Design and Analysis of Sequential Randomized Trials with Applications to Mental Health and Online Education

by

Timothy NeCamp

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
at The University of Michigan
2019

Doctoral Committee:

Professor Edward Ionides, Co-Chair
Assistant Professor Zhenke Wu, Co-Chair
Associate Professor Daniel Almirall
Associate Professor Srijan Sen
Associate Professor Ambuj Tewari

Timothy NeCamp

tnecamp@umich.edu

ORCID iD: 0000-0002-0192-6230

ACKNOWLEDGEMENTS

There are many people who have been vital to my development as a statistician and human over the past 5+ years.

I learned many things from my advisor Ed Ionides. He was constantly making sure I was progressing both in my research pursuits and as a long-term career statistician. Ed gave me the freedom to pursue my interests while ensuring my path did not go too far off course.

Zhenke Wu, my co-advisor, was also critical to my success. He always left his door open to me and was happy to work with me anytime I was stuck. He listened attentively to my ideas, making me feel valued as a PhD student. Also, through example, he demonstrated the importance of enjoying other aspects of life (e.g., skiing, climbing) outside of statistics. This kept me sane throughout graduate school.

I've also had the pleasure of working closely with other members of my committee, for which I'm grateful. Danny Almirall was my first 'advisor', and he is the reason I published my first first-author publication. He opened my eyes to the world of sequential randomized trials and applied, impactful statistics. Danny also demonstrated how a statistician can be both a great researcher and caring human being.

Ambuj Tewari is one of the most passionate statisticians I have met. Whether we are discussing a theoretical concept or an interesting application area, my meetings with Ambuj were always full of energy and excitement.

Without Srijan Sen, I would have never had the opportunity to work on the rich, complex, and fascinating Intern Health Study datasets. I'm extremely grateful that I could piggyback off of the years of hard work required to get the Intern Health Study in its current state. Working on this project has allowed me to work in an application area I care deeply about.

Outside of my committee, there are so many others who have played a critical role in my PhD. Liza Levina has been a great PhD Chair and is always willing to talk. The Statistics Department staff members (Bebe Zuniga-Valentino, Gina Cornacchia, Andrea Nashar, and Judy McDonald-my stats mom) have provided so much support and made my last 5 years free of hassle. I've collaborated with some excellent students, including Mengjie Shi, Jessica Liu, Julia Lynn. Josh Gardner and Chris Brooks helped me apply my research ideas to an application area I care deeply about, (online) education.

The D3 lab (especially Billie Nahum-Shani, Nick Seewald, Brook Luers, Stephanie Carpenter, and Shawna Smith) and its brainstorms have always kept me passionate about statistics and its power to solve important problems. Presenting to the lab has been critical to ensuring I do high-quality research.

My fellow PhD friends have provided numerous shoulders to lean on. Andrew Schwartz and Lizz Ultee have been caring and supportive STEM friends/frawgs. Joey Dickens has been a great office neighbor and friend/collaborator the last 5 years. Kimberly Chuang and Elena Malik have saved my life many of times at the local climbing gym.

I've been fortunate enough to have funding opportunities that have given me the ability to further explore my research interests. The Rackham Merit Fellowship (and Emma Flores-Scott, in particular) has provided both funding and support throughout my entire PhD. The NSF Graduate Research Fellowship has also been critical to my success.

Service work and helping those in need is a major passion of mine. Working with Statistics in the Community (STATCOM)—with Evan Reynolds, Steve Salerno, Emily Morris, Ina Conrado, Sharanya Chandran, and Lauren Beriont—has allowed me to pursue that passion and use my statistical knowledge to impact my community. I also love to teach and thanks to Brenda Gunderson I had many valuable teaching opportunities throughout my

PhD. Through example, Brenda also helped me take my teaching abilities to the next level.

There were numerous people from my pre-PhD life that I'm grateful for. Laura Kubatko and Steve MacEachern, my statistics professors at Ohio State, planted the seed of inspiration that led me to a statistics PhD. I'm grateful for my 7th and 8th grade mathematics students in Las Vegas that I taught through Teach for America. They inspired me as a human being and showed me the importance of ensuring my work benefits society. My research would not have been as impactful without their inspiration.

I'd also like to thank my family (Mom, Dad, Stephen, Jon, Lindsey) for keeping the bottom 3 tiers of my Hierarchy of Needs met (especially belongingness and love) throughout my entire life, and especially the last 5 years. This has given me the ability and freedom to pursue my passion for statistics.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

Dynamic treatment regimes, also called adaptive interventions, guide sequential treatment decision-making in a variety of fields, including healthcare and education. Dynamic treatment regimes accommodate differences between individuals and changes in individuals over time. Sequential randomized trials are a specific type of trial design useful for developing high-quality dynamic treatment regimes. Sequential randomized trials utilize re-randomization of individuals over time in order to discover how to sequence, time, and personalize treatments. Two of the most commonly used sequential randomized trial designs are sequential multiple assignment randomized trials and micro-randomized trials.

In this thesis, we contribute to both the design and analysis of sequential randomized trials. We describe design considerations for sequential randomized trials in online education. We present the design and analysis for a sequential randomized trial developed to reduce dropout in a massively open online course. We also develop statistical methodology and sample size formulae for sequential multiple assignment randomized trial designs which include cluster-level randomization. The techniques are inspired by a trial aiming to develop high-quality dynamic treatment regimes for mental health clinics. Lastly, we illustrate the design, describe the analysis, and present results of a large micro-randomized trial aiming to develop mobile health interventions for improving medical interns' mental health.

CHAPTER I

Introduction

An intervention (or treatment) is anything provided or manipulated in order to improve an outcome for an individual or group of individuals. Interventions are found in a variety of fields. In health, interventions (such as psychiatry appointments or mindfulness apps) can help individuals overcome depression and adopt healthy behaviors. In education, interventions (such as motivational messaging or gamification) can help prevent learner dropout and improve course engagement.

In these fields, there is often heterogeneity in intervention effectiveness. Heterogeneity exists both between individuals (i.e., different individuals respond differently to interventions) and within an individual over time (i.e., an individual responds differently to an intervention at different times). To accommodate this heterogeneity, interventions should be adaptive. Dynamic treatment regimes, also called adaptive interventions or adaptive treatment strategies, are sequences of rules which specify how to adapt (or re-adapt) interventions across individuals over time (Murphy et al., 2001). The adaptation can provide different interventions to different types of individuals (to account for heterogeneity between individuals) and can provide different interventions to the same individual over time (to account for heterogeneity within individuals). In order to substantially improve individuals' outcomes, developing high-quality dynamic treatment regimes is critical.

When developing high-quality dynamic treatment regimes, questions often arise about timing, personalizing, and sequencing interventions: When are certain interventions most effective? Which variables should be used to decide future intervention delivery? What is the best sequence of interventions to provide? **Sequential randomized trials** are specifically designed to answer these questions and, in turn, create high-quality dynamic treatment regimes (Almirall et al., 2018). In sequential randomized trials, a subject is randomized *multiple* times to different intervention options throughout the experiment. Sequential randomized trials are advantageous over single-randomized trials (e.g., 2 arm randomized control trial); through re-randomization, researchers can answer important questions regarding timing, personalizing, and sequencing of interventions. Two of the most common types of sequential randomized trials are sequential multiple assignment randomized trials (SMARTs) (Murphy, 2005) and micro-randomized trials (MRTs) (Klasnja et al., 2015).

SMARTs are a type of sequential randomized trial in which there are typically a small number (one or two) of re-randomizations and future randomizations often depend on response to prior intervention. For example, in a prototypical SMART, individuals are first randomized to one of two different intervention options. Then, at a pre-specified decision point following the initial intervention, users who did not respond to the initial intervention are re-randomized to different intervention options. Users who did respond to the initial intervention are not re-randomized and continue with the efficacious initial intervention (Pelham et al., 2016).

MRTs are a type of sequential randomized trial which uses a large number (hundreds or thousands) of re-randomizations in order to understand the short-term (proximal) effects of interventions. MRTs are useful for interventions that can be delivered quickly and frequently, such as mobile health interventions (i.e., text messages, notifications, or other interventions delivered through mobile devices) (Klasnja et al., 2015). The high frequency of intervention delivery permits the large number of re-randomizations.

1.1 Gaps in the Design and Analysis of Sequential Randomized Trials

Due to the complexity introduced through re-randomization, researchers need to think carefully about the design and analysis of sequential randomized trials. Design of sequential randomized trials needs to thoughtfully account for both statistical and domain-science considerations, especially when being conducted in novel domains. Statistical methodology needs to be developed which can account for the complex randomization schemes in sequential randomized trials. Lastly, statistical analyses need to be thorough in order to exploit the rich data provided by a sequential randomized trial.

In this thesis, I contribute to both the design and analysis of sequential randomized trials. I describe the design considerations for sequential randomized trials in a novel application area, online education. I develop statistical methodology and sample size formulae for a novel sequential randomized trial design which includes cluster-level randomization. Lastly, I illustrate the design, describe the analysis, and present results of a large sequential randomized trial involving over 1,000 individuals being randomized for 6 months. This sequential randomized trial aims to develop mobile health interventions to improve individuals' mental health.

1.2 Overview of Thesis

In this section, I detail three specific projects involving the design and analysis of sequential randomized trials. For each of these projects, I first provide background information and then describe the key contributions.

1.2.1 Sequential Randomized Trials in Scaled Digital Learning Environments

Though prevalent in healthcare, sequential randomized trials are rarely used to develop high-quality dynamic treatment regimes in education (Hedges, 2018; Almirall et al., 2018). In scaled digital learning environments, such as massively open online courses or intelligent tutoring systems, sequential randomized trials are even less prevalent (we do not know of any examples previously). Experimental designs in scaled digital learning environments are typically designed with a single randomization, such as A/B tests or single-randomized factorial designs (Kizilcec and Brooks, 2017).

Developing dynamic treatment regimes with sequential randomized trials can be valuable in scaled digital learning environments. Data (such as course clicks, videos watched, problem completion, answer submissions) are collected quickly and easily. These data can be used to both evaluate and adapt interventions in real-time. Scaled digital learning environments also permit easy intervention delivery; since interactions in these environments are virtual, intervention delivery can be automated. Lastly, these online courses often have a large number of users and typically restart every few weeks, resulting in increased statistical power to test hypotheses and making it possible for repeated experimental iterations (e.g., exploratory trial followed by a confirmatory trial, Collins et al. 2007).

1.2.1.1 Contribution

In our work (NeCamp et al., 2019), we designed, implemented, and analyzed the first sequential randomized trial developed for a scaled digital learning environment. As dropout is a major issue in online courses, the trial aimed to develop a high-quality adaptive email intervention to remind students to return to the course before dropping out. Our work highlights the particular benefits of using sequential randomized trials in scaled digital learning environments. It also discusses special design considerations that are necessary when using sequential randomized trials in this setting.

1.2.2 Estimation Methodology and Sample Size Formulae for Clustered Sequential Randomized Trials

Most dynamic treatment regime development occurs for dynamic treatment regimes at the individual-level (Methodology Center, 2016). Individuals are provided a sequence of treatments and decision rules for how to adapt that treatment. In turn, sequential randomized trials used to develop high-quality dynamic treatment regimes also occur at the individual-level. Sometimes, however, dynamic treatment regimes may need to be delivered at the cluster-level. For example, dynamic treatment regimes may be delivered to an entire school or mental health clinic, with the goal of helping the students or patients within those clusters (e.g., Kilbourne et al. 2014).

In order to develop high-quality cluster-level dynamic treatment regimes, researchers can utilize a particular type of sequential randomized trial, the cluster-randomized SMART. In a cluster-randomized SMART, sequential randomization occurs at the cluster-level, with outcomes observed at the individual-level. Due to the inherent dependence between outcomes of individuals within the same cluster, standard statistical methodologies and sample size formulae need to be modified. Sample size formulae and estimation methodologies which account for this dependence do not exist.

1.2.2.1 Contribution

This work (NeCamp et al., 2017) makes two contributions to the design and analysis of cluster-randomized SMARTs. First, a weighted least squares regression approach is proposed for comparing the cluster-level dynamic treatment regimes embedded in a SMART. The regression approach facilitates the use of baseline covariates, which is often critical in the analysis of cluster-level trials. Second, sample size calculators are derived for two common cluster-randomized SMART designs. The methods are motivated by the Adaptive Implementation of Effective Programs Trial (Kilbourne et al., 2014), which is, to our knowledge, the first cluster-randomized SMART in psychiatry. The trial aims to develop high-quality dynamic treatment regimes to improve the implementation of evidence based practices in mental health clinics.

1.2.3 Timing Mobile Health Interventions with Sequential Randomized Trials

Mobile devices, such as smart phones and wearables, are an ideal platform for dynamic treatment regimes (Nahum-Shani et al., 2017). Mobile devices have the capability of delivering interventions, such as notifications or text messages, quickly and frequently. Mobile devices can also collect real-time data (e.g., step count, heart rate) which can be used to determine optimal times for delivering interventions.

A major problem with the creation of high-quality mobile health dynamic treatment regimes is not knowing, apriori, when to send interventions. For example, would it be better to send a depression coping message when someone is currently depressed and in need or to send the coping message prior to the onset of depression when they may be more amenable to behavior change? MRTs are able to answer questions about intervention timing (Klasnja et al., 2015).

Since MRTs are a relatively new trial design, the design and analysis of MRTs are non-standard. Designing MRTs also requires context specific considerations in order to ensure: (1) the intervention and intervention delivery is practical and useful to the population of interest, and (2) the resulting trial data can be used to answer the questions of interest.

1.2.3.1 Contribution

In this work, we designed and analyzed an MRT aiming to develop a high-quality mobile health intervention. The intervention seeks to improve the mental health of individuals in stressful work environments. More specifically, our intervention provides notifications to help users improve their mood, increase their physical activity, and obtain sufficient sleep. The primary and secondary aims of the study focused on assessing real-time intervention moderators, variables measured throughout the trial which change the efficacy of treatment. These moderators can subsequently be used to determine the best times to deliver interventions.

To our knowledge, this is the largest and longest MRT run to date. The large size of our study gives us the ability to detect real-time moderators. Our design also has a unique nested structure in order to answer questions at different time scales and accommodate the needs of our study population. We present the design, analysis methods, and results of our trial. We also discuss the implications of the results for the development of high-quality mobile health interventions.

1.3 Organization of Thesis

In the next three chapters, I provide the details of the three contributions highlighted above. In Chapter II, I discuss the use of sequential randomized trials in scaled digital learning environments (NeCamp et al., 2019). This chapter also provides a thorough description of sequential randomized trials and their benefits for developing high-quality dynamic treatment regimes. In Chapter III , I illustrate our sample size formulae and estimation methodology for cluster-randomized SMARTs (NeCamp et al., 2017). In Chapter IV, I describe the design and analysis of an MRT used to discover how to time mobile health interventions for depression. Finally, in Chapter V, I discuss future work in the design and analysis of sequential randomized trials.

BIBLIOGRAPHY

Almirall, D., Kasari, C., McCaffrey, D. F., and Nahum-Shani, I. (2018). Developing optimized adaptive interventions in education. *Journal of Research on Educational Effectiveness 11*(1), 27–34.

Almirall, D., Nahum-Shani, I., Wang, L., and Kasari, C. (2018). Experimental designs for research on adaptive interventions: Singly and sequentially randomized trials. In *Optimization of Behavioral, Biobehavioral, and Biomedical Interventions*, pp. 89–120. Springer.

Collins, L. M., Murphy, S. A., and 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.

Hedges, L. V. (2018). Challenges in building usable knowledge in education. *Journal of Research on Educational Effectiveness 11*(1), 1–21.

Kilbourne, A. M., Almirall, D., Eisenberg, D., Waxmonsky, J., Goodrich, D. E., Fortney, J. C., Kirchner, J. E., Solberg, L. I., Main, D., Bauer, M. S., et al. (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.

Kizilcec, R. and Brooks, C. (2017). Diverse big data and randomized field experiments in massive open online courses. In *The Handbook of Learning Analytics*, pp. 211–222. SOLAR.

Klasnja, P., Hekler, E., Shiffman, S., Boruvka, A., Almirall, D., Tewari, A., and Murphy, S. (2015). Microrandomized trials: An experimental design for developing just-in-time adaptive interventions. *Health Psychology 34*(S), 1220.

Methodology Center (2016, May). Example SMART studies. 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. A., van der Laan, M. J., Robins, J. M., and CPPRG (2001). Marginal mean models for dynamic regimes. *Journal of the American Statistical Association 96*, 1410–1423.

Nahum-Shani, I., Smith, S. N., Spring, B. J., Collins, L. M., Witkiewitz, K., Tewari, A., and Murphy, S. A. (2017). Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine 52*(6), 446–462.

NeCamp, T., Gardner, J., and Brooks, C. (2019). Beyond A/B testing: Sequential randomization for developing interventions in scaled digital learning environments. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, pp. 539–548. ACM.

NeCamp, T., Kilbourne, A., and Almirall, D. (2017). Comparing cluster-level dynamic treatment regimens using sequential, multiple assignment, randomized trials: Regression estimation and sample size considerations. *Statistical Methods in Medical Research 26*(4), 1572–1589.

Pelham, W. E., Fabiano, G. A., Waxmonsky, J. G., Greiner, A. R., Gnagy, E. M., Pelham, W. E., Coxe, S., Verley, J., Bhatia, I., Hart, K., et al. (2016). Treatment sequencing for childhood ADHD: A multiple-randomization study of adaptive medication and behavioral interventions. *Journal of Clinical Child & Adolescent Psychology 45*(4), 396–415.

CHAPTER II

Beyond A/B Testing: Sequential Randomization for Developing Interventions in Scaled Digital Learning Environments

This work is originally published in NeCamp et al. (2019).

2.1 Introduction

In order to continually improve learner experience and maintain engagement, scaled digital learning environments (SDLEs), such as Massive Open Online Courses (MOOCs), intelligent tutoring systems, and open-ended digital educational games, utilize *interventions*. Interventions are modifications made to the learning environment or learners' experience of it, including changing current course content, prompting users to return to the course, or providing additional learning resources (e.g., Davis et al. 2018). It is common to find interventions which are technological, pedagogical, or programmatic in their implementation.

In SDLEs, there is typically diversity both *between* learners (e.g., different learners have different needs) (Kizilcec and Brooks, 2017), and *within* a learner over time (e.g., a learner's engagement may change throughout a course) (Kizilcec et al., 2013). To accommodate this diversity, interventions should be adaptive. *Adaptive interventions* can change based on the type of learner (to account for diversity between learners) and change as a learner progresses through the course (to account for diversity within a learner) (Almirall et al., 2018). For example, consider an intervention which introduces review problems throughout a course. The frequency of review problems might adapt based on timing (e.g., learners may not need to review as often in later weeks). The frequency of review might also adapt based on learner performance, building upon work on theories of spaced repetition (Reynolds and Glaser, 1964).

Due to the large number of users and ease of content manipulation in SDLEs, randomized controlled trials, such as A/B tests, are often used to evaluate intervention options. Typically, these experiments are single randomized trials where each subject is randomized once, and assigned to a single intervention option for the entire trial. However, when the goal is to design a high-quality adaptive intervention in SDLEs, researchers may have important questions about the sequencing, timing, and personalization of intervention options which cannot be answered by A/B tests.

In this work, we discuss and demonstrate the advantages of an experimental design for developing high-quality adaptive interventions in SDLEs: the *sequential randomized trial* (SRT). In SRTs, a subject is randomized several times to different intervention options throughout the experiment. Sequential randomization is beneficial over one-time randomization for several reasons. Firstly, by re-randomizing, subjects receive a variety of intervention sequences, and these various sequences can be compared to discover the optimal intervention sequence. Secondly, instead of only being able to assess the overall effect of receiving one particular treatment, sequential randomization lets researchers discover effects at smaller time-scales (e.g., treatment A does better in week 2 of the course, but treatment B does better in week 3 of the course). These discoveries inform at what time points certain interventions are most effective. Thirdly, re-randomization permits the discovery of important variables measured throughout the course for adapting and personalizing an intervention. As opposed to only being able to discover baseline personalization variables (e.g., treatment A works better for women), we can also discover mid-course personalization variables (treatment A works better for subjects who have been active in the course during the previous day).

The chapter is structured as follows: We provide an overview of related prior work on experimentation and personalized interventions in SDLEs in Section 2.2. In Section 2.3, we formally introduce SRTs and compare them to other common trial designs. In Section 2.4, to motivate the design we describe a novel SRT performed in MOOCs, the Problem-based Email Reminder Coursera Study (PERCS). This was an experiment aiming to improve student retention in an online course by developing a high-quality adaptive email intervention which leverages aspects of cultural diversity and inclusion (Aceves and Orosco, 2014). This case study both serves to illustrate the advantages of SRTs and provide context regarding implementation and analysis. Section 2.5 details three specific advantages of running SRTs. These advantages are exemplified by showing specific results from PERCS. We conclude by providing some practical recommendations for researchers designing their own SRTs in Section 2.6.

2.2 Prior Work

2.2.1 Adaptive Interventions

Adaptive interventions have been shown to be useful in SDLEs, and have been used extensively in web-based adaptive educational systems (Brusilovsky and Peylo, 2003). For instance, in Pardos et al. (2017), learners were provided personalized content recommendations based on their clickstream data, and in Davis et al. (2018), learners were provided personalized feedback on their learning plans. Adaptive sequences of interventions have also been developed in MOOCs. For example, in David et al. (2016), sequences of problems were adapted based on predictions of student knowledge acquisition. Similarly Davis et al. (2018) chose quiz questions based on course content accessed previously by the learner. While these are only a few examples of adaptive interventions in large scale learning environments, they motivate our desire to improve the process by which such interventions are developed.

In the work discussed, adaptive interventions were developed using learning theory and intention (Davis et al., 2018), or prediction and machine learning (Pardos et al., 2017; Yu et al., 2017; David et al., 2016). In all examples, experimentation was not used to develop the adaptive intervention. In some cases (Davis et al., 2018; David et al., 2016), experimentation was used for evaluating the adaptive intervention. In these cases, an A/B test is used to compare the designed intervention to a control.

Reinforcement learning techniques, such as multi-armed bandits and contextual bandits, are a type of adaptive intervention which combine exploration (often done through implicit experimentation) and optimization. They use real-time data to learn how to adapt and have also been shown to be useful in SDLEs (Liu et al., 2014; Clement et al., 2013; Segal et al., 2018; Rafferty et al., 2018).

2.2.2 Experimental Designs

Experimentation in SDLEs is a common tool for evaluating interventions. Unlike quasi- or natural experimental settings (González-Brenes and Huang, 2015; Mullaney and Reich, 2015), by randomly assigning interventions, effects of interventions can be separated from effects of confounders (variables that relate both to the type of treatment received and to subsequent outcomes).

A/B tests are a valuable experimental design for improving content and evaluating interventions in MOOCs (Savi et al., 2017). In Renz et al. (2016), for example, A/B tests evaluated emails and course on-boarding to improve learner engagement and prevent dropout. In Davis et al. (2016), A/B tests were used to test the effectiveness of self-regulated learning strategies. Kizilcec and Brooks (2017) survey prior work utilizing A/B tests in MOOCs to evaluate nudge interventions and test theory-driven course content changes.

There has also been considerable work in other types of experimental designs beyond

A/B testing in SDLEs. *Factorial designs* (Lomas et al., 2013) are common ways to evaluate multiple experimental factors simultaneously. *Automatic experimentation* (Liu et al., 2014), where algorithms are used to search through different intervention options, is another alternative to A/B testing. Though automatic exploration of intervention options may be more efficient, intervention options are still evaluated with single randomized trials. *Adaptive experimental designs* change the randomization probabilities to favor efficacious treatment, with the goal of both evaluating treatment and helping learners currently in the trial (Chow and Chang, 2008). Adaptive designs have been used in SDLEs (Williams et al., 2018). We address the differences between SRTs and these other kinds of designs in Section 2.3.2.

2.3 Sequential Randomized Trials

2.3.1 An Overview

SRTs are trials where an individual is randomized multiple times throughout the course of the trial. Suppose there are two intervention options, such as learners receiving videos taught by a female instructor (intervention A) or a male instructor (intervention B). The simplest example of a SRT would be: During week 1, users have a 50% chance of receiving intervention A and a 50% chance of receiving intervention B. During week 2, users are re-randomized to another treatment. They again have a 50% chance of receiving either intervention A or B, independent of their week 1 treatment or activity. Hence, about 25% of users will have received one of each sequence (A, A), (A,B), (B,A), or (B,B), where the parenthetical notation means (week 1 treatment, week 2 treatment).

This simple SRT can be modified for both practical and scientific reasons. Common modifications include using different time durations (e.g., re-randomize every month), increasing the number of time points (e.g., each person is randomized 10 times instead of twice), changing the number of treatments (e.g., A vs B vs C instead of A vs B, or A vs B in week 1 and C vs D in week 2), and altering the randomization scheme (e.g., not having uniform randomization probabilities each week).

SRTs (Lavori and Dawson, 2004, 2000) have become increasingly common in clinical settings (Lei et al., 2011) but are less common in educational settings (Almirall et al., 2018; Kasari et al., 2014) and are even rarer in SDLEs. Two of the most common types of SRTs are Sequential Multiple Assignment Randomized Trials (SMARTs) (Murphy, 2005), and Micro-randomized Trials (MRTs) (Klasnja et al., 2015).

SMARTs are often used in settings where, for either practical or ethical reasons, future randomization and treatment assignment should depend on response to prior treatment. For example, in a prototypical SMART, individuals are first randomized to one of two different treatment options. Then, at a pre-specified decision point following initial intervention, users who did not respond to initial treatment are re-randomized, while users who did respond to initial treatment are not re-randomized and continue with the efficacious initial treatment (Pelham Jr et al., 2016).

MRTs are useful for interventions that can be delivered quickly and frequently (such as delivering text messages or notifications to a subject's phone). Typically, the goal of an MRT is to estimate the short-term effect of these interventions and understand how that effect depends on time and context. MRTs have been mostly used in the mobile health space (Klasnja et al., 2015). Due to the high frequency of intervention delivery, users in an MRT are typically re-randomized hundreds or thousands of times.

2.3.2 Comparisons to Other Designs

2.3.2.1 Single Randomized Trials

Single randomized trials, such as A/B tests and factorial designs, are trials where subjects are randomized **one time**. In A/B tests (often called a 2-arm randomized controlled trial in healthcare and education), each subject is randomized one time to either intervention A or intervention B. Factorial designs are an extension of A/B tests where each subject is randomized one time to several intervention components simultaneously. SRTs differ from single randomized trials because in SRTs, subjects are randomized several times, causing them to receive different treatments at different times throughout the trial.

Single randomized trials can still be used to evaluate adaptive interventions. For example, if treatment A is defined as an adaptive intervention (e.g., a fixed sequence of different intervention options) and treatment B is defined as a control, an A/B test can compare the adaptive intervention to the control with a single randomization. However, this A/B test is limited in answering questions about sequencing, timing, and personalizing.

A/B tests are often used in confirmatory trials. Confirmatory trials are trials used to ensure strong evidence (or additional evidence) of a treatment's efficacy. In contrast, SRTs are useful as exploratory trials which explore a large number of possible treatment sequences and learn how to adapt those sequences. After running a SRT and developing a high quality adaptive intervention, the intervention can be confirmed in a simple A/B confirmatory trial (Collins et al., 2007).

SRTs can also be thought of as factorial designs, where users are sequentially randomized to each factor over time. In the simplified SRT example in Section 2.3.1, the design can be considered a 2×2 factorial design where factor 1 is week 1 treatment, and factor 2 is week 2 treatment (Almirall et al., 2014).

2.3.2.2 Online Optimization Designs

There are other designs (both experimental and not) which aim to optimize intervention delivery while simultaneously collecting data. In adaptive trial designs (Chow and Chang, 2008), randomization probabilities are changed throughout the trial in order to both provide efficacious treatment to users, and still obtain good estimates of treatment effects. Online reinforcement learning methods (such as multi-armed bandit and contextual bandit algorithms) can also be used for optimizing intervention delivery (Liu et al., 2014; Clement et al., 2013; Segal et al., 2018; Rafferty et al., 2018).

SRTs are distinctive from adaptive trial designs and online reinforcement learning methods since they do not use data collected during the trial to change randomization schemes. In reinforcement learning language, SRTs can be seen as pure exploration with no exploitation. There are advantages of not using earlier trial data to inform future treatment allocation:

1. Using online optimization techniques can cause bias in estimating treatment effects (Nie et al., 2017; Rafferty et al., 2018). These bias issues do not arise in SRTs.

2. SRTs provide rich exploratory data for discovering which variables are valuable for informing treatment decisions, making them useful when these variables are unknown (see Section 2.5.3). In contrast, many reinforcement learning algorithms, such as contextual bandits (Lan and Baraniuk, 2016), require these variables to be pre-specified.

3. SRTs can actually utilize reinforcement learning methods. Batch off-policy reinforcement learning algorithms (such as Q-learning) can be applied to SRT data to discover an optimal adaptive intervention, as in Zhao et al. (2009).

2.4 Applications of Sequential Randomized Trials

SRTs can inform a large variety of interventions including course content sequencing, course material manipulations, and learner nudges (such as encouraging messages and problem feedback). We highlight three examples. The first two examples are hypothetical scenarios to demonstrate different types of possible SRTs. The third example, PERCS, is a SRT run in a data science MOOC and is the main working example.

2.4.1 Video Optimization

For each video in a MOOC, there are two versions, one video which shows only slides and one which shows the instructor's head occasionally interspersed with the slides (Guo et al., 2014). Researchers are unsure about which sequence of videos are better. Learners might prefer only slides, some might prefer those with instructors, or some might prefer a variety of videos. Also, there may be learner characteristics that affect learner preference. For example, learners who initially performed poorly after watching an instructor video might be better off seeing a slide-only video. To provide insight into these questions and hypotheses, a researcher could run a SRT: When a learner enters a course, they are randomized initially to receive an instructor video or a slide only video. Then, after watching the first video, they are re-randomized to either instructor or slide video for the next video they watch. This continues through the entire course.

2.4.2 Content Spacing

Researchers are often unsure about optimal review problem sequencing to maximize knowledge retention while minimizing review time (Reynolds and Glaser, 1964). Learners may benefit from frequent review at the beginning, with less frequent review later. These benefits may also be dependent on certain learner characteristics. For example, poor-performing learners may benefit from more frequent review. A SRT can answer these questions: For a problem recommendation system, a learner starts the recommender for a time window (e.g., 50 problems). Then, after this time period, every learner is randomized to one of 3 groups: no review, minimal review, large review. The grouping determines how often they receive previously-seen problems (for review) during the next 20 problems. If a learner is in the no review, minimal review, or large review group they will receive previously-seen problems 0%, 5%, or 20% of the time, respectively. After completing the next 20 problems, each user will be re-randomized to one of the 3 groups. This randomization scheme continues. After every 20 problems completed, a user is re-randomized to one of the 3 groups.

