Does It Work for Me?

Supporting Self-Experimentation of Simple Health Behavior Interventions

by

Sayali Shekhar Phatak

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved April 2019 by the
Graduate Supervisory Committee:

Matthew Buman, Co-Chair
Eric Hekler, Co-Chair
Jennifer Huberty
Erik Johnston
Pamela Swan

ARIZONA STATE UNIVERSITY

May 2019

ABSTRACT

Many individual-level behavioral interventions improve health and well-being. However, most interventions exhibit considerable heterogeneity in response. Put differently, what might be effective on average might not be effective for specific individuals. From an individual's perspective, many healthy behaviors exist that seem to have a positive impact. However, few existing tools support people in identifying interventions that work for them, personally.

One approach to support such personalization is via self-experimentation using single-case designs. 'Hack Your Health' is a tool that guides individuals through an 18-day self-experiment to test if an intervention they choose (e.g., meditation, gratitude journaling) improves their own psychological well-being (e.g., stress, happiness), whether it fits in their routine, and whether they enjoy it.

The purpose of this work was to conduct a formative evaluation of Hack Your Health to examine user burden, adherence, and to evaluate its usefulness in supporting decision-making about a health intervention. A mixed-methods approach was used, and two versions of the tool were tested via two waves of participants (Wave 1, N=20; Wave 2, N=8). Participants completed their self-experiments and provided feedback via follow-up surveys (n=26) and interviews (n=20).

Findings indicated that the tool had high usability and low burden overall. Average survey completion rate was 91%, and compliance to protocol was 72%. Overall, participants found the experience useful to test if their chosen intervention helped them. However, there were discrepancies between participants' intuition about intervention effect and results from analyses. Participants often relied on intuition/lived experience over results for decision-making. This suggested that the usefulness of Hack Your Health in its current form might be through the structure, accountability, and means for self-

reflection it provided rather than the specific experimental design/results. Additionally, situations where performing interventions within a rigorous/restrictive experimental set-up may not be appropriate (e.g., when goal is to assess intervention enjoyment) were uncovered. Plausible design implications include: longer experimental and phase durations, accounting for non-compliance, missingness, and proximal/acute effects, and exploring strategies to complement quantitative data with participants' lived experiences with interventions to effectively support decision-making. Future work should explore ways to balance scientific rigor with participants' needs for such decision-making.

To my mother, Nirmala, for her endless support and encouragement, and for always believing in me. To Arvind mama, without whom I would have not had the opportunity to come to the United States for my graduate studies.

ACKNOWLEDGMENTS

My acknowledgements and thanks are not limited only to those who were involved directly in the research that I've included in this document. Over the last five years, I've had the opportunity to be involved in a variety of projects that have all helped shaped my thinking and my work. I want to thank all the colleagues, collaborators, and friends who have been with me through this journey.

Firstly, and most importantly, I want to thank my amazing mentor, Eric Hekler, who has tirelessly supported and advocated for me over the years. Thank you for believing in me, opening up a myriad of doorways, and always giving me the freedom to choose the path that I want. Thank you for pulling me up when I was sinking into moments of low confidence. Thank you for your invaluable advice, and all those meta, philosophical conversations that nudged me to think about the big picture. Thank you for inspiring me to not only be a better scientist, but a better person.

Matt Buman, for his support over the years. There was a day when you said to me "Remember, you don't have to answer all questions through one study", and while that may seem like a "duh" statement to some researchers, it wasn't for me. It stuck with me and has helped me get over my over-thinking and perfectionist tendencies. Your feedback and guidance have always been valuable. A special thanks to you for being incredibly supportive since Eric's transition. Marc Adams, whose research methods class I took back in 2014 piqued my interest in experimental design and single-case designs. Erik Johnston, Jennifer Huberty and Pamela Swan for their feedback on this project and serving on my dissertation committee over the years.

I would also like to acknowledge the wonderful collaborators of this project: Elaine Chen, Stephen Schueller, Richard Kravitz, Christopher Schmid, Ida Sim, and

Mike Seo. This would not have been possible without the advice, expertise and resources they all have continually volunteered for this project over the last two years.

I would also like to thank colleagues and collaborators from other projects that have inspired me: Dana Lewis, Daniel Rivera, John Harlow and Erik Johnston. Thank you for helping me see the world through different lenses.

I would also like to recognize and thank the various agencies and grants that have funded me over the years: The National Science Foundation, the Robert Wood Johnson Foundation, the Health Data Exploration Network, and the Precision Healthcare Initiative.

A big, big thank you to all my friends for keeping me sane through this journey. My wonderful friends here in San Diego who helped me have a life outside work and made this past year infinitely easier and enjoyable: Mason, Marcela, and James. My friends in Arizona: Jane, Mayra, Andrew, Abe, Chirag, and JD for all the hikes, camping trips, Rick & Morty binge sessions and hours and hours of the most random conversations. A special thank you to Jane for being a dear friend, and for listening to my rants about academia, and to Marcela for being my person, for our art parties, for our dance sessions, and for being patient with me in times of stress. My family: Dee, aai, and baba for always lifting my spirits and their unconditional love and encouragement.

Lastly, a huge shout out to all the artists whose music helped me get through the hours and hours of writing and slogging in front of a computer screen.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

Table Page

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background

Many physiological, cognitive, and behavioral interventions exist to improve one's health and well-being. There is considerable heterogeneity in how different individuals respond to the different interventions, i.e., one intervention does not work for everyone in the same manner[1–8]. Put differently, what might be an 'effective' intervention on average might not be effective for specific individuals[9]. Most of the evidence for popular interventions is from group-based studies. While these studies are important for understanding the effect of the treatment overall, they can rarely be used to predict if the intervention would work for a specific individual[10,11]. Unfortunately, while research and media often describe the benefits of different behavioral interventions that work on average, few highlight the variability in response that was observed, or put differently, how many individuals the intervention did not work for.

Behavior change is an adaptive problem that is complex and multifaceted, and whether or not an intervention will be effective for a given individual depends on attributes of the person, the context in which it is carried out, and time[12,13]. Previous work has indicated that there are many person-level and contextual factors that can influence whether or not a given intervention will work for a given person, such as personality[3,4,7,14], personal preference[15], perceived enjoyment[15], social context, time constraints, location, weather, etc. [7,15,16] Further, intervention characteristics such as medium of delivery[17] (e.g., in-person vs. telephonic, using an app vs. in-person), characteristics of person delivering the intervention[18], and intervention content and dosage could also impact whether or not an intervention works for a given person[19–21]. In addition, our previous work exploring individual differences in a physical activity intervention suggested that participants differed in the personal and contextual factors that affected their daily walking behavior (e.g., a few were affected by perceived stress, and several were affected by whether it was a weekend or a weekday)[22]. Moreover, many of these factors have time-varying characteristics, meaning, they change over time[23–25]. Knowledge of these complex factors that influence intervention efficacy and behavior further highlight limitations of applying group-level insights to make decisions at the individual level.

