is interested not only in determining whether the new intervention is more effective but also in whether

it is more cost-effective than the alternative choice. This scenario is summarized below and reviewed in more depth in tutorial articles by Petrou and Gray (2011) and Bensink and co-authors (2013), in a book by Drummond, Sculpher, Torrance, O'Brien, and Stoddart (2005), and in a review of best practices by the National Academies (2016).

Suppose that a new weight loss intervention is under consideration. In order to experimentally evaluate the effectiveness and cost-effectiveness of the intervention, participants will be randomized to receive either the new intervention or a control intervention. This choice of what intervention to compare with the new intervention is very important because it determines the precise question that the RCT answers and what claims can be made afterward. That is, an intervention can only be said to be more effective or more cost-effective compared to some defined alternative or "comparator," not just effective or cost-effective in the abstract. In an RCT, the comparator of interest is the treatment condition given to a control group. For the purposes of the current discussion, suppose that the comparator is an existing standard-of-care treatment, which is being used as a control group condition.

## *2.1 Operationally Defining Cost and Effectiveness*

It is necessary to specify an outcome variable of interest; the effectiveness of the intervention is essentially defined as the size of its effect on this outcome variable. The simplest choice for the outcome variable in the hypothetical weight loss example is to measure effectiveness as mean kilograms lost.[1]

### 2.1.1 Effectiveness Alone

In an RCT that ignores cost information, it is necessary only to measure the effectiveness variable in each group and to compare the group means (generally using some kind of significance test on that variable). If there is a statistically significant difference, then the treatment can be considered effective. However, in order to compare cost-effectiveness, rather than simply effectiveness, it is also necessary to measure the cost per participant of the experimental treatment and the control treatment. This includes the cost of paying the instructor or therapist, buying the equipment and medication, and so on. The cost might be a literal cost in money, but it might also be measured in something else such as time (participant time,

[1] Weight loss is actually a change score (value after minus value before). Change scores are not the only way to measure change from baseline. It has been argued that it is more statistically efficient to model posttreatment outcome adjusted for pretreatment status as a covariate, rather than using change scores directly (see, e.g., Vickers, 2001; Vickers & Altman, 2001). However, weight loss in kilograms is the simplest choice here and therefore the best choice for illustrative purposes.

provider time, or both). For now, assume there is a single effectiveness measure $E$ (weight loss in kilograms) and a single cost measure $C$ (in dollars). Let $E_0$ and $E_1$ designate the estimated effectiveness of the control and experimental interventions, and likewise $C_0$ and $C_1$ the estimated per-participant cost for the control and experimental interventions, respectively.

In the simplest case, the estimates of effectiveness may be the sample mean responses observed in each of the two conditions. Most often, there will be a pretest or other covariates, as well as missing data, and so the effectiveness estimates are likely to be fitted values from some model. Another option is to use a Bayesian approach; for example, Claxton (1999) focused on posterior means from a Bayesian model, which can take prior information into account, but in the absence of an informative prior, these simply reduce to frequentist means or fitted values. Methods for estimating cost are described in Bensink and colleagues (2013); Crowley and colleagues (2014); Drummond and colleagues (2005); Haddix, Corso, and Gorsky (2003); and Petrou and Gray (2011). However, for the purposes of this chapter, it is assumed that cost is a known quantity for each treatment.

If the goal of the study is to explore cost-effectiveness for future use in the field (i.e., regular clinical practice instead of well-controlled setting supervised by researchers), then the cost measure should be an estimate of future cost per person to deploy a treatment in the field, *not* the cost of including that condition in the current experiment. This is important, because costs and logistics may differ between an initial, well-controlled research study on the one hand and routine practice on the other. Because of differences in overhead resources available, the per-person cost of implementing a condition on a large scale in the future may not be the same as the average cost of implementing it in the sample, and the former is of more interest for decision-making.

## 2.2 The Cost-Effectiveness Plane

It is very helpful to think of each intervention as a point on a plane whose $x$-axis is $E$ and whose $y$-axis is $C$; this is called the cost-effectiveness plane (Black, 1990; see also Bensink et al., 2013; Petrou & Gray, 2011). As in Fig. 1, the control condition ($E_0$,$C_0$) can be pictured as the center (origin) of the plane. This does not imply that the control condition costs zero dollars and has zero effectiveness, but merely that it is the baseline (control or "comparator") to which the other condition will be compared. As illustrated in Fig. 1, valuable information for selecting a treatment is gained by considering which quadrant of the plane contains the point ($E_1$,$C_1$). If $E_1 > E_0$ and $C_1 < C_0$ (the "southeast" quadrant), then the new intervention is to be preferred because it is less expensive but more effective. If $E_1 < E_0$ and $C_1 > C_0$ (the "northwest" quadrant), then the standard intervention is better than the new, in that the new is both less effective and more expensive. In either of these quadrants, one

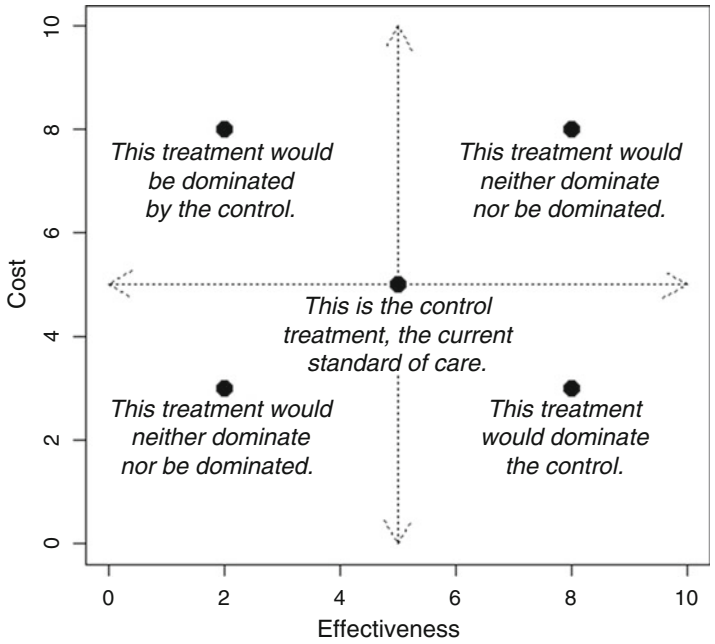


**Fig. 1** A sample cost-effectiveness plane (see Bensink et al., 2013; Black, 1990; Petrou & Gray, 2011). The *x*-axis represents effectiveness (e.g., mean weight loss per participant in kilograms for an obesity treatment program). The *y*-axis represents cost (e.g., total cost for time, provider salary, and supplies in units of $100 per participant). (Adapted from similar figures in Belsink et al., 2013; Black, 1990; Petrou & Gray, 2011)

treatment is said to *dominate* the other: it is better in at least one way and not worse in any other way, so the choice to be made is relatively unambiguous.[2]

In the other two cases, the data alone do not provide all of the information needed to make a decision, because one intervention is more expensive but more effective than the other, so a tradeoff is inevitable. If $E_1 > E_0$ and $C_1 > C_0$, the new intervention costs more, but delivers more. If $E_1 < E_0$ and $C_1 < C_0$, then the new intervention is less effective but less expensive. Thus, a decision still has to be made

[2]This chapter will use the term "dominate" in a somewhat informal way. The economics literature sometimes distinguishes between strong and weak dominance. Option A is said to strongly dominate Option B if A is both less costly and more effective. Option A is sometimes said to weakly dominate Option B if A is either equally costly but more effective or less costly but equally effective. This is not a large distinction for our purposes, first because it is unlikely that two treatments would have exactly the same cost or exactly the same benefit (Drummond et al., 2005) and second because it does not necessarily change the implied decision. The term weak dominance is also sometimes used when Option A delivers more effectiveness per unit cost than Option B (Gift et al., 2003). However, this kind of "dominance" does not necessarily imply that Option A is better in every sense, as Option B might still have less total cost or more total effectiveness. Therefore, the sense of the term "dominance" in this chapter is essentially that of strong dominance.

about whether to choose the more effective intervention or the less expensive one. Science can inform but cannot answer this question; financial, logistical, and ethical factors must be considered, and the answer is partially dependent on subjective judgment. Empirical tools to help inform the decision are the main focus of this chapter.

## *2.3 Options for Considering Cost and Effectiveness*

The following three subsections will describe the ramifications of three potential approaches to making cost-effectiveness tradeoffs. The first approach is to ignore cost and simply recommend the most effective intervention available. The second is to specify a fixed budget cap for the amount that may be spent per person and to recommend the most effective intervention among those that are expected to cost less than that amount. The third is to specify a ratio known as the willingness-to-pay parameter and use it as a standard way to compare the practical significance of the difference in effectiveness to that of the difference in cost.

### 2.3.1 Ignoring Cost

Suppose that the experimental intervention is more expensive but more effective than the standard-practice control intervention. What should the scientist recommend? In theory, the simplest approach could be to argue that the more effective treatment should always be provided, no matter the cost. Unfortunately, this is not always possible.

Many prevention scientists and health researchers are motivated by the desire to help people, save lives, or prevent suffering, and they are probably not viscerally motivated by the desire to save money for government bureaucracies or large insurance corporations. There is a very strong ethical and practical case to be made for increased societal investments in empirically supported interventions to prevent or treat adverse health conditions, especially for children and other people lacking the resources to provide adequately for themselves (see, e.g., Bierman, Henrichs, Welsh, Nix, & Gest, 2017; Bradley et al., 2016; Holzer, Whitmore Schanzenbach, Duncan, & Ludwig, 2008; Komro, Flay, Biglan, & Promise Neighborhoods Research Consortium, 2011; Prado et al., 2017; Shoemaker, Tully, Niendam, & Peterson, 2015). Well-delivered health interventions can provide great economic as well as human benefits (see, e.g., Belli, Bustreo, & Preker, 2005; Guyll et al., 2011). Furthermore, it could be socially unethical to allow patients to go without needed treatments for lack of money to pay (see American Medical Association, 1995; Williams, 2015). More also needs to be done to improve awareness and accessibility of existing resources (see Miller, Nowels, VanderWielen, & Gritz, 2016; National Academies, 2017). Unfortunately, however, no institution has infinite money, and no service provider has infinite time. Therefore, difficult decisions have to be made about

which programs to fund or which treatments to provide. Decisions would still have to be made even if society invested much more heavily and much more efficiently in human life and health than is currently the case. Even nonprofit foundations motivated purely by altruism have to decide between providing intensive programs to a few people in need and providing basic programs to many people in need; there may be a good argument for each, but it might not always be possible to do both. This is an inherent tradeoff that requires that financial cost information and scientific effectiveness information be considered together.

### 2.3.2 Specifying a Fixed Budget

The simplest way to make a cost-effectiveness decision is to have a known and fixed budget. Suppose it has been determined that the provider simply cannot afford to pay a cost of more than $C_{\max}$ units (dollars per participant, hours, etc.). Then the more effective treatment will be chosen only if its cost is less than $C_{\max}$. Otherwise, either the less expensive treatment will have to be used even though it is known to be less effective, or else more research will need to be done to find a more satisfactory—but still affordable—treatment.

While the perspective of a fixed budget may be useful for many researchers, it is not the main approach used in published cost-effectiveness studies. Perhaps the reason why health economists do not rely solely on a fixed maximum cost is that it involves somewhat dichotomous thinking: as long as the cost is less than $C_{\max}$, only effectiveness matters and the cost can be ignored; but if cost is even a penny greater than $C_{\max}$, then no gain in effectiveness is large enough to justify it. This approach may be too inflexible.

### 2.3.3 Specifying a Willingness-to-Pay Parameter

Instead of a fixed budget, economists are more likely to use the idea of willingness to pay. An insurer, granting foundation, or governmental body is willing to pay $\lambda$ units of cost, in exchange for one unit of effectiveness (see Bensink et al., 2013; Petrou & Gray, 2011). For example, in the weight loss example, perhaps they might be willing to pay $200 per kilogram expected to be lost.

There are different ways to think about the meaning of $\lambda$. Mathematically, it is a weight that converts effectiveness units into cost units so that both can be compared on a common scale; that is, the experimental treatment will be considered more cost-effective if $\lambda E_1 - C_1 > \lambda E_0 - C_0$. This may be easier to understand by first considering a situation in which the production of a physical commodity, rather than the health outcome from an intervention, is the process of interest. In an industrial or agricultural setting (e.g., Harrington, 1981), where methods of producing a product for sale are being compared, $E$ might be the expected sales value of product yield, and $C$ might be the investment required to produce the product. Thus, one could simply set $\lambda = 1$ and let $\lambda E - C$ be the net profit. It is then

easy to see that the manager of a farm or factory would want to choose the strategy with highest net profit, not necessarily with highest gross production alone or lowest cost alone. Similarly, in planning an intervention aimed at improving human health, it might sometimes be too simplistic to focus only on maximum effectiveness per participant or minimum cost per participant and might be necessary to consider both effectiveness and affordability. Thus, the estimated net benefit, $\lambda E - C$, generalizes the idea of profit in expressing the overall utility of a given treatment. In the case of the weight loss example, $\lambda$ is the number of dollars that a decision-maker would be willing to pay per expected kilogram lost.[3]

The $\lambda$ parameter can also be interpreted as a cutoff to be compared with another quantity called the incremental cost-effectiveness ratio (*ICER*). It is easiest to understand the *ICER* in the case where the experimental treatment is *more expensive but more effective* than the control treatment ($E_1 - E_0 > 0$ and $C_1 > C_0$), so that the analyst is trying to decide whether the cost increase is worthwhile. Then the *ICER* is defined as

$$ICER = \frac{C_1 - C_0}{E_1 - E_0}.$$

This is a measure of the gain in cost-effectiveness by using the experimental rather than the control treatment. More specifically, it is the increase in cost per unit of effectiveness gained when moving from the control to the new treatment. In this context, $\lambda$ is the cutoff being used to determine what counts as "large enough," because $\lambda E_1 - C_1 > \lambda E_0 - C_0$ if and only if $ICER < \lambda$. That is, $\lambda$ determines whether the cost-effectiveness of the new intervention (as measured by *ICER*) is better or worse than that of the old intervention. Geometrically, on the cost-effectiveness plane, an experimental treatment in the northeast quadrant must fall below a line with slope $\lambda$ passing through $(C_0, E_0)$ in order to be considered cost-effective. This is illustrated in Fig. 2.

Note that $\lambda$ is chosen subjectively or based on budgetary and legal standards. Thus, the conclusion in a cost-effectiveness analysis, unless one alternative dominates the other, will not be based solely on the observed data from the study. Although this reality may feel strange or uncomfortable to a scientist who wishes

[3]In practice the farm or factory owner might use a weight other than one; for example, they might discount (proportionally reduce) $E$, in order to reflect time delay in production and uncertainties in sales (see Harrington, 1981; Petrou & Gray, 2011). However, the basic intuition still applies: convert a gross gain to a net gain by subtracting off some measure of cost. This kind of discounting of future predictions can also be done in cost-benefit analysis in promoting human health. The discounted benefits minus discounted cost (essentially, $E$-$\lambda C$) is therefore called the "net present value" of an intervention (see Messonier & Meltzer, 2003). In the weight loss example, if a particular $\lambda$ is used to convert kilogram units into money units, and the decision-maker is trying to find out how much money to spend on weight loss versus other priorities, then cost-benefit analysis is being done, rather than cost-effectiveness analysis. The focus of this chapter is on cost-effectiveness rather than cost-benefit analysis, so the question of how to choose $\lambda$ is beyond the scope of this chapter.

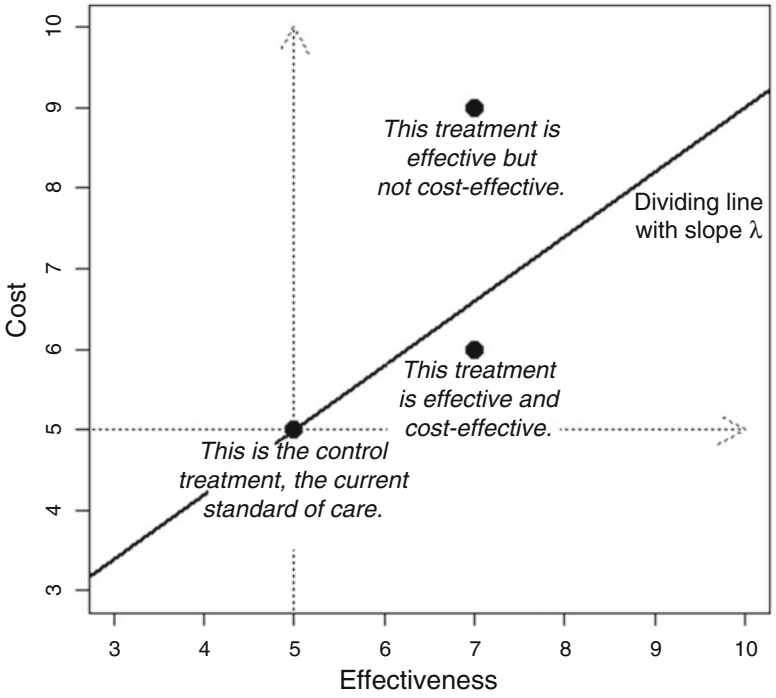


**Fig. 2** The northeast quadrant of a cost-effectiveness plane, reflecting a situation in which an experimental treatment is more expensive but more effective than the control treatment. As neither treatment dominates the other, it is necessary to introduce another parameter, typically a willingness-to-pay parameter $\lambda$. The new treatment is cost-effective if it is above the dividing line with slope $\lambda$ through the origin. Setting $\lambda = 0$ reflects a desire to save money at all costs (more cost-effective simply means less costly), while $\lambda = \infty$ suggests that cost is ignored (more cost-effective simply means more effective). (Adapted from similar figures in Belsink et al., 2013; Black, 1990; Petrou & Gray, 2011)

to be objective, it is unavoidable if a practical decision, and not just a theoretical conclusion, is to be made. This is because science can help people find the best way to achieve their goals, but cannot directly tell them what their goals should be or how they should weight competing goals.

The case described above, in which $E_1 > E_0$ and $C_1 > C_0$, is common and easy to imagine. Scientists are often trying to find a more effective alternative to existing intervention approaches and are accustomed to trying to establish that their intervention is more effective than the control treatment. It is also often true that experimental treatments cost more than the existing standard. However, it is also important to consider the opposite situation: a new treatment that is less expensive but less effective than the old one. The following subsection briefly considers this case.

### 2.4 What if the Experimental Treatment Is the Cheaper One?

From a purely mathematical standpoint, it does not matter whether the less expensive treatment is labeled as the control or the experimental one; all that matters is whether the point on the cost-effectiveness plane for the experimental treatment falls below the sloped line passing through the point on the plane for the control treatment. In other words, it is possible that the new intervention could be judged to be more cost-effective than the standard (because it is below the sloped line in Fig. 2), even though it is less effective than the standard (because it is to the left of the *y*-axis in Fig. 2). That is, an experimental treatment in any quadrant of Fig. 1 may potentially be chosen, except for the northwest quadrant, which can never be below the sloped line.

In the context of medical treatments for serious diseases, Bensink and colleagues (2013) recommend that an intervention that is less effective than the standard of care generally should not be chosen even if it is more cost-effective, because it could be unethical to restrict a patient to substandard treatment simply in order to save money. In this context, an experimental treatment must be both at least as cost-effective *and* at least as effective as the control in order to be chosen (i.e., it must both be below the sloped line and to the right of the *y*-axis in Fig. 1, thus eliminating both the northwest and the southwest quadrants from consideration).

However, there are other contexts, perhaps especially in the context of large-scale prevention programs, in which a *more cost-effective but somewhat less effective* intervention may still be desirable because given a fixed budget it can be given to more people. Suppose that the control condition is the best known treatment and has been shown to be effective in well-funded academic studies, but is too expensive and time-consuming to deploy at large scale in the field. In other words, suppose the control condition represents an expensive gold standard that only a few people are currently able to receive. In that case, if the new treatment is 90% as effective for 30% of the cost, then more lives may be saved by the new treatment than the old, simply because it can be given to more people. As an analogy, it has been found that the expansion of prescribing authority for nurse practitioners has increased the volume of care being accessed, even while somewhat reducing the cost per unit care (Muench, Coffman, & Spetz, 2016). Although one might prefer, all else being equal, to be treated by a physician with an M.D. degree rather than a nurse practitioner, the option of being treated by a nurse practitioner if a physician is unavailable can be beneficial for overall population health. This suggests that researchers should report both cost and effectiveness, so that those reading the research can make their own decisions based on the $C_{\max}$ or $\lambda$ defined by their own situation. In addition to reporting cost and effectiveness, it is usually important to report some measure of uncertainty, such as standard errors for the effectiveness estimates. This issue is discussed in the next section.

## 3 The Relationship Between Economic Decisions and Scientific Conclusions

As in other research settings, if statistical inference is to be made, it is also important to take into account the amount of uncertainty present in the estimates of effectiveness or cost. In this chapter so far, no standard error, $p$-value, or confidence interval has been mentioned; cost and effectiveness have been treated as known quantities, so the question has not been about whether the difference in effectiveness is statistically significant, but only whether it is practically significant in the sense of justifying cost. This is not realistic, because there is always uncertainty in real-world results. Clearly, one cannot ignore issues of statistical generalizability. However, issues of practical significance are just as important as statistical significance when making economic decisions. For example, if the null hypothesis $E_1 - E_0 = 0$ is strongly rejected ($p < 0.001$), but the difference on the raw scale is judged to be not large enough to justify the increased cost, then the control treatment will still be recommended. Conversely, suppose the null hypothesis is not quite rejected ($p = 0.07$), but there appears to be substantial benefit of the new treatment relative to the control ($E_1 > E_0$), and the increased cost is minimal ($C_1 \approx C_0$). The investigator probably cannot publish a scientific article recommending the new treatment as better in general, but also cannot reasonably describe the new treatment as having been discredited. The new treatment may still be very promising and might be supported with future studies.

### *3.1 Statistical Significance and Practical Significance in Four Hypothetical Examples*

For a graphical illustration of these ideas, consider Fig. 3. It shows four possible outcomes of a two-condition RCT comparing a more costly experimental treatment to a less costly control treatment. The error bars in the plot may represent either frequentist or Bayesian confidence intervals. For simplicity it is assumed that the cost for each treatment is fixed and known and did not have to be estimated using experimental data; otherwise, instead of error bars, there would be elliptical confidence regions reflecting uncertainty in both the $x$- and $y$-axes. For a comparison of methods of calculating these elliptical regions if they are required, see Stevens, O'Hagan, and Miller (2003).

In Fig. 3a, the difference in outcome between the control and experimental treatments would probably not be found statistically significant in a test because the confidence intervals appear to largely overlap. There is not enough evidence to claim that the experimental treatment is more effective than the control. That is, although the best point estimates suggest a provisional guess that it is slightly more effective, it may in fact be better, equal, or worse in effectiveness. Also, it is

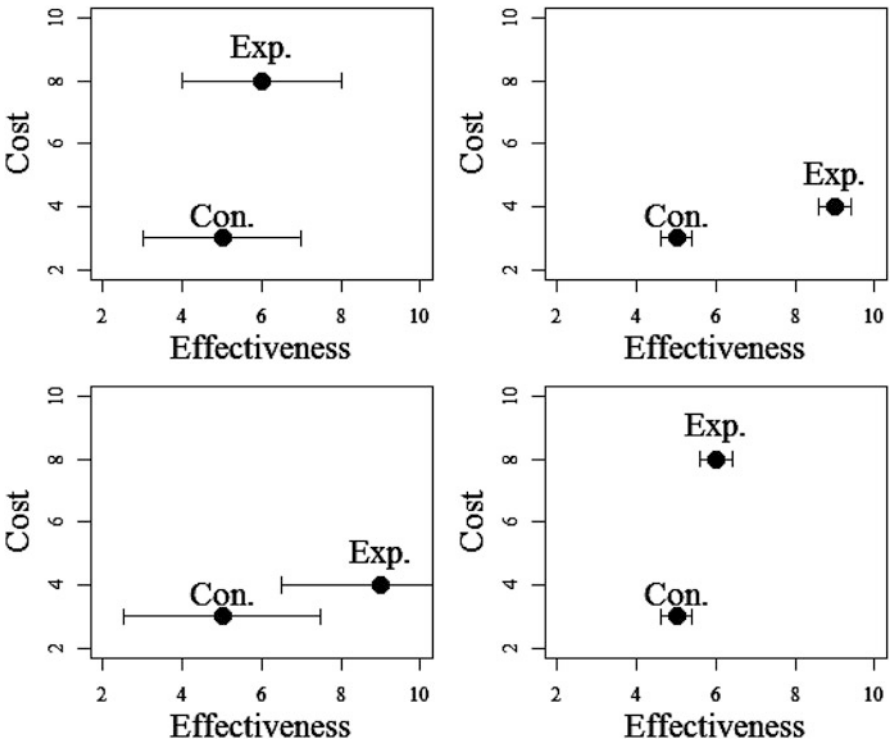


**Fig. 3** Four ways in which the statistical significance (in terms of confidence interval overlap) and the practical significance (in terms of cost-effectiveness) of two interventions may differ. All scenarios assume that the experimental treatment is estimated to be least somewhat more costly and at least somewhat more effective than the control treatment (i.e., they are all in the northeast corner of Figs. 1 and 2). "Con." denotes control and "Exp." denotes experimental. (**a**) Neither statistically significant nor cost-effective, (**b**) statistically significant and cost-effective, (**c**) not statistically significant but probably cost-effective, and (**d**) statistically significant but possibly not cost-effective

clear that it costs over twice as much as the control treatment. In most situations it would probably be decided to implement the control intervention rather than the experimental intervention in future practice.

In contrast, in Fig. 3b, the difference in outcome is highly statistically significant and also highly practically significant. Furthermore, the difference in cost is very small. These results would strongly support the use of the experimental intervention in future practice.

In Fig. 3c, the evidence is *not* conclusive enough to publish a scientific paper claiming that the experimental intervention has been reliably shown to be superior to the control intervention. However, it would be incorrect to say that the control and experimental interventions have been shown to be equivalent; the lack of statistical significance might be a result of an inadequate sample size or poor

measurement. Preliminary evidence suggests that the experimental intervention provides much better outcomes at a minimal increased cost, although this has not yet been scientifically established according to accepted standards. To rule out the experimental intervention from further consideration, based on a single and perhaps flawed trial, would be harmful both to science and to human health. Thus, the experimental intervention is at least worthy of further investigation, even though it cannot be conclusively argued to be superior.

In Fig. 3d, there is strong statistical evidence that the experimental treatment is more effective than the control treatment. However, the experimental treatment is much more expensive. In practice, consumers, insurers, or funding agencies would have to decide whether they could afford it (specify $C_{\max}$) and might also have to decide whether they are willing to pay the additional cost even if they can afford it (specify $\lambda$). Thus, the experimental treatment is more effective than the control, but whether it is judged more *cost*-effective depends on the budget $C_{\max}$ or the willingness-to-pay parameter $\lambda$ of the stakeholder who is considering paying for it. Without specifying $C_{\max}$ or $\lambda$, a conclusion that the experimental treatment is more effective can still be made, but a financial decision about whether or not to fund it in future practice cannot.