2.4.3 Problem-based Email Reminder Coursera Study (PERCS)

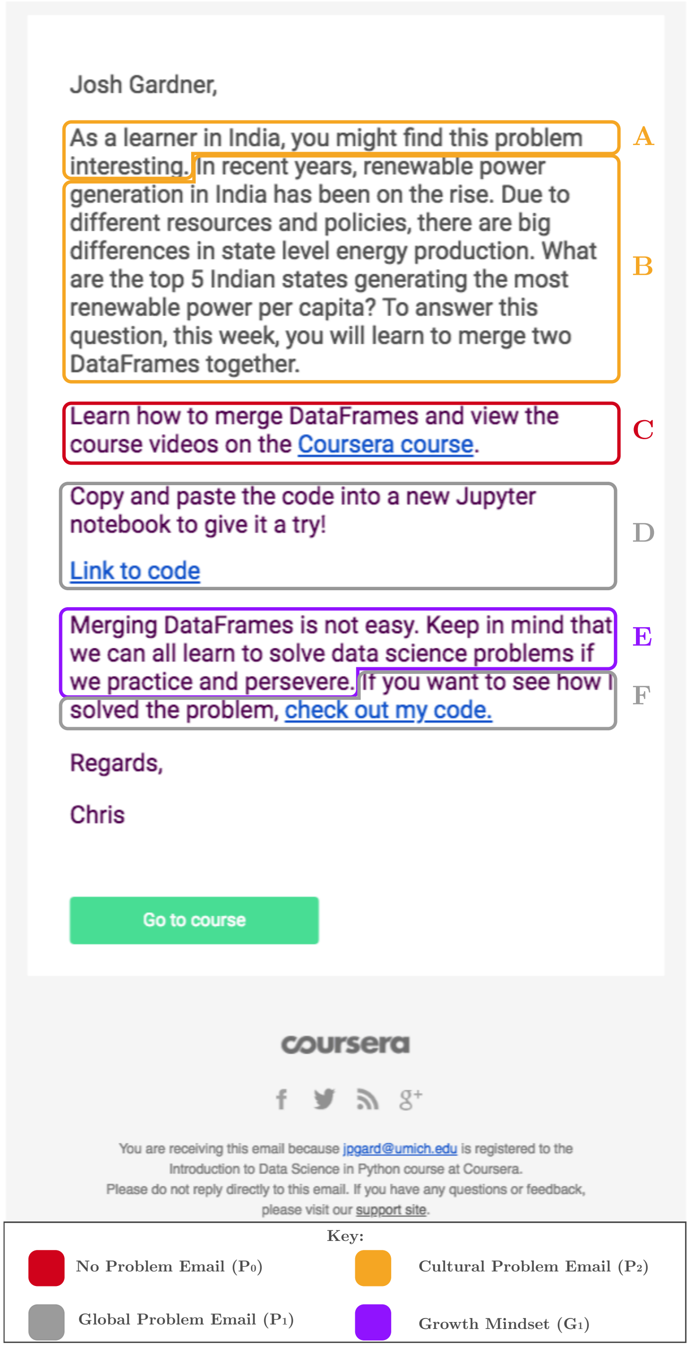
2.4.3.1 Motivation

A well-known challenge in MOOCs are the low completion rates. While there are many factors contributing to MOOC dropout (Greene et al., 2015), the goal of PERCS is to determine whether dropout can be ameliorated by using weekly email reminders to motivate learners to engage with course content. Our context of inquiry was the Applied Data Science with Python Coursera MOOC, taught by Christopher Brooks. Weekly emails were sent to learners and may have contained one or more of several factors intending to impact learner engagement (for an example, see Figure 2.1):

1. The email could have contained a motivating data science problem to challenge the user to learn the upcoming week's content. This factor was based on evidence from the problem-based learning literature suggesting that situating instruction in the context of problems is an effective way to engage learners (Barrows, 1985).

2. The email might also have contained a location-specific primer and a data science problem relevant to that user's specific culture (e.g., an Indian user might receive a problem about Bollywood or weather patterns in India). This factor was based on the work in the culturally relevant and culturally responsive pedagogy communities, where situating instruction in a manner that considers the local context is seen as beneficial (Aceves and Orosco, 2014).

3. The email may have utilized growth mindset framing (Dweck, 2008), a psychological framing method used to support learning. While growth mindset has been heavily studied, its effects are in dispute (Kaijanaho and Tirronen, 2018), and growth mindset framing has seen only limited application in SDLEs.



Josh Gardner,

As a learner in India, you might find this problem interesting. In recent years, renewable power generation in India has been on the rise. Due to different resources and policies, there are big differences in state level energy production. What are the top 5 Indian states generating the most renewable power per capita? To answer this question, this week, you will learn to merge two DataFrames together.

Learn how to merge DataFrames and view the course videos on the Coursera course.

Copy and paste the code into a new Jupyter notebook to give it a try!

Link to code

Merging DataFrames is not easy. Keep in mind that we can all learn to solve data science problems if we practice and persevere. If you want to see how I solved the problem, check out my code.

Regards,

Chris

Go to course

coursera

You are receiving this email because jpgard@umich.edu is registered to the Introduction to Data Science in Python course at Coursera.
Please do not reply directly to this email. If you have any questions or feedback, please visit our support site.

A
B
C
D
E
F
Key:
🟥 No Problem Email (P_0)
🟧 Cultural Problem Email (P_2)
⬜ Global Problem Email (P_1)
🟪 Growth Mindset (G_1)

Figure 2.1: Example email structure using content which is populated based on their assigned treatments. (A) Identity activation prompt. (B) Culturally-relevant problem using data related to geographic identity. (C) Reminder to return to course. (D) Link to problem code. (E) Growth mindset framing. (F) Link to problem solution. Elements of the same email type are grouped by color (see key).

Given these weekly email options, we hope to develop an adaptive intervention of weekly email reminders to increase engagement and reduce dropout. A high-quality adaptive intervention should sequence emails in a way that promotes engagement continually through the course. Since different emails may be more or less effective during different weeks in the course, each week, the intervention should send the most effective email for that week. Finally, since we might not expect the same email to work well for everyone, the intervention should also adapt to the learner's current course behavior. In order to develop a high quality adaptive intervention of emails, we need to answer the following research questions:

RQ1 **Sequencing**: Which sequence of emails most improves course activity in later weeks?

RQ2 **Timing**: Which email problem type is most effective, on average, for bringing learners back to the course during each week?

RQ3 **Personalization**: Are certain data science problem emails more or less effective for active learners?

2.4.3.2 Design

A sequentially randomized factorial trial design was an effective method to jointly address the main research questions of PERCS. At the end of weeks one, two, and three of the four-week long MOOC, learners were randomly assigned to receive one of four different email categories: an email message with a problem that reflects their geo-cultural situation based on IP address (*cultural problem email*), an email with a generic non-culture specific problem (*global problem email*), an email with *no problem*, or *no email* at all. Those who received an email were uniformly randomly assigned to have the email be framed with or without growth mindset.

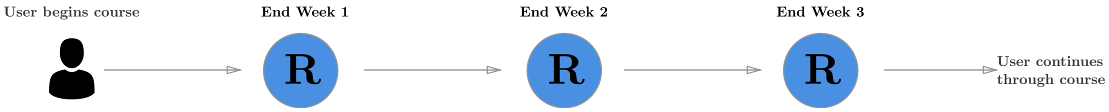


Figure 2.2: PERCS trial design. “R” indicates stages where randomization is conducted according to the probabilities in Table 2.1.

No Email E_0
0.14 (T_1)

Email E_1	**No Problem Email (P_0)**	**Global Problem Email (P_1)**	**Cultural Problem Email (P_2)**
No-Growth Mindset (G_0)	0.14 (T_2)	0.14 (T_3)	0.14 (T_4)
Growth Mindset (G_1)	0.14 (T_5)	0.14 (T_6)	0.14 (T_7)

Table 2.1: An overview of the probability learners will be assigned to a treatment T_n. Individual treatments are shown in white cells, while groups of treatments are referred to by the tab row or column headers (e.g., all cultural problem emails as P_2)

The growth mindset factor crossed with the three email categories makes each week of PERCS a 2×3 factorial design with an additional control condition of no email. See Table 2.1 for each week’s factorial design and randomization probabilities.

Figure 2.1 illustrates an example email. The emails were developed in a “cut and paste” format: When adding in a condition to an email (e.g., growth mindset framing or adding a problem), we do not change other aspects of the email, and simply insert text from the relevant condition into a consistent email template. Using the cut and paste across different conditions allows us to attribute treatment effects purely to the condition being added, and not other aspects of the email. Emails are delivered directly using the Coursera platform’s email capabilities for instructors.

The most novel aspect of PERCS is the sequential randomization. That is, a particular learner is not assigned to a single fixed email condition for all three weeks. Instead, as shown in Figure 2.2, a learner is re-assigned (with the same randomization probabilities) to a different email condition each week. The randomizations across weeks are independent, hence in PERCS, there are 7^3 different possible email sequences students may receive.

2.4.3.3 Notation

Throughout the rest of the chapter, we will be referring to specific treatments, groups of treatments, and sequences of treatments in PERCS. We introduce notation to ease our description. As shown in Table 2.1, $T_1, T_2, \ldots, T_7$ refers to a particular email. For example, T_5 is an email containing no problem, and growth mindset framing. To refer to groups of emails, we used labeling contained in the column and row headers in Table 2.1. E is used to group together conditions where any email was sent (E_1) vs no email being sent (E_0). So users in E_1 refer to any user receiving emails T_2 through T_7. G is used to group together growth mindset emails (G_1) and non-growth mindset emails (G_0). Users in G_0 refer to any user receiving emails $T_2, T_3,$ or T_4. P is used to group together no problem emails (P_0), global problem emails (P_1), and cultural problem emails (P_2). Thus users in P_1 are any users receiving emails T_3 or T_6. Lastly, we use parenthetical notation to describe the sequences of emails over the three weeks, i.e., (week 1 treatment, week 2 treatment, week 3 treatment). As an example, we would refer to users who received a global problem email in week 1, any email in week 2, and a no-growth mindset email in week 3 using (P_1, E_1, G_0).

2.4.3.4 Experimental Population

We focused our trial on the two largest populations of learners enrolled in the course as determined by IP address, Indian and US-based learners. All Indian and US learners who signed up for the Applied Data Science with Python Coursera MOOC between April 1 and June 10, 2018 participated in PERCS. A total of 8,681 unique learners (3,455 Indian, 5,226 US) were sent 22,073 emails.

2.4.3.5 Single randomized version of PERCS: PERCS-AB

To highlight the advantages of sequential randomization, PERCS will be compared to the single randomized version of PERCS, PERCS-AB. PERCS-AB has the exact same randomization probabilities as PERCS, however in PERCS-AB, learners are randomized only at the end of week 1 to one of the 7 email types. They are then sent that exact same email type in weeks 2 and 3. Hence, there are only 7 possible sequences (T_1, T_1, T_1), $(T_2, T_2, T_2), \ldots, (T_7, T_7, T_7)$.

2.5 Sequencing, Timing, and Personalizing Interventions through SRTs

.

Next we highlight how SRTs can provide answers to questions regarding sequencing, timing, and personalizing interventions. For each type of question, we provide (1) an overview of the question and contextualize it within PERCS, (2) evidence of why SRTs are beneficial for answering that question as illustrated through comparing PERCS and PERCS-AB, (3) further discussion, and (4) answers to the question in the context of PERCS.

2.5.1 Sequencing

2.5.1.1 Overview

When many different intervention options can be delivered at different times, proper sequencing of interventions is critical. An intervention that worked at one time may not work at a later time. Also, receiving the same intervention multiple times may be less effective than receiving a variety of interventions. Researchers may not know which sequence of intervention options will lead to the best outcomes for learners. For PERCS, RQ1 refers to sequencing– are there sequences of emails which improve course activity in the later weeks?

2.5.1.2 Advantages of SRTs

SRTs provide data that allows experimenters to compare different sequences of interventions. By re-randomizing learners, learners receive a variety of treatment sequences. In

PERCS, for example, with the large sample size, there are learners randomly assigned to each of the 7^3 possible sequences of treatments (e.g., (T_1, T_1, T_1), (T_1, T_1, T_2), (T_1, T_2, T_2), or (T_2, T_1, T_2)).

By randomizing learners to each possible treatment sequence, researchers are now able to compare these treatment sequences. For example, researchers might hypothesize that learners only need a cultural problem email in the first week to believe the content is inclusive. They could then compare the course activity of users initially receiving a cultural problem email followed by two global problem emails (P_2, P_1, P_1) vs users receiving a cultural problem email all three weeks (P_2, P_2, P_2). When thinking of PERCS as a $7 \times 7 \times 7$ factorial design (where each week's treatment is a different factor), comparing sequences is analogous to simple effects analysis in factorial designs.

In A/B tests, since learners are not re-randomized, only sequences where each learner received the same treatment every time can be compared. In PERCS-AB, learners are only randomized to 7 possible sequences and thus comparisons can only be done between these 7 sequences. If there is any benefit to receiving different interventions (and/or in a different order) then this could not be discovered by PERCS-AB. However one would note that all of the comparisons of PERCS-AB can be done with data collected through PERCS, coming at the cost of a reduced sample size. This tradeoff demonstrates again why SRTs are especially well suited for MOOC experimentation, where there is a large number of diverse participants (and thus a broad exploration might be suitable).

2.5.1.3 Discussion

Often times, researchers are not interested in such specific sequence comparisons. Instead, they are interested in questions about sequences of groups of interventions. In this case, one could perform similar comparisons, but combining users over all of these groups. For example, in PERCS, we can assess how often reminder emails should be sent. Is it better to space emails out weekly or bi-weekly?

To answer this question, we could compare course activity of users receiving the sequence which spaces emails out by 2-weeks (E_1, E_0, E_1), to the sequence which sends emails every week (E_1, E_1, E_1). The sample size for this comparison will be much larger than for the individual sequence comparisons.

In addition to comparing groups of sequences, researchers might only be interested in comparing sequences for a small number of time points. For example, we could use data collected in PERCS to understand if the effects of week two problem-based emails on course activity were different based on what type of problem-based email the user had received in week one. Specifically, does receiving no email vs cultural problem-based email in week one change the benefits of receiving a cultural problem email in week two? To answer this question, we compare course activity in week two of users receiving sequences (E_0, P_2,any) vs (P_2, P_2,any). Notice we are not concerned with week three assignment, so we include sequences with any treatment in week three (i.e., any includes $T_1, T_2, \ldots, T_7$).

2.5.1.4 Results from PERCS

We now present results addressing RQ1. Since there are a large number of potential sequences, here we focus on the highest-level comparison: understanding the proper sequence of any email and no email. To do the comparison we look at the proportion of students returning to the course in week 4 (the final week) after receiving a given sequence of email (E_1) and no email (E_0), as shown in Figure 2.3.

We note some important observations. For US learners, sequences with emails sent in the first week (i.e., those sequences which start with E_1, the last four columns in Figure 2.3), tend to be slightly more beneficial than sequences without. For Indian learners this is not true.

Overall, for both countries, the confidence intervals across sequences are mostly overlapping, indicating that differences in effects of various email sequences are not significant but only suggestive. The size of the confidence intervals change due to the probability of

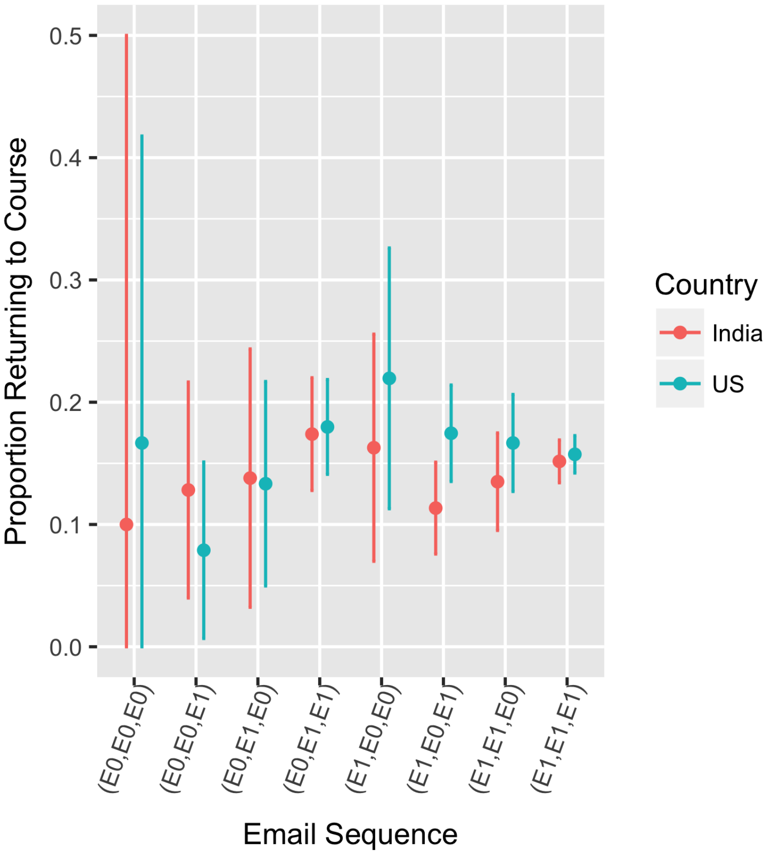


Figure 2.3: 95% confidence intervals for the proportion of students returning to the course in week 4 after receiving a given three week sequence of emails (E_1) and no email (E_0)

receiving an email (E_1) being larger than the probability of not receiving an email (E_0), and are not induced by the treatment itself.

2.5.2 Timing

2.5.2.1 Overview

Instead of being interested in a sequence of interventions, researchers might want to know about the effect of an intervention option at a particular time point. A treatment that was effective at the beginning of the course may be less effective towards the end of the course. By understanding effects of intervention options at a time point, designers can build adaptive interventions which deliver the optimal treatment at all times. For PERCS, RQ2 regards timing– which email type is most effective during each week?

2.5.2.2 Advantages of SRTs

Answering research questions about treatment timing is done by estimating average treatment effects, i.e., the average effect of a given treatment option at a given time point. Here, the averaging is over all prior treatment received, allowing the effect to be only for the particular time point of interest. When thinking of PERCS as a $7 \times 7 \times 7$ factorial design (where each week's treatment is a different factor), estimating average treatment effects is analogous to main effects analysis in factorial designs.

SRTs permit the estimation of average treatment effects at various time points. By re-randomizing individuals, we can separately estimate average treatment effects at each time point and eliminate dependence on treatment delivered previously.

To exemplify, in PERCS, to understand which email type is most effective in week 3, we compare the average effect of cultural problem emails in week 3 (any, any, P_2) compared to no email in week 3 (any, any, E_0). By doing this comparison, we average over all treatments delivered prior to week 3, and isolate the effect of interest to emails only in week 3. By re-randomizing, the individuals receiving a cultural problem email (P_2) in week 3 and individuals receiving no email (E_0) in week 3 both have, on average, the same distribution of treatments delivered prior to week 3. Hence, the comparison at week 3 is under the same prior treatment distribution for both groups in the comparison.

In PERCS-AB, such a comparison is impossible. Since individuals receive the same email all three weeks, the individuals receiving a cultural problem email in week 3 had a different prior treatment distribution compared to those receiving no email in week 3. This makes the comparison at week 3 implicitly dependent on these different prior treatment distributions.

In SRTs, average treatment effects can also be estimated for outcomes measured *after* the next re-randomization (sometimes called delayed effects Murphy 2005), by averaging over future treatment. For example, in PERCS (but not PERCS-AB), we are able to estimate the average effect of cultural problem emails in week 2 on week 4 course activity by averaging over week 3 treatment.

2.5.2.3 *Discussion*

Research questions on average treatment effects often seem similar to questions regarding sequence effects. Three different research questions in PERCS elucidate the differences:

(1) What is the effect of receiving a cultural problem email in week 3, after not receiving any email prior, (E_0, E_0, P_2) vs (E_0, E_0, E_0)?

(2) What is the average effect of receiving a cultural problem email in week 3, (any, any, P_2) vs (any, any, E_0)?

(3) What is the effect of receiving a cultural problem email every week until week 3, (P_2, P_2, P_2) vs (E_0, E_0, E_0)?

Questions 1 and 3 are questions of comparing sequences of treatments, while question 2 is about average treatment effects. Note that all three questions can be answered by PERCS, but only question 3 can be answered by PERCS-AB.

Also, one advantage of analyzing average treatment effects is that the sample size is typically larger than when comparing individual sequences of treatments, since the comparison does not restrict users based on what they received before or after the given week of interest.

2.5.2.4 *Results from PERCS*

To assess RQ2, for each week, we perform logistic regression with a binary outcome indicating whether the user clicks anything in the course during the week after receiving an email. The results are in Figure 2.4. Negative values indicate a reduced chance of returning to the course, compared to no email (E_0), while positive values indicate an increased chance of returning to the course.

The results in Figure 2.4 show that the impact of emails on Indian learners in weeks 2

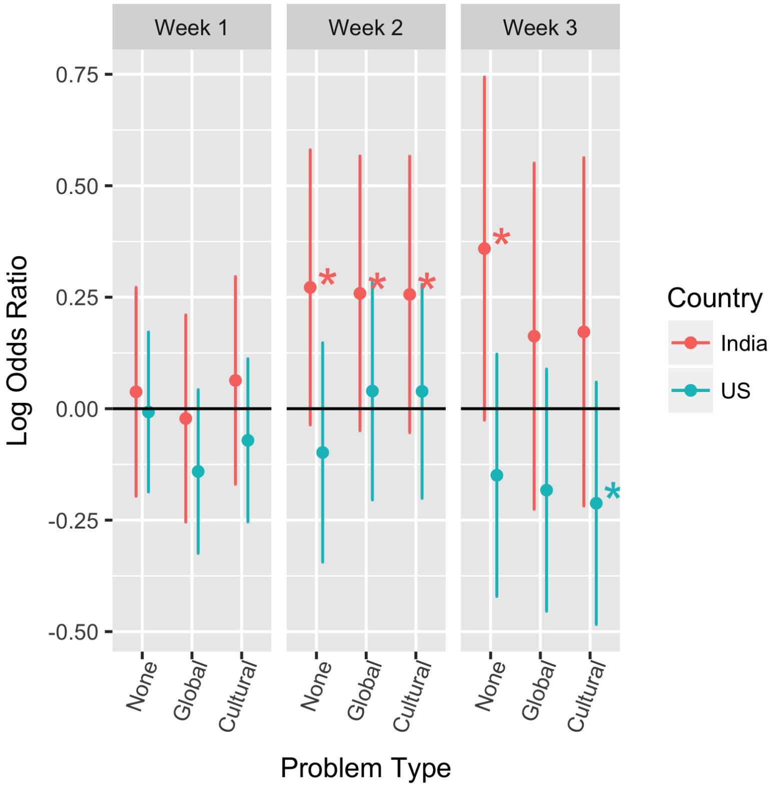


Figure 2.4: 95% confidence intervals of the log odds ratios of probability of returning to the course in the subsequent week for no problem (P_0), global problem (P_1), or cultural problem (P_2) emails, when compared to no email (E_0). * indicates a moderate effect with significance at $\alpha = .2$
.

and 3 is largely positive, but the impact of receiving a no problem email (P_0) is as good or better than receiving either global or cultural problem emails (P_1 and P_2, respectively). Also, in week 2, emails of all types were moderately effective, indicating this is a good time to send Indian users email reminders for this course.

For US learners, the effects are non-significant across all emails except for the cultural problem email (P_2) in week 3, and the log odds ratios are small and in many cases negative. This indicates that for all weeks, emails for US users did not impact their propensity to return to the course, and may even deter them from returning – a counter-intuitive, but important, insight regarding timing of communication with learners.

Emails were more effective for Indian learners compared to US learners, despite email open rates being significantly larger for US learners (41.7% open rate) than Indian learners (27.0% open rate, p-value < 0.001 for difference in proportions). This suggests that Indian learners may benefit even more if we could increase the email open rate. Also, note that open rates did not differ across email type, as all emails in a given week had the same subject line.

2.5.3 Personalizing

2.5.3.1 Overview

SDLEs are notable for the high degree of diversity both across and within learners (Kizilcec et al., 2013; Kizilcec and Brooks, 2017). Due to this diversity, we might expect treatment effects to vary along with relevant learner attributes. If an intervention works for a specific user at a given time, that intervention may not be effective for a different user, or even the same user at a different time. Personalizing treatment, by discovering when and for whom certain treatments are most effective, is critical. For PERCS, RQ3 regards personalization– are certain emails more effective for active learners?

In clinical trials, learning how to personalize treatment is synonymous with discovering *moderators* (Kraemer et al., 2002). Moderators are subject-specific variables which change the efficacy of an intervention. For example, if a MOOC intervention works better for older users than younger users, then age is a moderator and can then be used to personalize; one may only deliver the intervention to older learners. Answering RQ3 in PERCS is equivalent to understanding if previous course activity is a moderator of email effectiveness.

2.5.3.2 Advantages of SRTs

Both SRTs and A/B tests permit the discovery of *baseline moderators* – variables measured prior to the first treatment randomization (e.g., gender, location, age, scores on early assignments) which moderate the effect of treatments. Baseline moderators are important for personalizing treatment. However, understanding how to change treatment based on variables measured throughout the course – *mid-course moderators* – is critical. Discovering mid-course moderators tells practitioners how to personalize interventions throughout the course to account for heterogeneity both between users and within a user over time.

Unlike A/B tests, SRTs permit the discovery of mid-course moderators. Statistically, due to bias introduced when including post-randomization variables in analyses, one can only discover moderator variables measured prior to randomization (Kraemer et al., 2002). If all learners were only randomized once, potential moderators measured after the first treatment cannot be discovered. By re-randomizing in SRTs, moderators measured before each of the randomizations (which now includes mid-course data) can be discovered. Because users were randomized again at the end of week three, we can use PERCS data to answer RQ3 about week three emails. Specifically, we can assess how activity during week three moderates the effect of emails sent at the end of week three. In PERCS-AB, week three course activity cannot be assessed as a moderator since it is measured after the one and only randomization in week one.

2.5.3.3 Discussion

We can evaluate mid-course moderators that include previous treatment. For example, we may think that responsiveness to previous treatment could inform how to personalize future treatment. Those that were responsive should continue receiving the same treatment while those that were non-responsive should receive different treatment. In PERCS, for example, we might expect week two emails to benefit users who responded positively to emails in week one. One could assess this by comparing two groups: Group 1 are users who received an email (E_1) but did not click in the course afterwards (i.e., email non-responders). Group 2 are users who received an email (E_1) but did click in the course afterwards (i.e., email responders). We could compare the effect of emails in week two for Group 1 vs Group 2.

Also, as learning content optimization starts to happen in real-time (i.e., reinforcement learning), knowing which real-time variables to measure and use for online optimization is critical. SRTs permit this discovery.

2.5.3.4 Results from PERCS

For PERCS, we were most interested in the previous week's course activity as a mid-course moderator. We wanted to see how different types of emails effect active and inactive users differently. Active users are defined as users who had one or more clicks in the course during the previous week. Before the trial, we did not know if problem-based emails would encourage inactive users (because they need the motivational reminder) or discourage inactive users (because the problem may be too advanced). To assess this, in Figure 2.5 we plot the log odds ratios of the probability of returning to the course in the subsequent week for different email conditions, compared to the control of no email (E_0). A positive log odds ratio indicates users had a higher chance of returning to the course after receiving the corresponding email problem type (compared to the no email control).

In week one, for both US and Indian learners, reminder emails of all types performed better for active users than inactive users (higher log odds for blue than magenta, Figure 2.5). In week three, the sign of the moderation switched, as emails performed better for inactive users compared to active users (higher log odds for magenta than blue). This moderation was larger for Indian learners. Also, the log odds ratios for inactive users were positive, while the log odds ratios for active users were negative, suggesting that emails were beneficial for bringing back inactive users, but potentially harmful to active learners. To ensure we encourage inactive users to return to the course while not discouraging active users, these results suggest that email sending should adapt based on course activity and the course week. Note that the confidence intervals are mostly overlapping, indicating that differences are not significant but only suggestive.

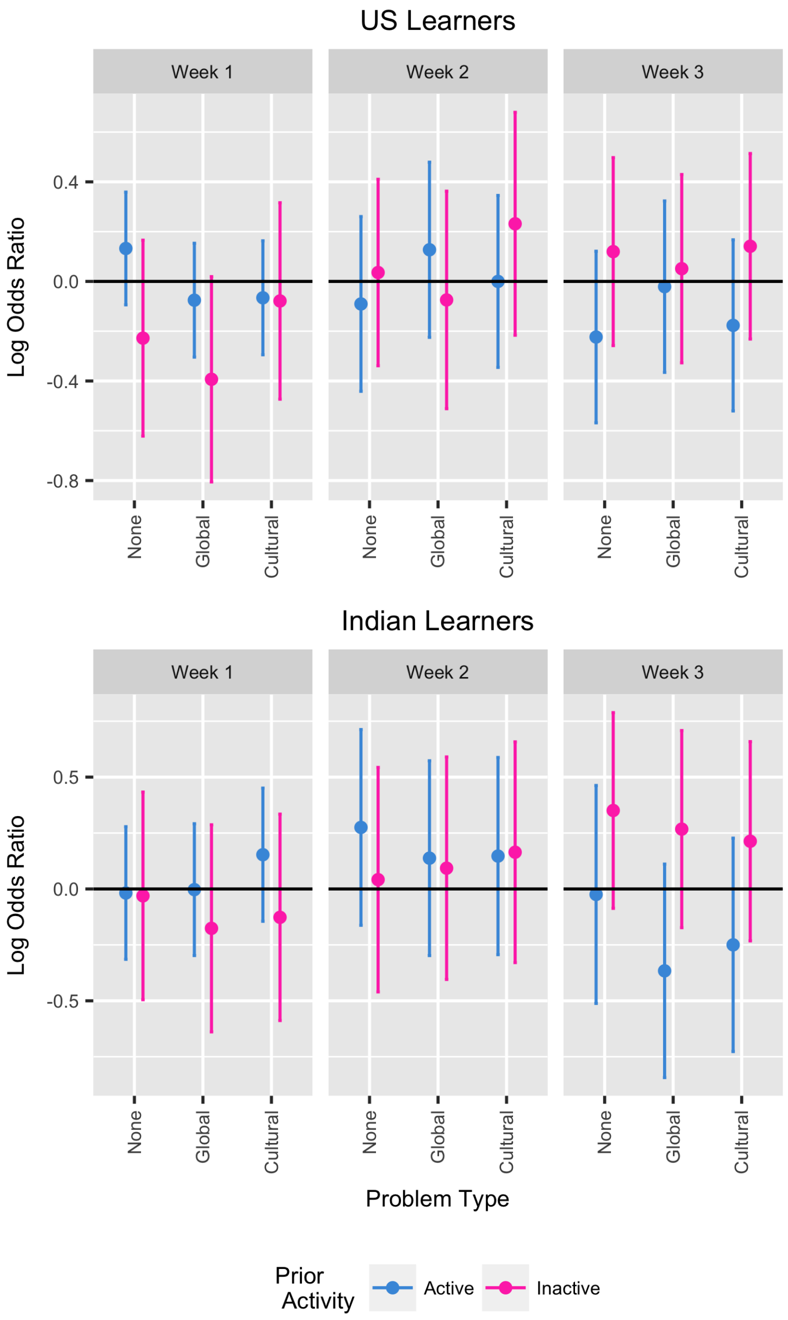


Figure 2.5: 95% confidence intervals of the log odds ratios of probability of returning to the course in the subsequent week for no problem (P_0), global problem (P_1), or cultural problem (P_2) emails, when compared to no email (E_0). The log odds ratio is calculated for users who had activity (active) or did not have activity (inactive) in the prior week.

2.6 Implications for Practice

This section outlines some useful considerations for researchers who are interested in designing and running SRTs for SDLEs. First, researchers should always remember that experiments are created to inform scientific understanding and answer questions about interventions. Deciding if a SRT is the correct trial design depends purely upon the scientific questions of interest. For example, if researchers are not interested in understanding how to sequence, time, and personalize interventions, running a SRT is unnecessary. In PERCS, if a researcher were interested in only comparing two sequences of interventions then it may be best to run an A/B test on those two sequences and forgo answers to other questions.

In situations where a SRT is appropriate, there are also important trial design considerations. All intervention sequences in a SRT should be useful, feasible, and scalable. Intervention sequences which could never be used in practice or which are knowingly deleterious to subjects should be avoided. For example, suppose a researcher was curious about optimal ordering of course content and, in turn, sequentially randomized learners to various sequences of content. If content B requires information taught in content A, none of the sequences in the experiment should place content B before content A. As another example, in PERCS, because the MOOC platform does not currently support fully-automated, scheduled delivery of email, messages were only sent once per week. We did not explore the possibility of sending emails many times per week (e.g., re-randomizing email sending every day) because, in practice, interventions which require manual daily email sending would not currently be feasible for instructors on this platform.

As in any other trial design, sample size and power calculations are important for SRTs. Sample size calculations should be based on the most important research questions the trial intends to answer. Since there are a larger number of possible treatment sequences in a SRT, calculating power and sample size requires researchers to consider which subset of users will be randomized to the sequences of interest. Then, sample size calculations are similar to A/B tests for both comparing interventions and discovering moderators (Oetting et al., 2011). Due to the larger space of potential treatment sequences, sample sizes usually need to be larger for SRTs (compared to A/B tests). To increase power, researchers may consider changing randomization probabilities to favor interventions of interest.

2.7 Conclusions, Limitations, and Future Work

Adapting, sequencing, and personalizing interventions is critical to meet the diverse needs of learners in SDLEs. When designing adaptive interventions, experimentation is useful for evaluating and comparing possible sequences. Single randomized A/B tests cannot answer questions about ways to adapt and sequence interventions throughout a course. In this work, we demonstrated how a new type of experimental design, SRTs, are valuable for developing adaptive personalized interventions in SDLEs. SRTs provide answers to questions regarding treatment sequencing, timing, and personalization throughout a course. Answers to these questions will both improve outcomes for learners and deepen understanding of learning science in scalable environments.

In this work, we provided a few examples of different SRTs. There are more variations of SRTs that may be useful for SDLEs. In PERCS, all users have the same randomization for all three weeks. However, if a learner is responding well to treatment, it may not make sense to re-randomize them and change their current treatment, as is done in Sequential Multiple Assignment Randomized Trials (SMARTs) (Murphy, 2005; Pelham Jr et al., 2016). Since many online courses are accessible to learners at any time, randomization timing could be based on when each user enters the course (randomized one, two, and three weeks from the day the user enrolls in a course). If treatment delivery timing is a research question, the delivery time can also be (re)randomized (e.g., each week of the course, re-randomize users to receive one weekly email or seven daily email reminders). Also, since many SDLEs provide learners with all course content at the beginning of the course, a trial aiming to re-randomize course content could use trigger-based re-randomizations–where users' future content is only changed after they have watched a certain video or completed a particular assignment–to prevent early exposure to future treatment.