Additionally, even from a user's perspective, multiple healthy behaviors and ways of doing them exist that all seem to have a positive impact in general. While a clinician can support a person in sifting through and selecting the "right" intervention for them, there are no comparable tools for behavioral/public health interventions. How then does an individual go about finding a behavioral/public health intervention that works for them, personally? *There is a need for tools that can assist individuals in decision-making of selecting behavioral/public health interventions to help them navigate the landscape of the numerous existing behavioral health interventions.*

In practice, when clinicians choose a particular treatment/intervention for a patient, they go through a variety of decisions based on factors like diagnosis, symptoms, preferences, comorbidities, and patient history to individualize the treatment for that patient using population-level knowledge[11,26], or use single-case or "N-of-1" experimentation to compare plausible treatment choices and choose the most suitable one[10,11,27]. N-of-1 trials are crossover experiments conducted with a single individual as the sole unit of observation[9,28,29]. Crossover experiments are a category of experimental designs in which participants are exposed to each intervention included in the study, or one active intervention and baseline or usual care. The effect of the treatment is estimated by comparing outcomes within person under each treatment (or treatment and usual care)[28,30]. Such experiments account for the heterogeneity in treatment response, and offer a unique opportunity to personalize interventions by allowing decision-making to be made at the individual level. N-of-1 trials share features that are common with group-based crossover designs, except that they attempt to establish causal inference for a single individual[28]. These designs are especially useful when the effects of the intervention are acute, and transient, i.e., there is low possibility of carry-over effects (effects of the intervention remain even after treatment has been removed)[30]. In spite of the increasing popularity over the years, scaling of such N-of-1 designs has been difficult, due to the complexity of designing such experiments and the need for an expert to oversee the experiment. Today, technological progress may enable much of the process to be automated, reducing researcher and participant burden and allowing such methods to be scalable[31–33].

In a public health setting, one way individuals can find out if a given intervention works for them is via self-experimentation. Self-experimentation is a data-driven approach in which the experimenter conducts the experiment on oneself using single-case or N-of-1 experimental designs (i.e., empirically tests whether the intervention 'worked' for them)[34]. Karkar et al.[35] describe self-experimentation as consisting of three phases: hypothesis formulation, hypothesis testing using N-of-1 study designs, and using results to gain insights to inform behavior change. The focus is not on gaining generalizable knowledge, but to aid self-knowledge and self-discovery in a scientifically grounded way, beyond relying on intuition[34–36]. The inherently malleable nature of behavioral interventions makes them especially well-suited to be tested using self-experimentation as their actual, real-world operationalization and impact ultimately depends on how users actually perform them, accept them as valid, and modify them as needed[37]. Tools that support self-experimentation can enable individuals to be involved in their own behavior change process, and iterate and modify intervention options until they find one that suits their life situation at the time[38,39]. Relevant to public health, this enables individuals to determine if a public health recommendation is valuable for them, without the need to work with a clinician.

Over the years, the Quantified Self movement has gained momentum, with growing interest in self-tracking and tools that aid self-discovery, such as self experimentation[40,41]. However, prior work suggests that people often lack scientific rigor in their methods [data collection, analysis, experimental design][36,42,43]. Numerous tools have been designed to help individuals track different aspects of their life from heart rate to finances[44], and even some that support diagnostic self-experimentation in specific clinical settings, such as TummyTrials developed by Karkar et al.[31] to help those suffering from Irritable Bowel Syndrome gain insights on the foods that trigger their symptoms. Lee et al.[39] designed a self-experimentation tool for behavior change that focused on user's self-creation of behavior change plans to improve sleep behavior.

In spite of this growing interest, very few tools are flexible enough to support self-experimentation by a wider user-base or the general population to test common behavioral interventions. In addition, while tools exist to facilitate habit formation[45,46] rarely do they help users decide whether a given activity is worth pursuing for them. Given this background**, the overarching research question** that we aimed to address through this dissertation was:

How might we design tools that can help individuals make decisions about which behavioral intervention to adopt, in a scientific and systematic way that also meets their personal needs and motivations?

1.2 Our Approach

We designed 'Hack Your Health', a tool that can support individuals in carrying out simple N-of-1 self-experiments to test if healthy activities that work in general and/or often associated with heterogeneity in response (specifically, meditation[8,47], vigorous exercise[48], gratitude journaling[49,50], or blocking distracting websites/apps[51,52],) actually improve aspects of their own psychological well-being (their perceived stress, energy, focus or happiness). This is done via comparison of participants' state on days they perform the activity vs. days they maintain their usual routine. Participants also track their perceived enjoyment of the activity as well as the fit of the activity into their daily routine. Psychological well-being was chosen as the outcome of interest because various components of psychological well-being, such as positive mood states, perceived stress, cognitive abilities and happiness have been associated with better health outcomes, such as better cardiovascular health[53], longevity[54], increased productivity[55], and lower medical morbidity[53]. In addition, psychological well-being is subjective and can be assessed through self-report in a less burdensome manner as compared to physiological outcomes, and offers a more scalable approach.

In Hack Your Health, the onus of performing/not performing the activity and self-reporting the data is on the participants, while the tool provides the scientific support in terms of experimental design, data analysis and providing feedback. In addition, all aspects of the tool are designed to be flexible enough to support self-experimentation by a diverse sample of users. Once the experiment ends, participants are provided personalized results. This tool was designed in partnership with Elaine Chen (previously affiliated with WNYC, www.wnyc.org; and currently, The New York Times) to be able to reach a larger audience and a more diverse group of individuals.

1.3 Purpose of this Dissertation

In order to design a scalable tool that can effectively aid individual decision-making, it is important to gain a clearer understanding of users' perspectives on N-of-1 experimentation for behavioral interventions (outside of a clinical context). **The purpose of this work was to address our overarching research question by conducting a formative evaluation of Hack Your Health**.