## *3.2 Possible Priorities in Data Analysis*

Figure 3 suggests that a *conclusion-priority analysis* (aimed at establishing and publishing evidence of effectiveness) and a *decision-priority analysis* (aimed at choosing whether or not to invest in a more expensive treatment option) are related but not the same. In conclusion-priority analysis, the researcher may strive to be as objective as possible and simply make conclusions from the evidence; however, in decision-priority analysis, subjectivity and context are unavoidable. In a conclusion-priority analysis, one either rejects or fails to reject the null hypothesis. However, in a decision-priority analysis, one must choose one option or another, and there might not be an option to "fail to choose." In a conclusion-priority analysis, a Type I error (false positive) is considered worse than a Type II error (false negative); in a decision-priority analysis, which error is worse depends on the practical consequences of the error. The differences between conclusion-priority and decision-priority analyses are further discussed in the Collins (2018) companion volume. Many researchers are more familiar with conclusion-priority analyses, especially in the context of null hypothesis testing.

These differences in focus are not merely different interpretations of the same analyses but can suggest a need for different methods of analysis. In conclusion-priority analysis, it is often customary to focus on presenting the results of null hypothesis significance tests in order to judge whether the treatments reliably differ in effectiveness. In a decision-priority analysis, significance tests are much less central, although they can still play a role. In a decision-priority analysis, the most important quantities are the costs $C_1$ and $C_0$, the estimated effectiveness $E_1$

and $E_0$, and the willingness-to-pay $\lambda$; quantities such as $p$-values are much less important (Claxton, 1999; Sullivan & Feinn, 2012). Once the cost, effectiveness, and willingness to pay are specified, a utility or net benefit (i.e., $\lambda E - C$) can be estimated for each treatment, and whichever treatment has the higher utility is preferred. It would not be reasonable to automatically choose the experimental option just because the difference in effectiveness is statistically significant in a null hypothesis test, because the difference in cost is also important. It would also not be reasonable to automatically choose the control option just because the difference was not statistically significant. That is, giving the benefit of the doubt to the older or less expensive treatment, just because it is associated with the "null hypothesis" condition rather than the alternative, might lead to poor treatment decisions, especially if policy-makers interpret a failure to reject the null hypothesis as proof that the null hypothesis is true. Tests and confidence intervals can indicate a degree of confidence in a choice or can suggest that more data are needed before making a final decision, but they do not, in themselves, directly tell us what choice to make.

Fortunately, an approach to scientific inference using significance tests and an approach to practical decisions using cost-effectiveness methodology can be complementary. Researchers can study the questions of both effectiveness and cost-effectiveness, even in the same trial, and both kinds of questions are valuable (Bensink et al., 2013; Ramsey et al., 2005). A simple focus on statistical significance ignores the scale of the $y$-axis and, hence, the difference between panels 3a and 3c and the difference between 3b and 3d in Fig. 3. A simple focus on point estimates of efficacy and cost ignores the width of the error bars. Considering both perspectives allows an investigator to avoid having to ignore these vital features of the findings. That is, a researcher can do both a conclusion-priority analysis and a decision-priority analysis on the same dataset and report the results of both. Thus, it is possible to make nuanced statements like "We established that the new treatment is more effective, but also found that it was unexpectedly costly and difficult, so we suspect that it may be too expensive to use at scale in our setting," or "Although the difference was not statistically significant, the new treatment still has great potential for public health due to its high estimated effectiveness and low estimated costs, and it is worth further study in a larger trial."

Also, the distinction between conclusion-priority and decision-priority analyses should not be considered absolute. In particular, there are analyses that could be seen as combining elements from a conclusion-priority analysis of effectiveness and a decision-priority analysis of cost-effectiveness, in order to provide a conclusion-priority analysis of cost-effectiveness. These analyses are very important in the cost-effectiveness and cost-benefit literature; it is desirable and important to be able to make a statistical inference with high confidence that one intervention is more cost-effective than another. Unfortunately, making a generalizable claim that an intervention is more cost-effective than a less expensive one is likely to require at least as large a sample size as making a generalizable claim that the intervention is more effective (to see this, consider that cost-effectiveness differences become equivalent to effectiveness differences if the costs are equal).

Therefore, in the optimization phase of MOST, there will probably not be adequate information to provide statistically significant evidence for pairwise differences of cost-effectiveness between individual conditions. Thus, the results from the analysis might be better considered to be primarily decision-priority (see Chapter 3 in Collins, 2018, companion volume). In the evaluation phase of MOST, conclusion-priority comparisons of cost-effectiveness might become more feasible because only one condition (the optimized intervention) is being compared to another (the control or standard of care). Conclusion-priority analyses of cost-effectiveness go beyond the scope of this chapter, but are briefly described in the following subsection, which may be skipped without loss of continuity.

## *3.3 Statistical Inferences About Cost-Effectiveness Measures*

Techniques for taking uncertainty into account when making conclusions about cost-effectiveness analyses are reviewed elsewhere (see, e.g., Bensink et al., 2013; Fenwick, O'Brien, & Briggs, 2004; Ramsey et al., 2005; Stevens et al., 2003). As an example of conclusion-priority analysis of cost-effectiveness, a researcher might provide a confidence interval for the incremental cost-effectiveness ratio. Alternatively, the investigator could present curves giving, for each possible value of $\lambda$, some measure of degree of certainty that the experimental treatment is or is not the best choice according to that $\lambda$. The most natural form of these curves, although only available in a Bayesian context, is a plot of the posterior probability that the experimental treatment has a higher net benefit than the control, as a function of $\lambda$. At the $\lambda = 0$ end of the curve, the plotted value is the estimated probability that the experimental treatment is less expensive, ignoring the effectiveness. At the $\lambda \to \infty$ end of the curve, the plotted value is the estimated probability that the experimental treatment is more effective, ignoring the cost. At intermediate values, the plotted values are the probability that the experimental treatment should be favored for a particular $\lambda$. The graph can even be used to find the $\lambda$ for which the probability is 50% (i.e., the least $\lambda$ at which the more expensive treatment begins to look better). These probabilities are intended to be used to inform practical decisions and determine whether further study is warranted, but they are not intended to be evaluated in a binary way as in a null hypothesis significance test. Drummond and colleagues (2005) recommend either (1) constructing joint confidence intervals for cost and effectiveness or confidence intervals for net benefit, (2) fitting regression models directly to obtain estimates and standard errors for net benefit, or (3) estimating cost-effectiveness acceptability curves. Drummond and colleagues (2005) and Gift et al. (2003) also point out the usefulness of doing sensitivity analyses to explore the effects of various choices and assumptions made during the analysis. Ramsey and colleagues (2005) summarize the findings of an international task force that studied how best to conduct cost-effectiveness analyses in the context of RCTs. They recommend reporting confidence intervals for the cost-effectiveness acceptability curve and encourage measurements of cost on a per-person basis. They

favor an intent-to-treat approach and also favor the use of bootstrapping, multiple imputation, and sensitivity analyses to get better measurements of uncertainties about costs and benefits. The National Academies (2016) report also recommends reporting uncertainty when making published claims about cost-effectiveness and discusses the role of Monte Carlo simulations and sensitivity analyses. Some of these advanced methods may require consulting with an economist and/or a statistician. Although useful for assessing generalizability, these advanced methods may not be strictly necessary in order to make a cost-effectiveness decision.

## *3.4 Summary and Further Resources*

In summary, cost and effectiveness for each condition in an experiment can be measured. Cost-effectiveness techniques help integrate both into a final decision. Making an economic decision about what treatment to provide is related to, but not exactly the same as, making a scientific decision about what treatment is effective. The economic decision requires a subjective statement of maximum budget, willingness to pay, or both. Further reviews of cost-effectiveness analyses and their use in RCTs can be found in Bensink et al. (2013); Gift et al. (2003); Ramsey et al. (2005); Sculpher, Claxton, Drummond, and McCabe (2006); and Weinstein, Siegel, Gold, Kamlet, and Russell (1996). They describe best practices, limitations, and issues to consider. Claxton, Lacey, and Walker (2000) describe the topic in a somewhat different way, using a Bayesian, decision-theoretic point of view. Two recent real-world examples of cost-effectiveness analyses of RCTs regarding weight loss are Fuller and colleagues (2013) and Tsai and colleagues (2013); see Saha, Gerdtham, and Johansson (2010) for a further review and discussion. In theory, cost-benefit analysis can even be used to plan whether it is worthwhile to do a particular research study or not; this is called "value of information" analysis (see Eeren, Schawo, Scholte, Busschbach, & Hakkaart, 2015; Tuffaha, Gordon, & Scuffham, 2014), but applying it to MOST could be complicated.

## *3.5 Relevance to MOST*

Cost-effectiveness methods are highly relevant to MOST when cost-effectiveness is the optimization criterion on which the investigator wishes to focus. There are at least two situations during the process of implementing MOST in which the investigator may wish to estimate cost-effectiveness: using data from a component-screening factorial or fractional factorial experiment during the optimization phase and using data from an RCT during the evaluation phase (see Collins, 2018, companion volume).

In the evaluation phase, the cost-effectiveness analysis would be of a confirmatory nature. It might be desirable to demonstrate that the selected ("optimized")

intervention package was more cost-effective than a standard intervention (in terms of *ICER*). Alternatively, it might be desirable to show that the selected package was more effective than a standard package while also meeting some cost-related constraint (such as costing less than a budget $C_{\max}$). Both of these goals are very similar to goals for cost-effectiveness analyses done often in the field of health economics. Thus, the ideas reviewed in this section can be applied directly to the evaluation phase.

In contrast, in the optimization phase, multiple factors are generally being considered at the same time, and, therefore, many potential interventions are implicitly under consideration: that is, each of the $2^k$ combinations of the $k$ factors is being considered to decide whether it is the optimal intervention according to the goals and optimization criterion at hand. The researcher may be considering only $k$ candidate intervention components, but the combinations of these choices represent $2^k$ potential interventions. If a complete factorial design is being used, this also represents $2^k$ treatment conditions. The basic ideas of the cost-effectiveness plane still apply, but a more complicated approach is required, which will differ somewhat from what is usually done in a two-condition setting. Thus, it will be necessary to consider some new questions and ideas. However, instead of jumping directly from a single-factor experiment with two conditions to a multiple-factor experiment with $2^k$ conditions, it is helpful to first consider an intermediate case: an experiment with, say, three or four conditions, not arranged in a factorial way. This is a sort of stepping stone that will introduce the comparison of multiple conditions at once, without dealing with the question of how to take factorial structure into account.

## 4 Randomized Controlled Trials with Three or More Conditions

In the previous section, some of the basic terminology and ideas of cost-effectiveness were discussed in the context of an RCT with two conditions (sometimes called "treatments" or "arms," hence the term "two-arm RCT"): a standard (control) intervention and an experimental (new) intervention. Although some of the underlying issues and available methods were potentially complicated, the two essential questions were simple: Is the new intervention more effective than the control, and is the new intervention more cost-effective than the control? However, some experiments involve assessing the comparative effectiveness and/or cost-effectiveness of three or more different conditions (arms), each representing a different treatment (intervention). Very often, one condition will be designated as the control, but this is not always the case. For example, three different treatments might be compared against each other without a control; or two controls, one inactive (attention only, wait list, or placebo) and one active (standard care), might both be compared against a new treatment. The multiple-condition RCT is still less complicated than a large factorial experiment, but it is more complicated than the two-condition RCT described above.

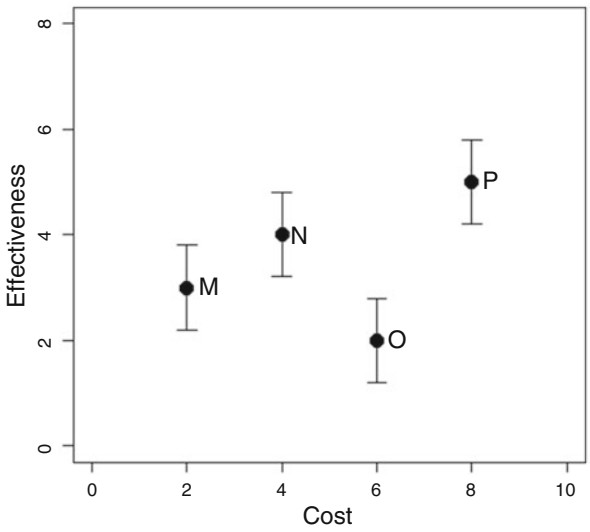


**Fig. 4** The costs of four different hypothetical interventions are plotted against point estimates and confidence intervals for their effectiveness. Note that cost is the *x*-axis and effectiveness is the *y*-axis, unlike Figs. 1, 2, and 3. It can be seen that intervention O is dominated by M and N

## *4.1 Comparing Multiple Points on a Cost-Effectiveness Plane*

There are several possible approaches to analyzing the results of this kind of study. To make them more concrete, consider the scenario in Fig. 4. It illustrates the effectiveness (in kilograms of weight loss) observed in a hypothetical study of four potential weight loss interventions, labeled M, N, O, and P. It is still assumed that the cost (perhaps in hundreds of dollars) per person to provide each potential intervention is known.

To avoid possible confusion, it is useful to point out that Fig. 4 is drawn somewhat differently from Figs. 1, 2, and 3. First, unlike Figs. 1 and 2, none of the treatments are put at the origin, because no single treatment is necessarily the standard of comparison. Second, unlike Figs. 1, 2, and 3, the *x*- and *y*-axes have been transposed, making effectiveness the *y*-axis. The previous figures had been drawn according to the convention in the field of cost-effectiveness, to facilitate comparison with other papers in that field. However, for reasons that will become clear later, Fig. 4 is drawn in a way that may be more familiar for researchers who are accustomed to studying comparative effectiveness without explicitly considering cost in the analysis. In such analyses, it is common to see treatment as the independent variable (hence the *x*-axis by general convention) and effectiveness as the dependent variable (hence the *y*-axis by general convention).

In Fig. 4, it appears that N performs significantly better than O and that P performs significantly better than M and O. These are interesting scientific conclusions, but they do not consider cost, so they alone are not enough to tell a decision-maker which condition to select as best for future evaluation or implementation. Suppose that beyond simply reporting the comparisons, it was also necessary to choose a single intervention to recommend, implement, or fund. This decision depends on

the circumstances and on one's goals; as discussed in the previous section, it does not flow directly from the scientific results, although ideally it should have a strong basis in them. In particular, N, O, and P are progressively more expensive than M. This needs to be considered along with the differences in effectiveness.

## *4.2 The Data Alone Do Not Always Provide a Universal Best Decision*

One straightforward first step is to eliminate intervention O from consideration. It is dominated by (inferior to) M and N. That is, it costs more but delivers less, so there would be no reasonable criterion by which O would be a good choice. The set of non-dominated options is sometimes called the Pareto front or Pareto frontier (see Chapman, Lu, & Anderson-Cook, 2014). In this example, they are M, N, and P. The use of the term "frontier" may seem strange; it is a metaphor suggesting a border distinguishing between the conditions that are dominated because they are too expensive, and those that are dominated because they are not effective enough. Every treatment on the "frontier" is the most effective option available for some budget, and so any one of them could be a rational choice. In our example, when considering M, N, and P, it is possible to make a rational argument for any of them, depending on how one wishes to make the subjective tradeoff between cost and effectiveness.

One could try to ignore the tradeoff and make the decision on a statistical basis alone. For example, one could somewhat naïvely argue that because M was not statistically significantly different from N, and N was not statistically significantly different from P, they therefore all had the same effectiveness. According to this argument, the less expensive one (i.e., M) should be recommended. However, this ignores the fact that P may still be statistically significantly more effective than M. More subtly, it also ignores the fact that failure to observe a statistically significant difference between two quantities does not prove that they are really the same. Thus, it would not make sense to recommend M solely on the basis of significance tests.

Another approach would be to argue as follows: because P is statistically significantly more effective than M and at least as effective as N, it is reasonable to recommend P. This argument would be compelling if resources were not limited, but in practice it ignores the fact that P is more expensive than M or N.

Someone else might offer treatment N as a common-sense compromise, but there does not seem to be a principled way to defend this. It has not been conclusively proven that N is better than M, and it has not been conclusively proven that P is not better than N (despite their overlapping confidence intervals). In some situations, the investigator might be permitted to refuse to make a recommendation at all until more data are available. However, in other situations this might not be feasible. Because the objective scientific findings alone do not translate directly into an unambiguous recommendation, the decision-maker must make the best decision available, keeping in mind what goal he or she is trying to achieve (or more

specifically, what optimization criterion is to be maximized) and what constraints there are on this choice. For example, suppose that the investigator is trying to optimize effectiveness subject to a constraint on cost, and P was judged to be too expensive. In that case, the investigator would presumably select N as the best bet available.

## *4.3 Comparing Possible Decision Goals in a Decision-Priority Analysis*

Using ideas from the section on two-condition RCTs, it is easy to think of several prototypical goals in choosing an intervention. They include:

1. **Maximize *E***: Find the most effective intervention overall.
2. **Minimize *C***: Find the least costly intervention overall.
3. **Maximize *E* with constraint on *C***: Find the most effective intervention that has at most some specified maximum cost.
4. **Minimize *C* with constraint on *E***: Find the least costly intervention that has at least some specified minimum effectiveness.
5. **Maximize E/C (or, equivalently, minimize C/E)**. Find the most effective intervention per unit cost.

In the terminology used in MOST, each of these is a possible optimization criterion (or optimization criterion and constraint, in the case of options 3 and 4 above). Less formally, however, it might be helpful to imagine these as the goals of different friends who ask the researcher what they should do.

**Maximize *E*** First suppose that a very wealthy friend, who wants very strongly to lose weight, were to approach the investigator and ask for a best guess at the most effective intervention, ignoring cost. In this case, the most rational recommendation is P. It would be honest to point out that the investigator is not confident that P is more effective than N (noting that their confidence intervals overlap in the figure), so that more study is needed. However, if pressed to give a single answer in the meantime, the investigator would probably reply that P appears to be the best bet in terms of the information currently available.

**Minimize *C*** Next suppose that a friend has limited resources and wants the intervention that costs the least. This is not an interesting goal in the context of the current example, because costs are assumed here to be known a priori, so there would be no need to do a statistical analysis to find the answer. The *y*-axis data would not be needed in order to reply that M was the cheapest. However, if cost and effectiveness were both subject-level observed variables, as in many formal cost-effectiveness analyses, this could be a nontrivial question. Still, this goal would be of only limited interest in the weight loss example, because if effectiveness is really irrelevant, then cost could be cut all the way to zero by simply doing nothing.

**Maximize *E* with Constraint on *C*** Now suppose a highly motivated friend, but one with modest resources, came to ask the investigator's advice. This friend also desperately wants to lose weight, but can only afford to pay a maximum cost of $C_{\max}$. The investigator can answer this question, but only if the friend is willing to tell what $C_{\max}$ is. Depending on $C_{\max}$, the best affordable intervention might be M, N, or P.

**Minimize *C* with Constraint on *E*** It is also possible that a friend might insist that he or she needed to lose at least $E_{\min}$ kilograms and wanted the cheapest intervention having an expected mean effectiveness of at least $E_{\min}$ per person. This requires, at a minimum, specifying $E_{\min}$. Suppose $E_{\min}$ is 3.5 kilograms. Even then, the question could still be interpreted in two ways. If it is only desired to have the estimated effectiveness be greater than 3.5, then condition N is sufficient. However, if it is desired to have the lower bound of the confidence interval be greater than 3.5, then condition P is required.

**Maximize E/C** Finally, suppose a very budget-conscious friend asked for the intervention that costs the fewest dollars per kilogram lost. This is different from the previous question. One option would be to calculate the ratio C/E for each intervention. Using the point estimates of E, this comes out to $3/2 = 1.5$, $4/4 = 1$, $2/6 \approx 0.333$, and $5/8 = 0.625$. Ironically, intervention M is the most cost-effective in this sense, even though it is the least effective, because it provides at least some effectiveness without much cost.

In this example, each intervention's cost and effectiveness can be compared to zero (spending zero money and losing no weight at all). In other examples, a zero point might not be meaningful, or it might be desirable for interpretational reasons to compare both cost and effectiveness to those observed in a control group rather than considering them in isolation. Thus, one would replace the cost-effectiveness ratio with the incremental cost-effectiveness ratio (*ICER*) as defined earlier. Specifically, a cost-effectiveness ratio (CER) is $C_k/E_k$ for a given treatment with cost $C_k$ and effectiveness $E_k$, while an incremental cost-effectiveness ratio (ICER) is $(C_k - C_0)/(E_k - E_0)$ for a given treatment $k$ versus a control condition having cost $C_0$ and effectiveness $E_0$. The investigator must decide which ratio is more important to optimize in his or her situation. The two ratios will probably differ from each other if the control condition is the usual standard of care, as it will almost certainly have some nonzero cost and probably have some nonzero effectiveness. The cost-effectiveness acceptability curve, mentioned earlier for the two-condition case, could also be extended in different ways to the comparison of multiple conditions.[4]

[4]One alternative could be to consider two conditions at a time and compute a confidence interval for the ICER, or for the cost-effectiveness acceptability curve at a given λ, for each of these pairwise comparisons. Another alternative would be to designate a control condition and compare every other condition with that one only. Furthermore, if one is using a Bayesian approach with posterior probabilities, one could calculate, for each candidate intervention, either its posterior probability of being having the best utility measure λE-C among the conditions available, or its posterior probability of having a utility measure above some threshold.

The diversity of possible goals may seem overwhelming, but of course it is not necessary for any given investigator to investigate all of them. Investigators have considerable freedom to specify a goal (i.e., an optimization criterion) that is of interest in their own situation (see also Collins, 2018, companion volume). The point of this section is not that there is a single correct answer or even a single correct question, but to illustrate the diversity of questions that can be meaningfully asked and answered using cost-effectiveness ideas.

## *4.4 Goals Involving Statistical Inference About Effectiveness*

In addition to the questions listed above, it is also possible to incorporate information on statistical significance into the statement of the goal. For example, someone could ask for *the least expensive intervention that has been shown to be more effective* than condition M (i.e., letting condition M serve as the control group). This is a reasonable question in light of the growing emphasis on evidence-based practice. However, it does not provide a way to avoid subjectivity and uncertainty. First, the answer might be different if condition N or condition O were designated as the control group. Second, both the sample size of the study and the desired level of confidence determine the length of the confidence intervals; therefore, they would both affect the answer to this question.

The situation would be somewhat different if a friend asks for *the least expensive intervention that has not been shown to be less effective* than M. This is because "not shown to be less effective" is not the same thing as "shown to be no less effective." In particular, the fact that two confidence intervals overlap is not sufficient reason to conclude that two means are known to be equal. They might have overlapped because of insufficient sample size or poor study implementation. That is, an absence of evidence is not evidence of a total absence of an effect. If the question of equivalency or non-inferiority is really important, then special methods must be used. Roughly speaking, they require choosing a large enough sample size so that the confidence interval for the difference in treatment effects will be narrower than the least clinically significant difference and then using special one-sided statistical inference techniques (see, e.g., Rothmann et al., 2003). These methods are important topics of current research in biostatistics (e.g., Zhang, Nie, Soon, & Zhang, 2014), but they have not yet been widely used in the social and behavioral sciences.

Finally, it is possible to construct confidence intervals for the utility or net benefit $\lambda E - C$ or to test contrasts between conditions on this quantity. In the weight loss example, one could measure the weight loss $Y_i$ and cost $C_i$ for each participant in the condition and use them to construct a new variable $\lambda Y_i - C_i$. (The notation $Y$ is used here for weight loss instead of $E$, because it is an individual's outcome rather than a group comparison.) Mean comparisons on this variable could be used to make conclusions about $\lambda E - C$. Of course, this requires specifying $\lambda$. Also, an investigator might consider using multiple-comparison techniques if it is desired to control experimentwise error rates when comparing multiple conditions. Therefore,

in the following subsection, multiple-comparison techniques that may be useful for optimization contexts are reviewed. They can be used either for comparing an outcome variable (estimating $E$) or comparing a cost-penalized outcome variable (estimating $E - \lambda C$). The following subsection may be skipped if the reader is not planning to implement multiple-comparison correction techniques; it is presented mainly as a possible aid to future research in this area.

## *4.5 Multiple-Comparison Techniques for Comparing Effectiveness or Utility*

Classic frequentist experimental design theory, much of it derived originally from agricultural studies, offers many methods for multiple comparisons. In a standard course on experimental design, a common approach for analyzing a study like that illustrated in Fig. 4 would be to compare the four intervention groups using a one-way analysis of variance (ANOVA) under the assumption of homoskedastic normal responses, or perhaps some alternative procedure such as a generalized linear model or nonparametric test. Assuming that its statistical assumptions were reasonably close to being met, the ANOVA could be used first to test the (admittedly rather uninteresting) omnibus null hypothesis that all interventions have the same mean effectiveness. The ANOVA would then be augmented by planned contrasts, pairwise comparisons, or specialized multiple-comparison techniques to make more specific statements about the response means of the individual interventions (see, e.g., Day & Quinn, 1989; Kuehl, 2000). This approach could presumably be combined in some way with cost information, perhaps by treating the cost-penalized outcome variable for some penalty weight $\lambda$, instead of the raw outcome itself, as the response variable. It could alternatively be done by comparing the effectiveness among only the treatments that were found to have average cost less than some fixed $C_{\max}$. In fact, there has been some research on incorporating multiple outcome dimensions into multiple-comparison testing (Hasler & Böhlendorf, 2013), although that is beyond the scope of this chapter.

The best known multiple-comparison procedures are those involving all pairwise comparisons of treatments. For example, in Fig. 4, tests or confidence intervals would be constructed for six contrasts: M versus N, M versus O, M versus P, N versus O, N versus P, and O versus P. Various techniques are available for performing these contrasts simultaneously at a fixed familywise Type I error rate. Of these, Bonferroni adjustment is the best known but not always the most efficient. An alternative would be to designate one intervention as the control and to compare each other intervention with it alone. For example, if M is the control, tests or confidence intervals would be constructed for only three contrasts: N versus M, O versus M, and P versus M. Because fewer comparisons are being done, the individual comparisons can be made slightly more powerful without affecting the familywise alpha level. However, another approach is more interesting for our purposes: Hsu's

multiple comparison with the best (Hsu, 1984; see also Bechhofer, Santner, & Goldsman, 1995; Ertefaie, Wu, Lynch, & Nahum-Shani, 2015). This procedure requires defining what constitutes a "best" intervention, generally the highest or lowest mean value on the response variable. The procedure then determines, for each other intervention, whether there is enough evidence to state with some confidence that it is not the best. The end product is a list of possible best interventions and a list of interventions that can be ruled out. For example, if only effectiveness were of interest, then N and P might be chosen as possible best interventions in Fig. 4. Unfortunately, requesting a very high degree of confidence or having a smaller sample size available might lead to a list of possible best interventions that is impractically long. The best-case scenario is that a single intervention is identified as best with high confidence; the worst-case scenario is that no interventions are ruled out, indicating very little confidence in which to choose. Technical details of these procedures are described further in standard experimental design textbooks such as Kuehl (2000). Ertefaie and colleagues (2015) consider how to implement a technique like Hsu's in the context of a sequentially randomized experiment, which is basically a form of factorial experiment.