Though useful, SRTs have limitations. SRTs are non-adaptive trial designs. If an experimenter wishes to both learn efficacious treatments and provide benefit to learners in the current trial, adaptive designs such as in Williams et al. (2018) would be more appropriate. Combining adaptive designs with SRTs may also be useful for both developing high-quality adaptive interventions and improving benefits for current learners (Cheung et al., 2015).

PERCS also has clear limitations. For one, since some users (10%) retook the online course multiple times, those users were repeated in multiple iterations of the trial. Secondly, it's important to note that PERCS was an exploratory trial not a confirmatory trial. The goal of PERCS was to explore several sequences of treatments and evaluate their efficacy for different learners. The next step for PERCS is to narrow down best treatment options based on current data (which may be different for Indian and US learners). Since the current evidence is not very strong, we would then run a second SRT with fewer treatments, acquiring more data on those treatments of interest. Once we have significant evidence to indicate which sequence of emails is optimal, we can then compare this learned optimal sequence of emails to a control in an A/B test confirmatory trial. Treatment A would be the hypothesized optimal adaptive sequence of emails and treatment B would be no email.

Additional analyses of the PERCS data would also be interesting. Using different outcomes other than course activity (such as course completion or assignment performance) could help further understanding of intervention efficacy. Also, email types had varying word lengths. Assessing how word length changes email efficacy could further elucidate the treatment effects. Lastly, using additional learner demographic information such as age, gender, or previous education as potential moderators would be useful. Although we currently cannot collect that information through the course, other studies have demonstrated how these characteristics can be inferred from available data (Brooks et al., 2018).

In the comparison of SRTs to A/B tests, we limited the comparison to one example design. There are many other possible comparators. For example, we could have compared

PERCS to a trial which performs a single-randomization in week 2, instead of week 1. This new design would then allow one to discover mid-course moderators (measured prior to week 2). The new design, however, would not be able to assess treatment effects in week 1. An important characteristic of SRTs is that all of the questions mentioned in Section 2.5 can be answered from data collected in one trial.

2.8 Acknowledgements

This work was supported in part under the Holistic Modeling of Education (HOME) project funded by the Michigan Institute for Data Science (MIDAS). We would also like to acknowledge Dr. Joseph Jay Williams and Dr. Daniel Almirall for valuable feedback on this work.

BIBLIOGRAPHY

Aceves, T. and Orosco, M. (2014). Culturally responsive teaching. http://ceedar.education.ufl.edu/wp-content/uploads/2014/08/culturally-responsive.pdf.

Almirall, D., Kasari, C., McCaffrey, D. F., and Nahum-Shani, I. (2018). Developing optimized adaptive interventions in education. *Journal of Research on Educational Effectiveness 11*(1), 27–34.

Almirall, D., Nahum-Shani, I., Sherwood, N., and Murphy, S. (2014). Introduction to smart designs for the development of adaptive interventions: with application to weight loss research. *Translational Behavioral Medicine 4*(3), 260–274.

Barrows, H. (1985). *How to design a problem-based curriculum for the preclinical years*, Volume 8. Springer Pub Co.

Brooks, C., Gardner, J., and Chen, K. (2018). How gender cues in educational video impact participation and retention. In *13th International Conference of the Learning Sciences*.

Brusilovsky, P. and Peylo, C. (2003). Adaptive and intelligent web-based educational systems. *Int. J. Artif. Intell. Ed. 13*(2-4), 159–172.

Cheung, Y., Chakraborty, B., and Davidson, K. (2015). Sequential multiple assignment randomized trial (SMART) with adaptive randomization for quality improvement in depression treatment program. *Biometrics 71*(2), 450–459.

Chow, S. and Chang, M. (2008). Adaptive design methods in clinical trials – a review. *Orphanet Journal of Rare Diseases 3*(1), 11.

Clement, B., Roy, D., Oudeyer, P., and Lopes, M. (2013). Multi-armed bandits for intelligent tutoring systems. *arXiv preprint arXiv:1310.3174*.

Collins, L. M., Murphy, S. A., and 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.

David, Y., Segal, A., and Gal, Y. (2016). Sequencing educational content in classrooms using bayesian knowledge tracing. In *Proc. LAK '16*, pp. 354–363.

Davis, D., Chen, G., van der Zee, T., Hauff, C., and Houben, G. (2016). Retrieval practice and study planning in MOOCs: Exploring classroom-based self-regulated learning strategies at scale. In K. Verbert, M. Sharples, and T. Klobučar (Eds.), *Adaptive and Adaptable Learning*, Cham, pp. 57–71. Springer.

Davis, D., Kizilcec, R., Hauff, C., and Houben, G. (2018). The half-life of MOOC knowledge: A randomized trial evaluating knowledge retention and retrieval practice in moocs. In *Proc. LAK '18*, pp. 1–10. ACM.

Davis, D., Triglianos, V., Hauff, C., and Houben, G. (2018). SRLx: A personalized learner interface for moocs. In *Lifelong Technology-Enhanced Learning*, Cham, pp. 122–135. Springer.

Dweck, C. (2008). *Mindset: The new psychology of success*. Random House Digital, Inc.

González-Brenes, J. and Huang, Y. (2015). "your model is predictive–but is it useful?" theoretical and empirical considerations of a new paradigm for adaptive tutoring evaluation. In *Proc. EDM '15*, pp. 187–194. ERIC.

Greene, J., Oswald, C., and Pomerantz (2015). Predictors of retention and achievement in a massive open online course. *American Educational Research Journal 52*(5), 925–955.

Guo, P., Kim, J., and Rubin, R. (2014). How video production affects student engagement: an empirical study of MOOC videos. In *L@S*, pp. 41–50. ACM.

Kaijanaho, A. and Tirronen, V. (2018). Fixed versus growth mindset does not seem to matter much: A prospective observational study in two late bachelor level computer science courses. In *Proc. ICER*, pp. 11–20. ACM.

Kasari, C., Kaiser, A., Goods, K., Nietfeld, J., Mathy, P., Landa, R., Murphy, S., and Almirall, D. (2014). Communication interventions for minimally verbal children with autism: A sequential multiple assignment randomized trial. *Journal of the American Academy of Child & Adolescent Psychiatry 53*(6), 635–646.

Kizilcec, R. and Brooks, C. (2017). Diverse big data and randomized field experiments in massive open online courses. In *The Handbook of Learning Analytics*, pp. 211–222. SOLAR.

Kizilcec, R., Piech, C., and Schneider, E. (2013). Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses. In *Proc. LAK '13*, pp. 170–179. ACM.

Klasnja, P., Hekler, E., Shiffman, S., Boruvka, A., Almirall, D., Tewari, A., and Murphy, S. (2015). Microrandomized trials: An experimental design for developing just-in-time adaptive interventions. *Health Psychology 34*(S), 1220.

Kraemer, H., Wilson, G., Fairburn, C., and Agras, W. (2002). Mediators and moderators of treatment effects in randomized clinical trials. *Archives of General Psychiatry 59*(10), 877–883.

Lan, A. and Baraniuk, R. (2016). A contextual bandits framework for personalized learning action selection. In *EDM*, pp. 424–429.

Lavori, P. and Dawson, R. (2000). A design for testing clinical strategies: biased adaptive within-subject randomization. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 163*(1), 29–38.

Lavori, P. and Dawson, R. (2004). Dynamic treatment regimes: practical design considerations. *Clinical Trials 1*(1), 9–20.

Lei, H., Nahum-Shani, I., Lynch, K., Oslin, D., and Murphy, S. (2011, 04). A SMART design for building individualized treatment sequences. *8*, 21–48.

Liu, Y., Mandel, T., Brunskill, E., and Popović, Z. (2014). Towards automatic experimentation of educational knowledge. In *CHI '14*, pp. 3349–3358. ACM.

Liu, Y., Mandel, T., Brunskill, E., and Popovic, Z. (2014). Trading off scientific knowledge and user learning with multi-armed bandits. In *EDM*, pp. 161–168.

Lomas, D., Patel, K., Forlizzi, J., and Koedinger, K. (2013). Optimizing challenge in an educational game using large-scale design experiments. In *CHI '13*, pp. 89–98.

Mullaney, T. and Reich, J. (2015). Staggered versus all-at-once content release in massive open online courses. In *L@S*, pp. 185–194. ACM.

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

NeCamp, T., Gardner, J., and Brooks, C. (2019). Beyond A/B testing: Sequential randomization for developing interventions in scaled digital learning environments. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, pp. 539–548. ACM.

Nie, X., Tian, X., Taylor, J., and Zou, J. (2017). Why adaptively collected data have negative bias and how to correct for it. *arXiv preprint arXiv:1708.01977*.

Oetting, A., Levy, J., Weiss, R., and Murphy, S. (2011). Statistical methodology for a SMART design in the development of adaptive treatment strategies. *Causality and Psychopathology: Finding the determinants*, 179–205.

Pardos, Z., Tang, S., Davis, D., and Le, C. (2017). Enabling real-time adaptivity in MOOCs with a personalized next-step recommendation framework. In *L@S*, pp. 23–32. ACM.

Pelham Jr, W. E., Fabiano, G., Waxmonsky, J., Greiner, A., Gnagy, E., III, W. P., Coxe, S., Verley, J., Bhatia, I., Hart, K., Karch, K., Konijnendijk, E., Tresco, K., Nahum-Shani, I., and Murphy, S. (2016). Treatment sequencing for childhood ADHD: A multiple-randomization study of adaptive medication and behavioral interventions. *Journal of Clinical Child & Adolescent Psychology 45*(4), 396–415. PMID: 26882332.

Rafferty, A., Ying, H., and Williams, J. (2018). Bandit assignment for educational experiments: Benefits to students versus statistical power. In *AIED*, pp. 286–290.

Renz, J., Hoffmann, D., Staubitz, T., and Meinel, C. (2016). Using A/B testing in MOOC environments. In *LAK '16*, pp. 304–313. ACM.

Reynolds, J. and Glaser, R. (1964). Effects of repetition and spaced review upon retention of a complex learning task. *Journal of Educ. Psych. 55*(5), 297.

Savi, A., Williams, J., Maris, G., and van der Maas, H. (2017, Feb). The role of A/B tests in the study of large-scale online learning.

Segal, A., David, Y., Williams, J., Gal, K., and Shalom, Y. (2018). Combining difficulty ranking with multi-armed bandits to sequence educational content. In *AIED*, pp. 317–321.

Williams, J., Rafferty, A., Tingley, D., Ang, A., Lasecki, W., and Kim, J. (2018). Enhancing online problems through instructor-centered tools for randomized experiments. In *CHI '18*, pp. 207:1–207:12. ACM.

Yu, H., Miao, C., Leung, C., and White, T. (2017). Towards AI-powered personalization in MOOC learning. *NPJ Science of Learning 2*, 1–5.

Zhao, Y., Kosorok, M. R., and Zeng, D. (2009). Reinforcement learning design for cancer clinical trials. *Statistics in Medicine 28*(26), 3294–3315.

CHAPTER III

Comparing Cluster-level Dynamic Treatment Regimens using Sequential, Multiple Assignment, Randomized Trials: Regression Estimation and Sample Size Considerations

This work is originally published in NeCamp et al. (2017).

3.1 Introduction

Interventions aimed at improving individual-level outcomes often occur at a cluster-level (Murray, 1998; Raudenbush and Bryk, 2002; Donner and Klar, 2010). Often, it may be necessary to use a tailored and dynamic approach to intervention in order to address cluster-level heterogeneity (Kilbourne et al., 2013). For example, due to differences in size, geography, or culture, some clusters may require more intensive or longer-duration intervention in order to improve patient-level outcomes.

Cluster-level dynamic treatment regimens (DTRs), also known as adaptive interventions, can be used to guide such sequential intervention decision-making at the cluster level. In a cluster-level DTR, the cluster-level intervention is potentially adapted (or re-adapted) over time based on changes in the cluster that could be impacted by prior intervention, e.g., adapting based on aggregate measures of the individuals that comprise it. A cluster-level DTR may also include intervention components dynamically tailored to the individuals within clusters.

Sequential multiple assignment randomized trials (SMARTs) represent an important data collection tool for informing the construction of DTRs (Murphy, 2005; Lei et al., 2012; Chakraborty and Moodie, 2013; Lavori and Dawson, 2014; Kosorok and Moodie, 2015). The focus of most SMARTs to date has been the development of individual-level DTRs to improve individual-level outcomes (e.g., see Methodology Center 2016).

There has been much less focus on analytic or design issues related to cluster-randomized SMARTs for developing cluster-level DTRs. In a cluster-randomized SMART, randomizations occur at the cluster level, yet outcomes are at the level of the individuals within the cluster. Using the Adaptive Implementation of Effective Programs Trial (Kilbourne et al., 2014) (ADEPT) as a motivating example, the focus of this chapter is on primary aim analysis and sample size considerations in cluster-randomized SMARTs. ADEPT, which is currently in the field, is to our knowledge the first-ever cluster-randomized SMART. The overarching goal of ADEPT is to develop a cluster-level DTR to improve the adoption of an evidence-based practice (EBP) for mood disorders in community-based mental health clinics and thereby improve patient-level mental health outcomes.

This chapter makes two contributions to the design and analysis of cluster-randomized SMARTs. First, we develop a regression approach for comparing the mean of a continuous patient-level outcome between the cluster-level DTRs embedded in a SMART. The regression approach is an extension of the estimator by Nahum-Shani et al. (2012) and first introduced by Orellana et al. (2010). The regression approach facilitates the use of individual- and cluster-level baseline (pre-randomization) covariates in the analysis of data from a cluster-randomized SMART.

Second, we develop sample size formulae (for the total number of clusters) to be used when the primary aim of the cluster-randomized SMART is a comparison of the mean of a continuous patient-level outcome between two DTRs beginning with different treatments. This is a common primary aim in SMARTs; see Oetting et al. (2011) (continuous end of study outcome) and Li and Murphy (2011) (survival outcome).

This regression approach can be used with any cluster-randomized SMART with repeated cluster-level randomizations. Sample size formulae are developed for two common types of two-stage SMART designs: the one used in ADEPT, in which only one group of non-responders are re-randomized, and the most popular type of SMART, in which all non-responders are re-randomized.

Consistent with the proposed regression approach, which facilitates the use of baseline covariates, the sample size formulae allow scientists to incorporate the correlation between a pre-specified baseline cluster-level covariate and patient-level outcomes, which leads to a reduction in the minimum number of clusters necessary (Spybrook et al., 2011). This chapter extends the work of Ghosh et al. (2015), which develops sample size calculators for a single type of cluster-randomized SMART in a non-regression context, i.e., without covariates.

3.2 SMARTs with Cluster-level Randomization

SMARTs are multi-stage randomized trial designs used explicitly for the purpose of building high-quality DTRs (Lavori and Dawson, 2000; Murphy, 2005). The multiple stages at which randomizations occur correspond to critical intervention decision points. At each decision point, randomization is used to address a question concerning the dosage (duration, frequency or amount), intensity, type, or delivery of treatment.

Here we consider SMARTs for developing cluster-level DTRs where the unit of randomization (and re-randomization) is a cluster and the outcomes are measured at the level of the individual.

3.2.1 Motivating Example: The ADEPT SMART Study

A schematic for the ADEPT trial (Kilbourne et al., 2014) is displayed in Figure 3.1. The overall aim of ADEPT is to develop a cluster-level DTR to improve the adoption of an EBP for mood disorders in community-based mental health clinics across Colorado and Michigan. The patient-level EBP is known as Life Goals (Kilbourne et al., 2012), a collaborative care, psychosocial intervention for mood disorders delivered to patients in six individual or group sessions. The primary outcome in ADEPT is a continuous, patient-level measure of mental health quality of life (MH-QOL).

ADEPT includes several interventions: the replicating effectiveness program (REP), REP plus External Facilitation (REP + EF), and REP plus External and Internal Facilitation (REP + EF + IF). REP is a cluster-level intervention focused on standardizing the implementation of the EBP into routine care through toolkit development, provider training, and program assistance. Facilitation is a cluster-level coaching intervention to help support the use of EBPs. EF is by phone and focuses on technical aspects of how to adopt the EBP; IF is in-person and involves working with a clinic manager to further embed the EBP.

ADEPT, which is currently in the field, involves community-based mental health clinics (approximately $N = 60$) that have failed to respond to an initial 6 months of REP (pre-randomization). During these 6 months, each clinic $i = 1, \ldots, N$ is expected to identify approximately $m_i = 10$ to 25 patients with mood disorders, all of which are followed for patient-level outcomes throughout the study. Clinics that enter the study (i.e., did not respond to REP at month 6) are randomized with equal probability to receive additional REP + EF or REP + EF + IF. After another 6 months, (i) REP + EF sites that are still non-responsive are randomized with equal probability to either continue REP + EF or augment with IF (REP + EF + IF) for an additional 12 months, (ii) REP + EF + IF sites that are still non-responsive continue REP + EF + IF, and (iii) facilitation interventions are discontinued for all sites that are responsive. A clinic is identified as “not responding” at months 6 and 12 if $< 50\%$ of the patients identified to be part of Life Goals during months 0-6 have received $\geq$3 Life Goals sessions.

By design, ADEPT has three DTRs embedded within it, which are displayed in Table 3.1. Each embedded DTR is labeled (a_1, a_2). For example, DTR $(1, -1)$ offers REP +

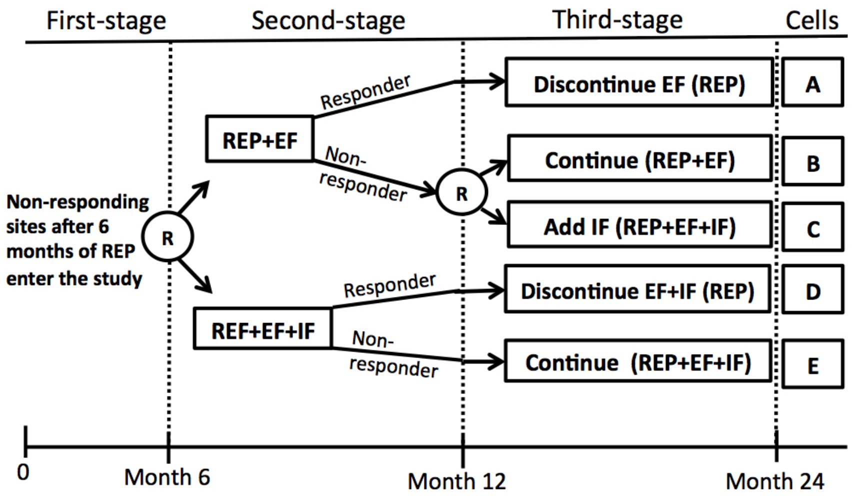


Figure 3.1: Schematic of ADEPT. The encircled R signifies randomization; cluster-level randomizations occurred at baseline and after 6 months of REP + EF or REP + EF + IF following identification of clinic responder status.

DTR Label	Second-stage	Status at end	Third-stage				Cell in	Known
(a_1, a_2)	Treatment	of second-stage	Treatment	A_1	R	A_2	Figure	IPW
$(1, 1)$	REP+EF	Resp	REP	1	1		A	2
		Non Resp	REP+EF	1	0	1	B	4
$(1, -1)$	REP+EF	Resp	REP	1	1		A	2
		Non Resp	REP+EF+IF	1	0	-1	C	4
$(-1, .)$	REP+EF+IF	Resp	REP	-1	1		D	2
		Non Resp	REP+EF+IF	-1	0		E	2

Table 3.1: The three DTRs embedded in ADEPT (Figure 3.1)

EF at month 6, then REP + EF is augmented with IF for clinics that remain non-responsive at month 12, whereas, EF is discontinued for clinics who are responsive at month 12.

3.2.2 The Prototypical SMART Design

In ADEPT, only clinics not responding to REP + EF were re-randomized at the next stage. This type of SMART (but with individual-level randomizations) has been previously employed in autism research, see Kasari et al. (2014) and Almirall et al. (2016).

Many other types of SMART designs are possible (see Methodology Center (2016)) for a comprehensive list with individual-level randomizations), including SMARTs where all units are subsequently re-randomized to the same set of next-stage intervention options (e.g., Chronis-Tuscano et al. 2016) and others where all units are re-randomized, but to different next-stage intervention options depending on response/non-response to first-stage intervention (e.g., Lu et al. 2016). Ultimately, the decision to choose a particular type of SMART is driven by scientific considerations.

By far the most common type of SMART is a two-stage design where (i) all units are randomized to two first-stage treatment options, (ii) a subset of units at the end of stage 1 (e.g., non-responders) are re-randomized to second-stage intervention options regardless of choice of first-stage intervention, and (iii) the remaining subset of units (e.g., responders) are not re-randomized. See Figure 3.2 for a generic example. We call this a "prototypical SMART design" given its popularity. Note that in the case of the prototypical SMART, there are four embedded DTRs; see Table 3.2.

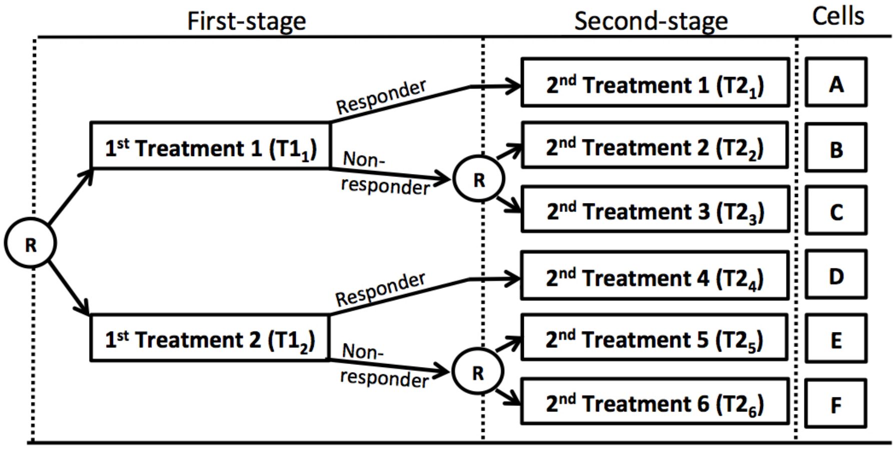


Figure 3.2: Schematic of a prototypical SMART design

Published examples of the prototypical SMART (with individual-level randomizations) include Pelham et al. (2016) in attention-deficit/hyperactivity disorder, Gunlicks-Stoessel et al. (2016) in adolescent depression, August et al. (2016) in conduct disorder prevention,

DTR Label (a_1, a_2)	First-stage Treatment	Status at end of first-stage	Second-stage Treatment	A_1	R	A_2	Cell in Figure	Known IPW
$(1,1)$	$T1_1$	Resp	$T2_1$	1	1		A	2
		Non Resp	$T2_2$	1	0	1	B	4
$(1,-1)$	$T1_1$	Resp	$T2_1$	1	1		A	2
		Non Resp	$T2_3$	1	0	-1	C	4
$(-1,1)$	$T1_2$	Resp	$T2_4$	-1	1		D	2
		Non Resp	$T2_5$	-1	0	1	E	4
$(-1,-1)$	$T1_2$	Resp	$T2_4$	-1	1		D	2
		Non Resp	$T2_6$	-1	0	-1	F	4

Table 3.2: The four DTRs embedded in a prototypical SMART (Figure 3.2)

Sherwood et al. (2016) and Naar-King et al. (2016) in weight loss, and McKay et al. (2015) in cocaine/alcohol use.

3.2.3 Common Primary Aims in a SMART

This chapter develops methods for comparing the mean of a continuous individual-level outcome between the DTRs embedded in a cluster-randomized SMART. This comparison can be conceptualized in various ways as a primary aim (Oetting et al., 2011; Almirall et al., 2014). (i) To compare first stage intervention options (averaging over the second stage intervention). In ADEPT, this is a comparison of DTR (-1,.) and the DTRs {(1,1), (1,-1)} (this was the primary aim in ADEPT; see Kilbourne et al. (2014)). (ii) To compare second stage intervention options (averaging over the first stage intervention). For example, in the prototypical design, this would be a comparison of DTRs {(1,1), (-1,1)} and DTRs {(1,-1), (-1,-1)} (e.g., see aim 3 in Pelham et al. (2016)). (iii) To compare the mean outcome between two DTRs beginning with the same first-stage treatment. In ADEPT, this is a comparison of DTR (1,1) and (1, -1). (iv) To compare the mean outcome between two DTRs that begin with different first stage treatments. In ADEPT, this is a comparison of (1,1) and (-1,.) or of (1,-1) and (-1,.).

The next section develops a regression estimator that can be used to address all of these primary aims using data from a cluster-randomized SMART. Following that, we derive sample size formulae for aim (iv). Simple extensions of standard sample size formulae may be used for primary aims (i), (ii), and (iii).

3.3 Methodology

3.3.1 Marginal Mean Model

For each SMART participant $j = 1, \ldots, m_i$ within each site $i = 1, \ldots, N$ we envision a primary end-of-study individual-level outcome Y_{ij}. Let the $p \times 1$ vector $\mathbf{X}_{ij}$ denote a pre-specified set of baseline covariates measured prior to the initial randomization. The baseline covariates, $\mathbf{X}_{ij}$, may be patient-level (e.g., age) or cluster-level (e.g., clinic location).

Denote $E_{a_1,a_2}(Y_{ij}|\mathbf{X}_{ij})$ as the marginal mean of Y_{ij} had the entire population been assigned to the DTR (a_1, a_2), conditional on baseline covariates, $\mathbf{X}_{ij}$ (Neyman et al., 1935; Rubin, 1978). The mean, $E_{a_1,a_2}(Y_{ij}|\mathbf{X}_{ij})$, averages over the response/non-response measure used in the DTR (a_1, a_2).

Let $\mu(\mathbf{X}_{ij}, a_1, a_2; \boldsymbol{\beta}, \boldsymbol{\eta})$ denote a marginal structural model (Robins, 1999; Hernán et al., 2000; Robins et al., 2000; Murphy et al., 2001; Orellana et al., 2010) for the mean $E_{a_1,a_2}(Y_{ij}|\mathbf{X}_{ij})$, which is linear in the unknown parameters $(\boldsymbol{\beta}, \boldsymbol{\eta})$. We provide examples below. We denote the causal effects between the DTRs by the $q \times 1$ vector $\boldsymbol{\beta}$, and denote associational effects between $\mathbf{X}_{ij}$ and Y_{ij} by the $p \times 1$ vector $\boldsymbol{\eta}$.

3.3.1.1 Example 1: ADEPT.

An example marginal mean model for the ADEPT study is

$$\mu(\mathbf{X}_{ij}, a_1, a_2; \boldsymbol{\beta}, \boldsymbol{\eta}) = \beta_0 + \beta_1 a_1 + \beta_2 a_2 I_{a_1=1} + \boldsymbol{\eta}^T \mathbf{X}_{ij}, \tag{3.1}$$

where $I_{a_1=1}$ is an indicator function which equals 1 when $a_1 = 1$.

Here we use a vector $\boldsymbol{\beta}$ with $q = 3$ to capture the causal effects for the 3 embedded DTRs. The covariates, $\mathbf{X}_{ij}$, could include, for example, the three baseline site-level variables used to stratify the initial randomization: US state (Colorado or Michigan), whether the site was a primary care or mental health site, and a site-average of individual MH-QOL scores. Using this model to address a primary aim of type (iv) above, the difference between the mean outcome had all clusters received DTR (1,1) and the mean outcome had all clusters received DTR (-1,.)—i.e., $E_{1,1}(Y_{ij}) - E_{\text{-}1,.}(Y_{ij})$—is given by $2\beta_1 + \beta_2$.

3.3.1.2 Example 2: Prototypical SMART.

In the prototypical SMART, we use a vector $\boldsymbol{\beta}$ with $q = 4$ to capture the causal effects for the 4 embedded DTRs.

$$\mu(\mathbf{X}_{ij}, a_1, a_2; \boldsymbol{\beta}, \boldsymbol{\eta}) = \beta_0 + \beta_1 a_1 + \beta_2 a_2 + \beta_3 a_1 a_2 + \boldsymbol{\eta}^T \mathbf{X}_{ij}. \tag{3.2}$$

Here, the difference between the mean outcome had all clusters received DTR (1,1) and the mean outcome had all clusters received DTR (-1,-1)—i.e., $E_{1,1}(Y_{ij}) - E_{\text{-}1,\text{-}1}(Y_{ij})$—is given by $2(\beta_1 + \beta_2)$.

3.3.2 Estimation

We now present an estimator for the unknown coefficients $(\boldsymbol{\beta}, \boldsymbol{\eta})$.

3.3.2.1 Notation.

Let $\mathbf{X}_i$ denote the $m_i \times p$ matrix $(\mathbf{X}_{i1}, \mathbf{X}_{i2}, \ldots, \mathbf{X}_{im_i})^T$ of covariates and $\boldsymbol{\mu}(\mathbf{X}_i, a_1, a_2; \boldsymbol{\beta}, \boldsymbol{\eta})$ be the $m_i \times 1$ vector of means $\left(\mu(\mathbf{X}_{i1}, a_1, a_2; \boldsymbol{\beta}, \boldsymbol{\eta}), \ldots, \mu(\mathbf{X}_{im_i}, a_1, a_2; \boldsymbol{\beta}, \boldsymbol{\eta})\right)^T$. Let $\mathbf{Y}_i$ be the $m_i \times 1$ vector of responses $(Y_{i1}, Y_{i2}, \ldots, Y_{im_i})^T$. Let A_{1i} denote the observed (i.e., randomly assigned) stage 1 treatment. In ADEPT, $A_{1i} = 1$ implies that cluster i received REP + EF as an initial treatment while $A_{1i} = -1$ implies cluster i received REP + EF +

IF. Let R_i, a binary variable, denote responder/non-responder status at the end of stage 1. In ADEPT, $R_i = 1$ if cluster i is a responder at the end of the first stage and $R_i = 0$ if cluster i is a non-responder. Let A_{2i} denote the observed (i.e., randomly assigned) stage 2 treatment. Note that, depending on the SMART design, A_{2i} may not be defined for some clusters i depending on the value of (A_{1i}, R_i). In ADEPT, A_{2i} is defined only for clusters with $A_{1i} = 1$ and $R_i = 0$. In the prototypical SMART, A_{2i} is not defined for clusters with $R_i = 1$. See Tables 3.1 and 3.2.

3.3.2.2 Estimator.

Building on Nahum-Shani et al. (2012), Orellana et al. (2010), and Lu et al. (2016), we obtain estimates of the coefficients $(\boldsymbol{\beta}, \boldsymbol{\eta})$ through solving an estimating equation.

In the estimator, the $m_i \times (q+p)$ matrix $\mathbf{D}(\mathbf{X}_i, a_1, a_2)$ is the derivative of $\boldsymbol{\mu}(\mathbf{X}_i, a_1, a_2; \boldsymbol{\beta}, \boldsymbol{\eta})$ with respect to $(\boldsymbol{\beta}, \boldsymbol{\eta})$; it can be thought of as the "design matrix" for DTR (a_1, a_2). For example, using the model in Equation 3.1 for ADEPT, the jth row of $\mathbf{D}(\mathbf{X}_i, a_1, a_2)$ is $(1, a_1, a_2 I_{a_1=1}, \mathbf{X}_{ij})$.

The $m_i \times m_i$ matrix $\mathbf{V}(a_1, a_2, \mathbf{X}_i)$ (abbreviated $\mathbf{V}_{i,a_1,a_2}$) is a working model for the covariance of $\mathbf{Y}_i$ conditional on $\mathbf{X}_i$ for DTR (a_1, a_2), $\text{Cov}_{a_1,a_2}(\mathbf{Y}_i|\mathbf{X}_i)$. In practice, the matrix $\mathbf{V}_{i,a_1,a_2}$ is unknown and must be estimated prior to solving Equation 3.3; see *Implementation* section.

The function $I(A_{1i}, R_i, A_{2i}, a_1, a_2)$ (abbreviated I_{i,a_1,a_2}) is a cluster-level indicator function which identifies whether (equals 1) or not (equals 0) cluster i was assigned to a sequence of treatments that is consistent with DTR (a_1, a_2). For example, in ADEPT, if $A_{1i} = 1, R_i = 0$, and $A_{2i} =$ -1, then cluster i is consistent only with DTR (1,-1); whereas if $A_{1i} = 1, R_i = 1$, then cluster i is consistent with both DTR (1,1) and (1,-1).