To be able to thoroughly examine participant experiences and perspectives on such self-experimentation, it is ideal that users first go through the entire experience of using the tool. As part of this study, users participated in an 18-day cross-over self-experiment to try an activity of their choice. For all 18 days, they tracked their psychological well-being, enjoyment and fit of the activity into their routine. Once their experiment ended, we shared their personalized results with them. Participants were also invited to provide quantitative and qualitative feedback in the form of follow-up surveys and interviews. We carried out the evaluation using an iterative and mixed methods approach and tested two versions of Hack Your Health across two waves of participants (Wave 1 and Wave 2) with the following specific aims:

Wave 1 Aims

1. Examine user burden and adherence to an 18-day self-experiment focused on assessing the impact of a behavioral intervention on the user's health and well-being
2. Evaluate the usefulness of Hack Your Health to support a person's decision-making related to continuing or not with simple behavioral interventions
3. Examine individuals' conceptual understanding of n-of-1 study methods for self-experimentation

4. Examine the heterogeneity of individual response to interventions in terms of psychological well-being, enjoyment, and fit into their life

In Wave 1 (N = 20), all participants tracked the same health outcomes (energy, focus, stress, happiness, enjoyment of activity, and fit of activity into their routine) in a quantitative manner. Based on insights and user feedback from Wave 1, we revised the daily surveys to incorporate more reflective and personalized components, specifically, qualitative tracking of participant experience with the activity, and tracking of self-selected outcomes that are personally relevant to them.

We then tested the revised version with a second wave of participants (Wave 2; N=8). In this study, beyond aims 1-3 from Wave 1, we also sought to examine the implications of including participant-chosen outcomes and open-ended reflection of experience with the activity on participant experience of using Hack Your Health. We only made minimal changes to the original version of the Hack Your Health system so that it would also allow us to comparatively examine themes and other findings from Wave 1.

Wave 2: Aims

1. Examine participant experience of tracking self-selected outcomes and of open-ended reflection of perceived impact of activity
2. Examine perceived utility of the 'usual routine' days

1.4 Contributions

Insights gained from this work highlighted the value of a structured self-experimentation approach in supporting decision-making about healthy behaviors. By including four distinct activities for participants to try, we were also able to begin exploring how user experience might differ in the context of different interventions.

Through the use of our mixed methods approach, we were able to uncover issues with using commonly used N=1 experimental designs in the context of decision-making related to behavioral interventions, specifically, when performing an activity in an experimental set up may *not* be appropriate. Insights from the second wave of the study, where we included qualitative tracking of experience with activity, revealed how participant experiences with the intervention were unique, varied over time and were often impacted by the context of their day and life. Additionally, we uncovered discrepancies between participants' lived experience and statistical analyses that such tools will likely need to account for in their design. Insights also indicated that participants may have ingrained beliefs about the impact of popular interventions and that intuition might be an important factor to consider when designing tools to support decision-making in the context of behavioral interventions. These findings were used to inform design implications for tools to support self-experimentation in this context.

1.5 Summary of Next Sections

In the following sections I first review existing literature that examines and illustrates the need for N-of-1 methods for individualization of interventions and to improve individual-level decision-making, as well as related work on tools that enable self-experimentation in health (Chapter 2). I then describe the user research that we conducted that informed the design of Hack Your Health (Chapter 3). In chapter 4, I describe the Hack Your Health system. In Chapters 5, I describe the two user studies we conducted to perform the mixed methods evaluation of Hack Your Health. Chapter 6 discusses findings from both studies, limitations and methodological considerations, and thoughts on future directions. In Chapter 7, conclusions, and contributions of this work are discussed.

CHAPTER 2

BACKGROUND AND RELATED WORK

Imagine Jane, a working, single mother who is looking for a way to lose weight. She remembers reading an article about diet X and how a study reported that people on diet X lost Y pounds over 12 weeks. Jane decides to test that diet. 12 weeks in, and she has lost some weight, but not as much as the article had mentioned. She did follow the diet well except a few cheat meals here and there. Why then did that happen? Was there some other reason?

This story sounds all too familiar, and brings up a few important points about how studies are conducted, reported, and interpreted.

2.1 Interventions That Work 'On Average' Are Often Not Well-matched to Individuals

From a statistical and causal inference perspective, statistical analyses used in group-based research designs such as randomized controlled trials (RCT) seek out an estimate of the 'average' effect i.e., the mean effect of intervention X on outcome Y (e.g., diet → weight loss). What often does not get highlighted in scientific journal articles, which then gets propagated in popular press articles about the work, is the variability in response to interventions. Individual differences in how people respond to interventions are common. Some diets work well for some, do not have an effect for others and, still for others, they actually gain weight during the trial (e.g., Gardner et al.[1]). The average response is the core focus of the results and much of the heterogeneity in response is not discussed in detail. The average insight is the one that gets translated to the public. There is good reason behind this because the RCT and, in particular, the statistics used, produces causal insights that, by definition, are only appropriate at the abstract level of a comparison between groups that people were randomly assigned to (e.g., intervention or control). The problem with this is that the RCT and statistics used, by definition of the method, do not provide insights appropriate for individuals in the study, nor about how individual differences may impact intervention effectiveness except sometimes via secondary analyses but even then, the insights are still gleaned about sub-populations within the group, and not at an individual level.

Put more simply, the current paradigm that includes RCTs as the apex is valuable for providing insights about the aggregate but comes at the cost of largely ignoring concrete manifestations in the real world (i.e., what works for a specific person) and including individual differences as random error in pursuit of the aggregate insight[56]. This subtlety is important but largely misunderstood when individuals themselves try to use evidence-based interventions, thus resulting in the real possibility that some people are engaging in evidence-based interventions but not gaining any real benefit from it. N-of-1 methods, including the statistics used for them, are a complementary approach that can provide insights for an individual about the possibility that an intervention is helpful, not just in general, but for them[11,33].

Related to external validity, scientific studies are often carried out in highly controlled settings (with an aim to establish internal validity), with relatively homogeneous samples (participants) defined by inclusion/exclusion criteria that may or may not be representative of the eventual population and setting that the treatment/intervention gets used in[9,37,57]. For example, in Jane's case, the studies she read about had probably recruited individuals who are obese, sedentary, and likely having no existing comorbidities (e.g., high blood pressure, or hypothyroidism) than the one of interest (overweight/obesity). Moreover, the study likely involved tracking food intake daily, and monthly check-ins with the study staff to maintain compliance to the diet protocol. This is different from most individuals, such as Jane, who hear results from the trial on popular media and then assume that the results are relevant to them when, technically, that extrapolation from the evidence is often not warranted. Much prior work highlights this gap between research and practice[58–60]. In real-world settings, many complex factors interact to influence behavior and response to interventions over time.

2.2 Heterogeneity of Response to Behavioral Interventions

Literature on many popular behavioral interventions exhibits variability in response across participants. For this review, we focus on interventions that are included in Hack Your Health. The rationale for selecting these interventions is provided in Chapter 3, where we describe the user research that guided the design of the tool.