Bechhofer and colleagues (1995) describe related selection approaches. All of these approaches combine statistical inference with decision guidance, but they are derived in somewhat different ways. The authors distinguish between "indifference zone," "subset selection," and "multiple-comparison" techniques. In indifference zone techniques, one seeks a high probability of choosing the best condition, assuming that this condition is at least some quantity (denoted delta) higher than the second best (i.e., one is indifferent to the choice of treatment as long as it is at least near to being the best, but any treatment outside this zone is considered significantly inferior). (If the best and the second best may be arbitrarily close, such as a mean weight loss difference of one gram, then there is no way of having adequate confidence to separate them.) In subset selection techniques, one seeks to find a subset of conditions that contains the best condition with high probability of success; the goal is to make this subset as small as possible without making confidence too low. Finally, multiple-comparison techniques focus on a joint confidence region for differences among condition means. Bechhofer and colleagues also describe methods for sample size planning in order to obtain a desired probability of a useful outcome under a given set of assumptions; this is somewhat analogous to power planning for more traditional analyses that focus on null hypothesis testing. However, the techniques in their book have not yet been widely used, at least in the social and behavioral sciences, and may require further research and elaboration to be helpful in the cost-effectiveness context. Also, many of the techniques reviewed in Bechhofer's text are intended more for non-factorial RCTs with multiple conditions ("one-way" layouts), or at most two-way factorial experiments, rather than factorial designs with many factors. They could perhaps be extended to higher-order factorial experiments, but this has not been done yet. Therefore, they are not reviewed further here.

### 4.6 Relevance to MOST

Almost any of the comparison criteria described in this section (e.g., maximizing $E/C$, maximizing $E$ with a constraint on $C$) could be an optimization criterion during an optimization trial in the context of MOST. However, the multiple-condition RCT design considered in this section is unlikely to be realistic in MOST. This is because one of the main limitations of the multiple-condition RCT is that often very few conditions can practically be included. As the total pool of participants is divided among more and more conditions, the power for making pairwise comparisons becomes smaller and smaller. This is one of the main reasons why factorial designs are used, instead of designs with many unrelated conditions (see Collins et al., 2009). Therefore, cost-effectiveness analyses with factorial designs will be explored in the following section.

## 5 Cost-Effectiveness Concepts in a Factorial Design

Analysis of data from a factorial experiment involves different issues from analyzing data from a standard multiple-condition RCT. To explore these issues in a relatively simple context, first consider a $2 \times 2$ factorial experiment comparing weight loss interventions. The first factor, $A$, compares a proposed intensive diet-focused intervention component (the "on" level) with a less expensive standard diet-focused intervention (the "off" level). The second factor, $B$, similarly compares an augmented exercise-focused intervention component (on) with a less expensive standard of care (off). Effect-coding notation will be used here (Kugler, Dziak, & Trail, 2018), denoting the standard level as $-1$ and the augmented level as $+1$ for each of the two components. Participants are assigned at random to four possible conditions: $(-1, -1)$, $(+1, -1)$, $(-1, +1)$, and $(+1, +1)$. Each condition represents a possible intervention package: neither factor on, only factor $B$ on, only factor $A$ on, or both factors on. The conditions are also known as "cells" because they can be listed in a table as combinations of the levels of the first factor (as rows) and the second (as columns). Factorial experiments are described further in Wu and Hamada (2009), Myers and Well (2003), and the Collins (2018) companion volume. Suppose that a researcher wants to do a cost-effectiveness analysis to choose which of the four cells mentioned above should be recommended for future use. The future use need not be direct implementation in general practice; it could be a confirmatory RCT in the evaluation phase of MOST.

### 5.1 The Role of Interactions in Estimating Cost-Effectiveness

There are two possible ways to interpret cost-effectiveness when analyzing a $2 \times 2$ factorial experiment. Is the investigator trying to answer two separate questions: Which level of factor $A$ is more cost-effective, and which level of factor

*B* is more cost-effective? Alternatively, is there only one question: Which of the four conditions defined by combinations of factor *A* and factor *B* is most cost-effective? The difference has to do with how one wishes to handle the possibility of interactions. In the first approach, the investigator would consider only the main effects of each factor, averaging over the levels of the other factor. In the second approach, the investigator would consider both their main effects and their possible interaction, in case the most cost-effective level of *A* depends on the level chosen for *B*. Interactions are introduced in many statistical textbooks, such as Myers and Well (2003) and Wu and Hamada (2009), and their implications for MOST are discussed at length in the Collins (2018) companion volume.

Interactions make analyses more complicated, but sometimes they cannot be ignored. First, *the factors could interact in determining effectiveness*. For example, it might be that some of the new educational content of the two experimental components overlaps, so their combination is slightly less effective than the sum of the parts. That is, there could be diminishing returns (the added component helps less in the presence of the other component than it would have helped in isolation) or even an iatrogenic interaction (the added component actually causes harm in the presence of the other component). Alternatively, they might work together very well, so that the effect of each factor is greater when the other is set to +1. This sort of interaction is familiar to readers with a background in experimental design. However, *the factors could also interact in determining cost*. For example, it might be that staff have to be hired to provide support for implementing the intensive level of either component, but the same staff member can handle both components. In this case, if it is decided to implement the augmented levels of both levels in future practice, this will cost somewhat less than the cost of augmenting *A* alone plus the cost of augmenting *B* alone, because the costs partly overlap. Note that when estimating "cost" here, the researcher should primarily consider the future marginal cost per patient in clinical practice, not necessarily the average cost per patient in the factorial experiment, because the goal is to make recommendations for future practice. For example, an experimenter should not include additional overhead resources or expenses charged by his or her research institution when calculating the costs to implement the final intervention.

If there are no practically significant interactions (neither in terms of cost or effectiveness), then decisions about the two components can be made independently. Separate tests of effectiveness and separate estimates of cost-effectiveness can be done for factor *A* (averaging over levels of factor *B*) and then for factor *B* (averaging over levels of factor *A*), in almost exactly the same way as if they had been two separate two-level RCTs. However, if there are practically significant interactions either in cost or in effectiveness, then a decision based solely on main effects will ignore some potentially important information and could lead to a poorer decision (Cox & Snell, 1989, pp. 84–92; Myers & Montgomery, 1995). It would thus be important to consider the simple effects of each factor conditional on the level of the other factor; that is, the researcher must directly compare the fitted values for estimated cost, effectiveness, or cost-effectiveness among all four of the cells (see Chapter 4 in Collins, 2018, companion volume for a greater description).

Furthermore, even if there are no interactions in the usual sense, a constraint may require that factors be considered together. Specifically, if a maximum cost constraint ($C_{\max}$) has been set, it may not be possible to choose a level for each factor without considering the level of the others. As an extreme example, suppose the investigator is asked for a recommended treatment to use in a resource-poor setting in which it is expected that there will be only enough money or time available to implement one of the two components in clinical work. In this case, even if the (+1,+1) condition had been found to be highly effective in the special setting of the factorial experiment, it cannot be recommended for general use in the resource-limited setting. For such a question, the only interesting conditions to be compared would be (−1, −1), (−1, +1), and (+1, −1). This is another scenario where a one-factor-at-a-time, main-effects-only approach would be inadequate for making a practical decision.[5]

## 5.2 Making a Bias-Variance Tradeoff

In the previous subsection, it was observed that ignoring interactions could lead to poor decisions when interpreting the results of a factorial experiment. However, the opposite approach, to consider both main effects and the interaction, has a serious disadvantage as well. To see this, consider the regression equation for modeling the expected value of the outcome $Y$ for a given cell in the $2 \times 2$ factorial. Let $X_A$ be the effect-coded representation of factor A and $X_B$ be the effect-coded representation of factor B. Then the saturated model is

$$\mathcal{E}(Y_i|X_A, X_B) = \beta_0 + \beta_A X_A + \beta_B X_B + \beta_{AB} X_A X_B.$$

Notice that four quantities (the cell means) are being represented by four quantities (the regression coefficients). This means that the model for the cell means is saturated, so the estimate for the expected value for each cell will be the observed average value in that cell. The estimated expected value for a particular cell in a saturated model is the sample mean of participants in that cell, and so it is informed only by the participants within each cell, as though the other cells were not present. The situation becomes more urgent as the number of factors increases. If fitted

[5] As a caveat, if it were known in advance that only three of the four conditions were of interest, then an incomplete factorial comparing only those three conditions might be more appropriate than a complete $2 \times 2$ factorial (see Collins et al., 2009 for advantages and disadvantages of incomplete designs). However, if, say, only four out of six components could affordably be implemented at a time in practice, there are still many affordable conditions, so a complete or fractional factorial experiment might still be appropriate. As another caveat, the challenge of creating special programs for resource-poor settings has raised important debates about equity and disparities which are beyond the scope of this chapter; for example, consider the criticism by Kozol (2005) and defense in Slavin (2006) of the use of a particular program for disadvantaged schools.

values from a $2^5$ factorial must be obtained in the same way as those from a 32-arm RCT, with each cell being studied on its own, then having power and precision for all pairwise comparisons could require tens of thousands of participants, more than most researchers can afford. This is because pairwise comparisons are simple effects, not main effects, and cannot combine the information from all participants as main effects can (see Collins et al., 2009, and the Collins, 2018 companion volume).

Using fewer parameters would allow information to be borrowed across cells, reducing standard errors. For example, suppose it was determined that the main effect of factor A was practically and statistically significant ($\beta_A \neq 0$) but the main effect of factor B and the interaction were not significant. Fitted values from the model

$$\mathcal{E}(Y_i|A = a, B = b) = \beta_0 + \beta_A X_A$$

would have lower standard errors than those from the saturated model because there are more constraints on this parsimonious model. Because both of the cells with $X_A = +1$ are constrained to have equal expected values, their estimate is the sample mean of all participants in each cell (possibly a weighted mean if cell sizes are unequal). Thus, each fitted value is being informed by twice as many individuals, leading to less random variability. Of course, there is a disadvantage to using a constrained model: the possibility of bias if the constraints do not accurately reflect the process or population being modeled.

This creates a dilemma. Trying to make cost-effectiveness decisions for one factor at a time risks bias by leaving some potentially important information out of the model, but trying to make decisions at the level of the cells while ignoring the factorial structure leads to too much sampling variance and not enough power. This dilemma was described but not fully resolved by Bechhofer et al. (1995). They describe methods for selecting the best level of each factor in a two-factor experiment, but they focus almost entirely on the case where interactions can be ignored. They briefly mention the possibility of interactions but conclude that in this case one must choose the best level of one factor within each level of the other factor (see also Wu & Cheung, 1994). If all of the factors represent intervention components (as opposed to existing population strata such as age or gender), then this essentially means choosing the best condition (cell). Thus, using their methods it is clear how to analyze a $2 \times 2$ factorial as either two artificially separated two-condition RCTs (one for factor $A$ and one for factor $B$), or how to analyze it as one single four-condition RCT, but it is still not clear whether or how it is possible to make specific use of the factorial structure to inform decisions.

One possible approach is to treat this dilemma as an example of a "bias-variance tradeoff." This is a very general idea that arises in many situations of statistical analysis, with variable selection in regression being a prominent example. An overly simple model can be biased, in the sense of being unable to detect certain effects and relationships. An overly complicated model can be subject to high sampling variability and overfit the available data. In other words, an overly simple model

misses some signal, and an overly complicated model misconstrues some noise as being part of the true signal. A model of moderate complexity can sometimes outperform either extreme by including enough detail to be useful but not enough to be misleading. When the amount of available data is limited, it may be necessary to fit a simpler model than one would otherwise wish (to "bet on sparsity") in order to have reasonable hope of obtaining interpretable and generalizable results. These ideas are further developed in statistical textbooks such as Hastie, Tibshirani, and Friedman (2009).

The idea of a bias-variance tradeoff suggests that the best answer is a compromise between extreme possibilities. Choosing which intervention is likely to give the best outcome, whether on the basis of effectiveness or cost-effectiveness, requires a model that provides fitted values for individual cells. However, the model must also be able to borrow information across cells as much as feasible by ignoring unimportant effects (especially nonsignificant higher-order interactions). One way to do this is to construct a parsimonious model that contains only statistically significant interactions and excludes statistically nonsignificant interactions and then use this model to obtain fitted values for effectiveness. For example, in a $2^k$ factorial experiment, a saturated model has $2^k$ parameters and is too large, while a main-effects-only model has $k$ parameters and is too small. A model with an intermediate number of well-chosen parameters could perform better than either. A method like this was employed by Harrington (1981) in one of the few published examples of a cost-effectiveness analysis on a factorial experiment with more than two factors. The advantage of this method is that it attempts to borrow information across cells wherever possible but not collapse cells where it is inappropriate to do so. The same procedure could be used for cost if that is an observed variable subject to variance.

The $2 \times 2$ design is the simplest factorial design, but factorial optimization trials in MOST are likely to include three or more factors (see Collins et al., 2009; Nair et al., 2008). Therefore, consider the following hypothetical example involving five factors.

## *5.3 A Hypothetical Five-Factor Example*

In order to demonstrate some of the ideas described above, consider the following hypothetical example of a five-factor experiment. In the simulated example considered here, five potential intervention components, labeled as R, S, T, U, and V here, are being tested for possible inclusion in a future highly effective and cost-effective weight loss MBI. For a somewhat similar real-world study that is currently being done, see Pellegrini, Hoffman, Collins, and Spring (2014, 2015a, 2015b). A simulated dataset for this hypothetical example is available online at https://methodology.psu.edu/mostbooks/dziak.

In this hypothetical example, each of the five components may be either provided or not provided; in addition, a limited amount of information and support is given

to all participants. For simplicity, assume that cost per person is measured in dollars and reflects the cost of supplies and payment for staff, although in practice it might also be important to consider participant burden. Also for simplicity, continue to assume that the effectiveness measure is weight loss in kilograms. The components are compared in a $2^5$ full factorial factor-screening experiment (as described in, e.g., Collins et al., 2009, 2014). There are $2 \times 2 \times 2 \times 2 \times 2 = 32$ cells (i.e., conditions; combinations of treatment factors) in the experiment, representing each possible combination of providing or not providing each of the five components. The factor *R,* which contrasts the inclusion versus exclusion of component R, is represented numerically by an effect-coded variable $X_R$, with $X_R = +1$ for inclusion (on) or $X_R = -1$ for exclusion (off); effect-coded variables $X_S$, $X_T$, $X_U$, and $X_V$ are defined similarly. Suppose that there are about 20 participants per cell, for a total size of about 640 participants. Also assume that the costs per component are known but unequal, as follows:

- The base cost per participant is $100.
- Turning factors R, S, or V on costs $200 each.
- Turning U on costs $400.
- Turning T on costs $600.

Thus, the possible cost per participant ranges from $100 (the basic treatment only, all components off) to $1700 (all components on).

Results of a multiple linear regression analysis are as shown in Table 1. Three- through five-way interactions were also tested, but none were significant, so only main effects and two-way interactions are shown. Factors *R*, *T*, and *V* are found to have statistically significant positive main effects. Factor *U* has a significant negative main effect. There are statistically significant positive interactions between components *R* and *V* and components *U* and *V*, as shown in Fig. 5. Following the reasoning of Collins et al. (2009, 2014, 2018 companion volume), a researcher would probably choose the combination R, T, and V. This combination would then be recommended for inclusion in a confirmatory evaluation in the evaluation phase of MOST and then for eventual use in practice if the evaluation is favorable. Indeed, the combination of R, T, and V will be more effective than the full package of all five components, because it does not include the counterproductive component U. It will also have lower cost than the full package, because it does not include either the useless component S or the counterproductive component U. Because it is less costly and more effective than the full package, it would certainly be more cost-effective by any reasonable definition. However, it costs $1100, and in some settings that might be too expensive. Budgetary realities might impose a maximum cost constraint $C_{\max}$, a maximum willingness-to-pay $\lambda$, or both. Thus, a further consideration of cost-effectiveness would be necessary.

In theory, cost-effectiveness decisions could be made either at the level of the component (e.g., "Is component R more cost-effective?", "Is component T cost-effective?,") or the cell (e.g., "is the condition with only component R included cost-effective?" or "Is the condition with components R and T included cost-effective?"). Making decisions component by component, without explicitly

**Table 1** Coefficient estimates from a saturated model fit to simulated $2^5$ factorial experiment data

| Effect | Estimate | $t$ | $p$ |
|---|---|---|---|
| Intercept | 0.027 | 0.117 | 0.907 |
| $X_R$ | 0.613 | 2.698 | 0.007** |
| $X_S$ | 0.071 | 0.313 | 0.754 |
| $X_T$ | 0.744 | 3.273 | 0.001** |
| $X_U$ | −0.539 | −2.370 | 0.018* |
| $X_V$ | 0.512 | 2.253 | 0.025* |
| $X_R \times X_S$ | 0.065 | 0.287 | 0.774 |
| $X_R \times X_T$ | 0.093 | 0.410 | 0.682 |
| $X_R \times X_U$ | −0.215 | −0.947 | 0.344 |
| $X_R \times X_V$ | 0.492 | 2.164 | 0.031* |
| $X_S \times X_T$ | 0.030 | 0.130 | 0.897 |
| $X_S \times X_U$ | 0.132 | 0.580 | 0.562 |
| $X_S \times X_V$ | −0.156 | −0.686 | 0.493 |
| $X_T \times X_U$ | −0.086 | −0.377 | 0.706 |
| $X_T \times X_V$ | −0.049 | −0.218 | 0.828 |
| $X_U \times X_V$ | 0.541 | 2.379 | 0.018* |

Notes: Three-, four-, and five-way interactions were included in the model but were not statistically significant and are not shown here. Standard errors are approximately 0.23 for each effect coefficient; they are approximately equal because the model is linear and the design is close to balanced. The error standard deviation is 5.75. Thus, for comparison, a Cohen's $d$ of 0.2 would correspond to a regression coefficient of $0.2 \times 5.75/(+1-(-1)) = 0.575$. Note that the intercept in an effect-coded factorial regression model is not the expected response for the "none" (all −1) cell, but is instead the mean expected response for all cells. The two values are close in the current example, but this is coincidental.
*$p < 0.05$; **$p < 0.01$.

comparing individual cells, seems simpler. However, because of the interaction, as well as a possible cost constraint on the total amount spent, it may be necessary to consider all of the components together instead of considering each separately: that is, to compare the effectiveness of specific cells. This suggests plotting the cells on a figure similar to Fig. 4. It will not be necessary to explicitly consider all 32 cells; it is already known that cells with factors $S$ or $U$ set to "on" will not be adequately cost-effective compared with other cells. However, this still leaves eight possible cells with $S$ and $U$ off:

- $R$, $T$, and $V$ on ($1100)
- $T$ and $V$ on ($900)

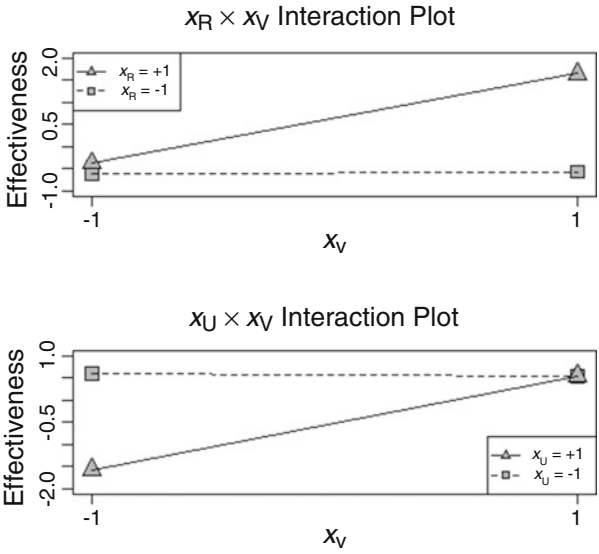


**Fig. 5** Interaction plots for simulated $2^5$ factorial example

- *R* and *V* on ($500)
- *V* on ($300)
- *R* and *T* on ($900)
- *T* on ($700)
- *R* on ($300)
- None on ($100)

Cost is assumed to be known for each cell, but an effectiveness estimate for each cell is still needed. Let $\hat{\mu}_k$ be the estimated effectiveness (the estimated expected value of weight loss) in cell *k*. The most obvious and straightforward (but not necessarily the best) option for this estimate is $\overline{y}_k$, the observed mean of the approximately 20 individuals in the cell. This is the same as the fitted value for the cell from the saturated linear model containing all possible main effects and interactions among the five components. However, this estimate does not borrow any information across cells nor take advantage of the factorial structure of the experiment in any way. Comparing any two cells in this way would make use of only about 40 participants and not use any information from the other 600. Therefore, it would not be a good choice unless the available sample size is huge.

It would probably be better to use information from a "parsimonious" (sparse or unsaturated) model, which borrows some information from similar cells when constructing the predicted value for a given cell. This is done by estimating the effectiveness for each cell not from the cell mean directly, but from its fitted value from a reduced regression equation that retains only significant effects. Substituting one regression equation for another may seem somewhat alarming because the

values and meanings of coefficients in a model may change as effects are added to or removed from the model; that is, in a multiple regression framework, the coefficient for a particular predictor variable expresses its degree of linear relationship with the response *after adjusting for the other predictor variables*, and not necessarily its *marginal* linear relationship with the response. However, in a linear model with effect-coded dichotomous factors near balanced allocation, removing one effect from the model has little effect on the coefficients for the other effects (see Collins, 2018, companion volume), and in any case it is only proposed to remove effects that are estimated to be fairly near zero.

## 5.4 Fitting a Parsimonious Predictive Model

How should the parsimonious model be constructed in this particular example? The nonsignificant main effect of factor $S$ and the nonsignificant interactions can be ignored, thus saving many degrees of freedom and reducing the risk of overfitting. However, even though it was already decided to exclude component U from the final intervention, it is not appropriate to exclude the corresponding factor $U$ from the model. This is because factor $U$ has a significant main effect and interaction and is therefore important to computing accurate fitted values.

The results from the selected and refitted unsaturated model are shown in Table 2, and the fitted values for the cells of most interest, using this model, are shown in Table 3. Note that because a linear model with effect coding is being used, the effects in Table 1 are near orthogonal; therefore, the coefficients in Table 2 are not very different from the corresponding coefficients in Table 1, even though some of the effects in Table 1 have been removed from the model and the coefficients have been re-estimated (Kugler et al., 2018; Myers & Well, 2003).

**Table 2** Coefficient estimates from parsimonious model fit

| Effect | Estimate | $t$ | $p$ |
|---|---|---|---|
| Intercept | 0.038 | 0.169 | 0.866 |
| $X_R$ | 0.614 | 2.749 | 0.006** |
| $X_T$ | 0.730 | 3.273 | 0.001** |
| $X_U$ | −0.537 | −2.408 | 0.016* |
| $X_V$ | 0.507 | 2.270 | 0.024* |
| $X_R \times X_V$ | 0.477 | 2.135 | 0.033* |
| $X_U \times X_V$ | 0.531 | 2.380 | 0.018* |

Notes: Standard errors are approximately 0.22 for each effect coefficient. The error standard deviation is 5.67. Thus, a Cohen's $d$ of 0.2 would correspond to a regression coefficient of $0.2 \times 5.67/(+1 - (-1)) = 0.567$
*$p < 0.05$; **$p < 0.01$

**Table 3** Fitted values from the saturated and parsimonious models

| Components | Effect-coded components | | | | | | Saturated model | | Parsimonious model | |
|---|---|---|---|---|---|---|---|---|---|---|
| Included | $X_R$ | $X_S$ | $X_T$ | $X_U$ | $X_V$ | Cost | Estimate | SE | Estimate | SE |
| $R+T+V$ | +1 | −1 | +1 | −1 | +1 | $1100 | 3.33 | 1.25 | 2.37 | 0.58 |
| $T+V$ | −1 | −1 | +1 | −1 | +1 | $900 | −0.35 | 1.39 | 0.19 | 0.60 |
| $R+V$ | +1 | −1 | −1 | −1 | +1 | $500 | 0.79 | 1.20 | 0.91 | 0.58 |
| $V$ | −1 | −1 | −1 | −1 | +1 | $300 | −0.64 | 1.29 | −1.27 | 0.59 |
| $R+T$ | +1 | −1 | +1 | −1 | −1 | $900 | 2.32 | 1.25 | 1.47 | 0.60 |
| $T$ | −1 | −1 | +1 | −1 | −1 | $700 | 0.73 | 1.20 | 1.19 | 0.58 |
| $R$ | +1 | −1 | −1 | −1 | −1 | $300 | −1.32 | 1.44 | 0.01 | 0.60 |
| None | −1 | −1 | −1 | −1 | −1 | $100 | 0.15 | 1.20 | −0.27 | 0.58 |

Notes: Estimate=estimated cell effectiveness and SE=standard error of the estimate. The SEs for the saturated model would be the same for each cell under balanced cell sizes, but in fact differ slightly due to different cell sizes in the simulated dataset, caused by simulated random missingness

For simplicity, Table 3 shows fitted values only for the cells that are most likely to be under consideration (those having factors $S$ and $U$ set to off). There are eight such cells in Table 3, reflecting two possibilities for factor $R$ (on or off), times two for $T$ and times two for $V$. The values in Table 3 are calculated from the fitted coefficients in Table 2. For example, the condition setting factors $R$, $T$, and $V$ to on has $x_R = +1$, $x_S = -1$, $x_T = +1$, $x_U = -1$, and $x_V = +1$. Thus, using the predictive model in Table 2 and recalling that factor $S$ is being ignored, the expected value for the response in this cell is

$$\begin{aligned}\hat{\mu}_{R,T,V} &= \beta_0 + \beta_R x_R + \beta_T x_T + \beta_U x_U + \beta_V x_V + \beta_{RV} x_R x_V + \beta_{UV} x_U x_V \\ &= 0.038 + (0.614)(+1) + (0.730)(+1) + (-0.537)(-1) + (0.507)(+1) \\ &\quad + (0.477)(+1)(+1) + (0.531)(-1)(+1) = 2.372.\end{aligned}$$

Some of the fitted values may seem surprising at first. Notice that though factor $V$ has a positive main effect (+0.507), the condition with only component $V$ included has a poorer expected response (−1.27) than the condition with no components included at all (−0.268). Specifically,

$$\begin{aligned}\widehat{\mu}_V &= 0.038 + (0.614)(-1) + (0.730)(-1) + (-0.537)(-1) + (0.507)(+1) \\ &\quad + (0.477)(-1)(+1) + (0.531)(-1)(+1) = -1.270\end{aligned}$$

but

$$\begin{aligned}\widehat{\mu}_{none} &= 0.038 + (0.614)(-1) + (0.730)(-1) + (-0.537)(-1) + (0.507)(-1) \\ &\quad + (0.477)(-1)(-1) + (0.531)(-1)(-1) = -0.268.\end{aligned}$$

This happens because of interactions (see the interaction plots in Fig. 5). Recall that ANOVA main effects are really averages of simple effects and can be very different from the individual simple effects when large interactions are present.