The weights $W(A_{1i}, A_{2i}, R_i)$ (abbreviated W_i) are the known cluster-level inverse probability weights (Orellana et al., 2010) (IPW), $W_i = 1/[f_{A_1}(A_{1i}) f_{A_2|A_1,R}(A_{2i}|A_{1i}, R_i)]$, where $f_{A_1}(a) = Pr(A_1 = a)$ and $f_{A_2|A_1,R}(a|b, c) = Pr(A_2 = a|A_1 = b, R = c)$ are probability mass functions. See Tables 3.1 and 3.2 for the known values of W_i in ADEPT and the prototypical SMART.

We obtain estimates $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\eta}})$ by solving for $(\boldsymbol{\beta}, \boldsymbol{\eta})$ in

$$0 = \sum_{i=1}^{N} \mathbf{U}_i(A_{1i}, R_i, A_{2i}, \mathbf{X}_i, \mathbf{Y}_i; \boldsymbol{\beta}, \boldsymbol{\eta}) \triangleq \sum_{i=1}^{N} \sum_{(a_1,a_2)} I_{i,a_1,a_2} W_i \cdot \mathbf{D}(\mathbf{X}_i, a_1, a_2)^T \mathbf{V}^{\text{-1}}_{i,a_1,a_2} (\mathbf{Y}_i - \boldsymbol{\mu}(\mathbf{X}_i, a_1, a_2; \boldsymbol{\beta}, \boldsymbol{\eta})). \quad (3.3)$$

The estimates $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\eta}})$ derived from solving Equation 3.3 are consistent and asymptotically normally distributed assuming the mean model (e.g., Equation 3.1 for ADEPT) is correctly specified. As in the generalized estimating equations literature (Liang and Zeger, 1986, 1993), there is no requirement that the working model $\mathbf{V}_{i,a_1,a_2}$ be a correct model for $\text{Cov}_{a_1,a_2}(\mathbf{Y}_i|\mathbf{X}_i)$. See Appendix A.3 for a sketch of the derivations.

3.3.2.3 Intuition for the Weights.

By design, in the observed data in a SMART, different clusters have different probabilities of being consistent with a specific DTR. For example, clusters assigned to cells A and B are consistent with DTR (1,1) (see Figure 3.1 and Table 3.1). However, clusters assigned to cell A had a 50% chance of being consistent with DTR (1,1), whereas clusters assigned to cell B had 25% chance of being consistent with DTR (1,1). Ignoring this known imbalance—i.e., using an unweighted average of observations in cells A and B to estimate the mean outcome had the entire population of clusters been assigned to DTR (1,1)—would cause the Cell A observations to have an unfairly larger influence on the estimator, leading to bias. The weights are designed to counteract this known imbalance and ensure that all clusters consistent with DTR (a_1, a_2) are represented equally. For example, in ADEPT, clusters in cell A are weighted by 1/0.5 = 2, whereas clusters in cell B are weighted by 1/0.25 = 4.

3.3.3 Implementation

Typically, in clustered settings, our working model for $\text{Cov}_{a_1,a_2}(\mathbf{Y}_i|\mathbf{X}_i)$, $\mathbf{V}_{i,a_1,a_2}$, is taken to be exchangeable and independent of $\mathbf{X}_i$, i.e., $\mathbf{V}_{i,a_1,a_2} = \sigma^{2*}_{a_1,a_2} \cdot \mathbf{Exch}_{m_i}(\rho^*_{a_1,a_2})$. Here $\sigma^{2*}_{a_1,a_2}$ and $\rho^*_{a_1,a_2}$ are scalars representing the conditional (on $\mathbf{X}_i$) variance and intra-cluster correlation (ICC) of the outcome under DTR (a_1, a_2), and $\mathbf{Exch}_{m_i}(\rho^*_{a_1,a_2})$ is an m_ixm_i exchangeable matrix (i.e., $[\mathbf{Exch}(\rho)]_{ii} = 1$ and $[\mathbf{Exch}(\rho)]_{ij} = \rho$ for $i \neq j$). Given this working model, the estimators $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\eta}})$ can be obtained using the following steps:

Step 1: Solve Equation 3.3 with $\mathbf{V}_{i,a_1,a_2}$ set to the identity matrix to obtain $(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\eta}}_0)$. For each embedded DTR (a_1, a_2) obtain residuals $\hat{\epsilon}_{ij,(a_1,a_2)}(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\eta}}_0) = Y_{ij} - \hat{\mu}(X_{ij}, a_1, a_2; \hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\eta}}_0)$.

Step 2: Estimate $\sigma^{2*}_{a_1,a_2}$ and $\rho^*_{a_1,a_2}$ using

$$\hat{\sigma}^{2*}_{a_1,a_2} = \frac{\sum_{i=1}^{N} [W_i I_{i(a_1,a_2)} \sum_{j=1}^{m_i} \hat{\epsilon}^2_{ij,(a_1,a_2)}]}{\sum_{i=1}^{N} W_i I_{i(a_1,a_2)} m_i} \text{ and } \hat{\rho}^*_{a_1,a_2} = \frac{\sum_{i=1}^{N} [W_i I_{i(a_1,a_2)} \sum_{j=1}^{m_i} \sum_{k \neq j}^{m_i} \hat{\epsilon}_{ij,(a_1,a_2)} \hat{\epsilon}_{ik,(a_1,a_2)}]}{\hat{\sigma}^{2*}_{a_1,a_2} \sum_{i=1}^{N} W_i I_{i(a_1,a_2)} m_i(m_i - 1)}. \tag{3.4}$$

Step 3: Solve Equation 3.3 with $\mathbf{V}_{i,a_1,a_2}$ set to $\hat{\boldsymbol{V}}_{i,a_1,a_2} = \hat{\sigma}^{2*}_{a_1,a_2} \cdot \mathbf{Exch}_{m_i}(\hat{\rho}^*_{a_1,a_2})$ to obtain $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\eta}}_1)$.

Step 4: Repeat Steps 2 and 3 with $\hat{\epsilon}_{ij,(a_1,a_2)}(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\eta}}_1)$ to obtain final estimates $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\eta}})$.

In simulations we do not find appreciable performance gains by iterating Steps 2 and 3 more than twice.

Steps 1-4 can be seen as extensions of standard GEE analysis (Liang and Zeger, 1986, 1993). Also, some analysts may choose to specify a working correlation structure which is equal for all DTR's. In this case, one could take a simple average of the estimates in Equation 3.4 across all regimens (a_1, a_2). Lastly, it is well known that by replacing the known W_i in each step above with estimated weights, statistical efficiency of the estimators may be improved (Robins et al., 1995; Hernan et al., 2002; Hirano et al., 2003; Bembom and van der Laan, 2007; Brumback, 2009; Williamson et al., 2014).

3.3.4 Standard Error Estimation

To estimate the variance of $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\eta}})$ we use the plug-in estimator, given by the $(q+p) \times (q+p)$ matrix $1/N \cdot \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\eta}}} = 1/N \cdot \hat{\boldsymbol{J}}^{\text{-}1}\hat{\boldsymbol{K}}\hat{\boldsymbol{J}}^{\text{-}1}$ where

$$\hat{\boldsymbol{J}} = \tfrac{1}{N}\sum_{i=1}^{N}\sum_{(a_1,a_2)} I_{i,a_1,a_2} W_i \mathbf{D}(\mathbf{X}_i, a_1, a_2)^T \hat{\boldsymbol{V}}^{\text{-}1}_{i,a_1,a_2}\mathbf{D}(\mathbf{X}_i, a_1, a_2),$$

$$\hat{\boldsymbol{K}} = \tfrac{1}{N}\sum_{i=1}^{N}\mathbf{U}_i(A_{1i}, R_i, A_{2i}, \mathbf{X}_i, \mathbf{Y}_i; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\eta}})\mathbf{U}_i^T(A_{1i}, R_i, A_{2i}, \mathbf{X}_i, \mathbf{Y}_i; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\eta}}).$$

See Appendix A.3 for an adjustment to the standard errors for the case when weights are estimated.

3.3.5 Hypothesis Testing

For any linear combination of $(\boldsymbol{\beta}, \boldsymbol{\eta})$, say $\mathbf{c}^T(\boldsymbol{\beta}, \boldsymbol{\eta})$ where $\mathbf{c}$ is a $(q+p)$-dimensional column vector, we use the univariate Wald statistic $Z = \sqrt{N}\mathbf{c}^T(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\eta}})/\sqrt{\mathbf{c}^T\,\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\eta}}}\mathbf{c}}$ to test the null hypothesis $H_0 : \mathbf{c}^T(\boldsymbol{\beta}, \boldsymbol{\eta}) = 0$. For example, in ADEPT, to test the difference in means had the entire population of clusters followed DTR (1,1) versus DTR (-1,.) (i.e., primary aim (iv) above) using the model in Equation 3.1, we set $\mathbf{c} = (0, 2, 1, \mathbf{0}_p)^T$. In large samples Z has a standard normal distribution under the null hypothesis. Hence, an α level test is "reject H_0 when $|Z| > z_{\alpha/2}$," where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of a standard normal distribution.

3.4 Sample Size Formulae

For both ADEPT and the prototypical SMART, we develop sample size formulae for the total number of clusters N for comparing the mean patient-level outcome between two embedded DTRs beginning with different stage 1 treatments. Specifically, for ADEPT, formulae are developed for testing null hypotheses of the form $H_0 : E_{1,b_2}(Y_{ij}) - E_{\text{-}1,.}(Y_{ij}) = 0$ for a fixed $b_2 \in (-1, 1)$ against alternate hypotheses of the form $H_1 : E_{1,b_2}(Y_{ij}) - E_{\text{-}1,.}(Y_{ij}) = \delta\sqrt{(\sigma^2_{1,b_2} + \sigma^2_{\text{-}1,.})/2}$. Here, δ is a standardized effect size (Cohen, 1988) and $\sigma^2_{b_1,b_2}$ is the outcome's marginal variance under DTR (b_1, b_2). For the prototypical SMART, formulae are developed for testing null hypotheses of the form $H_0 : E_{1,b_2}(Y_{ij}) - E_{\text{-}1,c_2}(Y_{ij}) = 0$ for a fixed $(b_2, c_2) \in (-1, 1)^2$ against alternate hypotheses of the form $H_1 : E_{1,b_2}(Y_{ij}) - E_{\text{-}1,c_2}(Y_{ij}) = \delta\sqrt{(\sigma^2_{1,b_2} + \sigma^2_{\text{-}1,c_2})/2}$. The formulae are based on using (3.3) to estimate the coefficients $\boldsymbol{\beta}$ in marginal models of the form (3.1) or (3.2) as follows: (i) with or without a pre-specified cluster-level covariate X_i, (ii) known weights W_i, and (iii) an exchangeable working covariance structure for $\mathbf{V}_{i,a_1,a_2}$. In addition, formulae are based on large sample approximations and a constant cluster size $m_i = m$ for all i. Extensions to the unequal cluster size case can be done as in Kerry and Bland (2001) or by conservatively setting m equal to the minimum cluster size. They also rely on the following *working population assumptions*.

1. **Equal exchangeable covariance matrices across regimens:** We assume the true marginal covariance matrices are equal for the two DTRs we are testing (e.g., $\text{Cov}_{1,b_2}(\mathbf{Y}_i) = \text{Cov}_{-1,c_2}(\mathbf{Y}_i) = \sigma^2 * \mathbf{Exch}(\rho)$ in the prototypical SMART)

2. **Conditional covariance inequality:** For a specific DTR, we assume non-responders do not vary from the marginal mean significantly more than responders. This assumption applies to different DTRs based on design, see below. A concern about this assumption should be raised only if the scientist, apriori, believed that, for a specific DTR, non-responders had significantly larger variances than responders or if the response rate was expected to be much larger than 0.5 (which is atypical for SMART designs). See Appendix A.1 for details.

3. **Correct marginal mean model:** We assume that $E_{a_1,a_2}(Y_{ij} \mid X_i) = \mu(X_i, a_1, a_2; \boldsymbol{\beta}, \eta)$ for the pre-specified cluster-level X_i, where $\mu(X_i, a_1, a_2; \boldsymbol{\beta}, \eta)$ is of the form (3.1) or (3.2). When X_i is not included in (3.1) or (3.2), this assumption is met trivially.

Each formula is a function of the cluster size m, the effect size δ, the outcome's ICC, ρ, the probability of a cluster responding after receiving a particular initial treatment, $p_1 = \mathbb{P}(R = 1|A_1 = 1)$ and $p_{\text{-}1} = \mathbb{P}(R = 1|A_1 = \text{-}1)$, and the standard normal quantiles $z_{\alpha/2}$ and z_β, where α is the size of our test and $1 - \beta$ is the power. We first provide formulae for estimation without covariates followed by the case when a cluster-level covariate X_i is used.

3.4.1 ADEPT Sample Size Formula

For ADEPT, working assumption 2 is: $E_{1,b_2}[(\mathbf{Y}_i - \boldsymbol{\mu}(1, b_2))(\mathbf{Y}_i - \boldsymbol{\mu}(1, b_2))^T|R = 0] \preceq E_{1,b_2}[(\mathbf{Y}_i - \boldsymbol{\mu}(1, b_2))(\mathbf{Y}_i - \boldsymbol{\mu}(1, b_2))^T] = \text{Cov}_{1,b_2}(\mathbf{Y}_i)$. Also, for ADEPT, working assumption 1 can be relaxed to $\sigma^2_{1,b_2} \leq \sigma^2_{-1,\cdot}$. Under these assumptions we obtain the sample size formula

$$N = \frac{4(z_\beta + z_{\alpha/2})^2}{m\delta^2} \cdot (1 + (m-1)\rho) \cdot (1 + \frac{1 - p_1}{2}). \tag{3.5}$$

3.4.2 Prototypical Sample Size Formula

For Prototypical SMART designs, working assumption 2 is: for both DTRs in our test, i.e., $(a_1, a_2) = (1, b_2)$ and $(-1, c_2)$, $E_{a_1,a_2}[(\mathbf{Y}_i - \boldsymbol{\mu}(a_1, a_2))(\mathbf{Y}_i - \boldsymbol{\mu}(a_1, a_2))^T|R = 0] \preceq E_{a_1,a_2}[(\mathbf{Y}_i - \boldsymbol{\mu}(a_1, a_2))(\mathbf{Y}_i - \boldsymbol{\mu}(a_1, a_2))^T] = \text{Cov}_{a_1,a_2}(\mathbf{Y}_i)$. Under these assumptions we obtain the sample size formula:

$$N = \frac{4(z_\beta + z_{\alpha/2})^2}{m\delta^2} \cdot (1 + (m-1)\rho) \cdot (1 + \frac{(1 - p_1) + (1 - p_{\text{-}1})}{2}). \tag{3.6}$$

Note this formula is identical to the formula in Ghosh et al. (2015) Also, note that we believe working assumptions 1-3 are implicit in their work.

The sample size formulae in Formula 3.5 and Formula 3.6 are intuitive. The first two terms in both formulae are identical; these terms compose the formula for the sample size for the difference in means in a 2-arm randomized control trial (RCT) with cluster-level randomization (Donner and Klar, 2010). The second term, in particular, is the expression for the variance inflation factor (VIF) arising from cluster-randomized trials. If ρ = 0 (i.e., VIF = 1), there is no inflation due to cluster randomization because we have no correlation within clusters. As ρ increases, each new observation within a cluster provides less unique information causing the VIF to increase. This, in turn, leads to an increase in sample size, N.

The third term, which is unique to SMARTs, is used to account for the fact that some clusters are re-randomized depending on response at the end of stage 1; hence, this last term is a function of the rate of response to first stage intervention. To understand this third term, it is useful to consider the following two extremes in the context of the prototypical SMART. If both response rates $(p_1, p_{\text{-}1})$ are 1, then there is no re-randomization and the design is analogous to a 2-arm cluster-randomized RCT (here, the third term is equal to 1). If, on the other hand, both response rates are 0, then all clusters are randomized twice; here, the third term is equal to 2. Note how the third term is different for ADEPT and the prototypical SMART due to the difference in randomization schemes. Also, the special case where response rates to initial treatments are equal (i.e., $p_1 = p_{\text{-}1}$) leads to a clustered version of the sample size formula in Oetting et al. (2011).

3.4.3 Sample Size Formulae with a Cluster-level Covariate

When including a cluster-level covariate in (3.1) or (3.2), working assumption 2 is similar for each corresponding design, except that it involves the conditional (on $\mathbf{X}_i$) marginal mean, i.e., $E_{a_1,a_2}[(\mathbf{Y}_i - \boldsymbol{\mu}(\mathbf{X}_i, a_1, a_2))(\mathbf{Y}_i - \boldsymbol{\mu}(\mathbf{X}_i, a_1, a_2))^T | R = 0] \preceq$ $E_{a_1,a_2}[(\mathbf{Y}_i - \boldsymbol{\mu}(\mathbf{X}_i, a_1, a_2))(\mathbf{Y}_i - \boldsymbol{\mu}(\mathbf{X}_i, a_1, a_2))^T]$. Also, our formulae depends on Cor(Y, X), which is the scalar correlation between the outcome Y_{ij} and the cluster-level covariate X_i under the DTRs in our test. Note that under assumptions 1 and 3, this correlation is constant across these DTRs (i.e., Cor$^2(Y, X) \triangleq$ Cor$^2_{1,b_2}(Y_{ij}, X_i) =$ Cor$^2_{-1,c_2}(Y_{ij}, X_i)$). We obtain the following sample size formula for ADEPT

$$N = \frac{4(z_\beta + z_{\alpha/2})^2}{m\delta^2} \cdot (1 + (m-1)\rho^*) \cdot (1 + \frac{1-p_1}{2}) \cdot [1 - \text{Cor}^2(Y, X)]. \tag{3.7}$$

For the prototypical SMART, the sample size formula is

$$N = \frac{4(z_\beta + z_{\alpha/2})^2}{m\delta^2}(1 + (m-1)\rho^*)(1 + \frac{(1-p_1) + (1-p_{\text{-}1})}{2})[1 - \text{Cor}^2(Y, X)],$$

where $\rho^* = (\rho - \text{Cor}^2(Y, X))/(1 - \text{Cor}^2(Y, X))$.

The use of a covariate leads to two changes in the sample size formulae. First, as expected (Spybrook et al., 2011), depending on the strength of the correlation between X and Y (i.e., $\text{Cor}^2(Y, X)$), the use of a covariate has the potential to reduce the minimum required sample size; this is because the use of covariates may improve the efficiency of our estimate of the coefficients $\boldsymbol{\beta}$. Second, there is a reduction in sample size due to the reduction in correlation, ρ^*, which, by definition, is always less than ρ.

3.4.4 Using the Sample Size Formula for the ADEPT study

To exemplify how the formula can be utilized in practice, we calculate how large of a difference between DTRs (1,-1) and (-1,.) we can detect in ADEPT. This difference would help us understand if it is better to give REP + EF + IF to non-responding clinics initially, or to delay REP + EF + IF until a clinic is non-responsive to REP + EF. In ADEPT, we expect the ICC of patient's MH-QOL to be $\rho = 0.01$ and the probability of responding when initially receiving REP + EF to be $p_1 = 0.2$. Using the true sample size of $N = 60$, a common cluster size of $m = 10$, and performing an $\alpha = 0.05$ level test ($z_{\alpha/2} = 1.96$), by rearranging Formula 3.5, we conclude that at 80% power ($z_\beta = 0.84$) we can detect an effect size of δ = 0.282.

3.5 Simulations

Simulations were conducted to evaluate the developed formulae and understand their robustness to violations of the working assumptions. Specifically, we evaluate formulae under four scenarios: (1) satisfying all working assumptions, (2) violating working assumption 1, (3) violating working assumption 2, and (4) violating working assumption 3. For each scenario, we compare the nominal power of 0.9 with the estimated power (based on 1000 iterations). Here, we present results for ADEPT; results were similar for the prototypical SMART.

Details concerning the data generative model can be found in Appendix A.2. Data were generated to mimic the ADEPT study. We considered different data generative scenarios with varied standardized effect sizes (δ = 0.2 (small), 0.5 (moderate)), cluster sizes ($m = 5, 10, 20$), ICC (ρ or ρ^* = 0.01 or 0.1), and, when there is a cluster-level covariate, the correlation between X and Y ($\text{Cor}^2(Y, X) \in [0.04, 0.4]$). We also considered different scenarios constituting violations of the working assumptions (details below). For each scenario, the sample size was selected based on the proposed formulae with nominal power $(1 - \beta)$ = 0.9 and Type-I error rate $\alpha = 0.05$. 1000 data sets were generated for each scenario. Each data set was analyzed as in the *Implementation* and *Standard Error Estimation* sub-sections, using the marginal mean model in Equation 3.1.

ICC, ρ	Effect Size, δ	Cluster Size, m	Sample Size, N	Assumptions are correct	Violating Assumption 1	Violating Assumption 2
0.01	0.2	5	306	0.894	0.891	0.886
		20	88	0.917	0.890	0.876*
	0.5	5	49	0.909	0.898	0.880*
		10	26	0.906	0.878*	0.893
0.1	0.2	5	412	0.910	0.901	0.870*
		20	213	0.922*	0.902	0.891
	0.5	5	66	0.909	0.888	0.898
		20	34	0.915	0.913	0.889

*The proportion is significantly different from 0.9 at the 5% level.

Table 3.3: Power analysis of Formula 3.5

Table 3.3 describes simulation results for the sample size formula in Formula 3.5. To violate assumption 1, we made the response variance under DTR$(1, b_2)$ 1.5 times the response variance under DTR $(-1, .)$. We could have also violated this assumption by deviating from an exchangeable covariance structure, however, in cluster-randomized trials it is rare to use an other covariance structure (Eldridge et al., 2009). To violate assumption 2, we made non-responders have significantly larger variance than responders under DTR $(1, b_2)$.

As expected, when no assumptions are violated (column 5), our estimated power is close to our pre-specified power, 0.9. When assumption 1 is violated (column 6) or assumption 2 is violated (column 7), we see that our power does not reduce dramatically. Hence, we conclude that our sample size formula is robust to violations of working assumptions 1 and 2. Also, the Type-I error rate does not depend on working assumptions 1 or 2. Hence, under each of these scenarios, with the effect size set to 0, the Type-I error rate is close to the nominal rate of 0.05.

Because working assumption 3 will always be true when there are no covariates, we run a second simulation, this time with a cluster-level covariate, to evaluate the robustness of the sample size formula in Formula 3.7 to a violation of this assumption. Specifically, to violate assumption 3, we deviate from the linear marginal mean in Equation 3.1 by generating data with $E_{a_1,a_2}(Y_{ij}|X_i) = \beta_0 + \beta_1 a_1 + \beta_2 a_2 I_{a_1=1} + \eta f_k(X_i)$ where $f_k(X_i) = X_i$ for $X_i \in [-k, k]$, $f_k(X_i) = k$ for $X_i > k$, and $f_k(X_i) = -k$ for $X_i < -k$ (i.e., the linear marginal mean is misspecified outside of $[-k, k]$). Here η is chosen to maintain the same values of Cor(Y, X). Setting $k = 2$ indicates a small violation (column 7) and setting $k = 1$ indicates a large violation (column 8). We still, however, analyze the data using the marginal mean model in Equation 3.1. The results are in Table 3.4.

As expected, when no assumptions are violated (column 6), our estimated power is close to our pre-specified power, 0.9. Note the reduction in sample size caused by the addition of a covariate. Under a small violation (column 7), we see the power is not significantly reduced. Under a large violation (column 8), we see our power is lowest when X

ICC, ρ^*	Effect Size, δ	Cluster Size, m	Cor$^2(Y, X)$	Sample Size, N	Assumptions	Small	Large
					are correct	Violation of	Violation of
						Assumption 3	Assumption 3
0.01	0.2	5	0.238	233	0.909	0.904	0.859*
		20	0.238	65	0.903	0.879*	0.777*
	0.5	5	0.043	47	0.891	0.901	0.902
		10	0.066	24	0.893	0.903	0.897
0.1	0.2	5	0.243	305	0.918	0.900	0.890
		20	0.243	159	0.915	0.922*	0.864*
	0.5	5	0.043	63	0.898	0.916	0.919*
		20	0.043	32	0.908	0.920*	0.899

*The proportion is significantly different from 0.9 at the 5% level.

Table 3.4: Power analysis of Formula 3.7

and Y are moderately correlated and the sample size is low. This is because when X and Y are weakly correlated, the overall influence of X is small, and hence misspecification of the relationship between X and Y will have little influence on our estimation and power. Also, once again, under these scenarios, with the effect size set to 0, the Type-I error rate is close to the nominal rate of 0.05.

3.6 Discussion and Future Work

This chapter presents a regression estimator and sample size formulae for comparing embedded DTRs using data arising from a cluster-randomized SMART. Methods were motivated by the ADEPT SMART, a study designed to develop a DTR (at the level of community-based mental health clinics) designed to improve mental health outcomes for patients clustered within those sites (Kilbourne et al., 2014). Sample size formulae were derived for both ADEPT and for a more common type of SMART.

There are a number of directions for future research in the analysis of cluster-randomized SMARTs. First, relatively staightforward applications of the estimator in Equation 3.3 with different link functions can be used to analyze, for example, binary, count, or zero-inflated outcomes.

Second, in practice, many cluster-randomized SMARTs will collect longitudinal (i.e., repeated measures) research outcomes at the patient-level. A natural next step is to combine the estimator presented here with methods for the analysis of longitudinal SMART outcomes (Lu et al., 2016) in order to accommodate two levels of clustering: repeated measures within patients within clusters.

Third, future work could also consider the use of variance components models, i.e., mixed effects or random effects models (Raudenbush and Bryk, 2002; Hedeker and Gibbons, 2006), which are now-standard in the analysis of randomized trials.

Fourth, while this chapter focuses on the analysis of primary aims in a SMART, in the DTR literature there is much interest in the development and application of analysis methods designed to generate hypotheses about more individually-tailored DTRs (Qian and Murphy, 2011; Linn et al., 2014; Moodie et al., 2014; Laber and Zhao, 2015; Zhang et al., 2015; Zhao et al., 2015; Zhou et al., 2015). Much of this literature has focused on identifying optimal DTRs at the individual level. Such methods could be extended for the analysis of data arising from cluster-randomized SMARTs to develop optimal cluster-level DTRs.

There are also a number of interesting methodological issues related to the design of cluster-randomized SMARTs (with implications for analysis methods). First, the sample size formulae derived here were limited to cases where our data contains a single cluster-level covariate. Future work may provide extensions to data containing multiple covariates and individual-level covariates.

Second, in this chapter we focus on SMARTs that are useful for developing of cluster-level DTRs where the initial and subsequent decisions are all at the cluster-level. However, there is currently much interest by educational scientists in SMARTs aimed at developing DTRs where sequences of intervention decisions are made at *both* the cluster and individual level. For example, we are currently involved in the conduct of a trial where the first stage intervention is at the level of classrooms with children with autism (such classrooms often include 1 to 3 children with autism), and the subsequent stages of intervention are at the level of the children themselves (Kasari et al., 2016).

3.7 Acknowledgements

R code for implementing the weighted least squares regression estimator and the sample size formulae for both ADEPT and the prototypical SMART are provided on the first author's website. This research is supported by the following NIH grants: R01MH099898 (Kilbourne & Almirall), P50DA039838 (Almirall), R01HD073975 (Almirall), R01DA039901 (Almirall). We also would like to thank Xi (Lucy) Lu for the helpful comments.

BIBLIOGRAPHY

Almirall, D., DiStefano, C., Chang, Y.-C., Shire, S., Kaiser, A., Lu, X., Nahum-Shani, I., Landa, R., Mathy, P., and Kasari, C. (2016). Longitudinal effects of adaptive interventions with a speech-generating device in minimally verbal children with asd. *Journal of Clinical Child & Adolescent Psychology 45*(4), 442–456.

Almirall, D., Nahum-Shani, I., Sherwood, N., and Murphy, S. (2014). Introduction to SMART designs for the development of adaptive interventions: with application to weight loss research. *Translational Behavioral Medicine 4*(3), 260–274.

August, G. J., Piehler, T. F., and Bloomquist, M. L. (2016). Being "smart" about adolescent conduct problems prevention: executing a smart pilot study in a juvenile diversion agency. *Journal of Clinical Child & Adolescent Psychology 45*(4), 495–509.

Bembom, O. and van der Laan, M. (2007). Statistical methods for analyzing sequentially randomized trials. *Journal of the National Cancer Institute 99*(21), 1577–82.

Brumback, B. A. (2009). A note on using the estimated versus the known propensity score to estimate the average treatment effect. *Statistics & Probability Letters 79*(4), 537–542.

Chakraborty, B. and Moodie, E. (2013). *Statistical methods for dynamic treatment regimes*. Springer.

Chronis-Tuscano, A., Wang, C. H., Strickland, J., Almirall, D., and Stein, M. A. (2016). Personalized treatment of mothers with ADHD and their young at-risk children: A SMART pilot. *Journal of Clinical Child & Adolescent Psychology 45*(4).

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

Donner, A. and Klar, N. (2010). *Design and analysis of cluster randomization trials in health research*, Volume 1. Wiley.

Eldridge, S. M., Ukoumunne, O. C., and Carlin, J. B. (2009). The intra-cluster correlation coefficient in cluster randomized trials: A review of definitions. *International Statistical Review 77*, 378–394.

Ghosh, P., Cheung, Y., and Chakraborty, B. (2015). Sample size calculations for clustered SMART designs. In M. R. Kosorok and E. E. Moodie (Eds.), *Adaptive Treatment Strategies in Practice: Planning Trials and Analyzing Data for Personalized Medicine*, Chapter 5, pp. 55–70. Alexandria, Virginia: SIAM.

Gunlicks-Stoessel, M., Mufson, L., Westervelt, A., Almirall, D., and Murphy, S. (2016). A pilot SMART for developing an adaptive treatment strategy for adolescent depression. *Journal of Clinical Child & Adolescent Psychology 45*(4), 480–494.

Hedeker, D. and Gibbons, R. D. (2006). *Longitudinal data analysis*, Volume 451. John Wiley & Sons.

Hernán, M., Brumback, B., and Robins, J. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology 11*(561-70).

Hernan, M., Brumback, B., and Robins, J. (2002). Estimating the causal effect of zidovudine on CD4 count with a marginal structural model for repeated measures. *Statistics in Medicine 21*, 1689–1709.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica 71*(4), 1161–1189.

Kasari, C., Gulsrud, A., and Almirall, D. (2016). Getting SMART about social and academic engagement of elementary aged students with autism spectrum disorder. http://ies.ed.gov/funding/grantsearch/details.asp?ID=1758.

Kasari, C., Kaiser, A., Goods, K., Nietfeld, J., Mathy, P., Landa, R., Murphy, S., and 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*(6), 635–646.

Kerry, S. M. and Bland, M. J. (2001). Unequal cluster sizes for trials in English and Welsh general practice: Implications for sample size calculations. *Statistics in Medicine 20*(3), 377–390.

Kilbourne, A. M., Abraham, K. M., Goodrich, D. E., Bowersox, N. W., Almirall, D., Lai, Z., and 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.

Kilbourne, A. M., Almirall, D., Eisenberg, D., Waxmonsky, J., Goodrich, D. E., Fortney, J. C., Kirchner, J. E., Solberg, L. I., Main, D., Bauer, M. S., et al. (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., Goodrich, D. E., Lai, Z., Clogston, J., Waxmonsky, J., and Bauer, M. S. (2012). Life goals collaborative care for patients with bipolar disorder and cardiovascular disease risk. *Psychiatric Services 63*(12), 1234–1238.

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

Laber, E. B. and Zhao, Y. (2015). Tree-based methods for individualized treatment regimes. *Biometrika 102*(3), 501–514.

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

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

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

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika 73*(1), 13–22.

Liang, K.-Y. and Zeger, S. L. (1993). Regression analysis for correlated data. *Annual Review of Public Health 14*(1), 43–68.

Linn, K. A., Laber, E. B., and Stefanski, L. A. (2014). Interactive Q-learning for probabilities and quantiles. *arXiv preprint arXiv:1407.3414*.

Lu, X., Lynch, K. G., Oslin, D. W., and Murphy, S. (2016). Comparing treatment policies with assistance from the structural nested mean model. *Biometrics 72*(1), 10–19.

McKay, J. R., Drapkin, M. L., Van Horn, D. H., Lynch, K. G., Oslin, D. W., DePhilippis, D., Ivey, M., and Cacciola, J. S. (2015). Effect of patient choice in an adaptive sequential randomization trial of treatment for alcohol and cocaine dependence. *Journal of Consulting and Clinical Psychology 83*(6), 1021.

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

Moodie, E. E., Dean, N., and Sun, Y. R. (2014). Q-learning: Flexible learning about useful utilities. *Statistics in Biosciences 6*(2), 223–243.