Physical Activity

Literature on physical activity indicates that considerable inter-individual variation exists in terms of the response to regular physical activity and even highly standardized training programs[5]. For example, a study by Bouchard et al.[61] looking at responses to a standardized training program found high individual differences, with changes in VO_2max ranging from no gains to 100% gains. Moreover, a study by Bouchard et al.[62] also indicated that as many as 8% of participants had adverse responses related to cardiovascular- and diabetes-related risk factors to regular exercise. Meta-analyses examining the impact of exercise on aspects of cognitive function such as attention and processing speed, and positive affect indicate modest improvements as a result of aerobic activity, but with high variability within and across studies[63,64]. A meta-analysis by Reed and Ones[64] of the effects of acute aerobic exercise on positive affect also suggests that although overall exercise seems to have a positive impact on affect, the effects may be different based on the intensity of exercise (higher overall impact for lower intensity exercise than moderate or high intensity exercise), or depend on participants' baseline fitness status, and even exercise duration. They also point out that much variation remains to be explained, that could be attributed to individual differences (e.g., positive affect at baseline) and unexamined moderators.

Meditation

Meta-analyses and systematic analyses looking at the impact of meditation suggest that overall, meditation has a positive impact on health outcomes such as perceived stress, emotional functioning, happiness, etc[8,47]. However, they also indicate that substantial heterogeneity exists in terms of those effects [6,7,65]. Some have identified changes in cognitive distortions[66], perceived control, and neuroticism as potential mediators/moderators of the impact of meditation on indicators of well-being[67,68]. In addition, few studies also suggest that certain kinds of meditation could have an adverse effect for some individuals[69–71], such as those with certain psychological problems. For example, some of the adverse effects identified in a study by Otis[69] included increased anxiety, boredom, depression, confusion, etc. In addition, different types/styles of meditation can have differential impact on emotional wellbeing[19].

Gratitude journaling

Research examining the impact of positive psychology interventions like gratitude journaling suggests that gratitude journaling is efficacious in improving happiness and well-being, however, there is high variability in effect sizes across studies[3,49,50,72].

Lyubomirsky and Layous[3] also discuss the concept of 'person-activity fit', based on prior work that that person-factors such as motivation, preference, personality are important moderators of intervention response[2,14]. In addition, baseline characteristics like initial level of psychosocial distress, degree of self-selection, and intervention duration have also been shown to impact the response to interventions[73].

Blocking Digital Distractions

The impact of digital distractions such as social media on aspects like psychological well-being and productivity is a relatively new area of enquiry that has recently gained a lot of attention in both research and media. Prior work suggests that digital distractions and interruptions could have a negative impact on productivity and performance[51,52,74]. However, extensive empirical work examining the effect of blocking those distractions is lacking. While it might theoretically seem like blocking digital distractions could improve productivity, the few existing studies suggest mixed results. A recent study that looked at the effect of interventions involving blocking or reduction of the use of non-essential websites (e.g., YouTube, Facebook, Twitter) showed that some individuals exhibited an increase in stress while others exhibited a decrease in stress[74] as a result of blocking distractions while at work. This could be because people use non-essential websites and social media for different reasons. For example, prior work has suggested that people may use non-essential sites to distract themselves after task completion while at work or to distract themselves if they are not making progress in work, and not having access to non-essential websites may worsen their perceived stress and productivity[52,75]. While it might seem conceptually counterintuitive, is suggested that this could happen because for some people, the *act of resisting the temptation* to use those websites could in turn negatively impact their productivity[76]. Level of self-control, and the individual's level of involvement with technology and social media have also been shown to moderate the impact of blocking distractions on productivity and stress[52,74]. In addition, effects might be different based on whether the blocking is done autonomously by the user or has been imposed on the user[52].

Although empirical work is lacking, this domain, complicated by the interactions between types of users, potential positive effects of social media coupled with suggested detrimental impacts of distractions might make this a particularly important area to study using N-of-1 approaches to improve understanding of the impact of individual differences on response.

2.3 Examining Interventions in Terms of Enjoyment and Fit into Routine

Heterogeneity of response is larger when considering the behavioral target itself as an outcome of the intervention when attempting to translate these behaviors into real-world contexts. The goal in public health interventions is not merely to determine if a behavioral intervention impacts a health outcome but also to help people enact these behaviors in a sustained manner in real-world contexts[77,78]. This further reinforces the need for complementary tools to help individuals determine not only if an intervention works for them (i.e., improves their well-being) but also if it fits into their life. For example, time and logistical constraints (such as having no access to a quiet location for meditation) are a frequently reported barrier for adoption of healthy behaviors[79,80]. Recognizing that individuals have only a finite amount of time, this establishes the need for tools that could feasibly help individuals select intervention options that not only produce desired effects, but also fit into a person's life.

Another critical aspect to consider is the person's enjoyment of the intervention. Even if an intervention produces a desired effect and fits into their life, if the activity is not enjoyable/self-rewarding, it is quite plausible that the person will not continue to engage in the activity in a real-world context. This is, indeed, a common critique of meditation or vigorous activity as there are some people that can do it, but just do not want to do it[81–83], and attrition rates are often high[84,85]. Enjoyment is considered to be an important determinant of adoption of physical activity[86]. Even in positive psychology interventions, the degree to which people report enjoying the activity is associated with how often they actually perform it[2].

2.4 The Paradox of Choice[87]

Moving beyond the heterogeneity of response to interventions, there is also a wide variety of plausible interventions one could use to produce desired effects, such as improving well-being. In particular, prior work illustrates that if one is interested in improving their quality of life and psychological well-being, there is clear evidence that they could consider engaging in vigorous physical activity or mediation and there is emerging evidence that strategies from positive psychology such as gratitude journaling, or reducing social media use, could each be used to improve them[8,52,73,74,88]. Moreover, it is important to note that there are numerous ways in which these interventions could be operationalized. For example, vigorous exercise could be carried out through sprinting, playing tennis, in a group or individually; gratitude interventions could be performed through a gratitude visit, or by writing down things you’re grateful for. There is a strong need then, from the perspective of translating evidence from controlled studies into practice, to provide a scalable and systematic approach for individuals to engage with evidence-based interventions to examine and test which operationalization (if any), works for them or not.

2.5 Using N-of-1 Experimentation to Translate Evidence from the Aggregate to the Individual

In clinical settings, when physicians choose a particular treatment for an individual, they go through a variety of decisions (e.g., available resources, context, comorbidities, patient preferences) to attempt to individualize the treatment to a patient using population-level knowledge[10]. Some even use a more systematic approach of single-case experimentation to compare two plausible treatment choices[9,27]. Single subject studies consider the individual as the sole unit of observation when studying the impact of an intervention, thus allowing inferences to be drawn at the individual level[11,89,90]. The objective is to find the optimal intervention for the individual in a data-driven manner. This kind of experimentation is potentially well-suited to this problem of translating evidence from aggregated insights into individualized insights. However, as public health-oriented interventions are classically not bound to clinics, this establishes a gap in how evidence-based behavioral interventions can be scaled to individuals in the population. In particular, the clinician plays a foundational role in figuring out if an intervention is right for a particular person but, for individual-based health interventions (e.g., moderate to vigorous physical activity, meditation), this might not be the case. For these interventions, it is largely the individual's responsibility to choose and adapt an intervention to suit their needs.