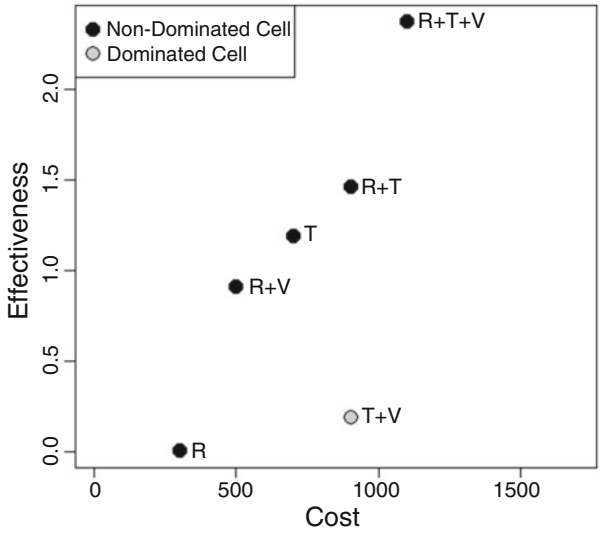


**Fig. 6** Plot of effectiveness (estimated from parsimonious model) by cost

## *5.5 Plotting Cell Predicted Effectiveness Against Cell Cost*

The fitted values from the parsimonious model are plotted against cost in Fig. 6. For simplicity, conditions that had negative estimated effectiveness or that were ruled out already are not shown. The $T + V$ condition is shown to be dominated, in that it is less effective than another condition that is no more expensive. The $V$ condition actually has a negative effectiveness estimate, so it is also dominated because even providing nothing would be less expensive and more effective. The other six cells are not dominated, so any of them could be a rational choice, depending on the researcher's exact goals.

If the goal is to find the most effective intervention that costs no more than $C_{\max}$, then one could draw a vertical line on the plot at cost $= C_{\max}$ and then choose the non-dominated point that is as close as possible to $C_{\max}$ but still to the left of $C_{\max}$. For example, if $C_{\max} = 500$, then one should choose the $R + V$ condition; if $C_{\max} = 1000$, then one should choose the $R + T$ condition.

If the goal is to find the most cost-effective intervention in terms of kilograms lost per dollar spent, then a different approach is required, as illustrated in Table 4. Arithmetically, one could simply divide the cost by the effectiveness for each cell to get the estimated price of losing 1 kilogram under each proposed treatment combination; this ratio is much like *ICER*, and the combination of components used in a given condition could be called cost-effective if the ratio is less than $\lambda$. In Table 4, the $R + T + V$ condition is found to be the most cost-effective in this sense because of its very large effectiveness, despite its relatively large cost. The

**Table 4** Cost-effectiveness and effectiveness-cost ratios

| Included | Cost (in hundreds of dollars) | Effectiveness (kg lost) | Cost/effectiveness | Effectiveness/cost |
|---|---|---|---|---|
| $R + T + V$ | 11 | 2.37 | 4.64 | 0.22 |
| $T + V$ | 9 | 0.19 | 47.4 | 0.02 |
| $R + V$ | 5 | 0.91 | 5.49 | 0.18 |
| $V$ | 3 | −1.27 | N/A | N/A |
| $R + T$ | 9 | 1.47 | 6.12 | 0.16 |
| $T$ | 7 | 1.19 | 5.88 | 0.17 |
| $R$ | 3 | 0.01 | 300 | 0.00 |
| None | 1 | −0.27 | N/A | N/A |

Notes: In this table, cost is measured in hundreds of dollars. Effectiveness is measured in kilograms lost. Therefore, the cost-effectiveness ratio is expressed in hundreds of dollars per kilogram lost, and the effectiveness-cost ratio is in kilograms lost per hundred dollars

increase in effectiveness is more than proportional to the increase in cost relative to other treatment options.

One could also divide effectiveness by cost; this ratio tells the estimated weight loss per dollar in a given cell. It is desirable for the ratio of effectiveness to cost to be large (or equivalently, for the ratio of cost to effectiveness to be small; either form makes sense, but of course it is important to be consistent when using them). Recall that Figs. 1, 2, and 3 designate cost to be the *y*-axis following economists' tradition, while Figs. 4, 5, and 6 designate effectiveness to be the *y*-axis following experimentalists' tradition. In Figs. 4, 5, and 6, the effectiveness-to-cost ratio is the slope of a line from the origin to a given point, and a decision-maker would want this slope to be as "steep" (i.e., as high in the positive direction) as possible to reflect high effectiveness per cost unit. In contrast, a steep slope from the origin (control) to the plotted (experimental) treatment point in Figs. 1, 2, and 3 would reflect a high incremental cost per incremental effectiveness unit, which is undesirable. One caution when using ratios is that subtracting a constant from the numerator and denominator of a ratio changes the overall ratio. Thus, it makes a difference whether the numerator and denominator are *total* effectiveness and *total* cost relative to doing nothing, or simply *incremental* effectiveness and *incremental* cost relative to the least expensive cell, treated as a kind of control condition. This rather subtle issue is described further in the following section.

## *5.6 Effectiveness Versus Control or Versus Nothing?*

As mentioned earlier, an intervention can only be effective or cost-effective in comparison to some alternative, not simply in a vacuum. In the context of the figures, if one wishes to draw a comparison line through the condition in order to compare its slope to λ, another endpoint is needed to draw the line. One might draw the line from

zero cost and zero effectiveness (representing not being treated at all). Alternatively, the other endpoint could be the cell with all components set to off (which might have some overhead cost and some effectiveness due to aspects of treatment that are given to all participants for practical or ethical reasons). For example, suppose that the fitted value for the "none" cell in Table 3 had been approximately 1 instead of approximately 0. That is, suppose that patients may lose a kilogram or so simply from being given the minimal level of education and positive attention, which was given to all participants in the study. If this loss is judged to be real, and not just a Hawthorne effect or regression to the mean, then it is conceivable that the "none" cell might be the most cost-effective. That is, it might be judged that a given budget could cause more total population weight loss by giving the minimal treatment to a very large number of people rather than giving any treatment augmentations to anyone. Simple or inexpensive interventions can sometimes have surprising and important effects, and even individually small effects may be meaningful when applied to large groups (e.g., the aspirin example in Rosnow & Rosenthal, 2003; see Sullivan & Feinn, 2012, for an opposing assessment).

For simplicity, the simulated example used in this section was constructed so that the "none" cell has very low cost and essentially zero effectiveness, so this decision does not have to be considered here. In other words, in this example, questions about total cost and absolute effectiveness (measured from zero) are practically the same as questions about *incremental* cost and effectiveness (measured from the "none" cell).

## 5.7 Interpreting Borderline Statistical Significance When Constructing the Parsimonious Model

This constructed example did not consider a possible case in which a given factor was not quite statistically significant, but suggested a beneficial effect. For example, suppose that the regression coefficient estimate for $X_S$ had been 0.40 with a $p$-value of 0.08, instead of 0.07 with a $p$-value of 0.75. In a conclusion-priority analysis, it could not be claimed that component S had been established effective at the 0.05 level. A researcher probably could not publish a paper arguing that S was shown to be effective, with only this level of evidence. However, in a decision-priority analysis within the context of MOST, it might make sense to keep factor S in the model. This is because if $X_S$ is not included in the parsimonious model, the contribution of component S will be assumed to be exactly zero for purposes of drawing the cost-effectiveness plot. Thus, every cell with S set to on will automatically be considered dominated, unless the cost of S is exactly zero. However, if S were very inexpensive or the willingness to pay was high, then eliminating it from consideration solely on the basis of the 0.05 threshold might not be rational (see Claxton, 1999). Instead, it might be wiser to include $X_S$ in the parsimonious model so that its estimated benefits can be openly compared to its cost.

A component whose effect-coded factor is included in the parsimonious model can still be eliminated from the final recommended intervention if it is judged not to be cost-effective; but a component omitted from the parsimonious model does not get another chance to be selected, even if it is very inexpensive. For this reason one should be somewhat careful about omitting the main effects of factors when constructing the parsimonious model in the optimization phase of MOST.

This reasoning would seem to suggest that borderline-significant *interactions* should also be included. However, there is some need for caution here. In a $2^5$ complete factorial design, there are 31 effects, including 5 main effects, 10 potential two-way interactions, 10 potential three-way interactions, 5 potential four-way interactions, and 1 potential five-way interaction. In a linear model with a balanced design, each of these effects has, in theory, an independent 5% chance of being declared significant even if its true effect is exactly zero. If a higher alpha level is used, then this chance will be even higher. Thus, some interactions will probably be found significant, likely one or more of the 16 high-order interactions (i.e., 3-, 4-, or 5-way), which are notoriously difficult to interpret. Allowing more and more interactions would bring the so-called parsimonious model closer and closer to being saturated and increase the risk of overfitting, especially since the same data are being used for model selection and estimation. This reflects the bias-variance tradeoff: leaving out an interaction risks bias, but including it risks increasing random noise. Further research is needed as to which decision rules are most reasonable here. Inspired by the ideas of the sparsity principle, hierarchical ordering principle, and heredity principle (Myers & Montgomery, 1995; Wu & Hamada, 2009), it might be reasonable to require a stricter significance standard for interactions whose parent main effects are not significant. For example, if the evidence is ambiguous, the investigator might be more skeptical about an $R \times S$ interaction than an $R \times T$ interaction, given that $S$ has a negligible main effect, but $R$ and $T$ have significant effects. The investigator would probably not be as skeptical about an $R \times U$ interaction because $R$ and $U$ are both significant, even though the effect of $R$ is negative. The investigator might also be more skeptical about an $R \times S \times T \times U$ interaction than an $R \times S$ interaction. Nonetheless, if there is solid evidence of a non-negligible interaction, then it must be included in the model in order to get reasonable predictions, and this may mean that its parent main effects and their lower-order interactions need to be included.

Cohen, Cohen, West, and Aiken (2003), in their textbook on regression analysis, give different advice about interactions for different situations. Each piece of advice is reasonable in some situations, but considered together they may be confusing. Depending on the relationship of the interactions to the hypothesis of most interest in the study, they might recommend testing a given interaction at a higher alpha level than normal (e.g., if the investigator wishes to be more confident in asserting that the interaction is not present; p. 373), a lower alpha level than normal (i.e., with a multiplicity correction; p. 296), or not at all (e.g., a high-order interaction between covariates of little interest in an observational study). A recommendation of how this differing advice applies to MOST requires further research; in particular, Monte Carlo simulation studies may be informative here.

## 5.8 *Options for Handling Uncertainty in the Selection of Terms or the Estimation of Effects*

A final limitation of the approach presented in this example is that statistical uncertainty, particularly model uncertainty, was not taken into account in making decisions. Even using the parsimonious model, the confidence intervals for the individual condition (cell) means are wide, as shown in Fig. 7. In fact, it could be argued that the confidence intervals shown in Fig. 7 should be drawn even wider, because they ignore possible error introduced by false negatives when choosing which coefficients to include in the parsimonious model. The issue of variable selection uncertainty and possible ways of correcting for it have been discussed in the statistical literature (see Chatfield, 1995; Claeskens & Hjort, 2008). One might also argue that adjustment for multiple comparisons should be made, as the researcher is basically making pairwise comparisons among the conditions of interest. However, these objections are beside the point, as even without adjustments it is clear that many of the pairwise differences are not statistically significant. The sample size for the factorial experiment had been reasonable for detecting small- to medium-sized main effects, but nowhere near large enough for making confident pairwise comparisons of 32 different cells. Nonetheless, it is a condition (i.e., cell), not a factor, which is being chosen for future implementation.

The best way to handle this dilemma is not yet clear and requires further research. Harrington (1981) collapses across nonsignificant interactions, as is recommended here, and also attempts to correct for possible overoptimism in experimental results by adjusting all of the effectiveness estimates downward by a fixed percentage. However, other than this he treats effectiveness estimates as point values, not

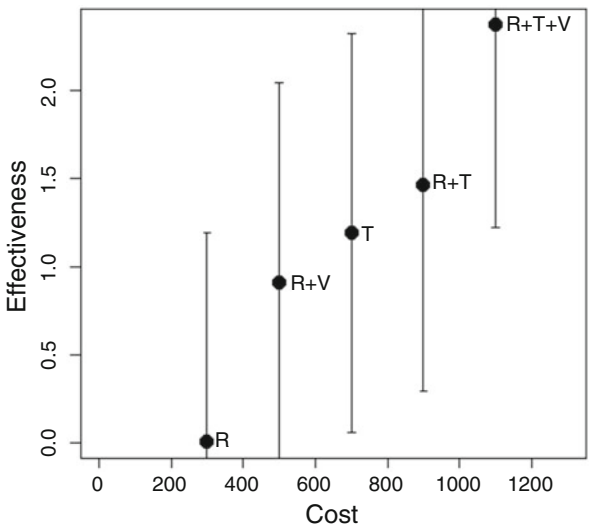


**Fig. 7** Effectiveness-cost plot for non-dominated conditions, with 95% confidence bands added

confidence intervals; that is, he basically ignores the standard errors of the fitted values. Although it seems naïve, this may actually be the most reasonable approach in some cases. It would be at least as naïve to argue that the least expensive cell available must be used, merely on the argument that the others have not been proven better in a pairwise sense. As mentioned earlier, failure to reject a null hypothesis in the conclusion-priority sense does not necessarily mean that the null hypothesis should be accepted as true when making decisions. It would be possible to say that the experiment was inconclusive and that a newer one with fewer factors or more participants needs to be done. This is mathematically reasonable but may or may not be practically feasible. For example, in the context of MOST, an investigator might or might not have sufficient funding to do multiple factorial experiments in the optimization phase before proceeding to the evaluation phase.

Chapman et al. (2014) point out that sampling variability may affect which conditions are estimated to be on the Pareto frontier. They suggest that in addition to creating a plot like Fig. 6 to show which cells are non-dominated according to their fitted values (which are the middle of their confidence intervals and hence best guesses at what their future effectiveness will be), one should also create a plot that compares worst-case estimates of their effectiveness (such as the lower endpoints of their confidence intervals). Generalizing this idea, one might also perform some kind of sensitivity analysis to account for the effects of model selection by randomly perturbing the data in some way or changing the model assumptions in some way, to see which conditions tend to remain non-dominated and which are sometimes dominated, but this is beyond the scope of this chapter.

It is reassuring that when a factorial experiment is being used as a screening experiment, it is likely to be treated as somewhat exploratory and followed up by further experimentation—either a more focused factorial experiment in the context of response surface optimization (see Myers & Montgomery, 1995) or a confirmatory two-condition RCT in the context of MOST (see, e.g., Collins et al., 2016). Thus, final recommendations will not be based on the factorial experiment alone. However, if the results of the factorial experiment are to be used in a different way from that envisioned in MOST (e.g., to determine which medical procedures should be permitted or reimbursed), then it becomes much more important to immediately acknowledge uncertainty, and it may be best to allow physicians and patients to select from several good options instead of claiming to have found a single best.

## *5.9 Application to MOST*

The approach of constructing a parsimonious model and using it to estimate condition effectiveness may be very important for MOST users for whom cost-effectiveness is the optimization criterion of interest. It is also a possible analytic approach even if effectiveness itself is the quantity to be optimized, without regard to cost. It also has some similarities to the approach recommended by Collins and

colleagues (2014), who emphasized making decisions based on plots of significant main effects and significant interactions (hence, indirectly, on fitted values from a parsimonious model).

No analysis method can guarantee that the condition selected by analyzing the factorial experiment will be optimal in the literal sense, that is, the very best possible out of all those that are available; the main challenge here is that there is not enough information in a realistically sized clinical trial to make confident pairwise comparisons among dozens of cells. However, the use of a parsimonious model is the most straightforward way in this situation to obtain the effectiveness estimates necessary to make principled choices about cost-effectiveness. Bayesian methods might provide an attractive alternative to significance testing for achieving a bias-variance tradeoff and making decisions about utility, but this has not yet been explored in the context of MOST; this is discussed further at the end of this chapter.

## 6 Dealing with Multiple Dimensions of Effectiveness or Cost

This chapter so far has been focused on research intended to find ways to balance two separate goals—high effectiveness and low cost. Considering two goals instead of one requires the introduction of new concepts. These concepts include identifying a Pareto frontier, as well as distinguishing between scientific questions about expected outcomes on the one hand and subjective decisions about weighing outcome versus cost on the other. However, in some cases there are more goals than just two. There might be more than one dimension of effectiveness, more than one dimension of cost, or both.

When effectiveness has more than one dimension, there are sometimes said to be multiple outcomes. For example, a healthy eating and exercise program may be hypothesized to improve mood and reduce weight. If the program is effective, it ought to do both things. However, one could imagine a program (perhaps involving grueling exercise and extreme caloric restriction) that reduces weight but makes mood worse. More realistically, some interventions might be more beneficial for one outcome than for another. For example, some character education or health education programs in schools are intended to help young people avoid more than one kind of unhealthy behavior. A particular intervention or component might have a significant effect on, for example, binge drinking—but little or no effect on smoking. In fact, two candidate intervention packages might each have statistically significant effects on a different outcome, but one is more effective for the first outcome, while the other is more effective for the second.

**Table 5** Effectiveness of two hypothetical interventions on two outcomes

| Intervention | Daily calorie reduction (kcal) | Daily activity increase (minutes) |
|---|---|---|
| A | 300 | 20 |
| B | 200 | 30 |
| Control | 10 | 2 |

## *6.1 An Example with Two Outcomes of Interest*

Because of the added complexity associated with discussing effectiveness on two outcome variables instead of one, let us simplify somewhat by returning to an RCT setting rather than a factorial experiment. Suppose that a three-condition RCT is comparing two interventions (A and B) with an information-only control condition. The outcomes of interest are reduction of calorie consumption and increase in daily physical activity, both evaluated as a change from baseline after 3 months. Suppose that the results were as shown in Table 5. Also suppose for simplicity that all of the pretest-to-posttest changes except those for the control condition were statistically significant at the 0.05 level and that all of the pairwise differences between conditions on the outcomes were also statistically significant. Last, suppose for now that difference in cost is negligible, and focus only on the two dimensions of effectiveness.

From a conclusion-priority point of view, the results in Table 5 would be sufficient for writing an interesting and informative scientific paper. Both interventions are effective on both outcomes, but Intervention A has more of an effect on calorie consumption, and Intervention B has more of an effect on daily activity. From a decision-priority point of view, however, the analysis is clearly not over. If someone were to ask the investigator to make a recommendation about which of the interventions would be best for them to use, what would be a reasonable reply?

Once again, the data alone cannot answer this question. If calorie consumption is more important, then Intervention A is better. If daily activity is more important, then Intervention B is better. It would be easy to reply that both outcomes are important. However, if it is considered necessary to make a single decision about which intervention to choose, then it is implicitly necessary to make some kind of statement about the relative importance of the goals. Because the relative value of different outcomes is now being considered, it may be necessary to move beyond cost-effectiveness into some form of cost-benefit analysis (for reviews of the latter, see Messonnier & Meltzer, 2003). This introduces many new possibilities and questions for analysis, but in the context of this chapter, it is only possible to introduce a few basic ideas.

Fortunately, as before, just because the decision is partly subjective does not mean it is entirely arbitrary; methods have been developed for combining information from multiple outcomes. These methods can be divided into two general approaches: *weighting* approaches (combining all of the outcomes into one common scale, based on some system of prioritization) and *partial ordering* or *set-valued* approaches (weeding out conditions that are clearly inferior and then providing

a list of all of the conditions that remain; Laber, Lizotte, & Ferguson, 2014). Both approaches have parallels in the previous sections about balancing cost and effectiveness. This is no coincidence; the question of cost-effectiveness is just a special case of the question of multiple outcomes, with cost as an outcome that is reverse-scaled (lower is better).

Let us first consider weighting approaches. In a way, they provide the most straightforward approach to an unambiguous decision. However, they raise the question of how to choose the weights. There are several ways to choose the weights; they are explored in the following three subsections.

## 6.2 *Standardized Weights for Combining Equally Important Variables*

The basic idea of the weighting approach is to turn two variables into one. In our example, let calorie reduction be $E_1$ and activity increase be $E_2$. The decision-maker must specify some weights $a_1$ and $a_2$ that reflect his or her own interpretation of the relative importance of the goals. Given these weights, the optimal decision for that decision-maker will be one that maximizes the quantity $a_1E_1 + a_2E_2$. This weighted sum is a new constructed variable that is used in place of $E_1$ and $E_2$. The advantage of this approach is that it applies a simple procedure that gives a straightforward answer. The disadvantage, of course, is the necessity of specifying $a_1$ and $a_2$.

The weights $a_1$ and $a_2$ must accomplish two tasks. First, they must standardize the outcomes somehow so that they can be made comparable even if they were originally on two different and incompatible scales of measurement. Second, they must express whether, and to what extent, one outcome is more or less important to the decision-maker.

To see the importance of standardization, consider the simplest approach to determining the weights: automatically making them equal (i.e., $a_1 = a_2 = 1$). In the example in Table 5, this approach would be illogical because the two outcomes are not on the same scale of measurement: one is in calories and the other is in minutes, and there is no meaningful way to add calories to minutes. Intervention A ($300 + 20 = 320$ points) appears to win over Intervention B ($200 + 30 = 230$ points), but only because of an arbitrary difference in scale of measurement. If activity had been measured in seconds, Intervention B would have won. It is necessary, therefore, to standardize the endpoints somehow. One option would be to divide each by its estimated standard deviation, although this arguably gives a disadvantage to outcomes that are more dispersed or harder to measure precisely. Another would be to scale each from 0 for the worst possible value on each outcome to 1 for the best. Defining the worst and best levels is not always possible, but when it is feasible, it can add flexibility to the analysis (see the discussion of additive versus multiplicative weighting in Chapman et al., 2014).

As important as standardization is, it is not the only function of the weights, because the variability or spread of a variable does not in itself express the practical worth of changing the variable. Even if the outcomes are expressed on the same scale, they might not be of equal interest: as an extreme example, risk of becoming infected with the common cold virus and risk of becoming infected with HIV can both be expressed as percentages or odds ratios. Thus, the choice to give both variables equal weight after standardization will sometimes be too simplistic.

Because the weight indicates not only statistical variability but also importance to the decision-maker, the optimization of a weighted combination of multiple different outcomes is essentially cost-benefit analysis. In this simple example, if being infected with HIV is considered to be $1,000,000 worth of undesirability, and being infected with a common cold is considered to be $100 worth of undesirability, then it would be rational to spend 10,000 times as much to prevent an HIV infection as to prevent a common cold. The difficulty, of course, is not in the math but in finding a rational way to measure the desirability or undesirability of outcomes on a common scale; methods for doing so are discussed in Messonier and Meltzer (2003).

## 6.3 *Using Weights from a Formula or Model to Construct a New Outcome Variable*

In some cases, there is a somewhat less subjective way to set the weights. As a simple example, a researcher might translate daily activity increase into calories. If the added activities burn an average of 240 kcal per hour (i.e., 4 per minute), the researcher could set $a_1 = 1$ and $a_2 = 4$. Then Intervention A will result in a total reduction of $300 + (4 \times 20) = 380$ calories per day, and Intervention B will result in a total reduction of $200 + (4 \times 30) = 320$ calories per day. From this perspective, Intervention A may be judged to be superior. This approach has at least the appearance of objectivity but still involves some risk of oversimplification. For example, exercise has other benefits besides burning calories (e.g., it may improve mood and attention), and these benefits are not being taken into account.

A more ambitious approach would be to weight both outcomes according to their predicted contribution to a later health outcome, based on a predictive model. This might involve using information from past literature to predict an outcome that was not measured during the duration of the study. For example, the investigator may have a predictive model that can estimate the 5-year risk of heart disease given the calorie intake and daily exercise estimates obtained from the study. Thus, if the overall goal is to reduce heart disease, the constructed variable of heart disease risk can be substituted for the two observed outcomes. Now there is only one effectiveness variable (i.e., estimated heart disease risk), and whichever intervention does a better job at improving it can therefore be considered the most effective intervention. Of course, this also involves making assumptions and simplifications. First, if the study used a 3-month follow-up to calculate 5-year risk, then the dubious

assumption is being made that behavior will remain constant over time. If it were feasible, it might be more defensible to have a longer follow-up and either actually observe whether heart disease occurs or not or at least do a more comprehensive assessment of risk. Second, there are many different *constructed* outcomes that could be chosen (e.g., risk of diabetes, cancer, depression, etc.), so an argument would need to be made for why only heart disease is being considered.

Another type of constructed outcome would combine multiple potential health conditions. For instance, the outcome might be death by any cause (whether predicted using a model or observed at a follow-up). However, even this may not be a full use of all of the information available. First, statistical power and precision for an analysis with death as an outcome may sometimes be very limited, due to the rare and binary nature of the response. Second, there are many health and social conditions that seldom directly cause death but do cause suffering and disability, and a truly comprehensive assessment would arguably take these into account somehow.

## *6.4 Quality-Adjusted Life-Years Are a Special Constructed Variable*

Because it is important both to save lives and to reduce suffering during life, cost-effectiveness and cost-benefit analyses in medical contexts often employ the idea of a quality-adjusted life-year (QALY). QALYs are an attempt to provide a combined measure of life and health. A treatment that provides an extra year of fully healthy life is considered to be worth more QALYs than a treatment that provides an extra year of life but with serious remaining mobility restrictions or pain. QALYs are a basic tool in many effectiveness and cost-effectiveness studies. Both the advantage and the difficulty in using QALYs are that they require that multiple aspects of life functioning and experience be considered together, not just one objectively observable outcome such as weight loss or smoking cessation. This requires setting numerical values on human experiences, a form of cost-benefit analysis. For instance, most people might intuitively agree that deafness is worse than acne but not as bad as death; but a quality-of-life scale requires actually assigning numbers to the undesirability of each. Considerable research has been done on using patients' reported preferences to create quality-of-life ratings (see Dasbach & Teutsch, 2003; National Academies, 2016; Petrillo & Cairns, 2008; Ramsey et al., 2005), but it is still a difficult and controversial endeavor, in part because it is difficult to take into account the fact that judgments may differ by person (see Holmes, 2013; Smith, 1987). However, QALYs are often used in comparative effectiveness and health economics (e.g., Li, Zhang, Barker, Chowdhury, & Zhang, 2010) because of their ability to incorporate many outcomes into a single measure for decision-making purposes. A PubMed search for "qalys or 'quality adjusted life years'" conducted on June 9, 2016, resulted in 11,624 articles. Nonetheless, QALYs will not be relevant or helpful to all researchers using MOST. Researchers in education or in primary

prevention may not have enough data to link their observable outcomes (such as short-term behavior changes) to disease states or may be interested in aspects of life functioning other than health per se (e.g., school performance). Thus, QALYs are certainly not the only valid form of weighted effectiveness measure for cost-effectiveness or cost-benefit analyses. Furthermore, in some situations, it will not be possible to find a weighting scheme that fully reflects the priorities and beliefs of all stakeholders. This possibility is described further in the following subsection:

## *6.5 When Value Judgments Are Unavoidable, There May Be No Consensus on Weights*

For some researchers who are comparing different candidate interventions, it will not make much practical difference whether the outcome is deaths prevented, years saved, or QALYs gained. For example, consider a school-based intervention to prevent the onset of tobacco use. In this example, all of the participants are young and are unlikely to develop severe consequences of tobacco use until many years after the end of the study. Thus, using any of these three overall health endpoints would require long-term predictions about how many students given a particular intervention would go on to develop smoking-related lung cancer, heart disease, or emphysema much later in life. Researchers might disagree on how best to *predict* such outcomes, even in the unlikely case in which they could agree upon how to weight them.

Alternatively, different stakeholders on the research team might have different ideas about what outcomes are important. One researcher might be interested in preventing smoking altogether (so that 1 cigarette per week is practically as bad as 50), while another is content with harm reduction (so that 1 is worse than none but better than 50).

For these reasons, the outcomes observed in a study might have to be treated as "apples and oranges": they are each important, but it is too difficult to get a consensus on which is more important or by how much. This suggests another approach to considering multiple outcomes, which is to abandon the search for a single best choice and instead simply provide a short list of reasonable choices.