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

Murphy, S. A., van der Laan, M. J., Robins, J. M., and CPPRG (2001). Marginal mean models for dynamic regimes. *Journal of the American Statistical Association 96*, 1410–1423.

Murray, D. M. (1998). *Design and analysis of group-randomized trials*, Volume 29. Oxford University Press, USA.

Naar-King, S., Ellis, D. A., Idalski Carcone, A., Templin, T., Jacques-Tiura, A. J., Brogan Hartlieb, K., Cunningham, P., and Jen, K.-L. C. (2016). 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., Waxmonsky, J., Yu, J., and Murphy, S. (2012). Experimental design and primary data analysis methods for comparing adaptive interventions. *Psychological Methods 17*, 457–477.

NeCamp, T., Kilbourne, A., and Almirall, D. (2017). Comparing cluster-level dynamic treatment regimens using sequential, multiple assignment, randomized trials: Regression estimation and sample size considerations. *Statistical Methods in Medical Research 26*(4), 1572–1589.

Neyman, J., Iwaszkiewicz, K., and Kolodziejczyk, S. (1935). Statistical problems in agricultural experimentation. *Supplement to the Journal of the Royal Statistical Society* (107-80).

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

Orellana, L., Rotnitzky, A., and 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.

Pelham, W. E., Fabiano, G. A., Waxmonsky, J. G., Greiner, A. R., Gnagy, E. M., Pelham, W. E., Coxe, S., Verley, J., Bhatia, I., Hart, K., et al. (2016). Treatment sequencing for childhood ADHD:

A multiple-randomization study of adaptive medication and behavioral interventions. *Journal of Clinical Child & Adolescent Psychology 45*(4), 396–415.

Qian, M. and Murphy, S. A. (2011). Performance guarantees for individualized treatment rules. *Annals of Statistics 39*(2), 1180.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical linear models: applications and data analysis methods*, Volume 1. Sage.

Robins, J. M. (1999). Association, causation, and marginal structural models. *Synthese 121*, 151–179.

Robins, J. M., Hernan, M., and Brumback, B. (2000). Marginal structural models and causal inference. *Epidemiology 11*(5), 550–560.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association 90*(429), 106–121.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics 6*, 34–58.

Sherwood, N. E., Butryn, M. L., Forman, E. M., Almirall, D., Seburg, E. M., Crain, A. L., Kunin-Batson, A. S., Hayes, M. G., Levy, R. L., and Jeffery, R. W. (2016). The BestFIT trial: A SMART approach to developing individualized weight loss treatments. *Contemporary Clinical Trials 47*, 209–216.

Spybrook, J., Bloom, H., Congdon, R., Hill, C., Martinez, A., and Raudenbush, S. (2011). Optimal design plus empirical evidence. http://wtgrantfoundation.org/resource/optimal-design-with-empirical-information-od.

Williamson, E. J., Forbes, A., and White, I. R. (2014). Variance reduction in randomised trials by inverse probability weighting using the propensity score. *Statistics in Medicine 33*(5), 721–737.

Zhang, Y., Laber, E. B., Tsiatis, A., and Davidian, M. (2015). Using decision lists to construct interpretable and parsimonious treatment regimes. *Biometrics 71*(4), 895–904.

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

Zhou, X., Mayer-Hamblett, N., Khan, U., and Kosorok, M. R. (2015). Residual weighted learning for estimating individualized treatment rules. *Journal of the American Statistical Association*.

CHAPTER IV

Assessing Real-time Moderation for Developing Adaptive Mobile Health Interventions for Medical Interns: A Micro-randomized Trial

4.1 Introduction

4.1.1 Background

According to the World Health Organization, depression is the leading cause of disease-associated disability in the world (World Health Organization, 2017). In the US, the burden of depression, including suicide, has continued to grow (Greenberg et al., 2015). In populations at high risk, prevention of depression may be an effective strategy. The U.S. National Academy of Medicine has highlighted the need to develop, evaluate, and implement prevention interventions for depression and other mental, emotional, and behavioral disorders (Hawkins et al., 2016).

Prevention interventions for depression are critical for individuals in stressful work environments. Stressful work environments can lead to increased rates of depression (Tennant, 2001). However, high stress can make individuals less receptive to intervention and behavior change (Baucom et al., 2015; Everett et al., 1995).

Unlike other recent advances, mobile technology has the potential to transform the delivery and timing of depression prevention interventions to meet the needs of highly stressed individuals. In contrast to more intensive treatment (such as therapeutic appointments), mobile health interventions (e.g., push notifications) can be delivered at low burden, which may be critical given individuals' high stress workloads. Mobile devices hold the power to deliver just-in-time adaptive interventions (JITAIs) (Nahum-Shani et al., 2017) to individuals during times when they are able to receive and respond to them. Lastly, mobile devices also collect objective measurements of an individual's context and behavior with minimal burden (such as step counts and sleep duration). This data may, in turn, be used to determine when to deliver interventions and evaluate intervention efficacy without bothering the individuals.

When initially designing a JITAI, these states of opportunity (Nahum-Shani et al., 2017)—times when individuals are receptive to positive behavior change—are not known. Timing is critical because poorly timed interventions can lead to loss of engagement with the intervention (Zhang and Elhadad, 2016). Timing interventions is also particularly important for individuals with in stressful work environments because poorly timed interventions could cause increased disengagement and treatment fatigue (Heckman et al., 2015).

Current behavioral theories lack the granularity and adaptivity necessary to inform the timing of the delivery of mobile health interventions (Riley et al., 2011; Spruijt-Metz and Nilsen, 2014). Many theoretical models are non-dynamic— they only consider treatment adaptation based on baseline characteristics (e.g., sex, depression history) (Riley, 2014). Timing and adapting treatment based on real-time variables is essential for developing high-quality JITAIs (Nahum-Shani et al., 2017).

This article takes a data-driven approach to inform dynamic timing of intervention delivery. Experimentation and data collection were used to provide empirical evidence for determining states of opportunity—the data will illustrate when interventions cause the positive behavior change in individuals, and when they do not.

In statistical terms, we formulate the task of empirically learning how to dynamically time interventions as *discovering time-varying moderators of causal treatment effects* (Klasnja et al., 2015). Time-varying moderators are *time-varying* because the moderators' values vary throughout the study (e.g., daily mood), and are *moderators* because they change—or moderate—the efficacy of subsequent treatments. For example, if push notifications containing sleep messages perform worse on individuals with little sleep in the previous night compared to individuals with high sleep, then previous night's sleep moderates the effect of sleep notifications. Discovering time-varying moderators informs treatment timing because treatment delivery can now be based on the observed values of these moderators. In the example, sleep notifications should now only be sent after individuals obtain a sufficient night's sleep.

We assessed time-varying moderators of mobile health interventions targeting three categories: mood, activity, and sleep. Stressful work environments can lead to sleep deprivation and physical inactivity (Kalmbach et al., 2018; Âkerstedt, 2006; Lallukka et al., 2004), two behaviors directly associated with depression (Kalmbach et al., 2018; Baglioni et al., 2011; Ströhle, 2009). To prevent depression in individuals experiencing high stress, it is critical to develop high-quality interventions which can help them maintain and improve their mood, either through targeting mood directly, or by indirectly improving activity and sleep (Kalmbach et al., 2018; Ströhle, 2009).

Our study population is medical interns. Medical interns (physicians in their first year of residency) experience stressful work environments throughout their entire internship year. Interns are known to suffer from depression at higher rates than the general public (Mata et al., 2015). Focusing on physician training, a rare situation where a dramatic increase in stress can be anticipated, provides an ideal experimental model to develop interventions for maintaining mental wellness during life and work stressors.

Our study, the 2018 Intern Health Study (IHS) (The Sen Lab, 2019), is 6-month long mobile health cohort study which tracks medical interns using phones and wearables. During the internship year, we conducted a micro-randomized trial (MRT)(Klasnja et al., 2015). Standard single-time point randomized control trials (RCTs) only inform moderation by baseline variables (Kraemer et al., 2002) and do not permit the discovery of time-varying moderators. The MRT was advantageous because it allows us to discover time-varying moderators of causal treatment effects (Klasnja et al., 2015).

During each week in the 6-month study, an intern was randomized to one of four possible treatments: a week of mood notifications, activity notifications, sleep notifications, or no notifications. The outcomes are average daily mood valence (measured through a one question survey), average daily steps (as a proxy for activity), and average daily sleep duration, where averages are 7 day averages of data collected during the week of treatment. The strongest moderators were hypothesized to be previous week's average daily mood, average daily steps, and average daily sleep, as these were the strongest predictors of the outcomes (based on previous years' IHS data (The Sen Lab, 2019)). We were only interested in a subset of combinations of outcomes, treatments, and moderators. These are specified next.

4.1.2 Study Aims

Here we highlight the primary and secondary aims of this paper. Below, the 'effect' (for which we are assessing moderation) corresponds to how a week of a certain notification category causally changes an outcome *compared to weeks with no notifications*.

The moderator aims listed below were not the only aims of the 2018 IHS. Main effects analyses were conducted prior to the analysis of moderator effects. This paper focuses on moderator analyses as those were the most interesting findings. Other study aims and results can be found in Appendix B.1.

4.1.2.1 Primary Aim

Our primary aim focuses on discovering how an intern's previous mood moderates the effect of notifications in general. Specifically, we examined: "Is the effect of a week of notifications (of any category) on average daily mood moderated by previous week's mood?' [**Outcome** = mood, **Treatment** = any (mood, activity, or sleep), **Moderator** =

mood]

Exploratory Sub-Aim

If we do find that mood moderates the effect of notifications, generally, we will assess if this moderation is consistent across all intervention categories. Specifically, we examined: “Is the effect of each individual category of notification on average daily mood moderated by previous week’s mood?” [**Outcome** = mood, **Treatment** = mood, activity, and sleep separately, **Moderator** = mood]

4.1.2.2 Secondary Aim 1

Secondary aim 1 focuses on discovering how an intern’s previous activity moderates the effect of notifications containing activity messages. Specifically, we examined: “Is the effect of a week of activity notifications on average daily step count moderated by previous week’s step count?” [**Outcome** = steps, **Treatment** = activity, **Moderator** = steps]

4.1.2.3 Secondary Aim 2

Secondary aim 2 focuses on discovering how an intern’s previous sleep moderates the effect of notifications containing sleep messages. Specifically, we examined: “Is the effect of a week of sleep notifications on average daily sleep moderated by previous week’s sleep?” [**Outcome** = sleep,**Treatment** = sleep, **Moderator** = sleep]

4.2 Methods

4.2.1 The Study App

Study participants were provided a Fitbit Charge 2 to collect sleep and activity data, and a phone appc downloaded to the intern’s phone. The app is able to conduct ecological momentary assessments (EMAs) (Shiffman et al., 2008), aggregate and visualize data, and deliver push notifications.

Since the primary aim of the study is focused on understanding the effects of interventions on intern mental health, we employ a daily EMA to measure mood valence (see Figure 4.1-ii). Daily mood is one of two cardinal symptoms of depression (Löwe et al., 2005). This daily mood EMA is used widely to track mood in depressed patients (Foreman et al., 2011). There are more widely used measurements of mental health other than mood valence (such as the Patient Health Questionnaire (Kroenke et al., 2001)), however these questionnaires are too time intensive to ask every day. Participants are prompted to enter their daily mood every day at 8pm.

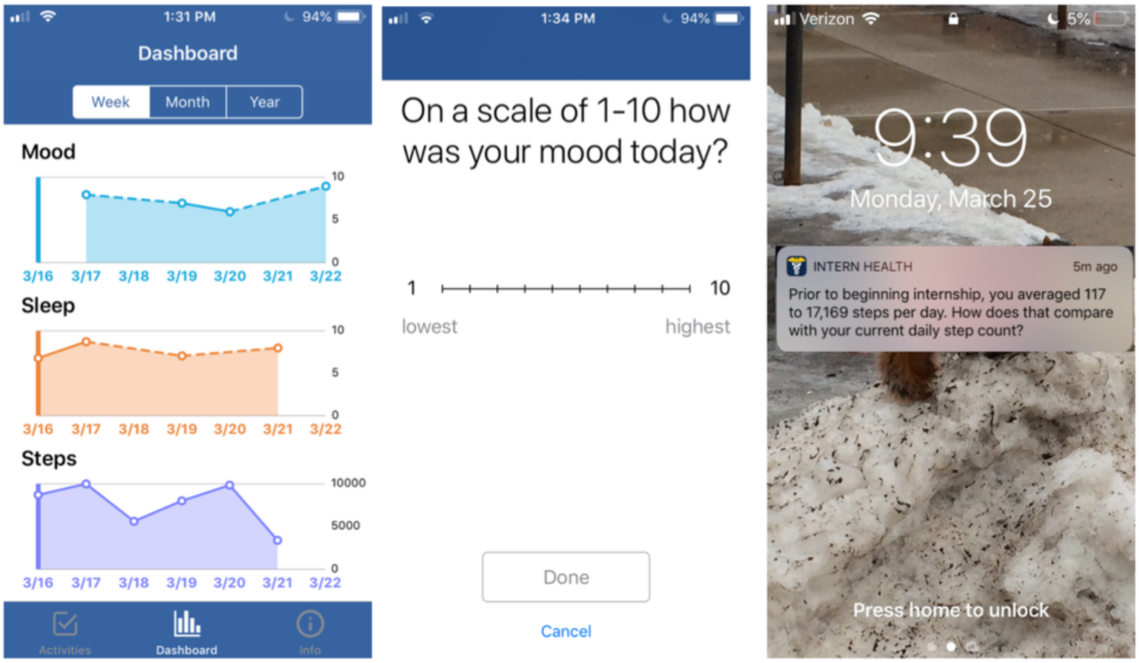
In addition to prompting and collecting EMA data, the study app aggregates and displays visual summaries of interns' historical data. The app aggregates raw step and sleep counts (collected through the Fitbit) and mood EMA data and display historical daily trends to the intern. See Figure 4.1-i. Displaying historical trends to the intern helps them self-monitor their mood, activity, and sleep trajectories, and could potentially lead to positive reactive behavior change (Korotitsch and Nelson-Gray, 1999). These displays are a type of 'pull' intervention —-interventions which are available only upon user request or user access. The 'pull' component was available to all participants at all times. Assessing its effects was not the focus of this study.

The IHS app also to delivers 'push' interventions—interventions delivered without user prompting. Evaluating the push notification intervention is the focus of this study.

4.2.2 Push Notification Intervention

Push notifications were provided to the interns through the study app, with the goal of improving healthy behavior in a target category of interest: mood, activity, and sleep. Mood notifications are intended to improve intern mood, activity notifications are intended to increase intern physical activity, and the sleep notifications are intended to increase intern sleep duration.

For all categories, there are two notification types: tips and life insights. Tips are non-



On a scale of 1-10 how was your mood today?

Prior to beginning internship, you averaged 117 to 17,169 steps per day. How does that compare with your current daily step count?

(i) App Dashboard

(ii) Mood EMA

(iii) Notifications

Figure 4.1: Screenshots of (i) app dashboard, (ii) mood EMA, and (iii) lock screen notifications

data-based notifications that provide support and advice for improving and maintaining healthy mood, activity, or sleep. Life insights are notifications based on the user's specific data. As many interns may not readily think to access the app and its data visualizations for self-monitoring, the life insights notifications are intended to provide interns a brief summary of their data directly without requiring them to access the app. By informing and motivating interns through feedback on their behavior, we aim to inspire healthy behavior change (DiClemente et al., 2001; Korotitsch and Nelson-Gray, 1999).

4.2.3 The Intern Health Study Micro-randomized Trial Design

In order to determine the best time to deliver notifications of different categories, we ran an MRT. The MRT design is pictured in Figure 4.2. The MRT design and protocol were approved by University of Michigan IRB (UM IRB Protocol #HUM00033029).

The main randomization was the weekly randomization to a specific notification category. We randomized an individual to one of three categories of notifications (mood, ac-

	Types	
Category	Life Insight	Tip
Mood	Your mood has ranges from 7 to 9 over the past 2 weeks. The average intern's daily mood goes down by 7.5% after intern year begins	Treat yourself to your favorite meal. You've earned it!
Activity	Prior to beginning internship, you averaged 117 to 17,169 steps per day. How does that compare with your current daily step count?	Exercising releases endorphins which may improve mood. Staying fit and healthy can help increase your energy level.
Sleep	The average nightly sleep duration for an intern is 6 hours 42 minutes. Your average since starting internship is 7 hours 47 minutes	Try to get 6 to 8 hours of sleep each night if possible. Notice how even small increases in sleep may help you to function at peak capacity & better manage the stresses of internship.

Table 4.1: Table of examples of 6 different groups of notifications

tivity, sleep) or to no-notification, which means the intern did not receive any notifications for the entire week. The no-notification week served as our baseline treatment comparator. That is, we were able to compare how a week of a certain notification category changed intern behavior when compared to a week of no notifications.

The randomization—and the ensuing analysis of effects—occurred at the weekly-level for two reasons. For one, the notifications are not intended to change the interns behavior in the next few hours, but over the next few days. Randomizing and analyzing effects at the weekly-level, as opposed to daily- or minute-level, permited discovery of notification effects over a longer period of time, which is valuable for our study population. Secondly, as interns are quite busy, they may not have significant behavior change after receiving a single notification. Instead interns received several notifications of the same category and had a consistent reminder about improving that category. This gave them the chance to

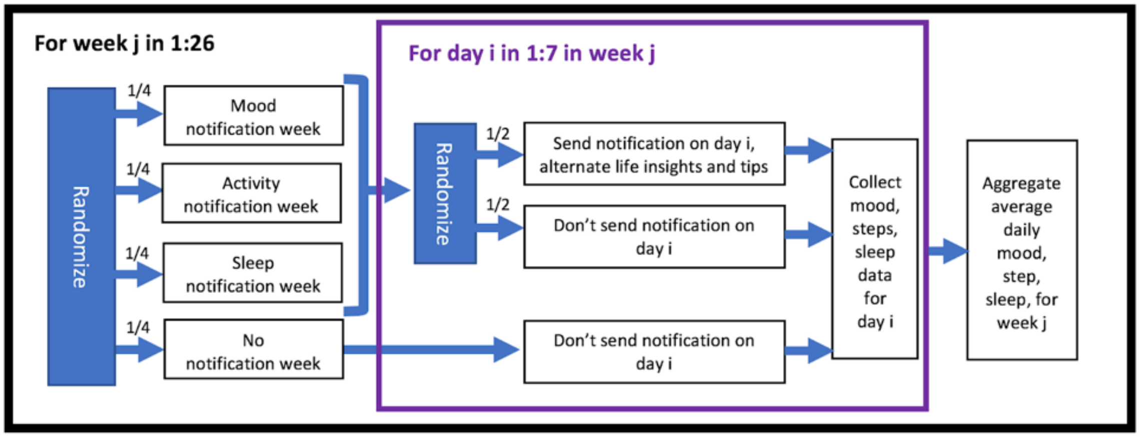

For week j in 1:26
Randomize
1/4
Mood notification week
1/4
Activity notification week
1/4
Sleep notification week
1/4
No notification week
For day i in 1:7 in week j
Randomize
1/2
Send notification on day i, alternate life insights and tips
1/2
Don't send notification on day i
Don't send notification on day i
Collect mood, steps, sleep data for day i
Aggregate average daily mood, step, sleep, for week j

Figure 4.2: Randomization scheme of the Intern Health Study MRT

register and remember the desired behavior change.

If a user was randomized to a week where they receive notifications, they were then randomized to receive a notification with 50% probability (i.e., for a mood notification week the user received, on average, 3.5 mood notifications that week). The purpose of this randomization is to balance delivering enough notifications to be noticeable and cause behavior change, but not too often that it leads to treatment fatigue (Heckman et al., 2015). Treatment fatigue is pervasive in mobile health (Nahum-Shani et al., 2017) and for individuals with heavy workloads (Heckman et al., 2015).

Another way to prevent treatment fatigue is through increased variability in notifications and the order they are received (Hockey, 2013). For each notification category, the notifications alternated between life insights and tips. Also, each notification was drawn randomly, without replacement, from a bucket of notifications. The bucket refilled once it was completely emptied. Alternating between life insights and tips increased the day to day variability of the notification framing. Drawing notifications without replacement ensured that users are not receiving repeats of the same notification. Under this scheme, on average, a user did not receive a repeated notification for 16 weeks.

4.2.4 Participants

Medical doctors starting their year-long internship in summer of 2018 were eligible to participate in the study. Interns were on boarded before the start of their internship (between April 2018-June 2018), in which they were instructed to download the study app, were provided Fitbits, completed a baseline survey, and were able to begin entering mood scores. Data collection began when they were enrolled in the study and continued until the end of the trial. Collecting data prior to the start of the internship provided baseline measurements of mood, step counts, and sleep which are valuable control variables in the analysis. The weekly randomizations and notification delivery began on June 30, 2018, one day prior to the start of interns' clinical duties. Interns were re-randomized every 7 days thereafter. During the study, notifications were sent at 3pm, mood EMAs were collected daily at 8pm, and sleep/step data were collected every minute. The interns received notifications for 6 months (26 weeks), and the trial ended on December 28th, 2018.

4.2.5 Statistical Analysis

4.2.5.1 Overview

To analyze the primary and secondary aims, we performed a moderator analysis for each of the outcomes, treatments, and moderators specified in Section 4.1.2. Here we first describe the general model and methods used. Then we provide the particular details for each aim of interest. Further details on the statistical methods can be found in Appendix B.3.

In the analysis, there were 4 sets of variables of interest:

1. The **outcome variables**, Y_t, corresponding to the treatment outcome of interest.

2. The **treatment indicator**, Z_t. For now, Z_t is binary (an indicator where $Z_t = 1$ implies it is a week where a user gets notifications of any category, and $Z_t = 0$ is a week where a user gets no notifications). The case where there are more treatments

of interest (e.g., mood, activity, and sleep notifications) will be described under the secondary aims.

3. The **moderator**, M_t, corresponding to the causal effect moderator of interest.

4. The last set of variables, X_t, are the **control variables**. The control variables are variables measured prior to each weekly randomization which are associated with Y_t. The purpose of the control variables is to reduce the variation in the outcome and reduce the standard error when estimating the treatment effect of interest. Based on analyses of previous years' IHS data (The Sen Lab, 2019), the control variables included data collected from the baseline survey (sex, Patient Health Questionnaire score (Kroenke et al., 2001), depression history, neuroticism), pre-internship data summaries (pre-internship average mood, step count, and sleep), and time-varying data (study week, previous week's average daily mood, step count, and sleep).

Note that the outcomes, treatment, and moderators correspond exactly to the outcomes, treatments, and moderators described in Section 4.1.2. Variables are indexed by time t, corresponding to each week of the study ($t = 1, \ldots, 26$). Since interns were randomized to different treatments each week, the outcomes, treatments, moderators, and control variables were aggregated at the weekly-level. Variables which are not time-varying (i.e., baseline and pre-internship data) remain constant for all values of t. Indexing M by t demonstrates our interest in assessing time-varying moderators.

A linear model was used as a working model for the moderator analysis. The model is a 'working' model, as indicated by "=", because the estimation methods do not require all parts of the model to be correctly specified. The outcome of interest (e.g., average daily mood), Y_t, was regressed on 4 sets of variables X_t, M_t, Z_t, and Z_tM_t, giving the linear working model the form:

$E(Y_t|X_t, M_t, Z_t)$ "=" $\alpha_0 X_t + \alpha_1 M_t + \beta_0 Z_t + \beta_1 Z_t M_t$.

In our model, the coefficient β_0 is interpreted as the treatment effect of notifications, compared to no notifications, when the moderator M_t is 0. The moderation effect of interest is the coefficient β_1 for the interaction of Z_t and M_t. This coefficient is interpreted as the change in treatment effect of treatment Z_t on Y_t for a 1 unit change in M_t. A positive value for β_1 indicates that the treatment works better after weeks when M_t is high, while a negative value indicates that the treatment works better after time points when M_t is low. Note that this moderation effect, β_1, is an average effect. It is average over time in the study and user-specific variables.

4.2.5.2 Estimation techniques

To estimate the coefficients, we used a weighted and centered least squares estimator described in Boruvka et al. (2018). The estimation method provides unbiased, robust estimates of the causal effect moderation of interest. The method is robust to misspecification of terms not interacted with treatment ($\alpha_0 X_t + \alpha_1 M_t$). The method also uses robust standard error estimation (i.e., sandwich estimator) to account for within-person dependencies in the data. The method was implemented in R using the package geepack (Halekoh et al., 2006).

4.2.5.3 Missing Data

Missingness will occured throughout the trial due to interns not completing self-reported mood survey or not wearing Fitbits. Multiple imputation (Little and Rubin, 2019), a robust method for dealing with missing data, was used to impute missing data at the daily level. Due to the complexity of the trial design and data structure, our imputation method combines imputation methods for longitudinal data (Bell and Fairclough, 2014) and sequentially randomized trials (Shortreed et al., 2014). Results were aggregated across multiple imputed data sets using Rubin's rules (Little and Rubin, 2019; Grund et al., 2016). Sensitivity analyses were performed to assess the sensitivity of the conclusions to missing data; the results were validated by re-analyzing smaller data sets, which only include interns or time points with minimal missingness. The results can be found in Appendix B.2

4.2.5.4 Primary Aim

The primary aim assesses previous week's average daily mood as a moderator of the effect of notifications on average daily mood. For this analysis the interpretation β_1 is "The change in treatment effect (for delivering a week of notifications compared to a week of no notifications) on average daily mood when the previous week's average daily mood increases by 1". A positive value for β_1 indicates that notifications have a better effect on mood (compared to no notifications) when the intern's previous mood was high, while a negative value for β_1 indicates that notifications have a better effect on mood when the intern's previous mood was low.

To evaluate if this effect is statistically significant, we performed a hypothesis test comparing the coefficient β_1 to 0, with a .05 type 1 error rate. We reported the estimated coefficient ($\hat{\beta}_1$), the standard error, and p-value of this test. Though estimating and testing the moderation effect is useful, it does not demonstrate the actual effects of treatment on the outcome of interest (e.g., are the notifications helping or hurting the interns). Hence, in addition to a hypothesis test, we also plotted the estimated treatment effect at various values of the moderator. We did this by using both the estimated slope ($\hat{\beta}_1$) and intercept ($\hat{\beta}_0$) of the moderation effect.

4.2.5.5 Secondary Aim 1

The first secondary aim assesses previous week's average daily step count as a moderator of the effect of activity notifications on average daily step count. For this aim, the treatment is no longer binary, since there are 4 possible notification categories. To evaluate the multivariate treatment, the treatment variable (Z_t) was encoded into 3 indicator variables: activity notification weeks ($Z_t = (1, 0, 0)$), sleep notification weeks ($Z_t = (0, 1, 0)$), mood notification weeks ($Z_t = (0, 0, 1)$), or no-notification weeks ($Z_t = (0, 0, 0)$). Here, the baseline is again no-notification weeks. The coefficients β_0 and β_1 are now 3-dimensional vectors as well, and we let β_{0i} and β_{1i} refer to the ith dimension of β_0 and β_1, respectively. See Appendix B.3 fo further details on the multivariate treatment model. The focus of inference for secondary aim 1 is on the first dimension of the moderation effect, β_{11} (i.e., the comparison between activity notification weeks and no-notification weeks).

The outcome of interest, Y_t, is average daily square root step count during the same week the subject receives notifications. The daily step count was square rooted because the raw step counts tend to have right skew. Square rooting reduced this skew, decreased outliers, and made our estimation more robust. Similarly, the moderator, M_t, of interest is the average daily square root step count of the previous week (i.e., week $t - 1$).

For this analysis, the focus is on the first dimension of the moderation effect, β_{11}. The interpretation β_{11} is "The change in treatment effect (for delivering a week of activity notifications compared to a week of no notifications) on average daily square root step count when the previous week's average daily square root step count increases by 1". A positive value for β_{11} indicates that activity notifications have a better effect on activity (compared to no notifications) when the intern's previous activity was high, while a negative value for β_{11} indicates that activity notifications have a better effect on activity when the intern's previous activity was low.

To evaluate the significance of this effect, we performed a hypothesis test comparing the coefficient β_{11} to 0, with a .05 type 1. We reported the estimated coefficient ($\hat{\beta}_{11}$), the standard error, and p-value. Again, in order to illustrate the actual size of the effect, we also used estimates $\hat{\beta}_{01}$ and $\hat{\beta}_{11}$ to plot the estimated treatment effect at various values of the moderator. For interpretability, this graph is on the re-transformed raw average daily step count scale.

4.2.5.6 Secondary Aim 2

The second secondary aim assesses previous week's average daily sleep count as a moderator of the effect of sleep notifications on average daily sleep count. Similar to secondary aim 1, the treatment here is no longer binary and we encoded the treatment vector the same way as Section 4.2.5.5. For this analysis, the focus of inference is on the second dimension (the indicator for sleep notification week), which compares sleep notification weeks to no-notification weeks. The outcome of interest, Y_t, is average daily square root sleep minutes during the same week the subject receives notifications. The daily sleep minutes was square rooted in order to reduce skew and decreases outliers. The moderator, M_t, of interest is also the average daily square root sleep minutes of the previous (i.e., week $t-1$).

For this analysis, the focus is on the second dimension of the moderation effect, β_{12}. The interpretation β_{12} is "The change in treatment effect (for delivering a week of sleep notifications compared to a week of no notifications) on average daily square root sleep minutes when the previous week's average daily square root sleep minutes increases by 1". A positive value for β_{12} would indicate that sleep notifications have a better effect on sleep (compared to no notifications) when the intern's previous sleep was high, while a negative value for β_{12} would indicate that sleep notifications have a better effect on sleep when the intern's previous sleep was low.

Again, we performed a hypothesis test comparing β_{12} to 0 with .05 type 1 error, and reported the estimated coefficient ($\hat{\beta}_{12}$), the standard error, and p-value. In order to illustrate the size of the effect, we used estimates $\hat{\beta}_{02}$ and $\hat{\beta}_{12}$ to plot the estimated treatment effect at various values of the moderator. This graph was re-transformed to the sleep minute scale.

4.2.5.7 Exploratory Sub-Aim

The exploratory aim assesses previous week's mood as a moderator of the effect of each notification category on average daily mood. For the exploratory aim, we performed a similar analysis as done in the Section 4.2.5.4, except the treatment was separated into 4 treatment categories (as in the Section 4.2.5.5). Since this aim is only exploratory, we did not calculate p-values. Instead we explored the estimated moderation effects visually. Specifically, for each notification category, we plotted the estimated treatment effect at various values of the moderator. This required making 3 separate lines using each dimension of $\hat{\beta}_0$ and $\hat{\beta}_1$, with $\hat{\beta}_{0i}$ providing the intercept and $\hat{\beta}_{1i}$ providing the slope. The moderator still corresponds to moderation of the effect of a certain notification category compared to no notifications.

4.3 Results

4.3.1 Participants

Participants were recruited through emails, which were sent to future interns from 47 different recruitment institutions between April 1st, 2018 and June 25th, 2018. The recruitment institutions comprised both medical schools, where emails were sent to all graduates, and residency locations, where emails were sent to all incoming interns. 5,233 future interns received the initial email inviting them to participate in the study. 2,134 (41%) interns downloaded the study app, completed the consent form, and filled out the baseline survey sometime before June 25th, 2018. The study app and study participation were restricted to interns using an Iphone, the phone brand used by a majority of interns. The 2,134 interns received a Fitbit Charge 2. Of the 2,134, 1,565 interns (73%) were randomly selected to participate in the MRT (see Appendix B.1 for an explanation of this initial randomization). These 1,565 interns were randomized according to Figure 4.2 starting on June 30, 2018 and continued in the MRT until December 28, 2018. Interns were incentivized to participate in the study by receiving the Fitbit and up to $125, distributed five times throughout the year ($25 each time) based on continued participation.

Of the 1,565 interns in the MRT, 56% were female, 49% had previously experienced

	1st Quartile	Median	Mean	3rd Quartile	Standard Deviation
Average daily mood	6.50	7.33	7.21	8.00	1.43
Average daily step count	6,193	7,983	8,274	10,050	3,285
Average daily hours of sleep	6.02	6.65	6.53	7.25	1.25

Table 4.2: Summary statistics of daily averages of mood, activity, and sleep during study, averaged over each week of the study. These are the primary outcomes and moderators used in the analyses of all study aims.

an episode of depression. The interns represented 321 different residency locations and 42 specialties. The study interns' baseline information closely resembled the known characteristics of the general medical intern population (Mata et al., 2015). Throughout the trial, we measure intern mood valence, steps, and nightly sleep. Summaries of the weekly-level averages of those data can be found in Table 4.2.

4.3.2 Main Findings

4.3.2.1 Primary Aim

We conclude that previous week's average daily mood is a statistically significant negative moderator of the effect of notifications on average daily mood. The estimate for the moderation is -0.052 (SE = 0.014 P = .001). The negative moderation implies that notifications performed better after weeks with low mood compared to weeks with high mood.