Previously, such N-of-1 experimentation was not scalable (high time, expert, as well as financial commitment), but with current technological and statistical advancements, it is becoming increasingly possible to conduct such trials simultaneously and on a large scale, and without the need for a human expert[27,31,32]. In addition, if a large number of users participate, individual-level insights can also be aggregated to gain generalizable insights about the kind of interventions that seem to work for whom, and under what context and circumstances[91,92].

2.6 Helping People Stop "shoulding" via Self-experimentation

One way individuals can find out if a given intervention works for them is via self-experimentation. The primary aim of self-experimentation in this context is to increase self-knowledge and support self-discovery. As Neuringer[34] suggests, it can help an individual create **personal** "IF-THEN" contingency statements about an intervention/activity [If I exercise, THEN I feel more energetic] as opposed to relying on **'should do'** style **impersonal** "IF-THEN" contingency statements, such as "IF one exercises, THEN they should feel more energetic]. As highlighted in the introductory chapter, this "should do" style contingency statement is, methodologically, an appropriate conclusion from the current RCT-paradigm of evidence-based practice (with all of the caveats about match to a specific person already highlighted). From the perspective of real-world constraints, particularly time, focus, and energy, the myriad "should do" suggestions available in our cultural could, in fact, be resulting in the unintended consequence of making people more stressed and reduce overall well-being via the paradox of choice. This is based on psychological research indicating that having too much choice can be stressful[87,93], that the feeling that one should do something can be stressful[94], and reduced agency in terms of making thoughtful decisions can have a detrimental impact on well-being. While this last part is conjecture and based on theory rather than data, one key first step to both test this conjecture and also to remedy the situation is by providing individuals tools to help move from "should do" contingency to personal IF-THEN contingencies. Some have also argued that such "researcher-as-subject" methods are important because they allow the person to gain "subjective experiential knowledge" and enable self-reflection, help them develop research questions more relevant to their context, and also improve individuals' mental models of how the world works beyond the context of the experiment[95,96].

While many existing tools support behavior change related to various behaviors from finance management to meditation, the focus is largely either on self-tracking or habit formation of specific behaviors. Very few tools take a step back and help individuals empirically test, through self-experimentation, *whether* and *which* intervention has the desired effect in the first place while also providing insights on if it fits into their lives and if they enjoy it.

2.7 Designing Tools to Support Scalable Self-experimentation

Recent work in Human-Computer Interaction (HCI) has explored the design of tools that support self-experimentation, often focusing on specific populations/domains and diagnostic self-experimentation. For example, TummyTrials[31] supports those suffering from Irritable Bowel Syndrome (IBS) in self-experimenting to find relationships between food triggers and IBS-related symptoms. SleepCoacher[32] identifies relationships between common factors that impact sleep quality. Lee et al.[39] designed a do-it-yourself (DIY) self-experimentation toolkit for behavior change, where users with sleep-related issues could self-create and test 'just-in-time' interventions, using self-experimentation not only to asking IF an intervention has an effect on a health outcome but also HOW to enact a behavioral plan that produces an effect, fits into a person's life, and was enjoyable enough to continue with. We built on this prior work by designing Hack Your Health, a self-experimentation tool for a broader audience, to choose and test simple behavioral interventions and test them to see if they have the hypothesized impact on their psychological well-being, and to self-track enjoyment as well as fit into their life.

Self-experimentation, based on the framework developed by Karkar et al.[35] involves three steps - hypothesis formulation, hypothesis testing, and using results to gain insights and inform behavior change. Each of these steps is complex and, classically, was handled through the patient-provider interaction within clinics. As our focus is on public health interventions, a central focus of this project is to work towards making this process of self-experimentation as simple and engaging as possible via digital technologies, by embedding many of these hard decisions and choices into the digital tool. Previous work has found that data analysis and scientific rigor while conducting the self-experiment are two common problems users face[36,42,43]. In addition, the process of self-experimentation can be time-consuming and, in particular, the effort may not be perceived as worthwhile compared to the benefit one receives. Approaches that combine user-led and system-led support and decision making can help overcome such barriers and reduce such participant burden[42]. Hack Your Health aimed to achieve that by providing system-led support for analysis and feedback [by providing personalized results] and experimental design, while at the same time, guiding the user through an experience of self-experimentation by providing flexibility of choice in terms of intervention-outcome combinations.

2.8 Rationale for this Work

Before this tool is made available to the general population, it is important to conduct a formative evaluation. In this context, it is important to understand whether the different elements of Hack Your Health are understood, accepted and perceived as useful by end-users, and to examine users' perspectives on such N-of-1 experimentation for behavioral interventions (specifically, outside of a clinical context). This is particularly important based on findings from our user research (Chapter 3) which highlighted some discrepancies between the research team and user's conceptualization of self-experimentation and the expectations from a self-experimentation tool such as Hack Your Health. For example, few participants expressed doubt about whether a given intervention would work for them, highlighting the possibility of unawareness, disinterest or discomfort with the uncertainty of intervention response. Karkar et al.[31] through their testing of TummyTrials and previous work on self-experimentation[43], found that self-experimentation (methodologically) was a difficult concept to understand for end-users, and also uncovered tensions between results from statistical analyses and people's own lived experiences. For example, participants had difficulty understanding results that were labelled 'no evidence' (Does it trigger symptoms or not?) to mean that the food did not trigger symptoms and they could continue to consume it. These tensions may be exacerbated when it comes to popular public health interventions, which are all expected to work, in general. As Kravitz et al.[91,97] suggest, there is a chance that learning that an intervention believed to be effective doesn't work may have a negative impact. Karkar et al[31]. also found that participants were often in situations where they needed to improvise to maintain fidelity of the experiments. Some were successful, and some weren't, highlighting the need to make sure that individuals have basic scientific understanding of the process of self-experimentation.

It is important to uncover such tensions, and gain an understanding of which aspects of self-experimentation users find useful, burdensome, confusing, or difficult to understand. This can help us understand how to frame and design the experience in a way that fulfills people's needs. In addition, building this tool also helps to foster education on how current scientific practices work today and, by extension, when and how to use the information to support individual decision-making.

User Burden and Adherence

User-burden can be defined as “the negative impact that computing systems might place on the user”. Gaining an understanding of the burden that a system places on the end-user is critical, because increased burden may have a negative impact on their ability to use, tolerate and continue use of the system[98]. An understanding of the system’s pain points provides insights about which aspects of the system need improvement. Measuring adherence to the intervention protocol is important to get a sense of whether the users are interacting with the system as desired. Adherence rates for different aspects of the system (adherence to experiment protocol, adherence to surveys) can help uncover elements of interaction with the system that users dislike or find burdensome.