The Pareto frontier (i.e., the set of non-dominated options) provides a conceptual and methodological framework for handling these situations (Chapman et al., 2014). As discussed earlier, a dominated option is one that is less effective in terms of one dimension of effectiveness than some other option, but not more effective on any other dimension of effectiveness. For example, in Table 5, row A does not dominate row B, and row B does not dominate row A, because they both have the advantage in one dimension of effectiveness or the other. They each dominate the control condition, because they each outperform it on both dimensions. However, the effectiveness of A cannot be directly compared with that of B without first making one's priorities clearer (e.g., setting weights). That is, condition A is better

**Table 6** Effectiveness of two hypothetical interventions on two outcomes

| Intervention | Daily calorie reduction (kcal) | Daily activity increase (minutes) | Cost to administer (hundreds of dollars) |
|---|---|---|---|
| A | 300 | 20 | 10 |
| B | 200 | 30 | 20 |
| Control | 10 | 2 | 5 |

than the control condition, and condition B is better than the control condition, but it is not possible to clearly state whether condition A is better, worse, or equal to B. This is an example of what mathematicians sometimes call a semiorder or partial ordering (see Luce, 1956): a set of objects, such that some pairs in the set can be put in order to determine which are greater or less, but other pairs in the set cannot necessarily be unambiguously compared.

Recall that in the examples shown in Figs. 4 and 6, with a single effectiveness measure and a single cost measure, one can first eliminate those options that are dominated by other options and then simply report the cost and effectiveness of the remaining options in the form of a table or plot. The job of choosing the desired tradeoff between cost and effectiveness is deferred to the final decision-maker (e.g., a physician, school superintendent, or health management organization director).

As noted earlier, effectiveness and cost can simply be seen as two different outcomes, the latter of which happens to be reverse-coded. Thus, the example in Table 5 is not really different from the preceding examples. In particular, the first two rows of Table 5 could be seen as a Pareto frontier, with the last row being dominated.

As mentioned earlier, one could combine a cost measure with two or more effectiveness measures (as in Table 5) or combine two or more cost measures (e.g., financial cost and participant burden) with one or more effectiveness measures. It would become more difficult to draw a plot like Fig. 4 or 6 because such a plot would now have to be three-dimensional. However, one could still provide a table. For example, Table 6 is an expanded version of Table 5, including a hypothetical cost measure.

## *6.6 Advantages and Disadvantages of Weights*

A philosophical advantage to a partial ordering approach is that it recognizes human subjectivity and does not pretend that a single answer will be best for everyone. Of course, this is also its disadvantage: it does not provide a single answer, and the work of making the final decision is left to someone in the future. The problem becomes more complex as more outcomes (i.e., more dimensions of effectiveness and/or cost) are considered. For example, in Table 6, because of the addition of an extra column, now none of the rows are dominated, not even the control condition (because, although it is less effective on both outcome measures, it still has the advantage

of being less expensive). A cynical assessment would be that nothing has been accomplished by such a study; the researcher began by being undecided between three options and ends in exactly the same place. However, a fairer assessment would be that much information has been gained but that different readers could prioritize this information differently in making their own choices. Graphical and statistical approaches are also available for exploring a set of choices further, such as by describing the different sets of weights that would be required in order to make one option emerge as the best (Chapman et al., 2014; Lizotte, Bowling, & Murphy, 2012).

In summary, some kind of prioritization, usually through weighting, is unavoidable in order to come to an unambiguous single answer. However, in the absence of such prioritization, one can still report some information in terms of a list of non-dominated conditions.

## 6.7 Constraints Instead of Weights

There are also other options besides weights for expressing one's priorities. One would be to try to optimize the effectiveness upon one outcome under the constraint that the effectiveness upon each of the other outcomes is in some region deemed acceptable. This is mathematically the same thing as maximizing a single effectiveness variable subject to a cost constraint. The difference is that instead of keeping cost below some $C_{\max}$, the constraint would be to keep the effectiveness on the constrained outcome above some $E_{\min}$. In this case, both outcomes are treated as important, but in different ways; one of them is considered a variable which *must be kept above a certain level*, while the other is considered a variable which *should be made as high as possible but does not have an absolute required cutoff*.

## 6.8 Importance of Multiple Outcomes

Multiple outcomes of interest are common in real-world empirical studies, and multiple goals or criteria to be optimized are common in real-world decisions. One general conclusion, then, is that the answer depends on the question: the optimal choice depends on the definition of what is to be optimized. Deciding exactly what to optimize is often at least as difficult as deciding how to optimize it. This is especially evident when deciding how to divide very scarce and very necessary resources. As a very extreme example, consider a famous controversy in medical ethics: how to determine the recipients of donated organs (discussed by American Medical Association, 1995; Courtney & Maxwell, 2009; Hoffmaster & Hooker, 2013, among others). From the perspective of maximizing population health in life-years (whether or not quality-adjusted), young and otherwise healthy people ought to be given priority over people who would not live very long even if they do

receive the organ. However, from the perspective of human equality, every patient ought to have an equal chance of receiving what they need to survive and should not be discriminated against on the basis of age or preexisting health conditions. If the optimization criterion could be unambiguously specified as either equality or effectiveness, then it would be easy to apply empirical data to compare different distribution strategies. However, in practice decision-makers need either to choose between one goal or the other or else find a workable way of compromising between them without perfectly optimizing either, and there is no numerical way to take the need for subjective human judgment out of this question.

The phenomenon of multiple outcomes can arise in unexpected ways in many contexts, even when the primary outcome may at first seem clear. For example, Laber et al. (2014) considered the problem of choosing a dosage of medication that balanced symptom relief with side effects. The side effects could be thought of as a second outcome, or equivalently as a non-monetary cost. Different patients might have different preferences for symptom relief versus side effect avoidance.

In summary, comparing candidate interventions across multiple outcomes involves an extension of the ideas and methods proposed earlier for cost-effectiveness. In fact, a study of cost and a single dimension of effectiveness can be reconceptualized as a special case of a study of two outcomes. Determining which intervention is best requires prioritizing the different outcomes—either by setting weights for each outcome or by designating one outcome to be optimized subject to a constraint on the other outcome. However, if no consensus can be reached about how to set the priorities, the researcher can still remove obviously inferior choices and provide a shortened list of best candidates to decision-makers. Let us now consider how these ideas apply to analysis of factorial experiments specifically.

## *6.9 Implementation in Factorial Designs and MOST*

How does this extend to analyzing the results of factorial experiments in the MOST framework? Presumably the basic process described earlier in the context of the simulated $2^5$ factorial experiment could be followed: obtain estimates of each dimension of effectiveness and cost for each cell, and use these to choose a cell. As before, the estimates should come from some kind of parsimonious model that takes advantage of the factorial structure by considering only significant main effects and significant interactions, rather than comparing the empirical means of each cell as the effectiveness estimates. It is not necessary that the same predictive model be used for each outcome. That is, a particular factor or interaction could be significant in predicting one outcome but not significant in predicting another. (It might be possible to borrow information across the different predictive models, but this has not yet been fully explored.) Once the predictive models are constructed, the non-dominated conditions can be reported in a table such as Table 6.

If there are many cells and several outcomes of interest, it might be difficult to weed out all of the dominated conditions manually. However, this can be done easily

with a computer. Compare each condition $c$ to each other condition $c'$, in order. If $c$ is equal to or worse than $c'$ on every outcome and is actually worse on at least one outcome, then $c$ is dominated (i.e., it can be ruled out as inferior to some other choice). Otherwise, c is not dominated (i.e., it is a viable choice).

These comparisons would presumably be made using the estimates from the parsimonious model, but would not necessarily have to be treated as pairwise tests for significant mean differences. That is, just as before, a researcher might have to give a tentative recommendation in favor of $c'$ even if it is not, strictly speaking, statistically significantly better than $c$. This is particularly likely to be the case in the optimization phase of MOST, because a condition must usually be chosen without having adequate power for all pairwise comparisons. Of course, even though a decision might have to be made, it is still healthy to remember that the uncertainty exists. That is, the decision to choose cell $c'$ in that situation should not necessarily be interpreted as a finding that cell $c'$ has been scientifically demonstrated to be definitively superior. Rather, it would simply be a case of doing the best one can with the information available while acknowledging that one's knowledge is not perfectly precise.

## 7 Discussion and Open Questions

This chapter has presented a framework, based on existing literature in engineering and health economics, for making decisions about cost-effectiveness for behavioral interventions using a factorial experiment. This is especially relevant for investigators planning a MOST investigation with cost-effectiveness as the optimization criterion, although many of the ideas are relevant in other research settings as well. Unfortunately, it is impossible at this time to provide a "cookbook" of procedures or a consensus on correct interpretations. This is partly because uncertainty and subjectivity are inextricably part of the decision-making process, and partly because little methodological research has been done on how to combine the factorial experiment literature with the cost-effectiveness analysis literature.

A recurring difficulty faced in this chapter was the necessity of choosing a condition (a cell in the factorial experiment) as most cost-effective, despite having very limited power for pairwise comparisons of cells. The recommended approach was to rely on effectiveness estimates from a parsimonious model containing only the significant main effects and interactions. This is not the only possible solution; another is Bayesian analysis, particularly with model averaging. This possibility is briefly described further at the end of this chapter.

Several open questions were encountered but not resolved in this chapter. Some of the most prominent are listed below. They may provide useful starting points to future methodological research.

## 7.1 How Should These Methods Be Adapted to Diverse Real-World Settings?

This chapter has not distinguished between efficacy (the usefulness of an intervention estimated under optimal conditions in a well-funded, well-controlled academic study) and effectiveness (the usefulness of an intervention estimated in realistic field conditions; see, e.g., Singal, Higgins, & Waljee, 2014). Ideally, cost-effectiveness decisions should be based on estimates of real-world effectiveness and cost (Drummond et al., 2005). However, these might not always be available. This chapter has also not focused on the estimation of mean cost in the realistic case in which treatment costs vary from participant to participant. For more information on such analyses, see Haddix et al. (2003), Petrou and Gray (2011), and Stevens et al. (2003). This chapter has also not considered how effectiveness (and thus cost-effectiveness) depends on personal characteristics (moderators; see, e.g., Simon & Perlis, 2010). The questions of what works best on average and of what works best for different individuals are distinct, but each line of research may provide relevant information for the other if understood correctly (see Garber & Tunis, 2009). Further research is also needed on how best to estimate cost from a factorial experiment, in which factors may interact not only in their effect on the outcome variable but also in their effect on cost (e.g., due to potentially overlapping overhead costs).

## 7.2 How and When Should Equity or Social Concerns Be Addressed?

A possible limitation of cost-effectiveness analysis is that it does not directly take into account equity, that is, fairness (National Academies, 2016; Smith, 1987). Some populations (e.g., very elderly people, people who have severe disabilities, or people who are addicted to drugs) may be more difficult to help than others, making interventions to help them appear to be less cost-effective than interventions for other populations; but this does not mean that they are less worthwhile. For reasons like this, it is often recommended that cost-effectiveness be considered together with other criteria, rather than on its own, when making funding decisions (Carroll, 2014; Drummond et al., 2005; National Academies, 2016). It has been argued by proponents of cost-benefit analysis that moral ideas such as equity can be incorporated by, for example, treating justice versus injustice as another category of benefit or cost (see, e.g., Zerbe, 2004); however, this question is beyond the scope of this chapter. Many empirical researchers, who will already have a specific target population in mind and are not entrusted with society-wide fiscal decisions, may not have to specifically address equity in their studies. However, exploratory analyses to compare cost-effectiveness among different subgroups in a sample (at least gender or broad age categories) might be worthwhile if sample size allows.

## 7.3 How Can Sample Size and Power Planning Be Done for Comparisons of Fitted Values?

Perhaps the largest limitation of the current chapter is that it is currently not possible to offer power or sample size guidance for cost-effectiveness comparisons in a factorial experiment. Sample size planning for factorial experiments is usually presented in terms of power for detecting statistically significant main effects and interactions. It is not immediately clear how this translates into the probability of finding the best or near-best cell, and more research is needed (the recent work of Ertefaie et al., 2015, may be a helpful starting point). Existing resources, usually in one-way ANOVA, suggest that the needed sample size depends on one's goal and assumptions (see, e.g., Bechhofer et al., 1995; Gupta & Hsu, 1978). A researcher who wishes to identify the best-performing cell with high confidence will need a large sample, especially if it is assumed that some cells might be very close to each other in importance. Therefore, the analyses described in this chapter might be considered more exploratory than confirmatory and aimed at finding a good intervention rather than the best intervention possible. This is relevant whether one is studying cost-effectiveness or simply effectiveness. It is sometimes argued in the literature (e.g., Clegg, Scott, Sidhu, Hewitson, & Waugh, 2001) that, where feasible, it would be preferable to base cost-effectiveness decisions on multiple studies rather than single ones.

## 7.4 What if the Factors Are Continuous Instead of Categorical?

This chapter only considers categorical (in fact, dichotomous) factors, rather than continuous ones. Factors with more than two categories are not very different conceptually, although they define an even larger number of conditions and have implications for statistical power and precision. Continuous numerical factors (such as the dose provided of various therapeutic agents) can also be involved in attempts to optimize an intervention or process for effectiveness or for cost-effectiveness, and they need to be treated in a different way. This often involves response surface designs and related analyses (see, e.g., Dillon, 1966; Dunn, 2016; Myers & Montgomery, 1995; Wu & Hamada, 2009). Dong et al. (2011) describe such an analysis in an agricultural study with multiple factors and outcomes. This author is unaware of any current examples in the social or behavioral sciences. However, continuous factors in the form of dosage levels are found in medical research, including some work that has focused on studying tradeoffs between multiple outcomes (such as effectiveness versus toxicity of cancer medications, with the latter being a kind of cost; see literature review in Thall & Cook, 2004).

## 7.5 What if the Outcomes Are Binary Rather than Continuous?

Consider a situation in which a researcher has a choice between multiple variables that are conceptually related. For example, the ultimate goal in a smoking cessation study is total cessation with no relapse. However, even a reduction in the number of cigarettes smoked would still be of some potential health value. Such reduction might be easier for the participants to achieve than total abstinence, as well as perhaps providing more statistical power because it is no longer a binary outcome. Therefore, it might be useful to consider both binary smoking status and daily cigarette count as two separate outcomes in a cost-effectiveness analysis, even though they are obviously related. How best to do this is a question for future methodological research. Also, it may be important not to take this idea too far. For example, a researcher might try to get by with a smaller sample size by simply establishing that a treatment reduces craving, rather than that it increases quit rate. Affecting a short-term predictor such as craving might be easier than preventing relapse itself. However, subjective craving in itself might not be enough to do a helpful cost-effectiveness analysis, if only because insurers' willingness-to-pay for simply reducing a subjective feeling might be low unless behavior change could also be demonstrated. If the actual outcome of interest is binary, then it is probably better to work with the binary outcome, even though the sample size required might be rather high. Drummond and colleagues (2005) further discuss the issues of handling intermediate endpoints and modeling their relationship to future endpoints.

## 7.6 What if Retention Varies by Condition or Is an Outcome of Interest in Its Own Right?

The idea of a Pareto frontier or of tradeoffs between multiple outcomes may provide a useful alternative way of looking at certain situations that appear at first to involve only a single outcome. For example, in studies involving cessation of smoking or other drug use, participants sometimes drop out of both treatment and the study before the scheduled end of treatment. Simply removing these individuals listwise from the analysis would likely give an overly optimistic estimate of remission rates, because relapse might be a cause or consequence of leaving the study. Different methods for imputing the final use state have been compared (Hedeker, Mermelstein, & Demirtas, 2007; Jackson, White, Mason, & Sutton, 2014). However, as an adjunct to these methods in analyzing this kind of data, it might be useful to consider retention in treatment and final cessation as two different outcomes of interest, rather than the former being simply a nuisance in analyzing the latter.

Admittedly, retention as a goal in itself is usually less impactful than cessation. After all, simply keeping a participant busy in a program that is not benefitting him or her is likely to be useless from the viewpoint of effectiveness, and actually

deleterious from the viewpoint of cost-effectiveness. Thus, from the perspective of evaluating the effectiveness of an existing intervention program, failure is failure, and it does not matter whether the addict relapses while in treatment or instead leaves treatment early and then relapses. However, from the perspective of understanding *how* the intervention works or fails and how it can be improved, the two ways of relapsing are different. For example, a very intrusive and restrictive treatment program might have higher dropout rates due to burden, but also higher eventual success rates for those who do not drop out, relative to a less rigorous program. Assuming that this finding was not simply due to selection bias (e.g., the rigor of the program merely causes unmotivated or unprepared individuals to leave, without actually improving the outcomes of the individuals who remain), it would provide some interesting information for a conclusion-priority paper. It might also provide, in an exploratory sense, ideas for future decision-priority research. For example, retention information might be useful in deciding whether a given intervention can best be improved by becoming less intensive or more intensive. In a factorial context, if it is found that one treatment component increases dropout but improves success rate for those who do not drop out, then this may suggest to a researcher that it needs to be augmented with a component aimed specifically at improving retention. There might also be psychological and medical situations in which program retention has some benefits even if patients do not recover on the primary endpoint; for example, being in the program might have some secondary benefit such as social support, monitoring, and the potential for crisis intervention

## 7.7 How Can Bayesian Methods Be Used to Help Estimate Cell Effectiveness or to Help Make Decisions?

Recall that in order to construct the parsimonious model for estimating the effectiveness of each cell in a factorial experiment, many interactions had to be tested. This created many opportunities for Type I or Type II errors, any of which could change the estimates, the standard errors, and the conclusions. An alternative option would be to use a shrinkage approach, whether informative Bayesian (Simon & Freedman, 1997), empirical Bayesian (Chen & Meeter, 1999), or frequentist (Li & Lin, 2009), perhaps in the form of model averaging (Claeskens & Hjort, 2008), in order to allow certain interactions to be included in a partial or restricted way rather than entirely in or entirely out. This might provide a more stable and interpretable answer. Instead of one supposedly best model arrived at after a series of error-prone dichotomous decisions, it is possible to work with a weighted combination of multiple possible models with differing posterior probabilities. Bayesian theory has natural connections with decision theory (see, e.g., Rice, 1995, for an introduction), especially because it can be used to estimate the posterior risk of each possible intervention, a measure which can combine cost information with the distribution of the expected outcome variable. Decision theory is considered relevant to engineering and design optimization in many fields (e.g., Davis, Kisiel, &

Duckstein, 1972). Some researchers have made suggestions about how to apply it to RCTs (Lindley, 1998; Longford, 2016; Manski & Tetenov, 2016), and presumably it could be extended to factorial experiments. However, this chapter has focused on the use of significance tests to create a parsimonious model, both because that approach will be somewhat familiar for many social and behavioral scientists and because it resembles the approaches used in some of the past published methodological research on optimization using factorial experiments (see Collins et al., 2014; Harrington, 1981; Taneja & Dudewicz, 1987). Readers may refer to Simon and Freedman (1997) for a discussion of potential benefits of using Bayesian shrinkage instead of dichotomous significance tests; Wu and Hamada (2009) also give a brief description of an approach to Bayesian model selection.

## 8 Conclusion

In conclusion, cost-effectiveness analyses of data from a factorial optimization trial can involve complexities and difficult subjective decisions but can provide a great deal of information. Researchers and practitioners regularly make decisions about what can be done with limited time and money, and these decisions should be made in a careful, mindful, and data-driven way. For this reason, further research on how to integrate data from factorial experiments into effectiveness and cost-effectiveness decisions is of great importance.

**Acknowledgment** The author thanks Dr. Linda Collins, Dr. Inbal Nahum-Shani, Dr. Daniel Max Crowley, and Dr. Kari Kugler for helpful discussions and comments on previous versions of this chapter and Amanda Applegate for proofreading assistance. Graphics were created using the R software package (https://www.r-project.org/).

## References

American Medical Association Council on Ethical and Judicial Affairs. (1995). Ethical considerations in the allocation of organs and other scarce medical resources among patients. *Archives of Internal Medicine, 155*, 29–40.

Appelbaum, B. (2011, February 16). As U.S. agencies put more value on a life, businesses fret. *New York Times*. Retrieved from http://www.nytimes.com

Bechhofer, R. E., Santner, T. J., & Goldsman, D. M. (1995). *Design and analysis of experiments for statistical selection, screening, and multiple comparisons*. New York, NY: Wiley.

Belli, P. C., Bustreo, F., & Preker, A. (2005). Investing in children's health: What are the economic benefits? *Bulletin of the World Health Organization, 83*, 777–784.

Bensink, M. E., Eaton, L. H., Morrison, M. L., Cook, W. A., Curtis, R. R., Kundu, A., … Doorenbos, A. Z. (2013). Cost effectiveness analysis for nursing research. *Nursing Research, 64*(4), 279–285.

Belsink, M. E., Eaton, L. H., Morrison, M. L., Cook, W. A., Curtis, R. R., Gordon, D. B., Kundu, A., … Doorenbos, A. Z. (2013). Cost effectiveness analysis for nursing research. *Nursing Research, 62*(4), 279–285.

Bierman, K. L., Henrichs, B. S., Welsh, J. A., Nix, R. L., & Gest, S. D. (2017). Enriching preschool classrooms and home visits with evidence-based programming: Sustained benefits for low-income children. *Journal of Child Psychology and Psychiatry, 58*, 129–137.

Black, W. C. (1990). The CE plane: A graphic representation of cost-effectiveness. *Medical Decision Making, 10*, 212–214.

Bradley, E. H., Canavan, M., Rogan, E., Talbert-Slagle, K., Ndumele, C., Taylor, L., & Curry, L. A. (2016). Variation in health outcomes: The role of spending on social services, public health, and health care, 2000–09. *Health Affairs, 35*, 760–768.

Carroll, A. E. (2014, December 15). Forbidden topic in health policy debate: Cost effectiveness. *New York Times*. Retrieved from http://www.nytimes.com

Chapman, J. L., Lu, L., & Anderson-Cook, C. M. (2014). Process optimization for multiple responses utilizing the Pareto front approach. *Quality Engineering, 26*, 253–268.

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society, Series A, 158*, 419–466.

Chen, C.-H., & Meeter, D. (1999). Empirical Bayes analyses of two-level factorials. FSU Statistics Report M920.

Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging*. New York, NY: Cambridge University Press.

Claxton, K. (1999). The irrelevance of inference: A decision-making approach to the stochastic evaluation of health care technologies. *Journal of Health Economics, 18*, 341–364.

Claxton, K., Lacey, L. F., & Walker, S. G. (2000). Selecting treatments: A decision theoretic approach. *Journal of the Royal Statistical Society, A, 163*(2), 211–225.

Clegg, A., Scott, D. A., Sidhu, M., Hewitson, P., & Waugh, N. (2001). A rapid and systematic review of the clinical effectiveness and cost-effectiveness of paclitaxel, docetaxel, gemcitabine and vinorelbine in non-small-cell lung cancer. *Health Technology Assessment, 5*(32), 1–195.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Collins, L. M. (2018). *Optimization of behavioral, biobehavioral, and biomedical interventions: The multiphase optimization strategy (MOST)*. New York, NY: Springer.

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*, 202–224.

Collins, L. M., Kugler, K. C., & Gwadz, M. V. (2016). Optimization of multicomponent behavioral and biobehavioral interventions for the prevention and treatment of HIV/AIDS. *AIDS and Behavior, 20*(Supplement 1), S197–S214.

Collins, L. M., Trail, J. B., Kugler, K. C., Baker, T. B., Piper, M. E., & Mermelstein, R. J. (2014). Evaluating individual intervention components: Making decisions based on the results of a factorial screening experiment. *Translational Behavioral Medicine, 4*, 238–251. https://doi.org/10.1007/s13142-013-0239-7

Courtney, A. E., & Maxwell, A. P. (2009). The challenge of doing what is right in renal transplantation: Balancing equity and utility. *Nephron Clinical Practice, 111*, c62–c68.

Cox, D. R., & Snell, E. J. (1989). *Analysis of binary data* (2nd ed.). London, UK: Chapman & Hall.

Crowley, D. M., Hill, L. G., Kuklinski, M. R., & Jones, D. E. (2014). Research priorities for economic analyses of prevention: Current issues and future directions. *Prevention Science, 15*, 789–798.

Crowley, M., & Jones, D. (2017). A framework for valuing investments in a nurturing society: Opportunities for prevention research. *Clinical Child and Family Psychology Review, 20*, 87–103.

Dasbach, E. J., & Teutsch, S. M. (2003). Quality of life. In A. C. Haddix, S. M. Teutsch, & P. S. Corso (Eds.), *Prevention effectiveness: A guide to decision analysis and economic evaluation* (pp. 77–91). New York, NY: Oxford University Press.

Davis, D. R., Kisiel, C. C., & Duckstein, L. (1972). Bayesian decision theory applied to design in hydrology. *Water Resources Research, 8*, 33–41.

Day, R. W., & Quinn, G. P. (1989). Comparisons of treatments after an analysis of variance in ecology. *Ecological Monographs, 59*, 433–463.

Dillon, J. L. (1966). Economic considerations in the design and analysis of agricultural experiments. *Review of Marketing and Agricultural Economics, 34*(2), 64–75.

Dong, W., Qin, J., Li, J., Zhao, Y., Nie, L., & Zhang, Z. (2011). Interactions between soil water content and fertilizer on growth characteristics and biomass yield of Chinese white poplar (Populas tomentosa Carr.) seedlings. *Soil Science and Plant Nutrition, 57*, 303–312.

Drummond, M. F., Sculpher, M. J., Torrance, G. W., O'Brien, B. J., & Stoddart, G. L. (2005). *Methods for the economic evaluation of health care programmes* (3rd ed.). New York, NY: Oxford University Press.

Dunn, K. (2016). *Process improvement using data* (online textbook). Retrieved from http://learnche.org/pid

Eeren, H. V., Schawo, S. J., Scholte, R. H. J., Busschbach, J. J. V., & Hakkaart, L. (2015). Value of information analysis applied to the economic evaluation of interventions aimed at reducing juvenile delinquency: An illustration. *PLoS One, 10*(7), e0131255.

Embry, D. D., & Biglan, A. (2008). Evidence-based kernels: Fundamental units of behavioral influence. *Clinical Child and Family Psychology Review, 11*, 75–113.

Ertefaie, A., Wu, T., Lynch, K., & Nahum-Shani, I. (2015). Identifying a set that contains the best dynamic treatment regimes. *Biostatistics, 17*(1), 135–148.

Fenwick, E., O'Brien, B. J., & Briggs, A. (2004). Cost-effectiveness acceptability curves - facts, fallacies and frequently asked questions. *Health Economics, 13*(5), 405–415.

Fuller, N. R., Colagiuri, S., Schofield, D., Olseon, A. D., Shrestha, R., Holzapfel, C., … Caterson, I. D. (2013). A within-trial cost-effectiveness analysis of primary care referral to a commercial provider for weight loss treatment, relative to standard care – An international randomised controlled trial. *International Journal of Obesity, 37*, 828–834.

Garber, A. M., & Tunis, S. R. (2009). Does comparative-effectiveness research threaten personalized medicine? *New England Journal of Medicine, 360*, 1925–1927.

Gift, T. L., Haddix, A. C., & Corso, P. S. (2003). Cost-effectiveness analysis. In A. C. Haddix, S. M. Teutsch, & P. S. Corso (Eds.), *Prevention effectiveness: A guide to decision analysis and economic evaluation* (pp. 156–177). New York, NY: Oxford University Press.