Figure 4.3 plots the estimated treatment effect at various values of the moderator. We see from Figure 4.3 that the effect of notifications (compared to no notifications) was positive for weeks when previous mood was low, but negative for weeks when previous mood was high. For example, when previous week's average daily mood was 3, we estimated that a week of notifications *increased* an intern's average daily mood by 0.19 (effect size = 0.13). However, when previous week's average daily mood was 9, we estimated that a week of notifications *decreased* an intern's average daily mood by 0.12 (effect size = -0.08). The

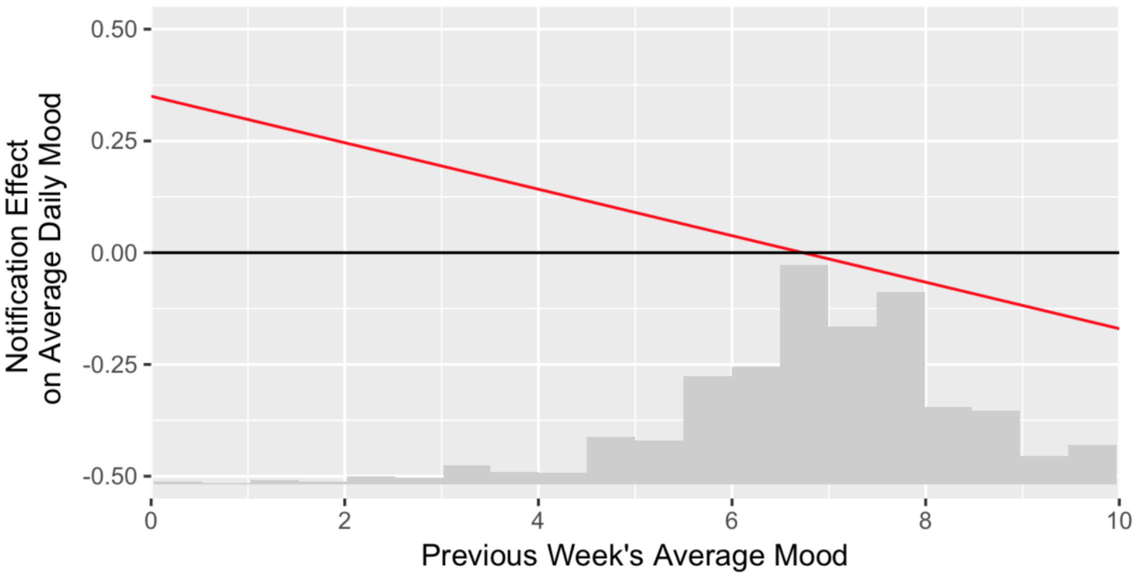


Figure 4.3: Estimated treatment effects (compared to no notifications) of notifications on average daily mood, at various values of previous week's mood. The x-axis also contains a scaled histogram of previous week's average mood.

positive treatment effect switched to a negative effect when the previous week's average daily mood was 6.7.

Exploratory Sub-Aim

For each notification, we plotted the estimated treatment effect at various values of the moderator. Essentially, we broke apart the moderation effect in Figure 4.3 into the 3 categories of notifications. The result is shown in Figure 4.4. We included the line for general notifications from Figure 4.3 for reference. Figure 4.4 demonstrates that the negative moderation by previous week's average daily mood was present in all 3 categories of notifications. Also, for all 3 categories of notifications, we see a positive treatment effect (on average daily mood) when an intern's previous mood is low, and a negative treatment effect when previous mood is high.

When previous week's average daily mood was 3, we estimated that a week of mood, activity, and sleep notifications *increased* an intern's average daily mood by 0.19, 0.16, 0.23 (effect sizes = 0.13, 0.11, 0.16), respectively. When previous week's average daily mood was 9, we estimated that a week of mood, activity, and sleep notifications decreased

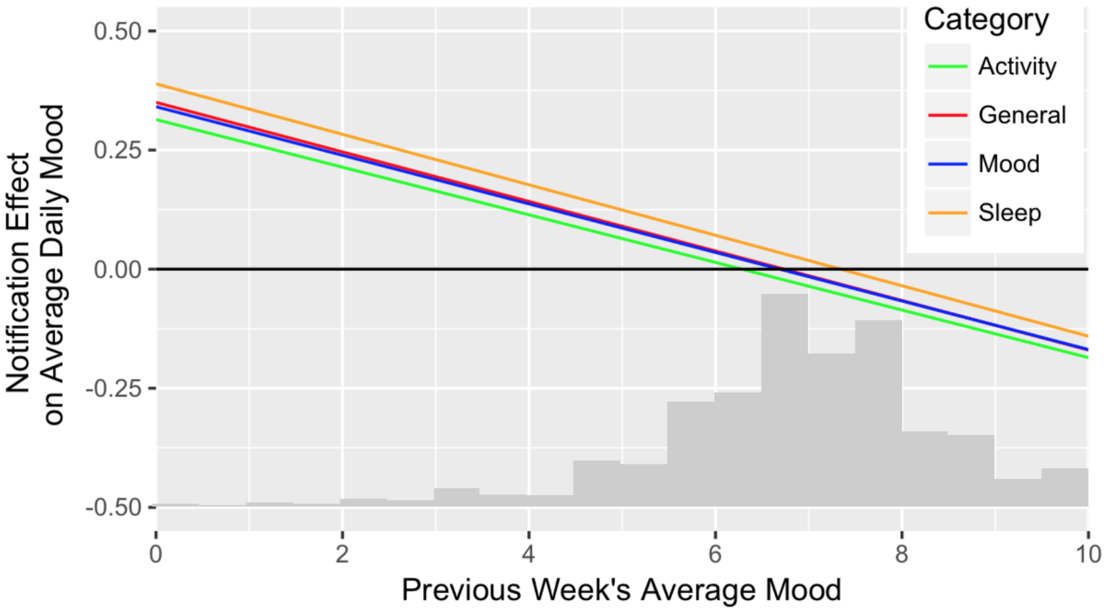


Figure 4.4: Estimated treatment effects (compared to no notifications) of different notification categories on average daily mood, at various values of previous week's mood. The x-axis also contains a scaled histogram of previous week's average mood.

an intern's average daily mood by 0.12, 0.14, 0.09 (effect sizes = -0.08, -0.10, -0.06), respectively.

4.3.2.2 Secondary Aim 1

We conclude that previous week's average daily steps is a statistically significant negative moderator of the effect of activity notifications on average daily steps. The estimate for the moderation is -0.039 (SE = 0.015, P = .013). The negative moderation implies that activity notifications perform better after weeks with low step counts compared to weeks with high step counts.

Figure 4.5 plots the estimated treatment effect at various values of the moderator. Note that in Figure 4.5, for interpretability, we re-transformed the moderation effect back from the analysis scale (square root step count) to the original scale (average daily step count). We see from Figure 4.5 that the effect of activity notifications (compared to no notifications) was positive for weeks when previous steps were low, but negative for weeks when previous

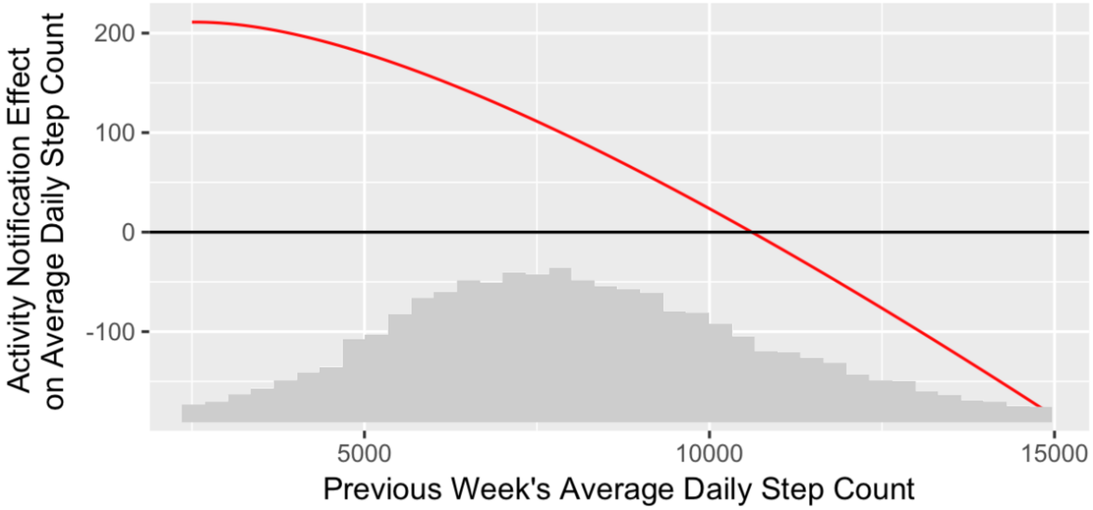


Figure 4.5: Estimated treatment effects (compared to no notifications) of activity notifications on average daily steps, at various values of previous week's step counts. The x-axis also contains a scaled histogram of previous week's average daily step count.

steps were high. For example, when previous week's average daily step count was 5,625, we estimated that a week of activity notifications *increased* an intern's average daily step count by 165 steps (effect size = 0.05). However, when previous week's average daily step count was 12,100, we estimated that a week of activity notifications *decreased* an intern's average daily step count by 60 steps (effect size = -0.02). The positive treatment effect switched to a negative effect at 10,614 steps.

4.3.2.3 Secondary Aim 2

We conclude that previous week's average daily sleep is a statistically significant negative moderator of the effect of sleep notifications on average daily sleep. The estimate for the moderation is -0.074 (SE = 0.018, P $<$ .001). The negative moderation implies that sleep notifications perform better after weeks with low sleep compared to weeks with high sleep.

Figure 4.6 plots the estimated treatment effect at various values of the moderator. Note that in Figure 4.6, for interpretability, we re-transformed the moderation effect back from analysis scale (square root sleep minutes) to the original scale (daily sleep minutes). Also

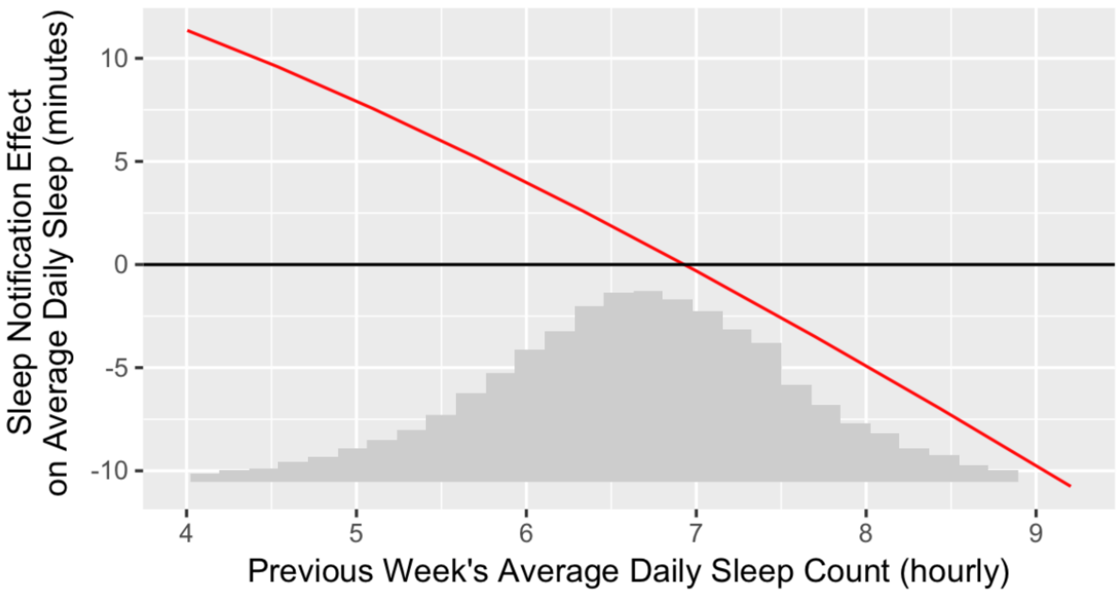


Figure 4.6: Estimated treatment effects (compared to no notifications) of sleep notifications on average daily sleep minutes, at various values of previous week's hourly sleep. The x-axis contains a scaled histogram of previous week's average daily sleep count.

for readability, the x-axis is on the hourly scale, while the y-axis is on the minute scale. We see from Figure 4.6 that the effect of sleep notifications (compared to no notifications) was positive for weeks when previous sleep was low, but negative for weeks when previous sleep was high. For example, when previous week's average daily sleep was 5 hours, we estimated that a week of sleep notifications *increased* an intern's average daily sleep by 8 minutes (effect size = 0.11). However, when previous week's average daily sleep was 8 hours, we estimated that a week of sleep notifications *decreased* an intern's average daily sleep by 5 minutes (effect size = -0.07). The positive treatment effect switched to a negative effect at 6.9 hours.

4.4 Discussion

4.4.1 Principal Findings

Overall, we found that the effects of notifications are negatively moderated by the subject's previous measurement of the outcome of interest. Specifically, we found that the effect of notifications on mood is negatively moderated by previous mood, with notifications producing a positive (beneficial) effect on mood when previous mood is low but a negative (harmful) effect when previous mood is high. For activity notifications, we found that the effect on step count is negatively moderated by previous step counts, with activity notifications having a positive (beneficial) effect on step count when interns were previously inactive but a negative (harmful) effect on step count when interns were active. Similarly, for sleep notifications, we found that the effect on sleep is negatively moderated by previous sleep duration, with sleep notifications having a positive (beneficial) effect on sleep when subjects are sleep deprived but a negative (harmful) effect on sleep when subjects have sufficient sleep.

4.4.2 Comparison with Other Studies

Other studies have also explored using real-time variables to determine the timing of mhealth interventions for mental health and stress. In many of these studies, the timing was based on proxies and predictors of stress and depression. For example, one study (Burns et al., 2011) used self-reported mood to send messages when a user's mood score was outside of their typical range. In another study (Smyth and Heron, 2016), interventions were delivered whenever self-reported stress or negative affect was high. In Smyth and Heron (2016), the authors validated the benefits of timing intervention delivery with a standard RCT. Our work differs from this work because we used experimentation to *learn* the best times to send interventions. We did not want to assume, beforehand, that interventions were only needed during periods of low mood or high stress. The MRT design allowed us to learn how to time intervention delivery.

There have been studies which sought to learn the best time to send interventions. Much of that work is focused on in-the-moment interruptibility, i.e., times when a user is open to interruption and willing to engage with a notification. For example, in one study (Pielot et al., 2017), the authors found that phone usage, time of day, and location were strong predictors of a user's willingness to engage with content (such as a game) provided via a push notification. Another study (Sarker et al., 2014) found that location, affect, current activity, time of day, day of week, and current stress are significant predictors of a user's willingness to respond to an EMA prompt. Another study (Bidargaddi et al., 2018) used an MRT to causally demonstrate that notifications (which ask users to self-monitor) are more effective when sent mid-day and on weekends. Our study also differs from this work. In our study, the outcome was not focused on short-term engagement with the notification but rather longer-term behavior change (i.e., improved weekly mood, activity, or sleep).

4.4.3 Implications

Our principal findings have implications for the development of mhealth interventions aiming to improve mood, activity, and sleep for individuals in stressful work environments. When developing high-quality mhealth interventions for this population, timing the delivery of notifications based on recent real-time data is essential. Delivering notifications when previous measurements of mood, sleep, and activity are low (i.e., when improvement is needed) can provide benefits to individual's mood and behavior. However, delivering notifications when those variables are high (i.e., when individuals are not in need of improvement) can potentially harm an individual's mood and behavior.

4.4.4 Study Strengths

Using an MRT design and repeatedly randomizing interns throughout the trial allowed us to assess causal effect moderation by time-varying measurements. With the large sample size (1,565 interns), we were able to detect the moderations of interest. The length of the study (6 months), demonstrates that our conclusions are valid beyond the first few months/weeks of the study. Finally, focusing the study on medical interns provided a unique opportunity to assess the efficacy of mobile health interventions on wellness during life and work stressors.

4.4.5 Limitations

There are other analyses that should be conducted to answer additional questions. In this paper, we only explored a linear moderation (β_1) for all moderators of interest. Exploring a more flexible moderation model (e.g., including non-linear terms or interactions between moderators) may demonstrate a more complex relationship. Also, the MRT design permits exploration of treatment effects and effect moderation at different time points in the study. Since the study is 6 months long, it may be useful to assess the moderation at each month in the study to understand how the moderation varies over time.

The results of the IHS MRT may not extrapolate to other populations because medical interns are different from the general population in average education level and socioeconomic status. Also, an intern's work schedule is another important potential moderator. Prior work (Sarker et al., 2014; Bidargaddi et al., 2018) has shown that mhealth message effectiveness does vary based on whether it is weekday or weekend. We could not assess this moderation, however, as work schedules were not reliably measured in this study. Another limitation in the study was the lack of message tailoring. Currently, the message framing and wording was the same, no matter the intern's current behavior. The messages (see Table 4.1) are framed towards improving mood, sleep, and activity. This framing may be frustrating to an intern who already has high mood or sufficient sleep/activity. Tailoring the wording of the messages (Krebs et al., 2010; Hawkins et al., 2008) could potentially eliminate the negative effect of messages when previous mood, sleep, or activity is high.

4.4.6 Future Iterations of the Intern Health Study

The IHS is an annual study which continues each year with a new cohort of interns (The Sen Lab, 2019). This provides multiple trial phases to continually update, optimize, and test interventions, and confirm findings from previous cohorts (Collins et al., 2007). Starting in the fall of 2019, we will run another study to test new hypotheses with improved interventions. Using the results and conclusions drawn from this study, in 2019 we plan to do the following:

1. Introduce tailored messages which are tailored based on an intern's previous mood, activity, and sleep (Krebs et al., 2010; Hawkins et al., 2008). For people with high previous measurements, the messages will be framed towards maintenance of healthy behavior, not improvement. The cutoffs that define 'high' and 'low' scores will be based on data collected from the 2018 study.
2. Collect work schedule information. This information will be used to compare message efficacy between work days and days off.

4.4.7 Conclusions

Overall, our study demonstrates the importance of real-time moderators for the development of high-quality mhealth interventions, especially for individuals in stressful work environments. There were times when the notifications were beneficial and times when the notifications were harmful to the study participants. Developers of mhealth interventions are encouraged to think deeply about the delivery of interventions and how real-time variables can be used to determine the best delivery times. The MRT design allowed us to discover real-time moderators. The design is useful for other app developers also aiming to learn when to deliver notification messages.

BIBLIOGRAPHY

Âkerstedt, T. (2006). Psychosocial stress and impaired sleep. *Scandinavian Journal of Work, Environment & Health 32*(6), 493–501.

Baglioni, C., Battagliese, G., Feige, B., Spiegelhalder, K., Nissen, C., Voderholzer, U., Lombardo, C., and Riemann, D. (2011). Insomnia as a predictor of depression: a meta-analytic evaluation of longitudinal epidemiological studies. *Journal of Affective Disorders 135*(1-3), 10–19.

Baucom, K. J., Queen, T. L., Wiebe, D. J., Turner, S. L., Wolfe, K. L., Godbey, E. I., Fortenberry, K. T., Mansfield, J. H., and Berg, C. A. (2015). Depressive symptoms, daily stress, and adherence in late adolescents with type 1 diabetes. *Health Psychology 34*(5), 522.

Bell, M. L. and Fairclough, D. L. (2014). Practical and statistical issues in missing data for longitudinal patient-reported outcomes. *Statistical Methods in Medical Research 23*(5), 440–459.

Bidargaddi, N., Almirall, D., Murphy, S., Nahum-Shani, I., Kovalcik, M., Pituch, T., Maaieh, H., and Strecher, V. (2018). To prompt or not to prompt? a microrandomized trial of time-varying push notifications to increase proximal engagement with a mobile health app. *JMIR mHealth and uHealth 6*(11), e10123.

Boruvka, A., Almirall, D., Witkiewitz, K., and Murphy, S. A. (2018). Assessing time-varying causal effect moderation in mobile health. *Journal of the American Statistical Association 113*(523), 1112–1121.

Burns, M. N., Begale, M., Duffecy, J., Gergle, D., Karr, C. J., Giangrande, E., and Mohr, D. C. (2011). Harnessing context sensing to develop a mobile intervention for depression. *Journal of Medical Internet Research 13*(3), e55.

Collins, L. M., Murphy, S. A., and 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.

DiClemente, C. C., Marinilli, A. S., Singh, M., and Bellino, L. E. (2001). The role of feedback in the process of health behavior change. *American Journal of Health Behavior 25*(3), 217–227.

Everett, K. D., Brantley, P. J., Sletten, C., Jones, G. N., and McKnight, G. T. (1995). The relation of stress and depression to interdialytic weight gain in hemodialysis patients. *Behavioral Medicine 21*(1), 25–30.

Foreman, A. C., Hall, C., Bone, K., Cheng, J., and Kaplin, A. (2011). Just text me: Using SMS technology for collaborative patient mood charting. *Journal of Participatory Medicine 3*, e45.

Greenberg, P. E., Fournier, A.-A., Sisitsky, T., Pike, C. T., and Kessler, R. C. (2015). The economic burden of adults with major depressive disorder in the United States (2005 and 2010). *The Journal of Clinical Psychiatry 76*(2), 155–162.

Grund, S., Robitzsch, A., and Lüdtke, O. (2016). mitml: Tools for multiple imputation in multilevel modeling. *Retreived from: https://cran.r-project.org/web/packages/mitml/index.html*.

Halekoh, U., Højsgaard, S., Yan, J., et al. (2006). The R package geepack for generalized estimating equations. *Journal of Statistical Software 15*(2), 1–11.

Hawkins, J. D., Jenson, J. M., Catalano, R., Fraser, M. W., Botvin, G. J., Shapiro, V., Brown, C. H., Beardslee, W., Brent, D., Leslie, L. K., et al. (2016). Unleashing the power of prevention. *American Journal of Medical Research 3*(1), 39.

Hawkins, R. P., Kreuter, M., Resnicow, K., Fishbein, M., and Dijkstra, A. (2008). Understanding tailoring in communicating about health. *Health Education Research 23*(3), 454–466.

Heckman, B. W., Mathew, A. R., and Carpenter, M. J. (2015). Treatment burden and treatment fatigue as barriers to health. *Current Opinion in Psychology 5*, 31–36.

Hockey, R. (2013). *The psychology of fatigue: Work, effort and control*. Cambridge University Press.

Kalmbach, D. A., Fang, Y., Arnedt, J. T., Cochran, A. L., Deldin, P. J., Kaplin, A. I., and Sen, S. (2018). Effects of sleep, physical activity, and shift work on daily mood: A prospective mobile monitoring study of medical interns. *Journal of General Internal Medicine 33*(6), 914–920.

Klasnja, P., Hekler, E. B., Shiffman, S., Boruvka, A., Almirall, D., Tewari, A., and Murphy, S. A. (2015). Microrandomized trials: An experimental design for developing just-in-time adaptive interventions. *Health Psychology 34*(S), 1220.

Korotitsch, W. J. and Nelson-Gray, R. O. (1999). An overview of self-monitoring research in assessment and treatment. *Psychological Assessment 11*(4), 415.

Kraemer, H. C., Wilson, G. T., Fairburn, C. G., and Agras, W. S. (2002). Mediators and moderators of treatment effects in randomized clinical trials. *Archives of General Psychiatry 59*(10), 877–883.

Krebs, P., Prochaska, J. O., and Rossi, J. S. (2010). A meta-analysis of computer-tailored interventions for health behavior change. *Preventive Medicine 51*(3-4), 214–221.

Kroenke, K., Spitzer, R. L., and Williams, J. B. (2001). The PHQ-9: validity of a brief depression severity measure. *Journal of General Internal Medicine 16*(9), 606–613.

Lallukka, T., Sarlio-Lähteenkorva, S., Roos, E., Laaksonen, M., Rahkonen, O., and Lahelma, E. (2004). Working conditions and health behaviours among employed women and men: The Helsinki health study. *Preventive Medicine 38*(1), 48–56.

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, Volume 793. Wiley.

Löwe, B., Kroenke, K., and Gräfe, K. (2005). Detecting and monitoring depression with a two-item questionnaire (phq-2). *Journal of Psychosomatic Research 58*(2), 163–171.

Mata, D. A., Ramos, M. A., Bansal, N., Khan, R., Guille, C., Di Angelantonio, E., and Sen, S. (2015). Prevalence of depression and depressive symptoms among resident physicians: a systematic review and meta-analysis. *Journal of American Medical Association 314*(22), 2373–2383.

Nahum-Shani, I., Smith, S. N., Spring, B. J., Collins, L. M., Witkiewitz, K., Tewari, A., and Murphy, S. A. (2017). Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine 52*(6), 446–462.

Pielot, M., Cardoso, B., Katevas, K., Serrà, J., Matic, A., and Oliver, N. (2017). Beyond interruptibility: Predicting opportune moments to engage mobile phone users. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 1*(3), 91.

Riley, W. T. (2014). Theoretical models to inform technology-based health behavior interventions. *LA, Marsch, SE, Lord, J. Dallery,(Eds), Behavioral Health Care and Technology: Using science-based innovations to transform practice*, 13–26.

Riley, W. T., Rivera, D. E., Atienza, A. A., Nilsen, W., Allison, S. M., and 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.

Sarker, H., Sharmin, M., Ali, A. A., Rahman, M. M., Bari, R., Hossain, S. M., and Kumar, S. (2014). Assessing the availability of users to engage in just-in-time intervention in the natural environment. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 909–920. ACM.

Shiffman, S., Stone, A. A., and Hufford, M. R. (2008). Ecological momentary assessment. *Annu. Rev. Clin. Psychol. 4*, 1–32.

Shortreed, S. M., Laber, E., Scott Stroup, T., Pineau, J., and Murphy, S. A. (2014). A multiple imputation strategy for sequential multiple assignment randomized trials. *Statistics in Medicine 33*(24), 4202–4214.

Smyth, J. M. and Heron, K. E. (2016). Is providing mobile interventions “just-in-time” helpful? an experimental proof of concept study of just-in-time intervention for stress management. In *2016 IEEE Wireless Health (WH)*, pp. 1–7. IEEE.

Spruijt-Metz, D. and Nilsen, W. (2014). Dynamic models of behavior for just-in-time adaptive interventions. *IEEE Pervasive Computing 13*(3), 13–17.

Ströhle, A. (2009). Physical activity, exercise, depression and anxiety disorders. *Journal of Neural Transmission 116*(6), 777.

Tennant, C. (2001). Work-related stress and depressive disorders. *Journal of Psychosomatic Research 51*(5), 697–704.

The Sen Lab (2019, April). Intern health study. `https://www.srijan-sen-lab.com/intern-health-study`.

World Health Organization (2017). Depression and other common mental disorders: global health estimates. Technical report, World Health Organization.

Zhang, S. and Elhadad, N. (2016). Factors contributing to dropping-out in an online health community: static and longitudinal analyses. In *Amia Annual Symposium Proceedings*, Volume 2016, pp. 2090. American Medical Informatics Association.

CHAPTER V

Future Work

With the goal of improving analysis methods and increasing applicability of sequential randomized trials, there are a number of directions for future research. In this chapter, we present a few of those possible directions.

5.1 Novel Designs of Cluster-randomized SMARTs

In the cluster-randomized SMART designs in Chapter III, all interventions and randomizations (and re-randomizations) occur at the cluster-level. It may be useful to have dynamic treatment regimes which include both individual-level and cluster-level interventions. For example, in education, a dynamic treatment regime which first intervenes at the classroom-level and then assesses response status and subsequently intervenes at the student-level may be useful and cost-effective. To evaluate such a dynamic treatment regime, the corresponding SMART would need to incorporate both individual-level and cluster-level randomizations. Extending the statistical methods and sample size formulae from Chapter III to such designs would be valuable. Some work in this area has already been done (Ktsanes, 2017).

5.2 Adaptively Randomized Sequential Randomized Trials in Scaled Digital Learning Environments

Online optimization techniques, such as reinforcement learning and multi-armed bandit algorithms, have become increasingly popular in scaled digital learning environments (Liu et al., 2014; Clement et al., 2013; Segal et al., 2018; Rafferty et al., 2018). Compared to the sequential randomized trials discussed in Chapter II, online optimization techniques are beneficial because they aim to provide efficacious treatment to users currently receiving interventions. However, in much of this online optimization work, intervention evaluation is absent. Combining online optimization techniques with sequential randomized trials can allow for broad exploration and evaluation of interventions, while simultaneously helping individuals within the trial. Adaptively randomized sequential randomized trials have been used in health (e.g., Cheung et al. 2015), but their use in scaled digital learning environments is minimal.

5.3 Incorporating Prediction into Dynamic Treatment Regimes

The development of prediction methods for high-dimensional data in mobile health and online education is a large area of research. In mobile health, deep learning algorithms are able to predict mood and stress using mobile sensor data (Taylor et al., 2017). In online education, real-time course data can be used to predict dropout (Gardner and Brooks, 2018). Incorporating predictions into dynamic treatment regimes could be useful. For example, a dynamic treatment regime could tailor future treatment delivery based on real-time predictions. Creating fast online prediction algorithms—that could be implemented on mobile devices and at scale—could improve the applicability of prediction in dynamic treatment regimes. Developing statistical methods and experimental designs to better evaluate the use of predictions in dynamic treatment regimes would also be useful.

5.4 Missing data in Micro-randomized Trials

Due to a variety of reasons including lack of self-report and not wearing wearables, missing data can occur at high rates in mhealth studies (e.g., Burke et al. 2012; Scherer et al. 2017, Figure B.3). Developing robust missing data techniques for mhealth MRTs is challenging because of both the large number of re-randomizations and the continual streams of sensor data which may be missing. There has been some work on developing multiple imputation techniques for SMARTs (Shortreed et al., 2014). Extending those methods to MRTs would be valuable.

There are many open research questions regarding the design and analysis of sequential randomized trials. This thesis aimed to answer some of those questions. By developing novel designs and analysis techniques for sequential randomized trials, data can be better used to develop high-quality dynamic treatment regimes for interns coping with depression, patients dealing with mental illness, and students hoping to learn through online courses. Through this work, I hope those interns, patients, and students are a step closer to receiving the care they need.

BIBLIOGRAPHY

Burke, L. E., Styn, M. A., Sereika, S. M., Conroy, M. B., Ye, L., Glanz, K., Sevick, M. A., and Ewing, L. J. (2012). Using mhealth technology to enhance self-monitoring for weight loss: a randomized trial. *American Journal of Preventive Medicine 43*(1), 20–26.

Cheung, Y. K., Chakraborty, B., and Davidson, K. W. (2015). Sequential multiple assignment randomized trial (SMART) with adaptive randomization for quality improvement in depression treatment program. *Biometrics 71*(2), 450–459.

Clement, B., Roy, D., Oudeyer, P., and Lopes, M. (2013). Multi-armed bandits for intelligent tutoring systems. *arXiv preprint arXiv:1310.3174*.

Gardner, J. and Brooks, C. (2018). Student success prediction in moocs. *User Modeling and User-Adapted Interaction 28*(2), 127–203.

Ktsanes, R. (2017). *Design and Analysis of Trials for Developing Adaptive Treatment Strategies in Complex Clustered Settings*. Ph. D. thesis, Northwestern University.

Liu, Y., Mandel, T., Brunskill, E., and Popovic, Z. (2014). Trading off scientific knowledge and user learning with multi-armed bandits. In *EDM*, pp. 161–168.

Rafferty, A., Ying, H., and Williams, J. (2018). Bandit assignment for educational experiments: Benefits to students versus statistical power. In *AIED*, pp. 286–290.

Scherer, E. A., Ben-Zeev, D., Li, Z., and Kane, J. M. (2017, Jan). Analyzing mhealth engagement: Joint models for intensively collected user engagement data. *JMIR Mhealth Uhealth 5*(1), e1.

Segal, A., David, Y., Williams, J., Gal, K., and Shalom, Y. (2018). Combining difficulty ranking with multi-armed bandits to sequence educational content. In *AIED*, pp. 317–321.

Shortreed, S. M., Laber, E., Scott Stroup, T., Pineau, J., and Murphy, S. A. (2014). A multiple imputation strategy for sequential multiple assignment randomized trials. *Statistics in Medicine 33*(24), 4202–4214.

Taylor, S. A., Jaques, N., Nosakhare, E., Sano, A., and Picard, R. (2017). Personalized multitask learning for predicting tomorrow's mood, stress, and health. *IEEE Transactions on Affective Computing*.