Usefulness

According to Nielson[99], usability is the quality attribute that assesses how easy user interfaces are to use, while utility refers to the design’s functionality, and whether the system does what the users need. Usability and utility together make up ‘usefulness’ of the system. Evaluating the usefulness of a system is critical to ensure that users actually use the system. If participants do not find the system useful, they may abandon it. The primary goal of Hack Your Health is to guide a user through trying out a healthy activity to figure out whether or not it helped improve their well-being as compared to their usual routine. In this context, we want to understand whether the tool was useful to participants in terms of deciding to continue or not with the activity they tried.

Understanding of N-of-1 Method

In the realm of health behavior change, the end-user needs to be actively involved in the process (to ‘test’ the activity requires them to perform it as per the protocol and experimental schedule), and so it's important that they have a conceptual understanding of the science of self-experimentation. In addition, a conceptual understanding is important for users to be able to appropriately interpret findings from such experiments, and acknowledge the uncertainty and biases that accompany self-experimentation. Assessing users' conceptual understanding of the methods used will help uncover misconceptions, aspects that users find confusing, and aspects that they find difficult to understand.

To our knowledge, very few studies have thoroughly examined the user burden, usefulness, and understanding of N-of-1 designs in a non-clinical context, hence justifying this study as a logical next step in this line of research.

CHAPTER 3

USER RESEARCH

3.1 Introduction and Background

In this chapter, I review user research we conducted[100] that guided the design of Hack Your Health[1] and helped us define the specific aims for this work.

We initially began designing this tool with Elaine Chen from WNYC (New York Public Radio), as an experience that would be offered to the WNYC audience, and hence the aim was to make it scalable enough to be used by a diverse group of up to 15,000 people. As of this writing, the tool may not be disseminated via WNYC due to funding and other issues, but we continue to work with Elaine Chen, and the aim to make it scalable still holds true. To design a large-scale tool such as this one, we worked under the following design constraints:

1. Include interventions and outcomes that are of interest to the WNYC audience and shown to be effective 'on average', but also often associated with heterogeneity of response.
2. Included interventions/activities should require minimal training, be short, easy to self-administer and perform.
3. The experience should have low participant burden.
4. Balance the desire to generate population-level insights while maintaining the spirit of self-discovery and self-experimentation.
5. Balancing the needs of three stakeholders: WNYC and the editorial team, researchers, and the end-users.

---

[1] Insights from this user research were presented as part of a poster at Pervasive Health conference 2018, and published as an extended abstract in the EU Digital Library

To inform the design of the tool, we needed to understand users' motivations and interest in participation and expectations from such a tool in the context of behavioral interventions. We conducted user research that focused on gaining an understanding of:

1. Which interventions and outcomes are of interest to an initial podcast audience (Only Human from WNYC)?
2. What are user perceptions related to self-experimentation and use of such a tool?
3. What are user motivations for engaging with such a tool, if any?
4. What are users' reasons behind choosing specific interventions?
5. What is the minimal information we need to provide to clearly define the experience of taking part in an n-of-1 study?

As part of this user research, we used an iterative design process where WNYC listeners (N=551) were surveyed and a subset (n=18) were interviewed. Recruitment was carried out via an email distributed to WNYC listeners of the podcast *Only Human* and also shared by the research team on social media. The email as well as survey introduction provided a short description of the tool. Participants were not compensated for completing the survey.

3.2 User Research Survey

The purpose of the survey was to understand the interventions and outcomes of interest to the WNYC audience to then be included in the final tool, and to gauge users' interest in a self-experimentation tool. We received 551 responses, and the majority were WNYC listeners living in the United States. Other demographic information about survey respondents is presented in Figure 3.1.

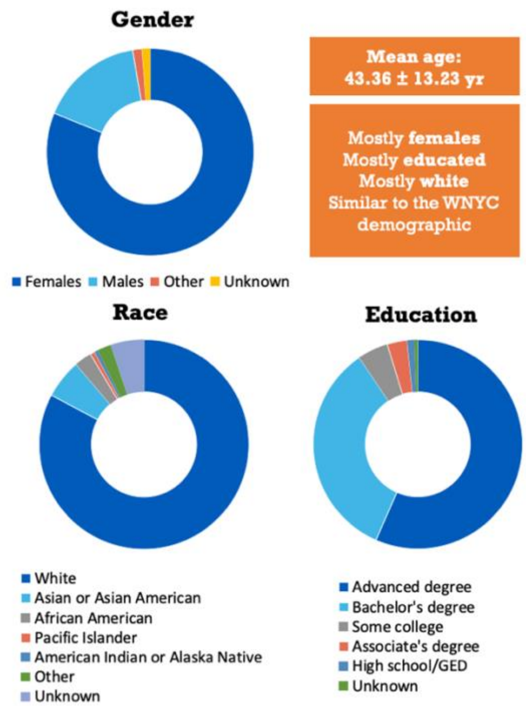


Figure 3.1. Demographic Characteristics of Survey Respondents from User Research

For this survey, we also included two outcomes (pain and bowel movement) in addition to psychological well-being based on interest from the WNYC editorial team. The survey consisted of a mix of multiple-choice questions and open-text responses [~19 questions].

Participants were asked to choose up to 3 outcomes that they would be interested in working on; four of the options were in the form of pairs [focus/distractions; stress/serenity; energy/fatigue; happiness/sadness; pain; bowel movements]. Out of the outcomes they chose that were pairs, they were then asked to choose the framing (positive/negative) that they prefer (e.g., focus vs. distractions). Outcome choices are listed in Table 3.1, from most popular to least popular. Positive framing was more popular amongst all outcomes other than stress. Based on these results, **energy, focus, stress and happiness were chosen as the outcomes to be included in the tool.**

Table 3.1. Outcomes Included in User Research Survey

Outcome	**N (551)**	**%**
Energy/Fatigue	436	79
Energy	304	70
Fatigue	132	30
Focus/Distraction	386	70
Focus	282	73
Distraction	104	27
Serenity/Stress	374	68
Stress	194	52
Serenity	180	48
Happiness/Sadness	241	44
Happiness	187	78
Sadness	54	22
Pain	96	17
Bowel Movements	57	10

For interventions, participants were asked to choose two interventions that they would like to try and compare to see which one (if any) improves the outcomes they selected. We initially included 11 behavioral interventions for them to choose from, with an aim to retain the top five, displayed in Table 3.2. The top 5 choices were similar across gender. As deep breathing and meditation were both similar and difficult to distinguish easily for the purposes of this tool, we combined them and included ‘deep breathing meditation' as one of the interventions. In addition, interviews indicated that digital distractions were one of the main reasons behind choosing 'focus' as the outcome to improve. Hence, we decided to include an intervention that focused on blocking digital distraction (e.g., Facebook, Twitter) as the fifth intervention in Hack your Health.