Gupta, S. S., & Hsu, J. C. (1978). Subset selection procedures with special reference to the analysis of two-way layout: Application to motor-vehicle fatality data. *SIGSIM Simulation Digest, 10*, 68–72.

Guyll, M., Spoth, R., & Crowley, D. M. (2011). Economic analysis of methamphetamine prevention effects and employer costs. *Journal of Studies on Alcohol and Drugs, 72*(4), 577–585.

Haddix, A. C., Corso, P. S., & Gorsky, R. D. (2003). Costs. In A. C. Haddix, S. M. Teutsch, & P. S. Corso (Eds.), *Prevention effectiveness: A guide to decision analysis and economic evaluation* (pp. 53–76). New York, NY: Oxford University Press.

Harrington, L. W. (1981). Economic analysis of $2^4$ factorial agronomic experiments. CIMMYT Economics Training Note, 1981. Retrieved from http://repository.cimmyt.org/xmlui/bitstream/handle/10883/855/25127.pdf

Hasler, M., & Böhlendorf, K. (2013). Multiple comparisons for multiple endpoints in agricultural experiments. *Journal of Agricultural, Biological, and Environmental Statistics, 18*, 578–593.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York, NY: Springer.

Hedeker, D., Mermelstein, R. J., & Demirtas, H. (2007). Analysis of binary outcomes with missing data: Missing = smoking, last observation carried forward, and a little multiple imputation. *Addiction, 102*, 1564–1573.

Hoffmaster, B., & Hooker, C. (2013). Tragic choices and moral compromise: The ethics of allocating kidneys for transplantation. *Milbank Quarterly, 91*(3), 528–557.

Holmes, D. (2013). Report triggers quibbles over QALYs, a staple of health metrics. *Nature Medicine, 19*, 248.

Holzer, H. J., Whitmore Schanzenbach, D., Duncan, G., & Ludwig, J. (2008). The economic costs of childhood poverty in the United States. *Journal of Children and Poverty, 23*(1), 1–4.

Hsu, J. C. (1984). Constrained simultaneous confidence intervals for multiple comparisons with the best. *Annals of Statistics, 12*, 1136–1144.

Jackson, D., White, I. R., Mason, D., & Sutton, S. (2014). A general method for handling missing binary outcome data in randomized controlled trials. *Addiction, 109*(12), 1986–1993.

Kolata, G. (1992, November 24). Ethicists struggle to judge the 'value' of life. *New York Times*. Retrieved from http://www.nytimes.com/

Komro, K. A., Flay, B. R., Biglan, A., & The Promise Neighborhoods Research Consortium. (2011). Creating nurturing environments: A science-based framework for promoting child health and development within high-poverty neighborhoods. *Clinical Child and Family Psychology Review, 14*, 111–134. https://doi.org/10.1007/s10567-011-0095-2

Kozol, J. (2005). Confections of apartheid: A stick-and-carrot pedagogy for the children of our inner-city poor. *Phi Delta Kappan, 85*, 265–275.

Kuehl, R. O. (2000). *Design of experiments: Statistical principles of research design and analysis* (2nd ed.). Pacific Grove, CA: Brooks/Cole.

Kugler, K. C., Dziak, J. J., & Trail, J. B. (2018). Coding and interpretation of effects in analysis of data from a factorial experiment. In L. M. Collins & K. C. Kugler (Eds.), *Optimization of behavioral, biobehavioral, and biomedical interventions: Advanced topics*. New York, NY: Springer.

Laber, E. B., Lizotte, D. J., & Ferguson, B. (2014). Set-valued dynamic treatment regimes for competing outcomes. *Biometrics, 70*, 53–61.

Li, R., & Lin, D. K. J. (2009). Variable selection for screening experiments. *Quality Technology and Quantitative Management, 6*(3), 271–280.

Li, R., Zhang, P., Barker, L. E., Chowdhury, F. M., & Zhang, X. (2010). Cost-effectiveness of interventions to prevent and control diabetes mellitus: A systematic review. *Diabetes Care, 33*, 1872–1894.

Lindley, D. V. (1998). Decision analysis and bioequivalence trials. *Statistical Science, 13*, 136–141.

Lizotte, D. J., Bowling, M., & Murphy, S. A. (2012). Linear fitted-Q iteration with multiple reward functions. *Journal of Machine Learning Research, 13*, 3253–3295.

Longford, N. T. (2016). Comparing two treatments by decision theory. *Pharmaceutical Statistics, 15*, 387–395.

Luce, R. D. (1956). Semiorders and a theory of utility discrimination. *Econometrica, 24*, 178–191.

Manski, C. F., & Tetenov, A. (2016). Sufficient trial size to inform clinical practice. *Proceedings of the National Academies of Science, 113*, 10518–10523.

Messonnier, M., & Meltzer, M. (2003). Cost-benefit analysis. In A. C. Haddiz, S. M. Teutsch, & P. S. Corso (Eds.), *Prevention effectiveness: A guide to decision analysis and economic evaluation* (pp. 127–155). New York, NY: Oxford University Press.

Miller, B. F., Nowels, M. A., VanderWielen, L. M., & Gritz, R. M. (2016). Mental health parity's limited impact on utilization and access for health plan beneficiaries. Washington, DC: Health Care Cost Institute. Executive summary retrieved from http://www.healthcostinstitute.org/files/HCCI-Issue-Brief-Mental-Health-Paritys-Limited-Impact.pdf

Muench, U., Coffman, J., & Spetz, J. (2016). Does independent scope of practice affect prescribing outcomes, healthcare costs, and utilization? Washington, DC: Health Care Cost Institute. Executive summary retrieved from http://www.healthcostinstitute.org/files/HCCI-Issue-Brief-Independent-Prescribing-Outcomes.pdf

Myers, J. L., & Well, A. D. (2003). *Research design and statistical analysis* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Myers, R. H., & Montgomery, D. C. (1995). *Response surface methodology: Process and product optimization using designed experiments*. New York, NY: Wiley.

Nair, V., Strecher, V., Fagerlin, A., Ubel, P., Resnicow, K., Murphy, S., ... Zhang, A. (2008). Screening experiments and the use of fractional factorial designs in behavioral intervention research. *American Journal of Public Health, 98*, 1354–1359.

National Academies of Sciences, Engineering, and Medicine. (2016). *Advancing the power of economic evidence to inform investments in children, youth, and families*. Washington, DC: The National Academies Press. https://doi.org/10.17226/2348

National Academies of Sciences, Engineering, and Medicine. (2017). *Communities in action: Pathways to health equity*. Washington, DC: The National Academies Press. https://doi.org/10.17226/24624

Pellegrini, C. A., Hoffman, S. A., Collins, L. M., & Spring, B. (2014). Optimization of remotely delivered intensive lifestyle treatment for obesity using the multiphase optimization strategy: Opt-IN study protocol. *Contemporary Clinical Trials, 38*(2), 251–259.

Pellegrini, C. A., Hoffman, S. A., Collins, L. M., & Spring, B. (2014). Optimization of remotely delivered intensive lifestyle treatment for obesity using the multiphase optimization strategy: Opt-IN study protocol. *Contemporary Clinical Trials, 38*(2), 251–259.

Pellegrini, C. A., Hoffman, S. A., Collins, L. M., & Spring, B. (2015b). Corrigendum to "Optimization of remotely delivered intensive lifestyle treatment for obesity using the Multiphase Optimization Strategy: Opt-IN study protocol". *Contemporary Clinical Trials, 45-B*, 468–469. https://doi.org/10.1016/j.cct.2015.09.001

Petrillo, J., & Cairns, J. (2008). Converting condition-specific measures into preference-based outcomes for use in economic evaluation. *Expert Reviews in Pharmacoeconomics and Outcomes Research, 8*(5), 453–461.

Petrou, S., & Gray, A. (2011). Economic evaluation alongside randomised controlled trials: Design, conduct, analysis, and reporting. *BMJ, 342*, d1548. https://doi.org/10.1136/bmj.d1548

Prado, E. L., Sebayang, S. K., Apriatni, M., Adawiyah, S. R., Hidayati, N., Islamiyah, A., … Shankar, A. H. (2017). Maternal multiple micronutrient supplementation and other biomedical and socioenvironmental influences on children's cognition at age 9–12 years in Indonesia: Follow-up of the SUMMIT randomised trial. *Lancet Global Health, 5*, e217–e228.

Ramsey, S., Willke, R., Briggs, A., Brown, R., Buxton, M., Chawla, A., … Reed, S. (2005). Good research practices for cost-effectiveness analysis alongside clinical trials: The ISPOR RCT-CEA task force report. *Value in Health, 8*(5), 521–533.

Rice, J. (1995). *Mathematical statistics and data analysis* (2nd ed.). Belmont, CA: Duxbury.

Rosnow, R. L., & Rosenthal, R. (2003). Effect sizes for experimenting psychologists. *Canadian Journal of Experimental Psychology, 57*, 221–237.

Rothmann, M., Li, N., Chen, G., Chi, G. Y. H., Temple, R., & Tsou, H.-H. (2003). Design and analysis of non-inferiority mortality trials in oncology. *Statistics in Medicine, 22*, 239–264.

Saha, S., Gerdtham, U.-G., & Johansson, P. (2010). Economic evaluation of lifestyle interventions for preventing diabetes and cardiovascular diseases. *International Journal of Environmental Research and Public Health, 7*, 3150–3195.

Sculpher, M. J., Claxton, K., Drummond, M., & McCabe, C. (2006). Whither trial-based economic evaluation for health care decision making? *Health Economics, 15*, 677–687.

Shoemaker, E. Z., Tully, L. M., Niendam, T. A., & Peterson, B. S. (2015). The next big thing in child and adolescent psychiatry: Interventions to prevent and intervene early in psychiatric illnesses. *Psychiatric Clinics of North America, 38*, 475–494.

Simon, G. E., & Perlis, R. H. (2010). Personalized medicine for depression: Can we match patients with treatments? *American Journal of Psychiatry, 167*, 1445–1455.

Simon, R., & Freedman, L. S. (1997). Bayesian design and analysis of two × two factorial clinical trials. *Biometrics, 53*, 456–464.

Singal, A. G., Higgins, P. D. R., & Waljee, A. K. (2014). A primer on effectiveness and efficacy trials. *Clinical and Translational Gastroenterology, 5*, e45. https://doi.org/10.1038/ctg.2013.13

Slavin, R. E. (2006). Shame indeed. *Phi Delta Kappan, 87*, 621–623.

Smith, A. (1987). Qualms about QALYs. *Lancet, 1*(8542), 1134–1136.

Stevens, A. W., O'Hagan, A., & Miller, P. (2003). Case study in the Bayesian analysis of a cost-effectiveness trial in the evaluation of health care technologies: Depression. *Pharmaceutical Statistics, 2*, 51–68.

Sullivan, G. M., & Feinn, R. (2012). Using effect size – Or why the p-value is not enough. *Journal of Graduate Medical Education, 4*, 279–282.

Taneja, B. K., & Dudewicz, E. J. (1987). Selection in factorial experiments with interaction, especially the 2 × 2 case. *Acta Mathematica Sinica New Series, 3*(3), 191–203.

Thall, P. F., & Cook, J. D. (2004). Dose-finding based on efficacy-toxicity trade-offs. *Biometrics, 60*, 684–693.

Tsai, A. G., Wadden, T. A., Volger, S., Sarwer, D. B., Vetter, M., Kumanyika, S., ... Glick, H. A. (2013). Cost-effectiveness of a primary care intervention to treat obesity. *International Journal of Obesity, 37*, S31–S37.

Tuffaha, H. W., Gordon, L. G., & Scuffham, P. A. (2014). Value of information analysis in healthcare: A review of principles and applications. *Journal of Medical Economics, 17*, 377–383.

Van Ryzin, M. J., Roseth, C. J., Fosco, G. M., Lee, Y.-K., & Chen, I.-C. (2016). A component-centered meta-analysis of family-based prevention programs for adolescent substance use. *Clinical Psychology Review, 45*, 72–80.

Vickers, A. J. (2001). The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient: A simulation study. *BMC Medical Research Methodology, 1*, 6.

Vickers, A. J., & Altman, D. G. (2001). Analysing controlled trials with baseline and follow up measurements. *BMJ, 323*, 1123–1124.

Weinstein, M. C., Siegel, J. E., Gold, M. R., Kamlet, M. S., & Russell, L. B. (1996). Recommendations of the panel on cost-effectiveness in health and medicine. *Journal of the American Medical Association, 276*, 1253–1258.

Williams, J. R. (2015). *Medical ethics manual* (3rd ed.). Ferney-Voltaire, France: The World Medical Association Retrieved from http://www.wma.net/en/70education/30print/10medical_ethics/

Wu, C., & Hamada, M. (2009). *Experiments: Planning, analysis, and optimization* (2nd ed.). New York, NY: Wiley.

Wu, K. H., & Cheung, S. H. (1994). Subset selection for normal means in a two-way design. *Biometrical Journal, 36*(2), 165–175.

Zerbe, R. O. (2004). Should moral sentiments be incorporated into benefit-cost analysis? An example of long-term discounting. *Policy Sciences, 37*, 305–318.

Zhang, Z., Nie, L., Soon, G., & Zhang, B. (2014). Sensitivity analysis in non-inferiority trials with residual inconstancy after covariate adjustment. *Applied Statistics, 63*, 515–538.

# Investigating an Intervention's Causal Story: Mediation Analysis Using a Factorial Experiment and Multiple Mediators


**Rachel A. Smith, Donna L. Coffman, and Xun Zhu**

**Abstract** Behavioral, biobehavioral, and biomedical interventions presume a causal story. This causal story is used to create a conceptual model of *why* the intervention caused the observed outcome. Methods are needed to investigate the extent to which changes in the outcome are due to exposure to an intervention. This chapter describes the use of a factorial experiment in which an intervention's components are varied, the mediating mechanisms and outcome are measured, and then an analysis is performed to investigate an intervention's causal story. To illustrate the procedures, a case study examines an intervention guided by the model of stigma communication (MSC; Smith, Commun. Theory 17:462–485, 2007; Stigma communication and health. In T. L. Thompson, R. Parrott, & J. Nussbaum (Eds.), *Handbook of health communication* (2nd ed., pp. 455–468). London, UK: Taylor & Francis, 2011; Commun. Monogr. 79:522–538, 2012). In the case study, we show how to conduct a mediation analysis when four components of an intervention have been manipulated in a $2^4$ factorial experiment ($N = 299$), and, in addition, four hypothesized mediators have been measured. We reflect on how the results provide more precise conclusions about the theory guiding the intervention and opportunities for refinement.

R. A. Smith (✉)
Department of Communication Arts & Sciences, Department of Human Development and Family Studies, and the Methodology Center, The Pennsylvania State University, State College, PA, USA
e-mail: ras57@psu.edu

D. L. Coffman
Department of Epidemiology and Biostatistics, College of Public Health, Temple University, Philadelphia, PA, USA
e-mail: donna.coffman@temple.edu

X. Zhu
Department of Communication Arts & Sciences, The Pennsylvania State University, State College, PA, USA
e-mail: xuz132@psu.edu



L. M. Collins, K. C. Kugler (eds.), *Optimization of Behavioral, Biobehavioral, and Biomedical Interventions*, Statistics for Social and Behavioral Sciences,
https://doi.org/10.1007/978-3-319-91776-4_8

## 1 Introduction

Behavioral, biobehavioral, and biomedical interventions (hereafter referred to primarily as interventions) presume a causal story: the outcome of interest changes as a result of exposure to an intervention. Put differently, exposure to the intervention *causes* the exposed audience to think, feel, or act differently than if they had not been exposed to the intervention.[1] An important corollary is that the use of an intervention presumes that we know *why* our intervention causes an observed change in the outcome.

As pointed out in Chap. 1 of the companion volume (Collins, 2018), the "why" may be built from various sources, such as theory, clinical experience, and practical experience; however, many interventions use theory to design and evaluate their interventions because theory provides "guides to understanding and predicting events in the world about them" (Jaccard & Jaccoby, 2010, p. 3). With respect to the multiphase optimization strategy (MOST), theory provides useful guidance for the preparation phase (see Chap. 2 of the companion volume, Collins, 2018, for more details): what should be included in an intervention, what outcomes can be expected, and why should the intervention produce the expected outcomes (Glanz & Bishop, 2010). For example, a theory may suggest that exposing people to health information describing the risk of contracting an antibiotic-resistant infection will cause them to wash their hands more often.

Often, theory provides more complicated stories, in which exposure to an intervention causes an outcome to change because of an intermediate psychological or physiological process. For example, the theory may state that exposing people to health information about their risk causes them to become afraid, and it is fear that causes people to wash their hands more often.

Some scholars (e.g., Glanz & Bishop, 2010) argue that an intervention is effective to the extent that the rationale (from theory and any other source) guiding the intervention correctly and sufficiently identifies what to include in the intervention, what outcomes to expect from exposure to it, and the causal process(es) by which the exposure results in the expected outcomes. One way to detail an intervention's causal story is to create a conceptual model as described in Chaps. 1 and 2 in the companion volume (Collins, 2018; also, see Chap. 1 in this volume).

A robust test of the intervention's causal story, including mediating mechanisms, has important implications for intervention science. If the intervention produces expected changes in an outcome, but the causal model for why the intervention should work is not supported, then intervention scientists have the opportunity to pause and investigate *why* an intervention is causing an effect on the outcome. There are many reasons why the causal story may not be supported. The theoretical rationale may not generalize to the intervention's context of interest. The theory may

[1]This story assumes that the intervention's outcome can change. For example, if an intervention's goal is to improve how often people wash their hands, then handwashing rates must be able to change.

be flawed and need refinement before it is applied in an intervention (Baranowski, Anderson, & Carmack, 1998). The theory may be valid and reasonable for the context of interest, but the abstract theoretical concepts may not have been instantiated well in the intervention, which requires a return to the preparation phase of MOST. Investigating these types of failures allows intervention scientists to advance theory, which, in turn, offers useful guidance for future practice. Regardless of the ultimate reasons for failure of an intervention's causal model, practically, it is better to know if an intervention may be making change for the wrong reasons before it is scaled up and distributed widely. The current state of the art, however, typically approaches effectiveness by demonstrating that an intervention has produced measurable change on a predefined outcome relative to a control group (Michie & Abraham, 2004), without demonstrating that the claims of mediating processes have been supported (Baranowski et al., 1998; Noar, 2007).

Evaluating the causal story of an intervention also provides an opportunity to investigate unintended iatrogenic effects. Critics may argue that an existing intervention causes an outcome to get worse or causes a different or an unintended negative outcome to occur. The effectiveness of an intervention is sometimes considered as "the level of good over harm that a program achieves under typical real-world conditions" (Flay, 1986, p. 451). A full investigation needs to demonstrate that the intervention does more good than harm in the targeted population and to reveal any paths through which unintended effects arise (Cho & Salmon, 2007; Flay, 1986). Existing theory may provide guidance on why exposure to a particular kind of intervention may result in iatrogenic effects, such as creating a health stigma. For example, a theory may explain why health-risk information written in a particular way causes the exposed audience to stigmatize those living with an antibiotic-resistant infection, instead of or in addition to causing them to wash their hands more.

Different methods may be needed for different purposes. One method might be used to show that an intervention has the intrinsic qualities it is claimed to possess (e.g., the intervention included scary images and content); another method might show that exposure to the intervention causes the outcome as predicted (e.g., exposure to intervention created fear, which increased handwashing); and a third method might investigate whether changes in the outcome are caused by exposure to the intervention (e.g., versus other alternative explanations). Previous research has focused on issues related to the intrinsic qualities of an intervention (e.g., O'Keefe, 2003). This chapter focuses on methods to investigate the causal story of how the intervention causes the outcome to change.

The methods needed to investigate the extent to which changes in the outcome are due to exposure to an intervention are not self-evident or trivial. In this chapter, we explore mediation analysis combined with a pretest/posttest experimental design as a powerful way to make causal inferences. For intervention scientists using MOST, there is also interest in optimizing the intervention. In many cases, we can decompose an intervention into multiple components. An intervention may include multiple components for two reasons. First, investigators may have included multiple components to trigger one intermediary psychological or physiological

process, which ultimately shapes the outcome of interest. For example, to increase perceived risk, investigators may include a frightening image, content about high rates of infection, and content about possible morbidity and mortality resulting from having an antibiotic-resistant infection in health information. Second, investigators may use different components to trigger different intermediary psychological or physiological processes that affect the outcome. For example, investigators may include one component to increase perceived risk of getting an antibiotic-resistant infection and a different component to increase perceived efficacy about a recommended strategy (e.g., handwashing) to avoid getting an infection. As described in MOST, decomposing an intervention into components makes it possible to use factorial experiments to investigate the components' relative effectiveness and interactions between components.

A holistic test of the intervention's causal model, then, can be carried out by conducting a factorial experiment of the components and then investigating the resulting effects with a mediation analysis. Although combining the benefits of factorial experiments with those of mediation analysis is intuitive, there currently is little guidance in the methodological literature about how it should be conducted. The purpose of this chapter is to discuss how to conduct a mediation analysis when several components of an intervention have been manipulated in a factorial experiment and several hypothesized mediators have been measured. This kind of procedure can be used to investigate whether an invention works as we expect and to investigate whether it causes unintended consequences. We will illustrate the procedures with a case study of an unintended consequence: a health stigma created as a result of a communication intervention. In this chapter, we review causal models and mediation analysis, review factorial designs, and then present an illustration of our suggested procedures with the case study.

## *1.1 Causal Models and Mediation Analysis*

Mediation analysis is a set of statistical procedures used to assess how well empirical data support a causal model that involves mediation (MacKinnon, 2008). In the simplest case, mediation analysis involves one independent variable (e.g., the intervention), one mediating variable, and one outcome variable. Figure 1 shows the relations among these three variables. As Fig. 1 shows, exposure to the intervention (the independent variable) causes change in a mediator (*a* path) that, in turn, causes change in the outcome (*b* path). The path represented by $c'$ is the effect of the intervention on the outcome that does not go through the mediator.

Estimating the paths in Fig. 1 enables researchers to address the following questions about an intervention's effects: (1) Did the mediator influence the outcome as hypothesized (*b* path)? (2) Did the intervention affect the mediator (*a* path)? (3) How much of the relation between the intervention and the outcome was *not* explained by the mediator ($c'$ path)? Addressing these three questions provides a detailed assessment of which parts of a causal model are supported by the empirical

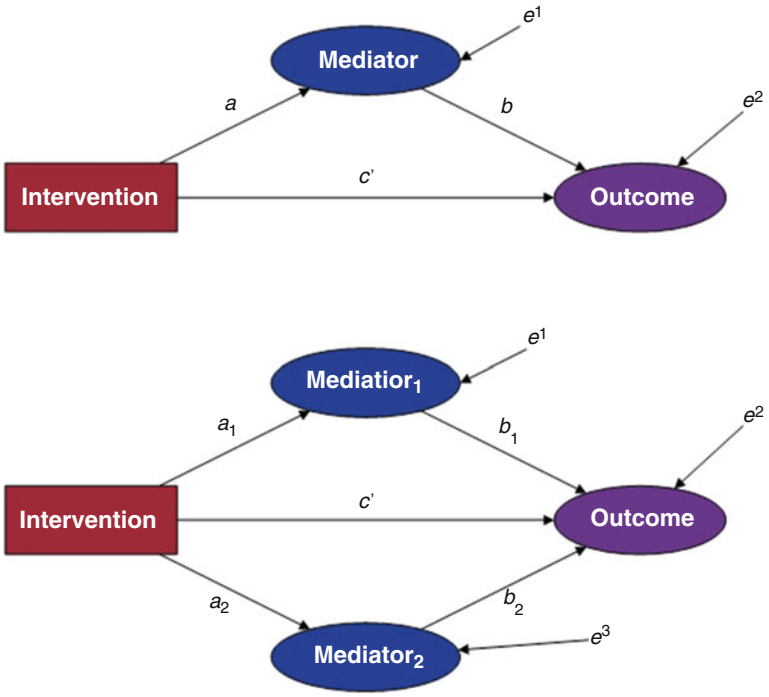


**Fig. 1** (**a**, **b**) The top figure shows the basic mediation model, with one intervention, one mediator, and one outcome. The bottom figure shows a mediation model with one intervention, two mediators, and one outcome. *Arrows* represent relations among variables: *a* represents relation of the intervention to the mediator (associated with action theory, Chen, 1990; MacKinnon et al., 2002); *b* represents relation of the mediator to the outcome adjusted for the intervention (associated with conceptual theory, Chen, 1990; MacKinnon et al., 2002); and $c'$ represents the relation of the intervention to the outcome adjusted for the mediator. The symbol $e^1$ (and $e^3$ in **b**) represents residuals in the mediator, and $e^2$ represents residuals in the outcome variable. The mediator and outcome are presented in an oval to suggest that they may be latent variables, whereas the intervention may be observed

data, which parts are not supported, and where the causal model may be incomplete. For example, a very small direct effect ($c'$ path) would suggest that the proposed mediating psychological or physiological processes are sufficient to explain the relation between exposure to the intervention and changes in the outcome of interest, whereas a large direct effect would suggest that additional mediators are needed to explain the observed relation and/or the intervention directly affects the outcome.

One approach to considering mediation has been discussed in reference to conceptual and action theories (Chen, 1990; MacKinnon, Taborga, & Morgan-Lopez, 2002). The focus is first on the *b* path, that is, on identifying mediators hypothesized to be causally related to the outcome (MacKinnon et al., 2002). This is referred to as the conceptual theory (Chen, 1990). The second focus is to create an intervention that causes change in the mediator (MacKinnon et al., 2002), referred to as action theory (Chen, 1990).

## 1.2 Multiple Mediators

An intervention may affect several mediators, which may have different influences on the outcome. For example, the desired effect on the outcome could be obtained through some mediators, and an effect in the opposite direction could be obtained through others.

There are currently two options for research scenarios involving an intervention and multiple mediators. The first option, which is by far the more common, is to refrain from breaking the intervention into components, instead manipulating a single independent variable with levels representing exposure to the entire *intervention package* (see Chap. 1 in the companion volume) or no exposure to the intervention package. In this approach an experiment would be conducted with random assignment to the intervention (or not), and all of the hypothesized mediators would be measured (MacKinnon et al., 2002). Figure 1b shows the relations with one intervention package, two mediators, and one outcome.

The results from mediation analyses of such an experiment can show which mediator is a more powerful predictor of the outcome (by comparing estimates from the *b* paths). Thus, the findings can enable researchers to make informed decisions about the mediators (MacKinnon et al., 2002). If the mediator effects (*b* paths) are not significant, but the intervention still influences the outcome (the $c'$ path), then attention needs to be paid to uncover an unmeasured mediational process (or processes; MacKinnon et al., 2002). The findings cannot, however, provide insights into which specific components of the intervention package caused changes in which mediators.