APPENDICES

APPENDIX A

Appendix of Chapter III

A.1 Derivation of Sample Size Formulae

Here we derive the sample size formulae. We begin by deriving the sample size formula for a prototypical design with no covariates. Then we make simple extensions to derive the formula under an ADEPT design and formulae when a cluster-level covariate is included. All sample size formulae are based on primary aim (iv), that is the marginal mean comparison of two DTRs that begin with a different initial treatment (i.e., $E_{1,b_2}(Y_{ij}) - E_{\text{-}1,c_2}(Y_{ij})$).

For simplicity, these derivations assume the marginal mean model is parameterized as follows.

For the Prototypical SMART we postulate a model of the form

$$\mu(\mathbf{X}_{ij}, a_1, a_2; \boldsymbol{\beta}, \boldsymbol{\eta}) = \beta_0 I\{a_1 = 1, a_2 = 1\} + \beta_1 I\{a_1 = 1, a_2 = -1\}$$

$$+ \beta_2 I\{a_1 = -1, a_2 = 1\} + \beta_3 I\{a_1 = -1, a_2 = -1\} + \boldsymbol{\eta}^T \mathbf{X}_{ij}.$$

For ADEPT we postulate a model of the form

$$\mu(\mathbf{X}_{ij}, a_1, a_2; \boldsymbol{\beta}, \boldsymbol{\eta}) = \beta_0 I\{a_1 = 1, a_2 = 1\} + \beta_1 I\{a_1 = 1, a_2 = -1\}$$

$$+ \beta_2 I\{a_1 = -1\} + \boldsymbol{\eta}^T \mathbf{X}_{ij}.$$

Fitting the re-parameterized models will yield the exact same conclusions as fitting the marginal mean models in Equations 3.1 and 3.2.

A.1.1 Prototypical Design without Covariates

For data arising from a prototypical SMART design, we derive the sample size formula for detecting a significant difference between mean outcomes from two treatment regimens, (b_1, b_2) and (c_1, c_2). Because we are interested in comparing two regimes starting with a different initial treatment, without loss of generality, we let $b_1 = 1$ and $c_1 =$ -1.

We are interested in testing the hypothesis

$$H_0 : E_{1,b_2}(Y_{ij}) - E_{\text{-}1,c_2}(Y_{ij}) = 0$$

against the alternative

$$H_1 : E_{1,b_2}(Y_{ij}) - E_{\text{-}1,c_2}(Y_{ij}) = \delta\sigma,$$

where $\sigma = \sqrt{(\sigma^2_{1,b_2} + \sigma^2_{\text{-}1,c_2})/2}$, $\sigma^2_{a_1,a_2} = \text{Var}_{a_1,a_2}(Y_{ij})$, and δ is the standardized effect size.

We make a series of assumptions to derive our sample size formulae. We highlight these assumptions throughout our derivation to illustrate their use in the calculation. Note that our sample size is developed for a fixed cluster size, m (extensions to the unequal cluster size case can be done as in Kerry and Bland (2001)).

Our test statistic used for the hypothesis test is

$$Z = \frac{\sqrt{N}(\hat{\mu}(1, b_2) - \hat{\mu}(\text{-}1, c_2))}{\sqrt{\hat{\tau}^2(1, b_2) + \hat{\tau}^2(\text{-}1, c_2) - 2\widehat{Cov}(\sqrt{N}\hat{\mu}(1, b_2), \sqrt{N}\hat{\mu}(\text{-}1, c_2))}}.$$

Here, $\hat{\tau}^2(a_1, a_2)$ is an estimate of the variance, $\tau^2(a_1, a_2)$ = Var$(\sqrt{N}\hat{\mu}(a_1, a_2))$ which can be calculated using the matrix, $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\theta}}}$, given in the supplementary material.

In large samples and under assumption 3 described in the *Sample Size Formulae* section, the distributions of $\hat{\mu}(1, b_2)$ and $\hat{\mu}(\text{-}1, c_2)$ can be approximated by a normal distribution, $\hat{\tau}^2(1, b_2) \approx \tau^2(1, b_2)$, $\hat{\tau}^2(\text{-}1, c_2) \approx \tau^2(\text{-}1, c_2)$, and $\widehat{Cov}(\sqrt{N}\hat{\mu}(1, b_2), \sqrt{N}\hat{\mu}(\text{-}1, c_2)) \approx Cov(\sqrt{N}\hat{\mu}(1, b_2), \sqrt{N}\hat{\mu}(\text{-}1, c_2)) = 0$ (the covariance is 0 due to the independence of estimators of marginal means with different initial treatments). Thus, Z approximately has a standard normal distribution under the null hypothesis.

Note that these calculations are exactly the same as those highlighted in the *Hypothesis Testing* section. Specifically, under the original parameterization, letting $\mathbf{c} = (0, 2, b_2 - c_2, b_2 + c_2, \mathbf{0}_p)^T$ then $\mathbf{c}^T(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\eta}}) = \hat{\mu}(1, b_2) - \hat{\mu}(\text{-}1, c_2)$ and $\mathbf{c}^T\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\eta}}}\,\mathbf{c} = \hat{\tau}^2(1, b_2) + \hat{\tau}^2(\text{-}1, c_2) - 2\widehat{Cov}(\sqrt{N}\hat{\mu}(1, b_2), \sqrt{N}\hat{\mu}(\text{-}1, c_2))$.

Under the alternative, our test statistic is normal with approximate mean $\sqrt{N}\delta\sigma / \sqrt{\tau^2(1, b_2) + \tau^2(\text{-}1, c_2)}$ and variance 1. Doing standard power calculations (Oetting et al., 2011) for a hypothesis test of size α, in order to obtain desired power of $1 - \beta$, we need to find N that satisfies

$$z_\beta \approx -z_{\alpha/2} + \frac{\delta\sigma\sqrt{N}}{\sqrt{\tau^2(1, b_2) + \tau^2(\text{-}1, c_2)}}$$

$$N = \frac{(z_\beta + z_{\alpha/2})^2(\tau^2(1, b_2) + \tau^2(\text{-}1, c_2))}{\delta^2\sigma^2}. \tag{A.1}$$

Everything in this formula can be explicitly found except $\tau^2(1, b_2)$ and $\tau^2(\text{-}1, c_2)$. Hence we now aim to derive upper bounds for these variables in order to write our sample size formula in terms of either known or easily elicited quantities.

Note that under the parameterization, for any DTR (a_1, a_2), $\tau^2(a_1, a_2) = \text{Var}[\sqrt{N}\hat{\mu}(a_1, a_2)] = [\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}}]_{(a_1, a_2)} = [\mathbf{J}^{\text{-}1}E[\mathbf{U}_i\mathbf{U}_i^T]\mathbf{J}^{\text{-}1}]_{(a_1, a_2)}$, with $\mathbf{U}_i$ and $\mathbf{J}$ defined in the supplementary material. Here, to simplify notation, in the 4x4 matrices, $\mathbf{M} = \boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}}$, $\mathbf{J}$, or E[$\mathbf{U}_i\mathbf{U}_i^T$], we define $[\mathbf{M}]_{(a_1, a_2)}$ as the diagonal element corresponding to DTR (a_1, a_2) (e.g., the (3,3) element for DTR $(-1, 1)$). Also, the vector $\mathbf{1}_m$ is defined as the mx1 vector of 1's.

Lastly, as defined in the *Sample Size Formulae* section, p_{a_1} is the probability of responding given the cluster had received initial treatment a_1.

After simplification, we find that $\mathbf{J}$ is a diagonal matrix with diagonal elements

$$[\mathbf{J}]_{(a_1,a_2)} = E[W_i I_{i(a_1,a_2)} \mathbf{1}_m^T \mathbf{V}_{i,a_1,a_2}^{\text{-}1} \mathbf{1}_m] = \mathbf{1}_m^T \mathbf{V}_{i,a_1,a_2}^{\text{-}1} \mathbf{1}_m.$$

For $E[\mathbf{U}_i \mathbf{U}_i^T]_{(a_1,a_2)}$, we perform the following simplification

$$E[\mathbf{U}_i \mathbf{U}_i^T]_{(a_1,a_2)}$$

$$= E[W_i^2 I_{i(a_1,a_2)} \mathbf{1}_m^T \mathbf{V}_{i,a_1,a_2}^{\text{-}1} (\mathbf{Y}_i - \boldsymbol{\mu}(a_1,a_2))(\mathbf{Y}_i - \boldsymbol{\mu}(a_1,a_2))^T \mathbf{V}_{i,a_1,a_2}^{\text{-}1} \mathbf{1}_m]$$

$$= \mathbf{1}_m^T \mathbf{V}_{i,a_1,a_2}^{\text{-}1} E_{a_1,a_2}[W_i (\mathbf{Y}_i - \boldsymbol{\mu}(a_1,a_2))(\mathbf{Y}_i - \boldsymbol{\mu}(a_1,a_2))^T] \mathbf{V}_{i,a_1,a_2}^{\text{-}1} \mathbf{1}_m$$

$$= \mathbf{1}_m^T \mathbf{V}_{i,a_1,a_2}^{\text{-}1} \Big[2E_{a_1,a_2}[(\mathbf{Y}_i - \boldsymbol{\mu}(a_1,a_2))(\mathbf{Y}_i - \boldsymbol{\mu}(a_1,a_2))^T | R=1] p_{a_1}$$

$$+ 4E_{a_1,a_2}[(\mathbf{Y}_i - \boldsymbol{\mu}(a_1,a_2))(\mathbf{Y}_i - \boldsymbol{\mu}(a_1,a_2))^T | R=0](1-p_{a_1}) \Big] \mathbf{V}_{i,a_1,a_2}^{\text{-}1} \mathbf{1}_m$$

$$= 2 * \mathbf{1}_m^T \mathbf{V}_{i,a_1,a_2}^{\text{-}1} \boldsymbol{\Sigma}_{a_1,a_2} \mathbf{V}_{i,a_1,a_2}^{\text{-}1} \mathbf{1}_m$$

$$+ 2(1-p_{a_1}) * \mathbf{1}_m^T \mathbf{V}_{i,a_1,a_2}^{\text{-}1} E_{a_1,a_2}[(\mathbf{Y}_i - \boldsymbol{\mu}(a_1,a_2))(\mathbf{Y}_i - \boldsymbol{\mu}(a_1,a_2))^T | R=0]$$

$$\cdot \mathbf{V}_{i,a_1,a_2}^{\text{-}1} \mathbf{1}_m.$$

To go from line 2 to 3, we assume Robin's consistency assumption holds, i.e., that the cluster's observed outcomes equal the cluster's potential outcomes under the observed DTR (Robins, 1997). Under this assumption we are able to switch from E, which is an expected value over observed data, to E_{a_1,a_2} which is the expected value had the entire population received DTR (a_1, a_2) (Neyman et al., 1935; Rubin, 1978).

For further simplification, we now make assumption 2. This assumption is equivalent to assuming, for a specific DTR (a_1, a_2) (we drop the a_1, a_2 from the subscripts for convenience) $|(\sigma_R^2 \rho_R - \sigma_{NR}^2 \rho_{NR}) p_{a_1} + (\mu_R - \mu_{NR})^2 (p_{a_1})(1 - 2p_{a_1})| \leq (\sigma_R^2 - \sigma_{NR}^2) p_{a_1} + (\mu_R -$

$\mu_{NR})^2(p_{a_1})(1-2p_{a_1})$, where $\mu_R, \sigma_R^2, \rho_R$ are the mean, variance, and ICC of responders had the whole population received DTR (a_1, a_2), (i.e., $\mu_R = E_{a_1,a_2}(Y_{ij}|R_i = 1), \sigma_R^2 = \text{Var}_{a_1,a_2}(Y_{ij}|R_i = 1), \rho_R = \text{Cov}_{a_1,a_2}(Y_{i1}, Y_{i2}|R_i = 1)/\text{Var}_{a_1,a_2}(Y_{ij}|R_i = 1)$), similarly defined for NR and non-responders (i.e., conditional on R = 0). Again, p_{a_1} is the probability of response, given initial treatment a_1.

For DTR (a_1, a_2), this condition is satisfied if the probability of response is less than or equal to 0.5 (which is typical for prototypical SMART designs), the non-responders of that regimen have a variance which is less than or equal to the variance of responders of the regimen, and both responders and non-responders have similar within cluster covariances.

Under assumption 2 and using our previous simplification we can bound our expression for $E[\mathbf{U}_i\mathbf{U}_i^T]_{(a_1,a_2)}$ by

$$E[\mathbf{U}_i\mathbf{U}_i^T]_{(a_1,a_2)} \leq$$

$$2(1+(1-p_{a_1})) * \mathbf{1}_m^T\mathbf{V}^{\text{-1}}_{i,a_1,a_2}\mathbf{\Sigma}_{a_1,a_2}\mathbf{V}^{\text{-1}}_{i,a_1,a_2}\mathbf{1}_m.$$

We next utilize the fact that our working covariance matrix, $\mathbf{V}_{i,a_1,a_2}$, is exchangeable. With some linear algebra, this assumption allows us to perform the following simplification

$$\frac{2(2-p_{a_1})\mathbf{1}_m^T\mathbf{V}^{\text{-1}}_{i,a_1,a_2}\mathbf{\Sigma}_{a_1,a_2}\mathbf{V}^{\text{-1}}_{i,a_1,a_2}\mathbf{1}_m}{(\mathbf{1}_m^T\mathbf{V}^{\text{-1}}_{i,a_1,a_2}\mathbf{1}_m)^2} = \frac{2(2-p_{a_1})\mathbf{1}_m^T\mathbf{\Sigma}_{a_1,a_2}\mathbf{1}_m}{m^2}.$$

Next, using assumption 1, we exploit the exchangeable population covariance structure (i.e., $\text{Cov}_{a_1,a_2}(\mathbf{Y}_i) = \sigma^2_{a_1,a_2} * \mathbf{Exch}(\rho_{a_1,a_2})$, where $\rho_{a_1,a_2} = \text{Cor}_{a_1,a_2}(Y_{i1}, Y_{i2})$). Putting everything together, we obtain

$$\tau^2(a_1,a_2) = \text{Var}[\sqrt{N}\hat{\mu}(a_1,a_2)] \overset{3}{=} [\mathbf{J}^{\text{-1}}E[\mathbf{U}_i\mathbf{U}_i^T]\mathbf{J}^{\text{-1}}]_{(a_1,a_2)} \overset{2}{\leq}$$

$$\frac{2(2-p_{a_1})\mathbf{1}_m^T\mathbf{V}^{\text{-1}}_{i,a_1,a_2}\mathbf{\Sigma}_{a_1,a_2}\mathbf{V}^{\text{-1}}_{i,a_1,a_2}\mathbf{1}_m}{(\mathbf{1}_m^T\mathbf{V}^{\text{-1}}_{i,a_1,a_2}\mathbf{1}_m)^2} = \frac{2(2-p_{a_1})\mathbf{1}_m^T\mathbf{\Sigma}_{a_1,a_2}\mathbf{1}_m}{m^2} \overset{1}{=}$$

$$\frac{2(2-p_{a_1})\sigma^2_{a_1,a_2}[1+(m-1)\rho_{a_1,a_2}]}{m}, \qquad \text{(A.2)}$$

where the numbering above the equalities and inequalities illustrates which assumption is being used.

We utilize the across regimen covariance equality of assumption 1 in order to simplify things further. Note that if one had good estimates of $\sigma^2_{a_1,a_2}$ and ρ_{a_1,a_2}, then you could easily obtain N by plugging in the values into Equation A.2 and then using these estimates in Equation A.1.

Using this equality, we combine Equation A.2 with Equation A.1 and simplify to obtain

$$N = \frac{4(z_\beta + z_{\alpha/2})^2}{m\delta^2} \cdot (1+(m-1)\rho) \cdot (1 + \frac{(1-p_1)+(1-p_{\text{-}1})}{2}).$$

A.1.2 ADEPT Design without Covariates

All the calculations done above are nearly identical for the ADEPT case. The only major difference arises from the lack of re-randomization of clusters receiving initial treatment $a_1 = -1$. This in fact makes the calculations simpler for $\tau^2(-1,.)$ In particular, we assume assumptions 1 and 3, however, assumption 2 only needs to be assumed for DTR (1,b_2). Under these assumptions we obtain

$$\tau^2(1,b_2) \le \frac{2(2-p_1)\sigma^2_{1,b_2}[1+(m-1)\rho_{1,b_2}]}{m}, \tau^2(\text{-}1,.) = \frac{2\sigma^2_{\text{-}1,.}[1+(m-1)\rho_{\text{-}1,.}]}{m}. \tag{A.3}$$

After utilizing the across regimen population covariance equality of assumption 1, we combine the Equation A.3 with Equation A.1 and simplify to obtain

$$N = \frac{4(z_\beta + z_{\alpha/2})^2}{m\delta^2} \cdot (1+(m-1)\rho) \cdot (1 + \frac{1-p_1}{2}).$$

For this sample size formula, we actually only need to assume $\sigma^2_{1,b_2} \le \sigma^2_{-1,.}$ (as opposed to the equality assumed in assumption 1) to ensure our power is larger than 1 - β.

A.1.3 With a Cluster-Level Covariate

In this section, we write the sample size formulae when adding a single cluster-level covariate to the model (i.e., the mx1 vector $\mathbf{X}_i \triangleq (X_{i1}, X_{i2}, \ldots, X_{im})^T = (X_i, X_i, \ldots, X_i)^T)$. First, for DTR (a_1, a_2), we define $\sigma^{2*}_{a_1,a_2} \triangleq E_{a_1,a_2}[\text{Var}_{a_1,a_2}(Y_{ij}|X_i)]$, and $\rho^*_{a_1,a_2} \triangleq E_{a_1,a_2}[\text{Cov}_{a_1,a_2}(Y_{i1}, Y_{i2}|X_i)]/\sigma^{2*}_{a_1,a_2}$, where the expectations are taken over X_i. Note in the homoscedastic case, the expectation is unnecessary since the conditional variances and covariances are constant for all X_i.

The key to extending our formulae to the covariate case is observing that the numerator of our sample size formulae will now be in terms of the average *conditional* (on X) variances and ICCs, $\sigma^{2*}_{a_1,a_2}$ and $\rho^*_{a_1,a_2}$, while the denominator remains in terms of the overall variance, $\sigma^2_{a_1,a_2}$. Since $\sigma^2_{a_1,a_2} = \sigma^{2*}_{a_1,a_2} + \eta^2 Var(X_i)$ and $\text{Cov}_{a_1,a_2}(Y_{i1}, Y_{i2}) = E_{a_1,a_2}[\text{Cov}_{a_1,a_2}(Y_{i1}, Y_{i2}|X_i)] + \eta^2 Var(X_i)$, then $\sigma^{2*}_{a_1,a_2}$ and $\rho^*_{a_1,a_2}$ must be less than or equal to $\sigma^2_{a_1,a_2}$ and ρ_{a_1,a_2}. Thus the numerator of our formulae is reduced while the denominator remains the same. With some algebra, this reduction is shown to be $1 - \text{Cor}^2(Y_{ij}, X_i)$. Note that this reduction can be shown to be the same reduction arising from including a cluster-level covariate in clustered RCTs, see Hedges and Rhoads (2009).

For simplification, we do all calculations assuming our covariate has mean 0 (this eliminates covariance between our marginal mean estimates). When our covariate does not have mean 0, one can show that mean centering our covariate does not change the value of our test statistic, and hence does not change our power. Thus our sample size formulae remains valid when the covariate does not have mean 0.

A.1.3.1 Prototypical Design

Using assumptions 1-3, the relationships above, and doing similar algebra as in the non-covariate case (except now everything is in terms of $\sigma^{2*}_{a_1,a_2}$ and $\rho^*_{a_1,a_2}$), we obtain for both DTRs of interest

$$\tau^2(a_1, a_2) \leq \frac{2(2-p_1)\sigma^{2*}_{a_1,a_2}[1+(m-1)\rho^*_{a_1,a_2}]}{m}. \tag{A.4}$$

Making assumption 1 (on marginal population variances and correlations) will lead to equality of expected conditional variances and correlations due to the simple relationship between the conditional and marginal variances and covariances highlighted above. Hence we define $\rho^* \triangleq \rho^*_{1,b_2} = \rho^*_{-1,c_2}$ and $\sigma^{2*} \triangleq \sigma^{2*}_{1,b_2} = \sigma^{2*}_{\text{-}1,c_2}$.

We also take advantage of the fact that $\text{Cor}^2_{1,b_2}(Y_{ij}, X_i) = \eta^2\text{Var}(X_i)/\sigma^2_{1,b_2} = \eta^2\text{Var}(X_i)/\sigma^2_{-1,c_2}$, i.e., is also equal across both regimes. We define $\text{Cor}^2(Y,X) \triangleq \text{Cor}^2_{1,b_2}(Y_{ij}, X_i) = \text{Cor}^2_{-1,c_2}(Y_{ij}, X_i)$.

Ultimately, we obtain

$$N = \frac{4(z_\beta + z_{\alpha/2})^2}{m\delta^2}(1+(m-1)\rho^*)(1+\frac{(1-p_1)+(1-p_{\text{-}1})}{2})[1-\text{Cor}^2(Y,X)].$$

A.1.3.2 ADEPT Design

Similar to the prototypical design, under assumptions 1-3, for DTR (1, b_2) we obtain the same bound as in Equation A.4. For DTR (-1, .), without utilizing assumption 2, we obtain

$$\tau^2(-1,.) = \frac{2\sigma^{2*}_{-1,.}[1+(m-1)\rho^*_{-1,.}]}{m}.$$

And, utilizing equality of population covariance across regimens, we end with

$$N = \frac{4(z_\beta + z_{\alpha/2})^2}{m\delta^2}(1+(m-1)\rho^*)(1+\frac{1-p_1}{2})[1-\text{Cor}^2(Y,X)].$$

Also, with some algebra, ρ^* can be expressed as $\rho^* = (\rho - \text{Cor}^2(Y,X))/(1-\text{Cor}^2(Y,X))$, allowing our two covariate sample size formulae to be a function purely of ρ and $\text{Cor}^2(Y,X)$.

A.2 Data-generative Models Used in Simulation Experiments

Below we describe how we generated data for our simulations.

A.2.1 Without Covariates

For Table 3.3, we generate data, $(A_1, R, A_2, \mathbf{Y})$, using the values in Table A.1 and, for each of the N clusters, doing the following:

1. Generate A_1 to be 1 or -1 with equal probability

2. Generate R to be 1 with probability p_{A_1} and 0 otherwise

3. Generate A_2 to be 1 or -1 with equal probability, for clusters with $A_1 = 1$, $R = 0$

4. Generate the mx1 vector $\mathbf{Y} = \mu_{A_1,R,A_2}\mathbf{1}_m + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim$ MVN$(\mathbf{0}_m, \boldsymbol{\Sigma}_{A_1,R,A_2})$, where $\boldsymbol{\Sigma}_{A_1,R,A_2} = \sigma^2_{A_1,R,A_2} \cdot \mathbf{Exch}_m(\rho_{A_1,R,A_2})$. Here μ_{A_1,R,A_2}, $\sigma^2_{A_1,R,A_2}$, ρ_{A_1,R,A_2} are the cell means, variances, and ICCs since they correspond to each cell in Figure 3.1.

Simulation	p_1	$p_{\text{-}1}$	$\mu_{1,1,.}$	$\mu_{1,0,1}$	$\mu_{1,0,-1}$	$\mu_{-1,1,.}$	$\mu_{-1,0,.}$	$\sigma^2_{1,1,.}$	
Table 3.3, Row 1, Col 5	0.2	0.3	34.71	32.71	28	32.7	31	63.36	
Table 3.3, Row 1, Col 6	0.2	0.3	34.71	32.71	28	32.14	31.44	63.36	
Table 3.3, Row 1, Col 7	0.2	0.3	33.36	33.05	28	32.7	31	1	
Simulation	$\sigma^2_{1,0,1}$	$\sigma^2_{1,0,-1}$	$\sigma^2_{-1,1,.}$	$\sigma^2_{-1,0,.}$	$\rho_{1,1,.}$	$\rho_{1,0,1}$	$\rho_{1,0,-1}$	$\rho_{-1,1,.}$	$\rho_{-1,0,.}$
Table 3.3, Row 1, Col 5	63.36	60	63.39	63.39	0.0	0.0	0.0	0.0006	0.0006
Table 3.3, Row 1, Col 6	63.36	60	43	43	0.0	0.0	0.0	0.0076	0.0076
Table 3.3, Row 1, Col 7	79.73	60	63.39	63.39	0.9	0.007	0.0	0.0006	0.0006

Table A.1: Pre-specified simulation values for Table 3.3

Under the specified means, variances, and ICCs in Table A.1, one can easily obtain the desired marginal (over R) means, variances, and ICCs under a specific DTR using the laws of total expectation and variation. For example, to obtain the marginal mean under DTR (1,1), one would calculate $\mu_{1,1,.}p_1 + \mu_{1,0,1}(1 - p_1)$. To calculate the variance under DTR (1,1), one would calculate $\sigma^2_{1,1,.}p_1 + \sigma^2_{1,0,1}(1 - p_1) + p_1(1 - p_1)(\mu_{1,1,.} - \mu_{1,0,1})^2$. To calculate the covariance under DTR (1,1), one would calculate $\sigma^2_{1,1,.}\rho_{1,1,.}p_1 + \sigma^2_{1,0,1}\rho_{1,0,1}(1-p_1) + p_1(1-p_1)(\mu_{1,1,.} - \mu_{1,0,1})^2$.

When no assumptions were violated (row 1 of Table A.1), the cell means and variances were first chosen to give marginal means and variances which are both similar to results expected in ADEPT and produce effect sizes matching Table 3.3. After obtaining the correct effect size, the cell ICCs were then chosen also to match values specified in Table 3.3. To violate assumptions (row 2 and 3 of Table A.1), the cell means, variances, and ICCs from row 1 were altered to create the correct violations.

A.2.2 With a Cluster-Level Covariate

To generate data for Table 3.4, we use a continuous cluster-level covariate. We generate data, $(X, A_1, R, A_2, \mathbf{Y})$, using the values in Table A.2 and, for each of the N clusters, doing the following:

1. Generate A_1 to be 1 or -1 with equal probability
2. Generate R to be 1 with probability p_{A_1} and 0 otherwise
3. Generate A_2 to be 1 or -1 with equal probability, for clusters with $A_1 = 1, R = 0$
4. Generate a single cluster-level covariate X from Normal(0,1)
5. a. Generate mx1 vector $\mathbf{Y} = (\mu_{A_1,R,A_2} + \eta X)\mathbf{1}_m + \boldsymbol{\epsilon}$ for Column 6

   b. Generate mx1 vector $\mathbf{Y} = (\mu_{A_1,R,A_2} + \eta f_k(X))\mathbf{1}_m + \boldsymbol{\epsilon}$ for Columns 7,8

   where $\boldsymbol{\epsilon} \sim$ MVN($\mathbf{0}_m, \boldsymbol{\Sigma}_{A_1,R,A_2}$), with $\boldsymbol{\Sigma}_{A_1,R,A_2} = \sigma^{2*}_{A_1,R,A_2} \cdot \mathbf{Exch}_m(\rho^*_{A_1,R,A_2})$. Here μ_{A_1,R,A_2}, $\sigma^{2*}_{A_1,R,A_2}$, $\rho^*_{A_1,R,A_2}$ are the cell means, conditional cell variances, and conditional cell ICCs since they correspond to each cell in Figure 3.1. Also, f_k is the same piecewise function defined in the *Simulations* section (i.e., which is non-linear outside of $[-k, k]$).

Simulation	k	p_1	$p_{\text{-}1}$	η	$\mu_{1,1,.}$	$\mu_{1,0,1}$	$\mu_{1,0,-1}$	$\mu_{-1,1,.}$	$\mu_{-1,0,.}$	$\sigma^{2*}_{1,1,.}$
Table 3.4, Row 1, Col 6		0.2	0.3	4.47	34.94	32.94	28	32.7	31	63.36
Table 3.4, Row 1, Col 7	2	0.2	0.3	4.69	34.95	32.95	28	32.7	31	63.36
Table 3.4, Row 1, Col 8	1	0.2	0.3	6.66	34.98	32.98	28	32.7	31	63.36
Simulation	$\sigma^{2*}_{1,0,1}$	$\sigma^{2*}_{1,0,-1}$	$\sigma^{2*}_{-1,1,.}$	$\sigma^{2*}_{-1,0,.}$	$\rho^{*}_{1,1,.}$	$\rho^{*}_{1,0,1}$	$\rho^{*}_{1,0,-1}$	$\rho^{*}_{-1,1,.}$	$\rho^{*}_{-1,0,.}$	
Table 3.4, Row 1, Col 6	63.36	60	63.39	63.39	0.0	0.0	0.0	0.0006	0.0006	
Table 3.4, Row 1, Col 7	63.36	60	63.39	63.39	0.0	0.0	0.0	0.0006	0.0006	
Table 3.4, Row 1, Col 8	63.36	60	63.39	63.39	0.0	0.0	0.0	0.0006	0.0006	

Table A.2: Pre-specified simulation values for Table 3.4

Under the specified conditional means, variances, and ICCs in Table A.2, one can again obtain the desired conditional (on X only) and marginal means, variances, and ICCs under a specific DTR using the laws of total expectation and variation. For example, to obtain the conditional variance under DTR (1,1), one would calculate $\sigma^{2*}_{1,1} = \sigma^{2*}_{1,1,.}p_1 + \sigma^{2*}_{1,0,1}(1 - p_1) + p_1(1 - p_1)(\mu_{1,1,.} - \mu_{1,0,1})^2$. For data generated as in 5a, to obtain the marginal variance under DTR (1,1), one would calculate $\sigma^2_{1,1} = \sigma^{2*}_{1,1} + \eta^2\text{Var}(X)$. For data generated as in 5b, we instead calculate $\sigma^2_{1,1} = \sigma^{2*}_{1,1} + \eta^2\text{Var}(f_k(X))$. Both Var$(X)$ and Var$(f_k(X))$ can be found using the known distribution of X.

The cell means, variances, and ICCs were chosen for the same reason as in the non-covariate case. The parameter, η, was chosen to match the Cor(Y, X) values in Table 3.4.

A.3 Asymptotic results for the estimator

This section shows consistency and asymptotic normality of the proposed estimator. These proofs are similar to those found in Lu et al. (2016) In Equation 3.3 the estimator was presented with fixed working covariance matrices, $\mathbf{V}_{i,a_1,a_2} := \mathbf{V}(a_1, a_2, \mathbf{X}_i)$. However, in practice $\mathbf{V}_{i,a_1,a_2}$ must be estimated. We represent $\mathbf{V}_{i,a_1,a_2}$ by $\mathbf{V}_{i,a_1,a_2}(\hat{\alpha})$, where $\hat{\alpha}$ arises from estimation of $\mathbf{V}_{i,a_1,a_2}$ (e.g., $\mathbf{V}_{i,a_1,a_2}(\hat{\alpha}) = \hat{\sigma}^{2*}_{a_1,a_2}\mathbf{Exch}_{m_i}(\hat{\rho}^{*}_{a_1,a_2})$ in the *Implementation* section).

Additionally, the known weights, W_i, can be estimated to improve efficiency (Robins et al., 1995; Hernan et al., 2002; Hirano et al., 2003; Bembom and van der Laan, 2007;

Brumback, 2009; Williamson et al., 2014). We also may want to allow weights to depend on baseline covariates, $\mathbf{X}_i$, information collected prior to first randomization, L_{0i}, and information collected between the first and second randomization, L_{1i}. Specifically, we allow $W_i = 1/[f_{A_1|\mathbf{X},\mathrm{L}_0}(A_{1i}|\mathbf{X}_i, \mathrm{L}_{0i})f_{A_2|\mathbf{X},\mathrm{L}_0,A_1,\mathrm{L}_1,R}(A_{2i}|\mathbf{X}_i, \mathrm{L}_{0i}, A_{1i}, \mathrm{L}_{1i}, R_i)]$, where $f_{A_1|\mathbf{X},\mathrm{L}_0}$ and $f_{A_2|\mathbf{X},\mathrm{L}_0,A_1,\mathrm{L}_1,R}$ are conditional probability mass functions for A_1 and A_2, respectively. We represent W_i by $W_i(\hat{\gamma})$, where $\hat{\gamma}$ arises from estimation and W_i.

We also allow for cluster sizes to be unequal across observations since this is typical in practice. Under these general settings, the estimating equation is

$$\begin{aligned} \mathbf{0} = \frac{1}{N}\sum_{i=1}^{N}\sum_{(a_1,a_2)} & I_{i,a_1,a_2}W_i(\hat{\gamma}) \\ & \cdot \mathbf{D}(\mathbf{X}_i, a_1, a_2)^T\mathbf{V}_{i,a_1,a_2}(\hat{\alpha})^{\text{-1}}(\mathbf{Y}_i - \boldsymbol{\mu}(\mathbf{X}_i, a_1, a_2; \boldsymbol{\beta}, \boldsymbol{\eta})). \end{aligned} \tag{A.5}$$

We first demonstrate the consistency of the estimator found by solving this equation.