Table 3.2. Interventions Included in User Research Survey

Intervention	N	%
Meditating for 10 minutes	214	39
Engaging in 10 minutes of vigorous activity	179	32
Taking 5 minutes for deep breathing	179	32
Writing 3 things you're grateful for	171	31
Performing one random act of kindness	88	16
Laughing for 5 minutes	72	13
Replacing an unhealthy snack with nuts	41	7
Not drinking any caffeinated beverages in the afternoon or evening	39	7
Having a 5-minute conversation with a stranger	35	6
Not drinking any sugar-sweetened beverages	35	6
Giving a hug to 3 different people	21	4

When asked about likelihood of participation, most indicated that they would be very likely (46%) or extremely likely (20%) to participate.

3.3 Interviews

We interviewed a subset of 18 of the survey respondents. The demographic information of the interview participants is provided in Table 2.3. Participants' survey responses were used to inform the exact questions asked in the semi-structured interviews. They were asked about (1) meaningful life changes if the outcomes were to improve; (2) why the interventions interested them; (3) prior experience with the interventions; (4) reasons behind choosing the intervention combination; (5) why the tool interested them, if it did.

Interviews were carried out over the phone, audio-recorded if participants consented. Participants were first provided a brief description of the tool and informed of the purpose of the interview. All interviews were transcribed verbatim. Data were analyzed using MAXQDA software (MAXQDA, VERBI Software – Consult – Sozialforschung GmbH, Berlin, Germany). Data was analyzed using a thematic analysis approach[101,102]. Thematic analysis was chosen as the method of analysis because our aim was to identify patterns across the data that could be important to factor in when designing the tool. Emerging themes were identified concurrently with the data collection process, and discussed with the entire research team. Data were then coded and re-coded (to ensure rigor) according to the identified themes. All interviews were then examined for the presence of the identified themes.

Findings from Interviews

Motivations for participation included an interest in forming habits out of the activities they chose (n=12), and wanting accountability (n=9) to aid habit formation: "*I am thinking of this as a springboard to get into meditation again*." (P4). Participants were also motivated to participate because of interest in contributing to science or a being part of a big project (n=5).

Many participants were interested in interventions they had been unable to successfully adopt in the past. For example, P2, when asked about reasons for choosing meditation in spite of previous unsuccessful attempts, said, *"..having had a bunch of false starts at it, it's always at the back of my mind like, knowing that I haven't fully gone through with it."*

Table 2.3. Demographic Characteristics of Interview Participants

Profession	Gender	Age	Education	Race
Research lab manager	M	31	Bachelor's degree	White/Pacific Islander
Computer-related work	M	35	Bachelor's degree	White
Yoga instructor	F	57	Some college	White
Computer-related work	F	28	Advanced degree	Asian/Asian American
Public speaking/acting coach	F	46	Advanced degree	White
Sales	F	54	Bachelor's degree	White
Social media manager	F	40	Bachelor's degree	Mixed European/Native American
University teacher	F	24	Associate's degree	Asian/Asian American
Chef	F	27	Bachelor's degree	White
Marketing Research	M	48	Advanced degree	White
Librarian	GNC*	32	Advanced degree	White
Manager	M	48	Advanced degree	African/African American
Entrepreneur	F	38	Bachelor's degree	White
Graduate student	F	25	Advanced degree	White
School teacher	F	34	Advanced degree	White
Pediatrician	F	40	Advanced degree	White
Consultant	F	48	Advanced degree	White
Software engineer	M	48	Advanced degree	White

*Gender non-conforming

Reasons behind intervention choice included choosing ones they think are beneficial (n=6), have been advised to try by someone (n=3), believe they 'should' be doing (n=7), or believe are backed by research (n=4). For example, P5, when asked about what she expects to happen if she tried gratitude journaling, said, "*Well, I know from the research that if you focus on things that you're grateful for, then you tend to be happier, one tends to be happier*. P6, who had been wanting to begin exercising regularly but hadn't been able to, said "*...I'm not finding the motivation to do it for some reason even though it's something I want to do and I know that it makes me feel better and I know it's something I should do*."

It is worth noting that contrary to our expectations (though also, in hindsight makes sense based on current portrayals of these interventions in popular media), few participants seemed to doubt whether a given intervention would be valuable for them: "*Well, I definitely think that, you know, deep breathing would definitely, well I would hope, would reduce my stress, well that's what I've heard from other people, and that just the act of reducing stress alone would increase my happiness*." (P11)

Although we believed that comparing two interventions made sense from a methodological perspective, participants did not intuitively understand the value in comparing two active interventions. P15, who chose gratitude journaling and deep breathing – "*In my life I would see those activities almost happening simultaneously. So I don't know that it would be necessarily easy to compare*."

Participants also chose activities with an intention to improve aspects of well-being other than/in addition to the ones they selected in the survey. Few were also interested in developing particular traits or skills through continued practice of the activity, rather than improving psychological wellbeing. For example, P17 wanted to perform acts of kindness in order to make generosity a "*fabric of who I am*".

As mentioned earlier, most participants wanted to try activities with an aim to form habits. They did not seem to identify the benefit of testing whether an intervention was effective for them. A reason for this could be the media's and research's lack of attention to the heterogeneity of response to these interventions that are thought to be beneficial in general, and people may not be aware of the existing variability.

3.4 Design Implications from User Research

Design implications that emerged from this work include designing the tool to inform participants' interest in habit formation, tracking multiple health outcomes instead of a single outcome, simplifying the experimental design (e.g., testing only one intervention against their usual routine), and highlighting existing individual differences to improve understanding of the logic of self-experimentation. These insights were then used to design and test low-fidelity prototypes with participants, and then a final high-fidelity version that was used for the formative evaluation (described in detail in Chapter 4).

These findings provided the basis of the aims of the subsequent studies, where we had participants use Hack Your Health to conduct a formative evaluation using mixed-methods to gain a clearer understanding of the user burden, conceptual understanding and perceived utility of N-of-1 self-experimentation in the context of behavioral interventions (Chapters 5 and 6).

CHAPTER 4

HACK YOUR HEALTH

This chapter describes the Hack Your Health system.

4.1 Overview

Hack Your Health is a simple tool that supports self-experimentation using N-of-1 crossover designs. Participants can try one activity through an 18-day experiment to test if it improves their psychological well-being. The tool is composed of a website where participants can sign up for their experiment, and provide baseline and contact information. Once they sign up, their experiment is set up in the backend. All experiment-related communication, such as daily reminders and links to daily surveys takes place via SMS.