## 1.3 Intervention Packages

Most interventions have multiple "content" components (see Chap. 1 of the companion volume, which provides a great example of a hypothetical intervention). This chapter focuses on health information as the intervention, such as information designed to persuade people to think or feel a certain way about a health topic (e.g., antibiotic-resistant infections are scary) or to engage in healthy behaviors (e.g., more handwashing). Health information may be designed with components composed of verbal (e.g., word and grammar) and nonverbal content (e.g., images, colors) to evoke the same mechanism. For example, as guided by theories of health-threat messaging (e.g., extended parallel process model; Witte, 1992), intervention scientists may include multiple forms of verbal content about the severity of antibiotic-resistant infections (e.g., morbidity, curability, and mortality) and the risk of contracting them (e.g., prevalence rates, especially among people similar to the target audience). They might also include images of infections and maps of prevalence rates. All of this verbal and nonverbal content instantiates a health-threat component in the health information, which is predicted to cause fear, which, in turn, is predicted to motivate people to take action to avoid the health threat

(Witte, 1992). Indeed, health messages created with this kind of guidance have been related to a variety of health topics (e.g., gun safety, sexually transmitted infections, cyberbullying; see Roberto, 2013 for a discussion). What is unclear for any particular intervention is how many instances of the verbal and nonverbal content were necessary to trigger the psychological process that induces action. Theories offer little information about whether more instances of the same kind of content (more verbal and nonverbal health-threat content) cause health outcomes to occur more quickly or more dramatically because they create a stronger psychological or physiological response—or whether there is a maximum amount of repeated exposure that could be offered before people become desensitized to it.

Clarifying what is needed to achieve a particular goal for a health outcome has many benefits. It would allow us to more accurately anticipate the effect size of an intervention as well as the speed at which the effect would occur. Practically, including more verbal and nonverbal content in an intervention can create greater demands on readers' time, attention, and processing, thereby increasing audience' fatigue toward an intervention (cf. So, Kim, & Cohen, 2017). Each word and image carry a cost in production (e.g., capturing an image, hiring speakers) and dissemination (e.g., longer and bigger press releases, brochures, pamphlets, and posters cost more money).

Clearly articulating the conceptual model for the health information also helps to refine the intervention's causal story. As highlighted in Chap. 1, "one component may be enhanced or undermined by the presence or level of one or more other components . . . A component may be effective only when certain other components are included or may be effective *unless* combined with certain other components" (p. 5). Some theories suggest that the intervention must include multiple components to create an effect. In health messaging, a theory may posit that multiple kinds of content are needed to create a particular sort of *frame* that privileges some explanations and outcomes over others (e.g., Entman, 1993); and without all of the content, the frame is not created. Furthermore, it is unclear whether each component causes an outcome to occur through the same or different mediating mechanisms. Knowing what parts of an intervention are needed, and in what amounts, has benefits to optimizing an intervention. It also has benefits for theory development. Through optimization, for example, we might discover nonlinear relationships in which too many components trigger a level of the mediating psychological and physiological mechanism that makes the health outcome worse. Research (e.g., Shen, 2016) has shown such a nonlinear relationship with fear, where there is a "sweet spot" in which fear best motivates behavioral action; too much or too little results in less action.

Instead of keeping the intervention package as a whole, researchers can choose to use a factorial experiment in which components of the intervention are manipulated, and participants are randomly assigned to a combination of them. The factorial experiment has several advantages over separate studies varying one component at a time. First, the factorial experiment treats the components in the intervention package as parts of a theory, instead of investigating each one in isolation. Jaccard and Jaccoby (2010) suggest that factorial experiments involving several independent variables can "make the relationships you have intuitively generated more explicit" (p. 122). Second, it allows for testing interactions among the intervention's compo-

nents, which reveals whether the effect of one component on a mediator or outcome varies depending on the level of another component. Third, the factorial experiment requires fewer subjects to maintain a comparable level of statistical power (Collins, Dziak, & Li, 2009) and thus can be relatively economical.

To summarize this section, mediation analysis is a powerful means by which to investigate an intervention's causal story. For an intervention that can be broken into components, a factorial experiment in which the components are manipulated and then evaluated with a mediation analysis provides even more powerful insights into why the intervention works (or does not work). Based on these insights, intervention scientists are better positioned to make informed decisions about which and how many components need to be included for the desired outcome to appear.

## 2 Case Study

The present study tests a potential unintended consequence: whether exposure to information about a health condition results in stigmatizing people living with the health condition. This kind of intervention is not uncommon when trying to inform the public, especially about health crises. For example, the world learned of the largest outbreak of Ebola in recorded history (Centers for Disease Control and Prevention [CDC], 2015) from health agencies through their press releases and interactions with media outlets. Although health agencies (and media outlets) probably intended to share the outbreak news to inform the public and promote behaviors to limit the epidemic's spread, the 2014 Ebola coverage has been described as stigmatizing (e.g., Cheung, 2015). This critique is not novel: media coverage has been thought to be responsible for stigmas associated with other infectious diseases (e.g., Eichelberger, 2007), mental illness (e.g., President's New Freedom Commission, 2003), and suicide (e.g., World Health Organization, 2014).

Until the publication of the model of stigma communication (MSC; Smith, 2007, 2011), little attention was paid to explaining why some messages provoke stigma-related outcomes, but others do not (Corrigan, Powell, & Michaels, 2013; Pescosolido, Martin, Lang, & Olafsdottir, 2008). This is surprising, because one claim in stigma research (e.g., Goffman, 1963; Link & Phelan, 2001; Pescosolido et al., 2008) is that stigmas are communicated to and among community members so that the public learns to recognize the stigmatized people and to stigmatize them (Smith, 2007, 2011).

The MSC provides guidance on why writing health information with a *stigma frame* (intervention) causes readers to experience cognitions and negative feelings (mediators) that increase the message recipients' stigmatization of persons infected with the contagious disease (outcome). Frames are considered distinct from health information's topic (Pan & Kosicki, 1993). Instead, frames provide cues (1) to define the problem at hand—to determine who is doing what, with what consequences; (2) to diagnose the cause of the problem; (3) to evaluate, morally, the problem, its generators, and its effects; and (4) to offer and justify remedies for

the problem (Entman, 1993). Thus, for intervention scientists, the MSC provides guidance about the content to avoid in order to prevent stigmatizing those with the health condition.

Stigmatization is a broad term encompassing many outcomes, including forming stigma beliefs, spreading stigma messages, creating social distance, and regulating stigmatized people's lives. This case study focuses on regulating the lives of stigmatized people in ways that may infringe upon their rights and create discrimination; regulation includes isolating infected persons from the general public, forcing treatment even if unwanted, monitoring by health agencies, quarantining by officials, providing separate transportation and separate medical shelters, registering with health departments, and creating real-time, publicly available maps of where stigmatized persons reside. (Hereafter this outcome will be referred as *regulation*.) For many health conditions, these types of actions do not curb the presence or spread of the health threat. For example, the risk of contracting a diseases caused by vectors, such as mosquitos, mice, or birds, does not improve by quarantining infected people.

In this chapter, we test the MSC in the context of providing health information about an infectious disease. The group at risk for stigmatization when deploying such an intervention is those who have contracted an infectious disease. According to the MSC, the relation between exposure to health information with a stigma frame and regulation of infected people is mediated by the following mechanisms: (1) belief that stigmatized individuals made choices that resulted in getting infected (*responsibility*); (2) perception of stigmatized people as a unique, distinct social group (*group entitativity*); (3) belief that the stigmatized group and its members are able and likely to harm the rest of the community by their presence and actions (*dangerous*); and (4) feeling anger, disgust, and fear toward the stigmatized people (*negative affect*).

## *2.1 Separating the Stigma Frame into Content Components*

According to the MSC (Smith, 2007, 2011, 2012), a *stigma frame* includes four intrinsic features (mark, label, etiology, and peril). A stigma frame, then, represents the type of multiple-component intervention described in Chap. 1 of the companion volume (Collins, 2018), in which the assumption is that all four features are needed to create the effect. (We treat this assumption as an empirical question that will be addressed in the factorial experiment.) As shown in Fig. 2, according to the MSC, exposure to health information with all four features results in stigmatization (e.g., stigma beliefs, social distance, or regulation), through four different mediators.

A *mark* describes ways to identify another person as a member of a stigmatized group (Smith, 2007). Marks are particularly effective when they include visible (Deaux, Reid, Mizrahi, & Ethier, 1995; Frable, 1993; Jones et al., 1984) and disgusting (Goffman, 1963; Haidt, McCauley, & Rozin, 1994; Jones et al., 1984) features, because they are easier to recognize and they evoke the stigma-appropriate action tendency to remove and isolate stigmatized persons.

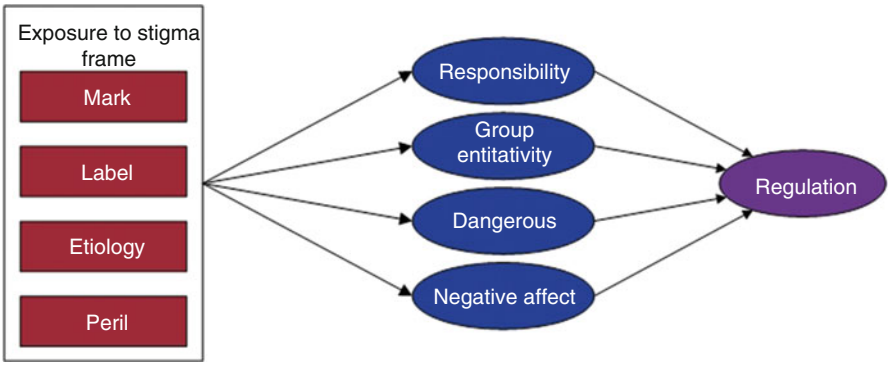


**Fig. 2** Conceptual model of the causal process by which presenting health information with a stigma frame results in stigmatizing those with the health condition

*Labels* are terms used to reference people with the health condition. Labeling a group by its stigmatizing issue keeps the group threat salient, which would encourage separating the stigmatized group away from a community. It accentuates intergroup differences (cf. Tajfel, 1959) and leads to depersonalizing people into embodiments of a group's attributes. Differences between reference terms can be subtle yet important. For example, one can refer to people who suffer from a disorder (e.g., people living with epilepsy), or instead label affected people as the disease (e.g., epileptics).

*Etiology* is the description of how a person becomes a member of the stigmatized group. Etiology content includes implicit or explicit references to the choice and control a person has over their stigmatized condition (Smith, 2007, 2012); it aligns with how practice links to negative consequences. Some infectious diseases have multiple forms of transmission, which provide intervention scientists with options about which form of transmission to highlight. A recent example is the Zika virus, which can be transmitted through mosquitos or human-to-human contact (Musso et al., 2015). Content highlighting human behavior implies choice and control over violating social contracts. People judge those who engage in taboo activities or act in ways that threatened the group as having a fundamental character flaw of immorality (Goffman, 1963; Jones et al., 1984).

Conceptually, *peril* content describes the danger the stigmatized group poses to the rest of the community (Smith, 2007). Drawing upon research in product hazards (DeTurck, 2002), peril content may include the source of the danger, recommendations to avoid dangerous people, and the consequences if one fails to avoid the danger (Smith, 2007). Operationally, peril content has been varied by describing the consequences of the infectious disease (Smith, 2012) as fatal, incurable, and causing great muscle pain and possible paranoia and aggression (high peril) or easily curable and rarely fatal, with mild, temporary discomfort (low peril).

The illustration tested herein includes (a) a description of visible symptoms of infection (mark); (b) the use of a group label when referring to those with the infection (label); (c) a description of the curability, pain, and suffering associated

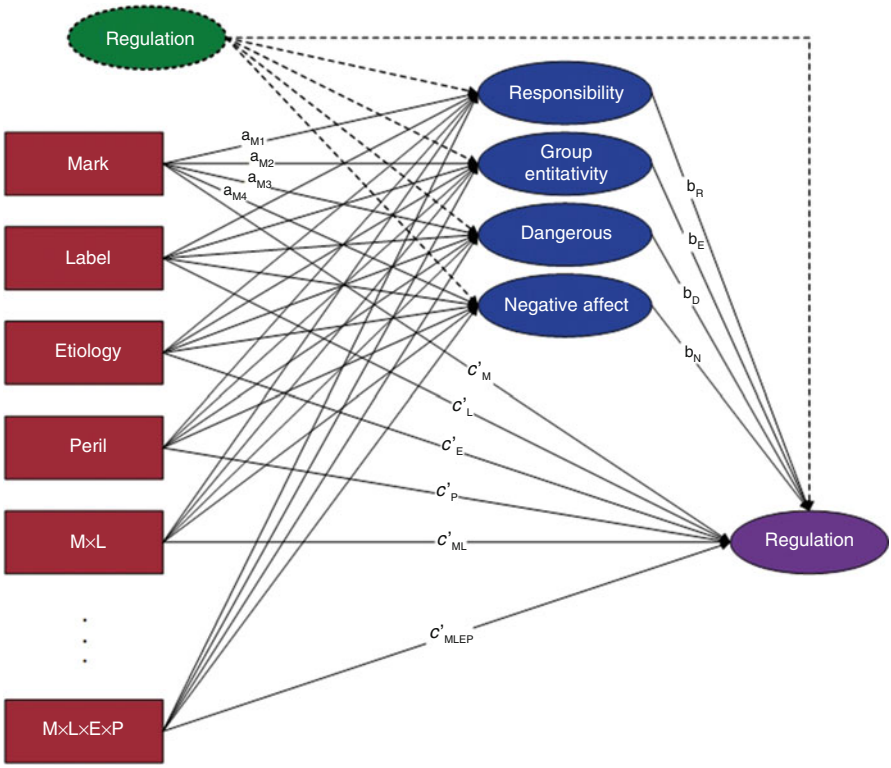


**Fig. 3** The graphic represents a mediation model with a factorial experiment, four mediators, and one outcome measured before and after exposure to the health information (intervention). *Dotted lines* represent correlated error terms and the pretest covariate

with the infectious disease (peril); and (d) the human means (vs. nonhuman, such as mosquito bites) by which people become infected (etiology).

Although it is possible to make one intervention package of health information with all of the theory-driven content and then evaluate the health information's influence on the outcome through the theorized mediators, this strategy leaves us with little insight into which content or mediator was most important (see Chap. 1). It is possible to vary the four content components in a $2^4$ factorial experiment (see Fig. 3). The factorial design allows us to examine the main effects of content components and their interactions. The MSC predicts that the devastating effects of stigmatization (damaging material, social, and psychological well-being; e.g., Goffman, 1963; Hatzenbuehler, Phelan, & Link, 2013; Link, Phelan, & Hatzenbuehler, 2014; Miller & Major, 2000) occur as a result of the presence of all four content components. For an interventionist, such results would suggest that the presence of one or two components such as describing visible symptoms to recognize someone who is infected (mark) and the consequences of infection (peril)

may be included without concern, as long as the other components (e.g., labels and etiology) are not included. The theory, however, could be wrong. It may be that only a subset of the content components creates stigmatization and they would be the most important components to avoid.

### *2.2 The Present Study*

A mediation analysis of a factorial experiment provides a comprehensive test of many aspects of the MSC. Importantly, it enables us to assess whether any of the mediators and content components could be removed from the MSC to produce a more parsimonious theory. We hope to answer three kinds of questions with the mediation analysis: (1) How well do the mediators (responsibility, group entitativity, dangerous, and negative affect) predict postexposure levels of regulation (*b* paths)? (2) How well does each type of content (mark, label, etiology, and peril) evoke each of the four mediators (*a* paths)? (3) How much of the relation between the intervention's content components and regulation (outcome) is explained by the four hypothesized mediators (the product of *a* and *b* paths)? (4) How much variability between the content components and regulation support is unexplained ($c'$ paths)?

## 3 Methods

### *3.1 Participants*

Participants ($N = 299$, 58% female, 0.3% unidentified) on average were 36 years old ($SD = 12.13$, $Minimum = 21$, $Maximum = 74$). Participants were recruited through Amazon's Mechanical Turk, which provides access to a diverse pool of adults who react similarly to those recruited through offline strategies (Mason & Suri, 2011). Participants identified their race as White (81%), African American (8%), Asian (6%), American Indian or Alaska Native (1%), and Native Hawaiian or Pacific Islander (1%); 3% did not report a racial identification. Five percent identified their ethnicity as Hispanic.

### *3.2 Design*

The experimental design was a $2^4$ ($2 \times 2 \times 2 \times 2$) between-subjects factorial, with factors (level) as follows: *MARK* (yes, no), *LABEL* (yes, no), *ETIOLOGY* (airborne [human], mouse-only [nonhuman] transmission), and *PERIL* (high, low). The intervention was adapted from Smith (2012). The intervention is based on CDC newsroom reports of hantavirus pulmonary syndrome but was given a fictitious

name—cautela acervusary virus (CAV; see Smith, 2012 for a full description of the intervention's conditions). The disease was described as creating open sores on arms and producing a loud, wet cough (mark) or having no visible symptoms (no mark). Infected people were referred to as Cavers (label) or as people infected with CAV (no label). Etiology was described as transmission either via contact with infected mice feces (vector-borne; nonhuman etiology) or directly between persons through sneezed or coughed droplets (airborne; human etiology). The disease was described as painful, incurable, and often fatal, with possible paranoid delusions and aggression (high peril) or as mildly uncomfortable, curable, nonfatal, with no mental changes (low peril). Regulation (outcome) was measured before and after exposure to the intervention.

## *3.3 Procedures*

An institutional review board approved the study. The procedures follow those in Smith (2012). Adults registered with Amazon's Mechanical Turk participated in this study. After giving consent, the online survey instructions stated that the participants would be presented with health information under consideration for future public health alerts and that the health condition is a fictitious representation of an existing illness. The participants were asked to consider the message's content (as opposed to delivery). Participants were shown 1 of the 16 possible versions of the health information through random assignment (about 19 participants per cell). Participants were asked to judge the health information (the intervention) on its believability, credibility, and importance; these judgments did not vary by condition ($F < 1$). After exposure to the health information, participants were asked to complete the scales listed below and then answer demographic questions. Of note, participants completed the scales capturing mediators before the scales capturing the outcome.

## *3.4 Measurement*

### 3.4.1 Mediators

**Dangerous.** Four items (adapted from Smith, 2012; $\alpha = .92$, asymptotically distribution-free (ADF) 95% CI [.90, .94; Maydeu-Olivares, Coffman, & Hartmann, 2007], *Skewness* $= -0.02$, *Kurtosis* $= -1.26$) were used to assess the perceived danger of infected persons (e.g., putting others in danger and dangerous to be around). Responses were marked on five-point scales (1 = *strongly disagree* to 5 = *strongly agree*), with higher scores indicating that participants perceived infected persons as more dangerous.

**Responsibility.** Four items (adapted from Smith, 2012; $\alpha = .70$, ADF 95% CI [.64, .76], *Skewness* = 0.23, *Kurtosis* = 0.23) were used to assess the perceived responsibility of infected persons for getting CAV (e.g., at fault for or responsible for). Responses were marked on five-point scales (1 = *strongly disagree* to 5 = *strongly agree*), with higher scores indicating that participants perceived infected persons as more responsible for their infection.

**Group Entitativity.** Six items (described in Smith, 2012; $\alpha = .88$, ADF 95% CI [.86, .91], *Skewness* = 0.29, *Kurtosis* = −0.40) were used to assess the degree to which those infected with CAV were a distinct social entity. Responses were marked on five-point scales (1 = *strongly disagree* to 5 = *strongly agree*), with higher scores indicating greater perceived entitativity.

**Negative Affect.** Four items ($\alpha = .83$, ADF 95% CI [.80, .87], *Skewness* = 0.90, *Kurtosis* = 0.10) were used to assess how strongly participants felt different emotions after reading the message: disgust, anger, fear (from Smith, 2012), and worry (added in this experiment). Responses were marked on five-point scales (1 = *not at all* to 5 = *strongly*). Higher scores indicate greater negative affect.

### 3.4.2 Outcome

**Regulation.** Eight items (based on Smith, 2012; pretest $\alpha = .90$, ADF 95% CI [.88, .92], *Skewness* = −0.46, *Kurtosis* = 0.04; posttest $\alpha = .94$, ADF 95% CI [.93, .95], *Skewness* = −0.30, *Kurtosis* = −0.96) were used to assess to what extent participants agree with eight different intervention options, if this were a real situation and this alert had been shared with them. These options included isolating infected persons from the non-infected public, treating infected persons even if they did not want to be treated, monitoring infected persons by health personnel, quarantining infected persons, providing separate transportation for infected persons, providing new medical shelters in which to treat infected persons, registering infected persons with the health department, and providing a way to see where infected people live (i.e., a map). Responses were marked on five-point scales (1 = *strongly disagree* to 5 = *strongly agree*), with higher scores indicating stronger support for regulation.

## 4 Results

### *4.1 Descriptive Statistics*

The measures showed acceptable levels of skewness and kurtosis; therefore, they were not transformed. Table 1 presents the means, standard deviations, and zero-order correlations among variables. The measured mediators—dangerous,

**Table 1** Descriptive statistics for variables ($N = 299$)

| | *M* | *SD* | *Range* | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|---|---|
| 1. Pre-regulation | 3.65 | 0.85 | 4.00 | | | | | |
| 2. Post-regulation | 3.23 | 1.21 | 4.00 | .45* | | | | |
| 3. Responsible | 2.63 | 0.73 | 4.00 | .23* | .13* | | | |
| 4. Group entitativity | 2.54 | 0.92 | 4.00 | .19* | .37* | .05 | | |
| 5. Dangerous | 2.83 | 1.29 | 4.00 | .15* | .74* | .09 | .25* | |
| 6. Negative affect | 2.03 | 0.98 | 4.00 | .07 | .39* | .06 | .29* | .40* |

Note: Responses were marked on five-point scales (e.g., 1 = *strongly disagree* to 5 = *strongly agree*)
*$p < .05$

responsibility, group entitativity, and negative affect—were averaged into composite scores before the descriptive statistics were computed. Table 1 shows that the correlations between some mediators were statistically significant.

Change scores were created to provide insight into whether exposure to the health information created change. Participants' willingness to regulate infected people changed after exposure to the intervention ($M_{change} = -0.42$, $SD = 1.12$, $Maximum = 3.50$, $Minimum = -4.00$). Participants with higher preexposure willingness to regulate showed less change, $r\ (297) = -.27$, $p < .001$. Figure 4 shows the mean changes in regulation support by experimental condition, arranged by size. The means show that regulation, the form of stigmatization investigated in this study, increased in only 4 of the 16 conditions. In Fig. 4, the on (e.g., high or present) condition is represented with a capital letter, whereas the off (e.g., low or not present) condition is represented with a lowercase letter. As expected, the version with all four content components (i.e., "MLEP") was one of the four conditions with increases in willingness to regulate infected people. The three other conditions (MlEP, mLEP, and mlEP) all had the etiology and peril components. This finding suggests the etiology and peril components may be particularly important components in a stigma frame.

## *4.2 Mediation Analysis*

This mediation analysis, like all mediation analyses, requires several assumptions. We assume that the mediators and residuals are independent (MacKinnon, Fairchild, & Fritz, 2007). We also assume no interaction between the experimental condition and the mediators; no misspecification of the causal order of the stimuli, mediators, and outcome (as would occur if, e.g., in reality content component → regulation → dangerous); and no misspecification due to unmeasured variables or imperfect measurement. These assumptions are challenging to test (MacKinnon et al., 2007).

Figure 3 illustrates the tested model, which included $a$, $b$, and $c'$ paths. This model included one dependent variable: posttest regulation. The model included the

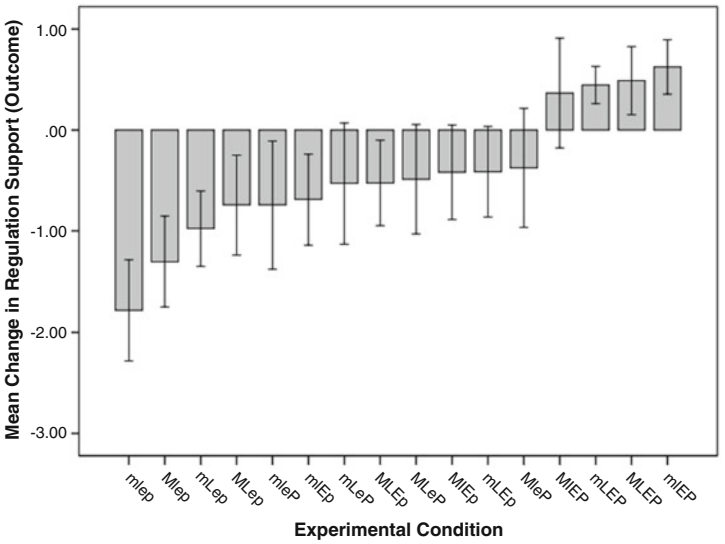


**Fig. 4** Average change in willingness to regulate infected people and 95% confidence intervals by experimental condition. The "on" (yes, person-oriented, or high) condition is represented with a capital letter; the "off" (no, nonhuman, or low) condition is represented with a lowercase letter

following independent variables: the pretest of regulation and the 15 vectors representing the four main effects, six two-way interactions, four three-way interactions, and one four-way interaction. Because effect (−1, 1) codes were used, the ANOVA effect vectors as well as the resulting effect estimates were uncorrelated. Because we used random assignment to experimental conditions, the pretest of regulation was not correlated with the ANOVA effect vectors in the model. In addition, there were (1) the four mediators, (2) paths from the pretest of regulation to the mediators, (3) paths from the ANOVA effect vectors (i.e., the *a* paths), and (4) paths from the mediators to the posttest of regulation (i.e., the *b* paths). Finally, the $c'$ paths were the paths from the ANOVA effect vectors to the outcome.

To conduct the mediation analysis, we fit a structural equation model (SEM) using maximum likelihood estimation in the software program AMOS 24. The variables represented in ovals in Fig. 3 were modeled as latent; the measured items were included in the measurement submodel tested in AMOS but are not shown in the figure. Neither the errors of prediction nor the errors of measurement were allowed to covary. A post hoc power analysis was conducted for the SEM; it includes main effects, interaction effects, mediators, and pretest regulation. With $N = 299$ and $df = 1052$ (1225 distinct sample moments-173 parameters to be estimated), the power to detect a close-fitting model (root mean-square error of approximation ($RMSEA$) = .05) versus a perfectly fitting model ($RMSEA = .00$) at $p < .05$ was greater than .999 (Preacher & Coffman, 2006). Bias-corrected confidence intervals

**Table 2** Effects of mediators on post-exposure regulation (*b* paths)

| | Unstandardized *b* path | Standardized *b* path | 95% CI |
|---|---|---|---|
| Responsible | 0.05 | .02 | −.08, .12 |
| Group entitativity | 0.16 | .07 | −.01, .15 |
| Dangerous | 0.68 | .73* | .62, .83 |
| Negative affect | 0.05 | .05 | −.05, .15 |

Notes: CIs are for the standardized estimate, with bias-corrected confidence intervals (Hayes & Scharkow, 2013) using bootstrapping procedures (2000 bootstrap samples)
*$p < 0.05$

(Hayes & Scharkow, 2013) were estimated using 2000 bootstrapped samples. The goodness-of-fit estimates for the SEM were $\chi^2(1058, N = 299) = 2459.34$, $p < .05$, $RMSEA = .07$ (CI: .06, .07), and $SRMR = .07$.

### 4.2.1 Mediators to Outcome

First, we assessed how well the hypothesized mediators predicted the outcome (*b* paths). Table 2 shows the *b* paths. Only one of the mediators, dangerous, was a statistically significant predictor of regulation.

### 4.2.2 Experimental Factors to Mediators

The next step was to assess the effects of the factors on the mediators (*a* paths). In this case, because dangerous was the only statistically significant mediator and the other mediator effects are close to zero, we have shown only how well the intervention's components predicted dangerousness (see Table 3). Two main effects and three interactions were statistically significant. As hypothesized, content describing the infectious disease as more perilous increased perceptions of the infected persons as dangerous in comparison to less-perilous content. As predicted, infected persons were perceived as more dangerous if the infection had airborne (human-based) instead of vector-borne (nonhuman) etiology.