Theorem 1.1. *Assume the marginal model is correctly specified, that is,* $E_{a_1,a_2}[Y_{ij}|\boldsymbol{X}_{ij}] = \mu(\boldsymbol{X}_{ij}, a_1, a_2; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0)$*, where* $(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0)$ *is the true value for the parameter* $(\boldsymbol{\beta}, \boldsymbol{\eta})$ *in the marginal mean model. Assume* Y_{ij} *is conditionally independent of* $\boldsymbol{X}_{ik}$ $\forall k \neq j$ *given* $\boldsymbol{X}_{ij}$*. Also assume that there exists* α^+, γ_0 *such that* $\sqrt{N}(\hat{\alpha} - \alpha^+) = O_p(1)$ *and* $\sqrt{N}(\hat{\gamma} - \gamma_0) = O_p(1)$ *(i.e., are bounded in probability), where* $W_i(\gamma_0) \equiv W_i$*, the true inverse-probability weight. Then the estimator* $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\eta}})$ *obtained by solving Equation* A.5 *is consistent for* $(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0)$.

Proof. Define $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\eta})$ to denote the marginal mean model parameters with true values $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0, \boldsymbol{\eta}_0)$. We denote the estimating equation in A.5 as

$\mathbf{0} = 1/N\sum_{i=1}^{N}\mathbf{U}_i(\mathbf{Z}_i; \boldsymbol{\theta}, \hat{\alpha}, \hat{\gamma})$, where $\mathbf{Z}_i$ is all observed covariates and responses for cluster i. It remains to show that $E[\mathbf{U}_i(\mathbf{Z}_i; \boldsymbol{\theta}_0, \alpha^+, \gamma_0)] = \mathbf{0}_{p+q}$, from which consistency can be established as done for the standard GEE estimator (Liang and Zeger, 1986). Here the expectation is over observed data (with respect to the distribution of the observed data, P_{obs}) as opposed to E_{a_1,a_2}, which is an expectation over data arising as if all clusters had received DTR (a_1, a_2) (with respect to the distribution P_{a_1,a_2}).

Note that $I_{i,a_1,a_2}/W_i$ is the Radon-Nikodym derivative between P_{obs} and P_{a_1,a_2}. And thus,

$$
\begin{aligned}
&E[\mathbf{U}_i(\mathbf{Z}_i; \boldsymbol{\theta}_0, \alpha^+, \gamma_0)] = \\
&\sum_{(a_1,a_2)} E_{a_1,a_2}[\mathbf{D}(\mathbf{X}_i, a_1, a_2)^T \mathbf{V}_{i,a_1,a_2}(\alpha^+)^{\text{-}1}(\mathbf{Y}_i - \boldsymbol{\mu}(\mathbf{X}_i, a_1, a_2; \boldsymbol{\theta}_0))] \\
&\sum_{(a_1,a_2)} E_{\mathbf{X}_i}[\mathbf{D}(\mathbf{X}_i, a_1, a_2)^T \mathbf{V}_{i,a_1,a_2}(\alpha^+)^{\text{-}1}] E_{a_1,a_2}[\mathbf{Y}_i - \boldsymbol{\mu}(\mathbf{X}_i, a_1, a_2; \boldsymbol{\theta}_0)|\mathbf{X}_i] \\
&= \mathbf{0}_{p+q}.
\end{aligned}
$$

The final equation equals zero due to the conditional independence and correct specification of the marginal mean model.

We next prove the asymptotic normality of our estimator obtained in Equation A.5. We borrow notation from the previous proof.

Theorem 1.2. *Assuming mild regularity conditions, the same assumptions as in Theorem 1.1, the cluster sizes are bounded, and that the weight parameter γ is obtained from maximum likelihood estimation for treatment assignment probabilities, with a score function $\boldsymbol{S}_\gamma$. Then $\sqrt{N}((\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\eta}}) - (\boldsymbol{\beta}_0, \boldsymbol{\eta}_0))$ is asymptotically multivariate normal with zero mean and covariance matrix $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\eta}}} = \boldsymbol{J}^{\text{-}1}(\boldsymbol{K} - \boldsymbol{C}\boldsymbol{B}^{\text{-}1}\boldsymbol{C}^T)\boldsymbol{J}^{\text{-}1}$, where $\boldsymbol{K}$, $\boldsymbol{B}$, $\boldsymbol{C}$, and $\boldsymbol{J}$ are given by*

$$
\begin{aligned}
\boldsymbol{J} = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} E \sum_{(a_1,a_2)} I_{i,a_1,a_2} W_i(\gamma_0) \\
\cdot \boldsymbol{D}(\boldsymbol{X}_i, a_1, a_2)^T \boldsymbol{V}_{i,a_1,a_2}(\alpha^+)^{\text{-}1} \boldsymbol{D}(\boldsymbol{X}_i, a_1, a_2),
\end{aligned}
$$

$$
\boldsymbol{K} = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} E[\boldsymbol{U}_i \boldsymbol{U}_i^T], \boldsymbol{B} = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} E[\boldsymbol{S}_{\gamma_0,i} \boldsymbol{S}_{\gamma_0,i}^T], \boldsymbol{C} = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} E[\boldsymbol{U}_i \boldsymbol{S}_{\gamma_0,i}^T],
$$

with $\boldsymbol{U}_i \triangleq \boldsymbol{U}_i(\boldsymbol{Z}_i; \boldsymbol{\theta}_0, \alpha^+, \gamma_0)$.

Proof: Using the same argument for GEE estimators (Liang and Zeger, 1986) we obtain

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \left[\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \frac{\partial \mathbf{U}_i(\mathbf{Z}_i; \boldsymbol{\theta}_0, \alpha^+, \gamma_0)}{\partial \boldsymbol{\theta}}\right]^{\text{-1}} \left\{\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \mathbf{U}_i(\mathbf{Z}_i; \boldsymbol{\theta}_0, \alpha^+, \gamma_0)\right.$$
$$\left. + \left[\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \frac{\partial \mathbf{U}_i(\mathbf{Z}_i; \boldsymbol{\theta}_0, \alpha^+, \gamma_0)}{\partial \gamma}\right] \sqrt{N}(\hat{\gamma} - \gamma_0)\right\} + o_p(\mathbf{1}).$$

Using the fact that $\mathbf{S}_\gamma$ is the score function for $\hat{\gamma}$ to express $\sqrt{N}(\hat{\gamma} - \gamma_0)$ as a sum. Also, using the fact that our cluster size is bounded combined with the Law of Large Numbers, we write all long-run averages of random variables as long run averages of expectations. Hence, we obtain

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \left[\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} E\left\{\frac{\partial \mathbf{U}_i}{\partial \boldsymbol{\theta}}\right\}\right]^{\text{-1}} \left\{\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \mathbf{U}_i - \right.$$
$$\left. \left[\lim_{N\to\infty} \frac{1}{N}\sum_{j=1}^{N} E(\mathbf{U}_j \mathbf{S}_{\gamma_0,j}^T)\right] \left[\lim_{N\to\infty} \frac{1}{N}\sum_{j=1}^{N} E(\mathbf{S}_{\gamma_0,j} \mathbf{S}_{\gamma_0,j}^T)\right]^{\text{-1}} \mathbf{S}_{\gamma_0,i}\right\} + o_p(\mathbf{1})$$
$$= \mathbf{J}^{\text{-1}} \frac{1}{\sqrt{N}}\sum_{i=1}^{N} (\mathbf{U}_i - \mathbf{C}\mathbf{B}^{\text{-1}}\mathbf{S}_{\gamma_0,i}) + o_p(\mathbf{1}) \to Normal[\mathbf{0}, \mathbf{J}^{\text{-1}}(\mathbf{K} - \mathbf{C}\mathbf{B}^{\text{-1}}\mathbf{C}^T)\mathbf{J}^{\text{-1}}].$$

Remark: Note that with unequal cluster sizes, $\mathbf{U}_i$ and $\mathbf{S}_{\gamma,i}$ are not identically distributed, and hence we must express our variances with long run averages. If cluster sizes were equal, averaging would not be necessary and we would obtain $\mathbf{K} = E[\mathbf{U}_i \mathbf{U}_i^T]$, $\mathbf{B} = E[\mathbf{S}_{\gamma_0} \mathbf{S}_{\gamma_0}^T]$, $\mathbf{C} = E[\mathbf{U}_i \mathbf{S}_{\gamma_0}^T]$, and $\mathbf{J} = E[\partial \mathbf{U}_i / \partial \boldsymbol{\theta}]$.

To obtain estimates for our standard error of $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\eta}})$ we use plug in estimates of **K**, **B**,

$\mathbf{C}$, and $\mathbf{J}$. Specifically, we set

$$\hat{\boldsymbol{J}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{(a_1,a_2)} I_{i,a_1,a_2} W_i(\hat{\gamma})$$
$$\cdot\, \mathbf{D}(\mathbf{X}_i, a_1, a_2)^T \mathbf{V}_{i,a_1,a_2}(\hat{\alpha})^{\text{-}1}\mathbf{D}(\mathbf{X}_i, a_1, a_2),$$

$$\hat{\boldsymbol{K}} = \frac{1}{N}\sum_{i=1}^{N}\hat{\mathbf{U}}_i\hat{\mathbf{U}}_i^T, \hat{\mathbf{B}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{S}_{\hat{\gamma},i}\mathbf{S}_{\hat{\gamma},i}^T, \hat{\mathbf{C}} = \frac{1}{N}\sum_{i=1}^{N}\hat{\mathbf{U}}_i\mathbf{S}_{\hat{\gamma},i}^T,$$

where $\hat{\boldsymbol{U}}_i = \boldsymbol{U}_i(\mathbf{Z}_i; \hat{\boldsymbol{\theta}}, \hat{\alpha}, \hat{\gamma})$.

Thus, the plug in estimator for $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}}$ is $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\theta}}} = \hat{\boldsymbol{J}}^{\text{-}1}(\hat{\boldsymbol{K}} - \hat{\mathbf{C}}\hat{\mathbf{B}}^{\text{-}1}\hat{\mathbf{C}}^T)\hat{\boldsymbol{J}}^{\text{-}1}$ and we obtain $\widehat{Var}(\hat{\boldsymbol{\theta}}) = 1/N \cdot \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\theta}}}$.

BIBLIOGRAPHY

Bembom, O. and van der Laan, M. (2007). Statistical methods for analyzing sequentially randomized trials. *Journal of the National Cancer Institute 99*(21), 1577–82.

Brumback, B. A. (2009). A note on using the estimated versus the known propensity score to estimate the average treatment effect. *Statistics & Probability Letters 79*(4), 537–542.

Hedges, L. and Rhoads, C. (2009). *Statistical power analysis in education research*. National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education. NCSER 2010-3006.

Hernan, M., Brumback, B., and Robins, J. (2002). Estimating the causal effect of zidovudine on CD4 count with a marginal structural model for repeated measures. *Statistics in Medicine 21*, 1689–1709.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica 71*(4), 1161–1189.

Kerry, S. M. and Bland, M. J. (2001). Unequal cluster sizes for trials in English and Welsh general practice: Implications for sample size calculations. *Statistics in Medicine 20*(3), 377–390.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika 73*(1), 13–22.

Lu, X., Nahum-Shani, I., Kasari, C., Lynch, K. G., Oslin, D. W., Pelham, W. E., Fabiano, G., and Almirall, D. (2016). Comparing dynamic treatment regimes using repeated-measures outcomes: modeling considerations in SMART studies. *Statistics in Medicine 35*(10), 1595–1615. sim.6819.

Neyman, J., Iwaszkiewicz, K., and Kolodziejczyk, S. (1935). Statistical problems in agricultural experimentation. *Supplement to the Journal of the Royal Statistical Society* (107-80).

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

Robins, J. (1997). Causal inference from complex longitudinal data. In *Latent Variable Modeling and Applications to Causality.*, Lecture Notes in Statistics. New York: Springer-Verlag.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association 90*(429), 106–121.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics 6*, 34–58.

Williamson, E. J., Forbes, A., and White, I. R. (2014). Variance reduction in randomised trials by inverse probability weighting using the propensity score. *Statistics in Medicine 33*(5), 721–737.

APPENDIX B

Appendix of Chapter IV

B.1 Additional Analyses

In this appendix we present additional analyses conducted for the 2018 Intern Health Study MRT.

B.1.1 Main effects of notification categories on weekly outcomes

The first additional analyses conducted were the main effects analyses of all the moderator analyses presented in the main paper. These analyses look at the non-moderated effects of different categories of notifications on various outcomes. The analysis methods are the exact same, except the model no longer contains an interaction between the treatment and moderator (eliminating $\beta_1 Z_t M_t$). The outcome variables are still aggregated at the weekly-level. These analyses answer the following questions:

1. What is the effect of notifications (of any category) on average daily mood compared to no notifications?
2. What is the effect of mood notifications on average daily mood compared to no notifications?

		Outcome		
		Mood	**Step**	**Sleep**
Notification Category	General	-0.029 (P = .003) d = -0.020		
	Mood	-0.023 (p = .153) d = -0.016		
	Activity		0.693 (P = .023) d = 0.044	
	Sleep			0.051 (P = .073) d = 0.036

Table B.1: Effects (p-values) and effect sizes, d, of various notification categories on different outcomes. Effects are compared to a baseline of no notification.

3. What is the effect of activity notifications on average daily step count compared to no notification?

4. What is the effect of sleep notifications on average daily sleep compared to no notifications?

The results are presented in Table B.1. There is strong evidence of a negative effect of notifications on mood. There is weak evidence of a negative effect of mood notifications on mood. There is strong evidence of a positive effect of activity notifications on step counts. Lastly, there is moderate evidence of a positive effect of sleep notifications on sleep. The effect sizes for all of these effects are small.

B.1.2 Comparing Life Insights and Tips

In addition to comparing notification categories, we were also interested in comparing notification types (life insights or tips). In the 2018 IHS MRT, life-insights and tips were not randomly assigned, but were instead alternated deterministically (see Figure 4.2). However, the decision between sending a notification and not sending a notification on a given day was randomly assigned (with 50% probability). Hence, for this analysis, the efficacy of different notification types is evaluated by comparing life-insight notification days to no-notification days and comparing tip notification days to no-notification days. Since these

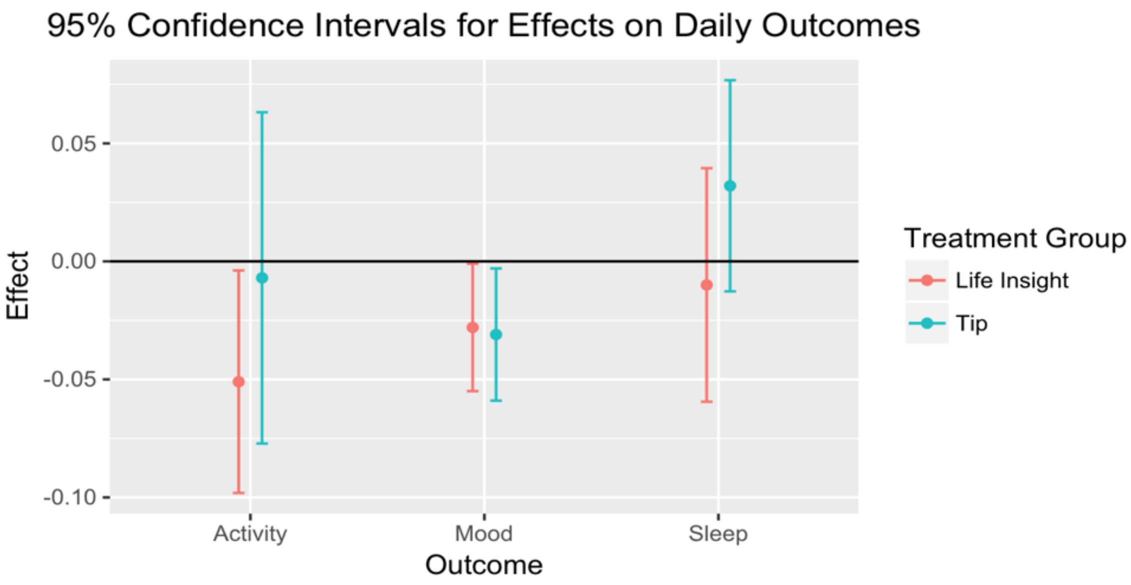


Figure B.1: 95% confidence intervals and point estimates of the effects of life insights (compared to no notification) and tips (compared to no notification) on daily step count, mood, and sleep.

randomizations were done at the daily level, the outcomes of interest are also at the daily level. The outcome is daily step count, daily mood, or nightly sleep duration on the day a particular notification type was sent. Again, this analysis uses a weighted and centered least squares estimator. Figure B.1 presents 95% confidence intervals and point estimates of these effects.

Figure B.1 demonstrates there is moderate evidence that tips perform slightly better than life insights for daily steps and sleep. There is moderate evidence of a positive effect of tips on daily sleep. There is strong evidence of a negative effect of life insights on daily step count. For daily mood, the effects of both life insights and tips are negative, and there does not appear to be a difference between the two types.

The deterministic alternating between life insights and tips does make causal interpretation of these effects difficult. Nonetheless, the analyses are informative and provide some evidence of the different effects of notification types.

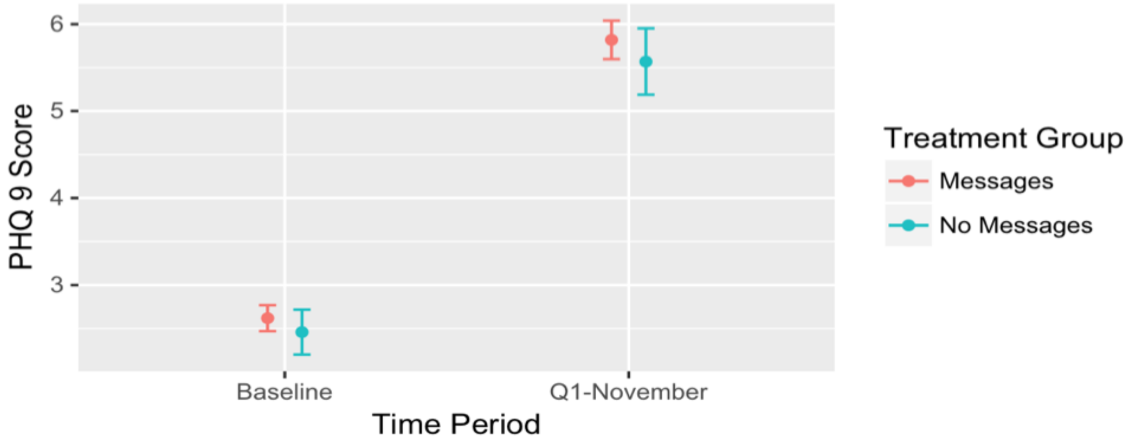


Figure B.2: 95% confidence intervals and point estimates of the average PHQ-9 score at baseline and 4 months into internship (quarter 1). Lower score corresponds to a lower frequency of depressive symptoms.

B.1.3 Long-term Effects of Notifications on Mental Health

We were also interested in the long-term effects of notifications on intern mental health. To assess the long-term effects, we included an additional baseline randomization prior to the start of the internship. For this randomization, 25% of interns are randomized to receive no notifications for the entire internship, while the other 75% would enter the MRT and receive notifications under the scheme shown in Figure 4.2. To assess the long-term mental health of interns, we use the PHQ-9 (Kroenke et al., 2001). The PHQ-9 score of each intern is measured at baseline (prior to internship) and in November (4 months into the internship). In the 2018 IHS, 546 interns were randomized to not receive notifications for the entire internship, while 1,565 interns were randomized to receive notifications during the internship. Point estimates and 95% confidence intervals of the average PHQ-9 score of each group are shown in Figure B.2. For PHQ-9, a lower score corresponds to a lower frequency of depressive symptoms.

Figure B.2 demonstrates there is no evidence of a positive effect (i.e., lower PHQ-9) of notifications on average PHQ-9 score. In fact, the average PHQ-9 score for the notification group is slightly larger than for the no notification group. This difference is not statistically significant.

B.2 Missing Data and Sensitivity Analyses

Missing data occurred throughout the intern health study. The primary aim's outcome (mood) was missing because interns failed to self-report. The other two secondary outcomes (sleep and step count) were missing as collection required interns to wear their Fitbits. Figure B.3 displays the percentage of interns with at least one non-missing sleep, step, or mood observation for each week in the study. There was a downward trend in percentage of users with non-missing data. It is known that attrition over time is a major issue in mobile health studies (Eysenbach, 2005). In this appendix, we explore the sensitivity of the main results in the paper to missingness.

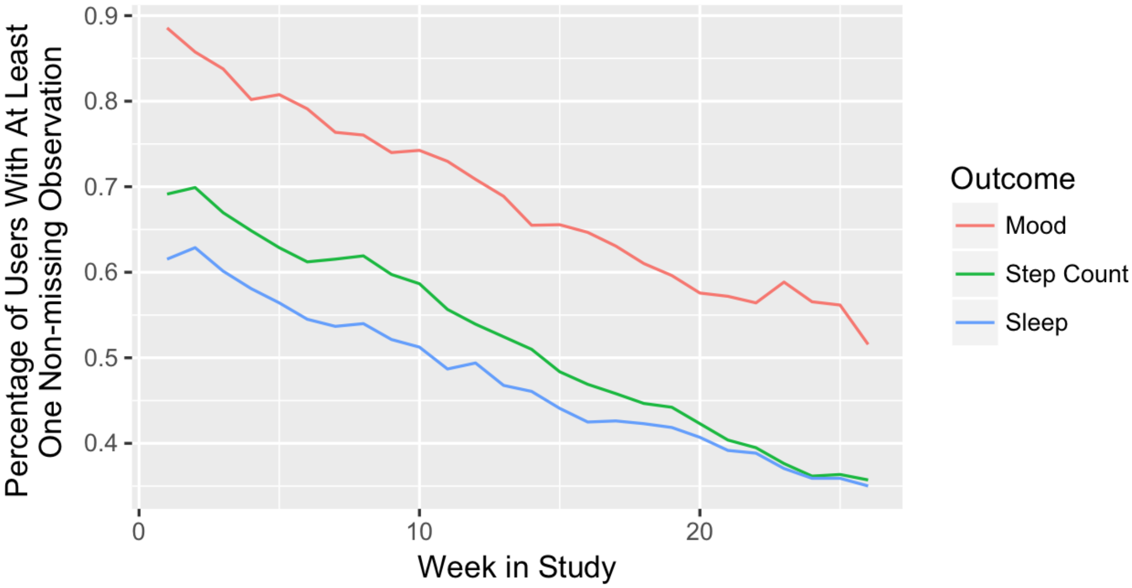


Figure B.3: Percentage of interns with at least one non-missing sleep, step, or mood observation for each week in the study

We provided sensitivity analyses for two different types of missingness in the outcome of interest: dropout and weekly missingness. For the dropout sensitivity analyses, we eliminated imputed data from users who dropped out from the study early. That is, if a user stopped entering mood scores after November 1st, then for the mood outcome analyses, we eliminated that user's imputed data from November 1st onward. For weekly missingness sensitivity analyses, we eliminated weeks with a large percentage of missing data in the outcome of interest.

B.2.1 Sensitivity of the Primary Aim Results

We evaluated the sensitivity of the estimate of the moderator, -0.052 (SE = 0.014, P = .001), of previous week's mood on the effect of notifications (of any category) on average daily mood.

B.2.1.1 Dropout Sensitivity

We eliminated all imputed data for users after they have dropped out. For example, if a user has stopped entering mood scores after November 1st, we removed all data for that user after November 1st from the analysis.

The new estimate of the moderator is -0.039 (SE = 0.014, P = .006).

B.2.1.2 Weekly Missingness Sensitivity

In our analysis, we eliminated all weeks where more than 5 daily mood scores are missing.

The new estimate of the moderator is -0.024 (SE = 0.013, P = .076).

B.2.1.3 Conclusions

The primary aim conclusions are mildly sensitive to missingness. The size of estimated moderation for the primary aim was reduced when eliminating dropouts or weeks with a large amount of missingness. The sign of the moderation remained negative, matching the conclusions made in the paper.

B.2.2 Sensitivity of Secondary Aim 1 Results

We evaluated the sensitivity of the estimate of the moderation, -0.038 (SE = 0.015, P = .013), of previous week's step count on the effect of activity notifications on average daily step count.

B.2.2.1 Dropout Sensitivity

We eliminated all imputed data for users after they have dropped out. For example, if a user has no step count data after November 1st we removed all data for that user after November 1st from the analysis.

The new estimate of the moderator is -0.003 (SE = 0.020, P = .874)

B.2.2.2 Weekly Missingness Sensitivity

In our analysis, we eliminated all weeks where more than 5 daily step counts are missing.

The new estimate of the moderator is 0.004 (SE = 0.021, P = .858)

B.2.2.3 Conclusions

The secondary aim 1 conclusions are very sensitive to missingness. For both dropout and weekly missingness, the moderation effect is now very close to 0.

B.2.3 Sensitivity of Secondary Aim 2 Results

We evaluated the sensitivity of the estimate of the moderation, -0.074 (SE = 0.018, P = .001), of previous week's sleep time on the effect of sleep notifications on average daily sleep.

B.2.3.1 Dropout Sensitivity

We eliminated all imputed data for users after they have dropped out. For example, if a user has no sleep data after November 1st, we removed all data for that user after November 1st from the analysis.

The new estimate of the moderator is -0.034 (SE = 0.025, P = .173)

B.2.3.2 Weekly Missingness Sensitivity

In our analysis, we eliminated all weeks where more than 5 daily sleep times are missing.

The new estimate of the moderator is -0.044 (SE = 0.022, P = .044)

B.2.3.3 Conclusions

The secondary aim 2 conclusions are mildly sensitive to missingness. The size of moderation for secondary aim 2 was reduced when eliminating dropouts or weeks with a large amount of missingness. The sign the moderation remained negative, matching the conclusions made in the paper.

B.2.4 Overall conclusions

Overall, this analysis has demonstrated some sensitivity of the conclusions to missingness in the data. The conclusions of the primary aim and secondary aim 2 seem to be robust to missingness. The conclusions for secondary aim 1, however, are very sensitive.

The reduction in effect size after dropping imputed data from the analysis could indicate a few things. In the worst case, it could indicate that the large effect size is an artifact of the imputation model itself. That is, the methods used to overcome the missing data are biasing the estimates away from 0. On the other hand, the reduction in effect size could indicate that the effect is strongest for interns with a large amount of missingness. In this case, dropping the imputed data would bias the estimates towards 0. One of the challenges of dealing with missing data is not knowing the truth because the data needed to distinguish these two scenarios is missing.

B.3 Further Details on the Statistical Methods

In this appendix, we provide further details on the statistical model, methodology, and implementation used in the main paper. In this appendix, **boldface** is used to indicate multi-dimensional column vectors. $\mathbf{X}'$ indicates vector transpose.

B.3.1 Statistical Model

A linear model was used as a working model for the moderator analysis. The model is a 'working' model, as indicated by "=", because the estimation methods do not require correct specification of parts of the model not interacted with treatment, such as $\boldsymbol{\alpha_0'}\mathbf{X_t} + \alpha_1 M_t$ below.

In all aims $\mathbf{X_t}$ is an 11-dimensional vector of control covariates. The control covariates used in all analyses are baseline sex, baseline PHQ-9 score, baseline depression history, baseline neuroticism, pre-internship average daily mood, pre-internship average daily square root step count, pre-internship average daily square root sleep minutes, previous week's average daily mood, previous week's average daily square root step count, previous week's average daily square root sleep minutes, and study week. $\boldsymbol{\alpha_0}$ is the corresponding 11-dimensional vector of coefficients for the 11 control covariates. M_t is the 1-dimensional moderator of interest. Y_t is the outcome of interest. Z_t is the treatment indicator. In the primary aim, Z_t is 1-dimensional, with $Z_t = 1$ indicating a notification week of any category and $Z_t = 0$ indicating a no notification week. In the secondary aims and exploratory sub-aim, the treatment is no longer binary since there are 4 possible notification categories. Z_t is now a 3-dimensional vector which encodes 3 indicator variables: activity notification weeks ($\mathbf{Z_t} = (1, 0, 0)'$), sleep notification weeks ($\mathbf{Z_t} = (0, 1, 0)'$), mood notification weeks ($\mathbf{Z_t} = (0, 0, 1)'$), or no-notification weeks ($\mathbf{Z_t} = (0, 0, 0)'$). Below are the working models used in analyses.

B.3.1.1 Primary Aim

$E(Y_t|\mathbf{X_t}, M_t, Z_t)$ "=" $\boldsymbol{\alpha_0'}\mathbf{X_t} + \alpha_1 M_t + \beta_0 Z_t + \beta_1 Z_t M_t =$

$\alpha_{01}X_{t1} + \alpha_{02}X_{t2} + \cdots + \alpha_{0,11}X_{t11} + \alpha_1 M_t + \beta_0 Z_t + \beta_1 Z_t M_t$

In the primary aim, Y_t is average daily mood and M_t is average daily mood of the previous week.

B.3.1.2 Secondary Aims and Exploratory Sub-aim

$E(Y_t|\mathbf{X_t}, M_t, \mathbf{Z_t})$ "=" $\boldsymbol{\alpha_0'}\mathbf{X_t} + \alpha_1 M_t + \boldsymbol{\beta_0'}\mathbf{Z_t} + \boldsymbol{\beta_1'}\mathbf{Z_t} M_t =$

$\boldsymbol{\alpha_0'}\mathbf{X_t} + \alpha_1 M_t + \beta_{01}Z_{t1} + \beta_{02}Z_{t2} + \beta_{03}Z_{t3} + \beta_{11}Z_{t1}M_t + \beta_{12}Z_{t2}M_t + \beta_{13}Z_{t3}M_t$

In secondary aim 1, Y_t is average daily square root step count and M_t is average daily square root step count of the previous week. In secondary aim 2, Y_t is average daily square root sleep count and M_t is average daily square root sleep count of the previous week. In the exploratory aim, Y_t is average daily mood and M_t is average daily mood of the previous week.

B.3.2 Methodology

To estimate the coefficients of interest, we used the weighted and centered least squares estimator outlined in Boruvka et al. (2018). The method is robust to misspecification of parts of the model not interacted with treatment.

The methods developed in Boruvka et al. (2018) are useful for robust estimation when treatment assignment probabilities are time-varying (for example, an MRT where the probability of treatment assignment is based on data collected throughout the trial). In the IHS MRT, the treatment assignment probabilities were constant across weeks. Because of this, in the estimating equation, all weights were equal to 1 and the centering term, ρ, was constant ($\rho = 0.75$ in primary aim, $\boldsymbol{\rho} = (0.25, 0.25, 0.25)'$ in secondary/exploratory aims).

The method also uses the standard sandwich estimator for robust standard error estimation (Huber et al., 1967). As mentioned in Boruvka et al. (2018), an independent working correlation matrix was used to prevent biased estimation of coefficients.

The estimating equation approach with robust standard error estimation is advantageous because it does not require distributional assumptions on the continuous outcomes. The approach also permits arbitrary dependencies between observations in the data, as was expected with the repeatedly measured outcomes.

B.3.3 Implementation

The method was implemented in R using the package geepack (Halekoh et al., 2006). The method was implemented using the standard geeglm function with a centered treatment indicator. That is, for the primary aim Z_t was transformed to $Z_t - \rho$, and for the secondary aims and exploratory aims $\mathbf{Z_t}$ was transformed to $\mathbf{Z_t} - \boldsymbol{\rho}$.

Since multiple imputation was used to deal with missingness, the coefficients and standard errors were estimated for each imputed data set. The coefficients and standard errors were combined across multiple imputations using Rubin's rules. The testestimates function in the mitml R package (Grund et al., 2016) was used to combine estimates.

Code will be made available on the first author's website.

BIBLIOGRAPHY

Boruvka, A., Almirall, D., Witkiewitz, K., and Murphy, S. A. (2018). Assessing time-varying causal effect moderation in mobile health. *Journal of the American Statistical Association 113*(523), 1112–1121.

Eysenbach, G. (2005). The law of attrition. *Journal of Medical Internet Research 7*(1), e11.

Grund, S., Robitzsch, A., and Lüdtke, O. (2016). mitml: Tools for multiple imputation in multilevel modeling. *Retreived from: https://cran.r-project.org/web/packages/mitml/index.html*.

Halekoh, U., Højsgaard, S., Yan, J., et al. (2006). The R package geepack for generalized estimating equations. *Journal of Statistical Software 15*(2), 1–11.

Huber, P. J. et al. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1, pp. 221–233. University of California Press.

Kroenke, K., Spitzer, R. L., and Williams, J. B. (2001). The PHQ-9: validity of a brief depression severity measure. *Journal of General Internal Medicine 16*(9), 606–613.