Three interactions were statistically significant: *MARK* × *LABEL* and *LABEL* × *ETIOLOGY*, which are antagonistic interactions (see Chap. 4 in the companion volume, Collins, 2018) and *ETIOLOGY* × *PERIL* (which is a synergistic interaction). The pattern of means showed weaker perceptions of the infected people as dangerous in the no mark/no label condition versus the others, and in the no label/nonhuman etiology condition versus the others. In contrast, perceiving infected people as dangerous was particularly common in the human etiology/high-peril condition versus the others. Presenting both human etiology for a disease and significant consequences caused people to support regulating infected people.

**Table 3** Parameter estimates for relations from message components to dangerous

| | Unstandardized *a* path | Standardized *a* path | 95% CI |
|---|---|---|---|
| Pretest regulation | 0.58* | .23 | .13, .33 |
| *MARK* | 0.02 | .01 | −.09, .11 |
| *LABEL* | 0.08 | .05 | −.04, .15 |
| *ETIOLOGY* | 0.76* | .52 | .44, .61 |
| *PERIL* | 0.44* | .31 | .22, .40 |
| *MARK* × *LABEL* interaction | −0.14* | −.10 | −.19, −.01 |
| *MARK* × *ETIOLOGY* interaction | 0.01 | .00 | −.09, .10 |
| *MARK* × *PERIL* interaction | −0.06 | −.04 | −.14, .05 |
| *LABEL* × *ETIOLOGY* interaction | −0.15* | −.11 | −.20, −.01 |
| *LABEL* × *PERIL* interaction | −0.05 | −.03 | −.12, .06 |
| *ETIOLOGY* × *PERIL* interaction | 0.17* | .12 | .02, .21 |
| *MARK* × *LABEL* × *ETIOLOGY* interaction | 0.06 | .04 | −.05, .13 |
| *MARK* × *LABEL* × *PERIL* interaction | 0.02 | .01 | −.08, .10 |
| *LABEL* × *ETIOLOGY* × *PERIL* interaction | 0.03 | .02 | −.08, .10 |
| *MARK* × *ETIOLOGY* × *PERIL* interaction | 0.10 | .07 | −.02, .16 |
| *MARK* × *LABEL* × *ETIOLOGY* × *PERIL* interaction | 0.10 | .07 | −.02, .16 |

Notes: Analysis of the separate *a* paths, provided by the factorial experimental design, shows the variation in the main effects of factors and interactions among factors. CIs are for the standardized estimate, with bias-corrected confidence intervals (Hayes & Scharkow, 2013) using bootstrapping procedures (2000 bootstrap samples)

*$p < .05$

### 4.2.3 Unexplained Effects

The final step was to assess the main effects of the factors and the interactions between factors on the posttest outcome ($c'$ paths, i.e., direct effects). Table 4 shows that there is a statistically significant, direct effect of *PERIL* on regulation. This finding suggests that additional mediators need to be included to completely understand how peril content shapes regulation. Table 4 also shows the indirect effects of the factors on regulation via the mediators. The tests of the indirect effects consider the mediators together, which may be reasonable when the mediators are correlated (VanderWeele & Vansteelandt, 2014).

## 5 Discussion

This chapter explores issues involved in testing an intervention's causal story and demonstrates how to conduct a mediation analysis in situations in which an intervention has been decomposed into several components and has been theorized to

**Table 4** Parameter estimates for direct and indirect effects of factors on postexposure regulation

| | Direct | Std. direct | 95% CI | Indirect | Std. indirect | 95% CI |
|---|---|---|---|---|---|---|
| Pretest regulation | 0.56* | .24 | .14, .34 | 0.45* | .19 | .19, .28 |
| *MARK* | 0.04 | .03 | −.04, .10 | 0.01 | .01 | .01, .08 |
| *LABEL* | 0.04 | .03 | −.04, .10 | 0.06 | .05 | .05, .12 |
| *ETIOLOGY* | −0.01 | −.01 | −.09, .08 | 0.52* | .39 | .39, .48 |
| *PERIL* | 0.12* | .09 | .01, .17 | 0.34* | .25 | .25, .35 |
| *MARK* × *LABEL* | 0.04 | .03 | −.04, .10 | −0.11* | −.08 | −.08, .00 |
| *MARK* × *ETIOLOGY* | −0.03 | −.02 | −.09, .05 | 0.00 | .00 | .00, .08 |
| *MARK* × *PERIL* | −0.04 | −.03 | −.10, .05 | −0.04 | −.03 | −.03, .04 |
| *LABEL* × *ETIOLOGY* | 0.05 | .04 | −.03, .10 | −0.11* | −.08 | −.08, −.01 |
| *LABEL* × *PERIL* | −0.04 | −.03 | −.10, .04 | −0.03 | −.02 | −.02, .05 |
| *ETIOLOGY* × *PERIL* | −0.05 | −.04 | −.11, .03 | 0.12* | .09 | .09, .16 |
| *MARK* × *LABEL* × *ETIOLOGY* | −0.01 | −.01 | −.07, .06 | 0.04 | .03 | .03, .10 |
| *MARK* × *LABEL* × *PERIL* | 0.01 | .00 | −.07, .08 | 0.02 | .01 | .01, .09 |
| *LABEL* × *ETIOLOGY* × *PERIL* | 0.04 | .03 | −.04, .10 | 0.02 | .01 | .01, .08 |
| *MARK* × *ETIOLOGY* × *PERIL* | −0.06 | −.05 | −.12, .02 | 0.07 | .06 | .06, .13 |
| *MARK* × *LABEL* × *ETIOLOGY* × *PERIL* | 0.00 | .00 | −.07, .07 | 0.06 | .05 | .05, .12 |

Notes: Std. = standardized. CIs are for the standardized estimate, with bias-corrected confidence intervals (Hayes & Scharkow, 2013) using bootstrapping procedures (2000 bootstrap samples)
*$p < .05$

cause an outcome to occur through multiple mediators. The methods demonstrated herein tested the data from a factorial experiment with a mediation analysis.

## 5.1 Insights into the Case Study

As a reminder, the case study illustrating these issues explored an iatrogenic effect, creating a health stigma, as a consequence of presenting health information about a health condition (the intervention) with a *stigma frame* (Smith, 2007). The stigma frame is theorized to include four intrinsic features, representing the assumption that all four features are needed to cause stigma-related outcomes (see Chap. 1 for multiple-component interventions with different assumptions). We separated the intrinsic features into four content components and examined them using a factorial experiment that included a factor corresponding to each of the four components (i.e., 16 different versions of the health information).

### 5.1.1 The Intervention Package

Conservatively, if a stigma frame worked as one package, then stigmatization should only have increased when all of the content components were present (labeled as "MLEP"). As seen in Fig. 4, regulation, the form of stigmatization investigated in this study, increased in 4 of the 16 conditions. The four conditions were ones in which *ETIOLOGY* was set to person-oriented (airborne), *PERIL* was set to high, and *MARK* and/or *LABEL* were present. This finding confirms the findings from the main effects: the etiology and peril components may be particularly important components in a stigma frame. Without the factorial design, there would have been no way to glean any of these theoretical insights.

From a practical perspective, the finding that etiology and peril are such important content components calls for careful attention. Intervention scientists who want to present health information that highlights a human etiology for contracting a disease and significant physical consequences associated with it may need to find strategies to avoid creating new or increasing existing health stigma as they optimize their intervention for delivery.

### 5.1.2 The Causal Process

The factorial experimental design and focus on mediators provided information that allowed for more precise conclusions about the theoretical model and opportunities for refinement. The mediation analysis revealed several statistically significant indirect effects of the experimental factors on stigmatization through the hypothesized mediators. One consequence of multiplying the estimates for the two paths together is that the product term (i.e., indirect effect) does not allow an assessment of the two component paths of the mediated effect separately. Furthermore, the current test of the indirect effects considers the mediators to be correlated (VanderWeele & Vansteelandt, 2014). Indeed, Table 1 showed that a few of the mediators were correlated, but one (responsibility) was not correlated with any of them. Instead of relying only on indirect effects, we decided to attend to the effects of the mediators on the outcomes (*b* paths) explicitly.

The mediation analysis showed that perceiving infected people as dangerous was the only supported mediator for understanding why the health information caused stigmatization to occur. We tested one type of stigmatization: regulation support. If the results generalize, this finding suggests that a more parsimonious theoretical model with the elimination of group entitativity, responsibility, and negative affect may be needed. The findings suggest that perceiving those with the health condition described in health information as dangerous is the central mechanism among the theorized contenders.

We then investigated which content components caused infected people to be perceived as dangerous. Two factors, *ETIOLOGY* and *PERIL*, showed statistically significant main effects. Significant interactions between other factors, including *MARK* and *LABEL*, appeared. These findings suggest that content components, like etiology, produce the mediating effects by themselves, as evidenced by main effects,

and in combination with other components, as evidenced by interactions. These interactions again provide general support for the idea that these multiple intrinsic features create a stigma frame that causes an outcome together. Without the factorial design, there would have been no way to glean any of these theoretical insights.

Data from a factorial experiment allowed us to investigate important claims in the mediation analysis. This is a case of mediated moderation, which is when the interaction between two predictor variables has an effect on the outcome through a mediator. In our case, the two predictor variables are both randomly assigned, which means that their interaction effect can be interpreted as causal. If the two predictor variables had not been randomly assigned, or if only one of them had, then the effect could not be interpreted as causal without further assumptions.

### 5.1.3 Implications for Theory

Good theories are parsimonious (Jaccard & Jaccoby, 2010). In an attempt to build communication theory, researchers often include multiple message variables and multiple mediators to explain particular message effects or outcomes. The illustration showed how the relatively new MSC (Smith, 2007) benefits from the rigorous analysis proposed in this chapter. The findings suggest that a more parsimonious theoretical model might remove group entitativity, responsibility, and negative affect and keep only dangerous. We note that the use of multiple mediators is one way to deal with omitted confounders of the M to Y relation. If some potential mediators show evidence for a mediated effect and others do not, then this provides evidence for the specificity of that mediator (for more discussion of these issues see MacKinnon & Pirlott, 2015 and Pirlott & MacKinnon, 2016).

The direct effects of the content components on the outcome (Table 4) showed that the presence of peril content increased stigmatization, even after accounting for the mediators. This finding suggests that further research is needed to identify an additional mediator (or mediators) to explain why presenting different types of peril in health information cause the exposed audience to stigmatize infected people. One option may be to include a measure of perceived threat of the infectious disease. It is possible that participants are responding to differing levels of threat, and perceived threat of the infection is a more important mediator than perceived dangerousness of infected people.

Notably, the model explains 78% of the variance in postexposure regulation; there is still 22% of the variance left to explain. The work to refine the MSC will not be complete until the mediators significantly and sizably predict its stigma-related outcomes, the stimuli evoke the mediators significantly and sizably, and the residual variance is eliminated, with the outcomes completely (100%) predicted by the theoretical model. This target for completion may be more aspirational than practical. On a practical level, even though the peril component caused people to perceive infected people as more dangerous, which caused greater stigmatization, it is not clear what precise, fine-grained aspect of the manipulation led to the effect. More investigations can be done to increase precision.

### 5.1.4 Limitations to the Case Study

The findings reported here are limited in a few ways. The mediators were measured immediately after exposure to the message. It is possible that the cognitive mediators (responsibility and group entitativity) would have demonstrated more or less measurable change after more time had elapsed. More work is needed on the role of time in mediation, and behavioral theories should be more specific about the expected time lag between cause and effect (Collins, 2006). Our findings are also limited by the sample's demographics.

A different manifestation of the intervention components (e.g., different wording or imagery) might have evoked the mediators such as responsibility more strongly. Perceived responsibility, then, might have affected the outcome measurably. Notably, participants did not perceive infected people as responsible, and the standard deviation was small. Varying the stimuli, timing, and measures would provide confidence in the findings and support for eliminating mediators from the MSC. Importantly, the findings reported here are limited to regulation, which is just one of three stigma-related outcomes described in the MSC. Further studies are needed to examine the MSC's parsimony in explaining the other two outcomes: stigma-belief formation and social distancing.

This case was conducted in a setting with a large participant pool (through Amazon's Mechanical Turk) and with content components that we could easily manipulate. For field interventions in other settings (e.g., schools, doctor's offices) with other forms of delivery (e.g., in-person sessions; literature in doctor's offices), the factorial design may be more challenging (although, see the Piper et al.'s (2018) chapter in this volume).

## 6 Future Directions

A future direction is to extend the model to include moderation effects between observed pre-intervention variables and intervention components. Although we included interactions among our experimentally manipulated variables (referred to as treatment interactions in the epidemiology literature; VanderWeele, 2009), we did not include moderators, such as personality traits. For example, it is possible that some people are more susceptible or resistant (e.g., cynics, Smith, 2012) to the health information. Audience variables like these could be important moderators. Moderation effects of any type in combination with mediation processes are quite complex, and there is a growing literature on these complexities (e.g., Edwards & Lambert, 2007; Muller, Judd, & Yzerbyt, 2005; Preacher, Rucker, & Hayes, 2007).

Factorial experiments designed to assess the effects of individual intervention components enable testing hypotheses not only about which mediators have an effect on the outcome, but also about which components of the intervention package have (or do not have) an effect on particular mediators and about whether there is a direct effect for each component. This information can be extremely valuable

in intervention development and in advancing intervention theory (Lipsey, 1993). Although factorial experiments are still relatively rare in behavioral intervention science, applications are increasing, as seen in the other chapters of this book. Although the factorial experiment has clear benefits, guidance is needed for how to deal with many orthogonal tests and controlling experiment-wise error rates. The illustration included 60 *a* path estimates, 4 *b* path estimates, and 16 $c'$ path estimates. In addition, we had two levels of each component (e.g., high or low). We could have presented more levels (e.g., high, medium, and low), which may have resulted in more tests at some stage. The issues of correcting for error rates are not trivial, and corrections for type I error can come at the cost of creating type II errors (e.g., Smith, Levine, Lachlan, & Fediuk, 2002). In intervention science, we may want to hold type II error rates as most critical as we discover what is important as we create an intervention. Control of type I error rates may become more critical as we move through the optimization process. Correction for experiment-wise error rates in multiple mediator models has rarely been addressed (c.f., MacKinnon, 2000) and needs attention.

## *6.1 Conclusion*

We all benefit by learning how to innovate and create, as well as test and prune, our ideas to their most elegant forms. We hope that this chapter provides one step in that direction.

**Acknowledgments** Our thanks go to John Dziak, two anonymous reviewers, and researchers at the Methodology Center for feedback on earlier drafts. This project was supported by several National Institutes of Health (NIH) awards: R21 HG007111 from the National Human Genome Research Institute, P50 DA010075 from the National Institute on Drug Abuse, P50 CA143188 from the National Cancer Institute, and R01 DK097364 from the National Institute of Diabetes and Digestive and Kidney Diseases. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or its institutes.

## References

Baranowski, T., Anderson, C., & Carmack, C. (1998). Mediating variable framework in physical activity interventions: How are we doing? How might we do better? *American Journal of Preventive Medicine, 15*, 266–297. https://doi.org/10.1016/S0749-3797(98)00080-4

Centers for Disease Control and Prevention. (2015). *2014 Ebola outbreak in West Africa*. Retrieved March 31, 2015, from http://www.cdc.gov/vhf/ebola/outbreaks/2014-west-africa/index.html

Chen, H. T. (1990). *Theory-driven evaluations*. Newbury Park, CA: Sage.

Cheung, E. Y. L. (2015). An outbreak of fear, rumours and stigma: Psychosocial support for the Ebola virus disease outbreak in West Africa. *Intervention, 13*, 70–76.

Cho, H., & Salmon, C. T. (2007). Unintended effects of health communication campaigns. *Journal of Communication, 57*, 293–317. https://doi.org/10.1111/j.1460-2466.2007.00344.x

Collins, L. M. (2006). Analysis of longitudinal data: The integration of theoretical model, temporal design and statistical model. *Annual Review of Psychology, 57*, 505–528. https://doi.org/10.1146/annurev.psych.57.102904.190146

Collins, L. M. (2018). Conceptual introduction to the multiphase optimization strategy (MOST). In L. M. Collins & K. C. Kugler (Eds.), *Optimization of behavioral, biobehavioral, and biomedical interventions*: *Advanced topics*. New York, NY: Springer.

Collins, L. M., Dziak, J. R., & Li, R. (2009). Design of experiments with multiple independent variables: A resource management perspective on complete and reduced factorial designs. *Psychological Methods, 14*, 202–224. https://doi.org/10.1037/a0015826

Corrigan, P. W., Powell, K. J., & Michaels, P. J. (2013). The effects of news stories on the stigma of mental illness. *Journal of Nervous and Mental Disease, 201*, 179–182. https://doi.org/10.1097/nmd.0b013e3182848c24

Deaux, K., Reid, A., Mizrahi, K., & Ethier, K. A. (1995). Parameters of social identity. *Journal of Personality & Social Psychology, 68*, 280–291. https://doi.org/10.1037/0022-3514.68.2.280

DeTurck, M. A. (2002). Persuasive effects of product warning labels. In J. P. Dillard & M. Pfau (Eds.), *Persuasion: Developments in theory and practice* (pp. 213–232). Thousand Oaks, CA: Sage.

Edwards, J. R., & Lambert, L. S. (2007). Methods for integrating moderation and mediation: A general analytical framework using moderated path analysis. *Psychological Methods, 12*, 1–22. https://doi.org/10.1037/1082-989X.12.1.1

Eichelberger, L. (2007). SARS and New York's Chinatown: The politics of risk and blame during an epidemic of fear. *Social Science & Medicine, 65*, 1284–1295. https://doi.org/10.1016/j.socscimed.2007.04.022

Entman, R. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication, 43*, 51–58. https://doi.org/10.1111/j.1460-2466.1993.tb01304.x

Flay, B. R. (1986). Efficacy and effectiveness trials (and other phases of research) in the development of health promotion programs. *Preventive Medicine, 15*, 451–474. https://doi.org/10.1016/0091-7435(86)90024-1

Frable, D. E. (1993). Dimensions of marginality: Distinctions among those who are different. *Personality and Social Psychology Bulletin, 19*, 370–380. https://doi.org/10.1177/0146167293194002

Glanz, K., & Bishop, D. B. (2010). The role of behavioral science theory in development and implementation of public health interventions. *Annual Review of Public Health, 31*, 399–418. https://doi.org/10.1146/annurev.publhealth.012809.103604

Goffman, E. (1963). *Stigma: Notes on the management of spoiled identity*. Englewood Cliffs, NJ: Prentice Hall.

Haidt, J., McCauley, C. R., & Rozin, P. (1994). Individual differences in sensitivity to disgust: A scale sampling seven domains of disgust elicitors. *Personality and Individual Differences, 16*, 701–713. https://doi.org/10.1016/0191-8869(94)90212-7

Hatzenbuehler, M. L., Phelan, J. C., & Link, B. G. (2013). Stigma as a fundamental cause of population health inequalities. *American Journal of Public Health, 103*, 813–821. https://doi.org/10.2105/AJPH.2012.301069

Hayes, A. F., & Scharkow, M. (2013). The relative trustworthiness of inferential tests of the indirect effect in statistical mediation analysis: Does method really matter? *Psychological Science, 24*, 1918–1927. https://doi.org/10.1177/0956797613480187

Jaccard, J., & Jaccoby, J. (2010). *Theory construction and model building skills: A practical guide for social scientists*. New York, NY: Guilford Press.

Jones, E. E., Farina, A., Hastorf, A. H., Markus, H., Miller, D. T., & Scott, R. A. (1984). *Social stigma: The psychology of marked relationships*. New York, NY: W. H. Freeman.

Link, B. G., & Phelan, J. C. (2001). Conceptualizing stigma. *Annual Review of Sociology, 27*, 363–385. https://doi.org/10.1146/annurev.soc.27.1.363

Link, B. G., Phelan, J. C., & Hatzenbuehler, M. L. (2014). Stigma and social inequality. In J. D. McLeod, E. Lawler, & M. Schwalbe (Eds.), *Handbook of the social psychology of inequality* (pp. 49–64). Dordrech, The Netherlands: Springer.

Lipsey, M. W. (1993). Theory as method: Small theories of treatments. In L. B. Sechrest & A. G. Scott (Eds.), *Understanding causes and generalizing about them* (pp. 5–38). San Francisco, CA: Jossey-Bass.

MacKinnon, D. P. (2000). Contrasts in multiple mediator models. In J. S. Rose, L. Chassin, C. C. Presson, & S. J. Sherman (Eds.), *Multivariate applications in substance use research: New methods for new questions* (pp. 141–160). Mahwah, NJ: LEA.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. New York, NY: Lawrence Erlbaum Associates.

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology, 58*, 593–614. https://doi.org/10.1146/annurev.psych.58.110405.085542

MacKinnon, D. P., & Pirlott, A. G. (2015). Statistical approaches for enhancing causal interpretation of the M to Y relation in mediation analysis. *Personality and Social Psychology Review, 19*, 30–43.

MacKinnon, D. P., Taborga, M. P., & Morgan-Lopez, A. A. (2002). Mediation designs for tobacco prevention research. *Drug & Alcohol Dependence, 68*, 1–25. https://doi.org/10.1016/S0376-8716(02)00216-8

Mason, W., & Suri, S. (2011). Conducting behavioral research on Amazon's mechanical Turk. *Behavioral Research, 44*, 1–23. https://doi.org/10.3758/s13428-011-0124-6

Maydeu-Olivares, A., Coffman, D. L., & Hartmann, W. M. (2007). Asymptotically distribution-free interval estimation for coefficient alpha. *Psychological Methods, 12*, 157–176. https://doi.org/10.1037/1082-989X.12.2.157

Michie, S., & Abraham, C. (2004). Interventions to change health behaviours: Evidence-based or evidence-inspired? *Psychology & Health, 19*, 29–49. https://doi.org/10.1080/0887044031000141199

Miller, C. T., & Major, B. (2000). Coping with stigma and prejudice. In T. F. Heatherton, R. E. Kleck, M. R. Hebl, & J. G. Hull (Eds.), *The social psychology of stigma* (pp. 243–272). New York, NY: Guilford Press.

Muller, D., Judd, C. M., & Yzerbyt, V. Y. (2005). When moderation is mediated and mediation is moderated. *Journal of Personality and Social Psychology, 89*, 852–863. https://doi.org/10.1037/0022-3514.89.6.852

Musso, D., Roche, C., Robin, E., Nhan, T., Teissier, A., & Cao-Lormeau, V. M. (2015). Potential sexual transmission of Zika virus. *Emerging Infectious Diseases, 21*, 359–361. https://doi.org/10.3201/eid2102.141363

Noar, S. M. (2007). An interventionist's guide to AIDS behavioral theories. *AIDS Care, 19*, 392–402. https://doi.org/10.1080/09540120600708469

O'Keefe, D. J. (2003). Message properties, mediating states, and manipulation checks: Claims, evidence, and data analysis in experimental persuasive message effects research. *Communication Theory, 13*, 251–274. https://doi.org/10.1093/ct/13.3.251

Pan, Z., & Kosicki, G. M. (1993). Framing analysis: An approach to news discourse. *Political Communication, 10*, 55–75. https://doi.org/10.1080/10584609.1993.9962963

Pescosolido, B. A., Martin, J. K., Lang, A., & Olafsdottir, S. (2008). Rethinking theoretical approaches to stigma: A framework integrating normative influences on stigma (FINIS). *Social Science & Medicine, 67*, 431–440. https://doi.org/10.1016/j.socscimed.2008.03.018

Piper, M. E., Schlam, T. R., Fraser, D., Oguss, M., & Cook, J. W. (2018). Implementing factorial experiments in real-world settings: Lessons learned while engineering an optimized smoking cessation treatment. In L. M. Collins & K. C. Kugler (Eds.), *Optimization of behavioral, biobehavioral, and biomedical interventions: Advanced topics*. New York, NY: Springer.

Pirlott, A. G., & MacKinnon, D. P. (2016). Design approaches to experimental mediation. *Journal of Experimental Social Psychology, 66*, 29–38.

Preacher, K. J., & Coffman, D. L. (2006, May). *Computing power and minimum sample size for RMSEA* [Computer software]. Available from http://quantpsy.org/

Preacher, K. J., Rucker, D. D., & Hayes, A. F. (2007). Assessing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate Behavioral Research, 42*, 185–227. https://doi.org/10.1080/00273170701341316

President's New Freedom Commission on Mental Health. (2003). *Achieving the promise: Transforming mental health care in America*. Bethesda, MD: U.S. Department of Health & Human Services.

Roberto, A. J. (2013). Editor's note for the extended parallel process model: Two decades later. *Health Communication, 28*, 1–2. https://doi.org/10.1080/10410236.2013.743748

Shen, L. J. (2016). Putting the fear back again (and within individuals): Revisiting the role of fear in persuasion. *Health Communication*. Advanced online publication. https://doi.org/10.1080/10410236.2016.1220043

Smith, R. A. (2007). Language of the lost: An explication of stigma communication. *Communication Theory, 17*, 462–485. https://doi.org/10.1111/j.1468-2885.2007.00307.x

Smith, R. A. (2011). Stigma communication and health. In T. L. Thompson, R. Parrott, & J. Nussbaum (Eds.), *Handbook of health communication* (2nd ed., pp. 455–468). London, UK: Taylor & Francis.

Smith, R. A. (2012). An experimental test of stigma communication features with a hypothetical infectious disease alert. *Communication Monographs, 79*, 522–538. https://doi.org/10.1080/03637751.2012.723811

Smith, R. A., Levine, T. R., Lachlan, K. A., & Fediuk, T. K. (2002). The high cost of complexity in experimental design and data analysis: Type 1 and Type II error rates in multi-way ANOVA. *Human Communication Research, 28*, 515–530. https://doi.org/10.1111/j.1468-2958.2002.tb00821.x

So, J., Kim, S., & Cohen, H. (2017). Message fatigue: Conceptual definition, operationalization, and correlates. *Communication Monographs, 84*, 5–29. https://doi.org/10.1080/03637751.2016.1250429

Tajfel, H. (1959). Quantitative judgement in social perception. *British Journal of Psychology, 50*, 16–29. https://doi.org/10.1111/j.2044-8295.1959.tb00677.x

VanderWeele, T. J. (2009). On the distinction between interaction and effect modification. *Epidemiology, 20*, 863–871. https://doi.org/10.1097/EDE.0b013e3181ba333c

VanderWeele, T. J., & Vansteelandt, S. (2014). Mediation analysis with multiple mediators. *Epidemiologic Methods, 2*, 95–115. https://doi.org/10.1515/em-2012-0010

Witte, K. (1992). Putting the fear back into fear appeals: The extended parallel process model. *Communications Monographs, 59*, 329–349. https://doi.org/10.1080/03637759209376276

World Health Organization. (2014). *Preventing suicide: A global imperative*. Geneva, Switzerland: World Health Organization.

# Index





L. M. Collins, K. C. Kugler (eds.), *Optimization of Behavioral, Biobehavioral, and Biomedical Interventions*, Statistics for Social and Behavioral Sciences,
https://doi.org/10.1007/978-3-319-91776-4