Participants first go to the Hack Your Health website (hack-your-health.org), which provides detailed information about the tool and the available activities they can try and the outcomes they will be tracking as part of the experiment. If interested, and once they know which activity they want to try, they can proceed to complete the sign-up survey for that activity. Participants are nudged to choose an activity that they don't already perform on a regular basis. The steps in the signup process for each participant are as follows:

1. Complete an electronic informed consent form.
2. Complete the signup survey
    i) Provide other relevant demographic and other information.
    ii) Provide their phone number and email address.
    iii) Complete baseline questionnaires (this step is relevant to this particular research project)

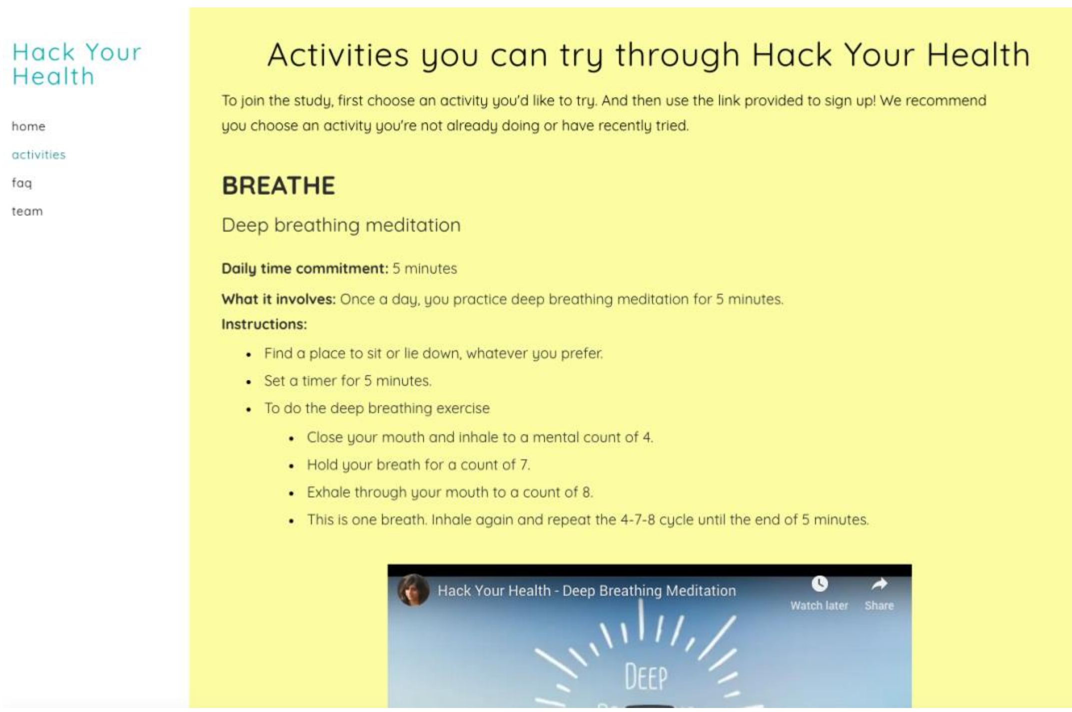
3. Once survey is complete, participants are set up in the system. They are also sent an email providing more details about the self-experiment protocol (hack-your-health.org/next-steps/) such as the fact that they will be performing the activity on a few days and other days will be rest days where they have to maintain their usual routine, and that they have to complete the daily surveys even on rest days.
4. At the backend, each participant is randomly assigned an experimental schedule (1 out of the 12 choices listed in Table 4.2).
5. Once their experiment begins, on all 18 days, participants receive a daily morning SMS reminding them to do the day's assignment (either to do the activity, or stick to their usual routine), and an evening SMS with a link to the daily self-report outcomes survey. The evening SMS also informs them of their upcoming day's assignment to give them a heads up and some time to plan it into their schedule.
6. Once the experiment is completed, and provided there is enough data collected, their data is analyzed and insights are shared with them in the form of a PDF sent via email.

4.2. Interventions included in Hack Your Health

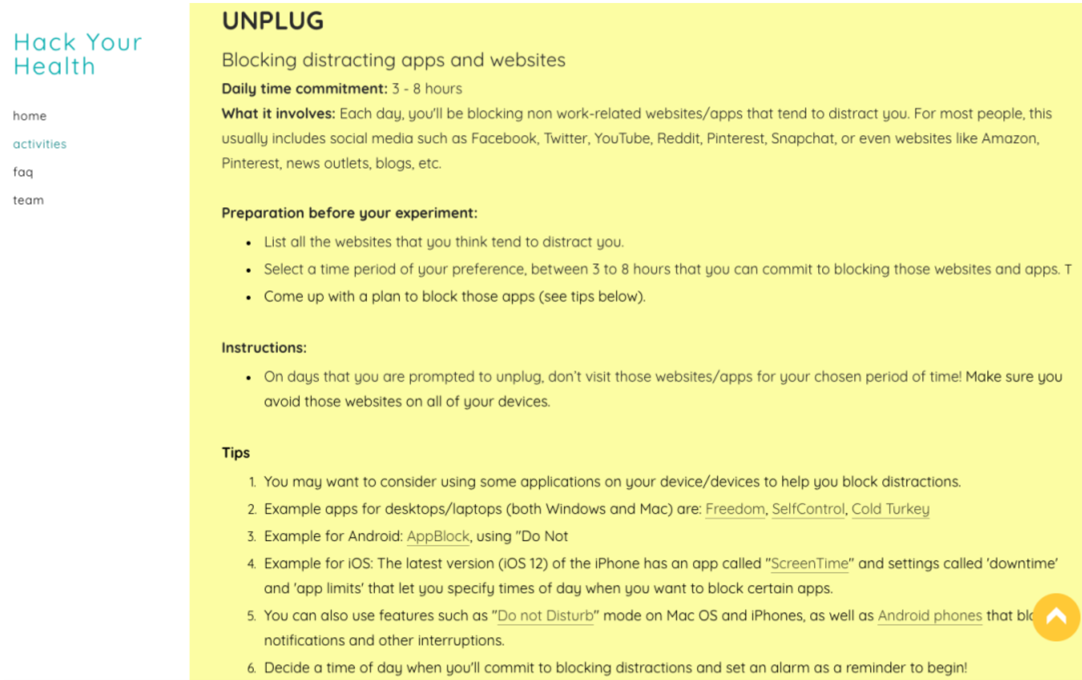
The behavioral interventions included as part of Hack Your Health are:

1. Deep breathing for 5 minutes
2. Performing vigorous physical activity for 10 minutes
3. Gratitude journaling (writing three things you're grateful for)
4. Blocking digital distractions (distracting websites/apps for a self-selected duration of 3-8 hours)

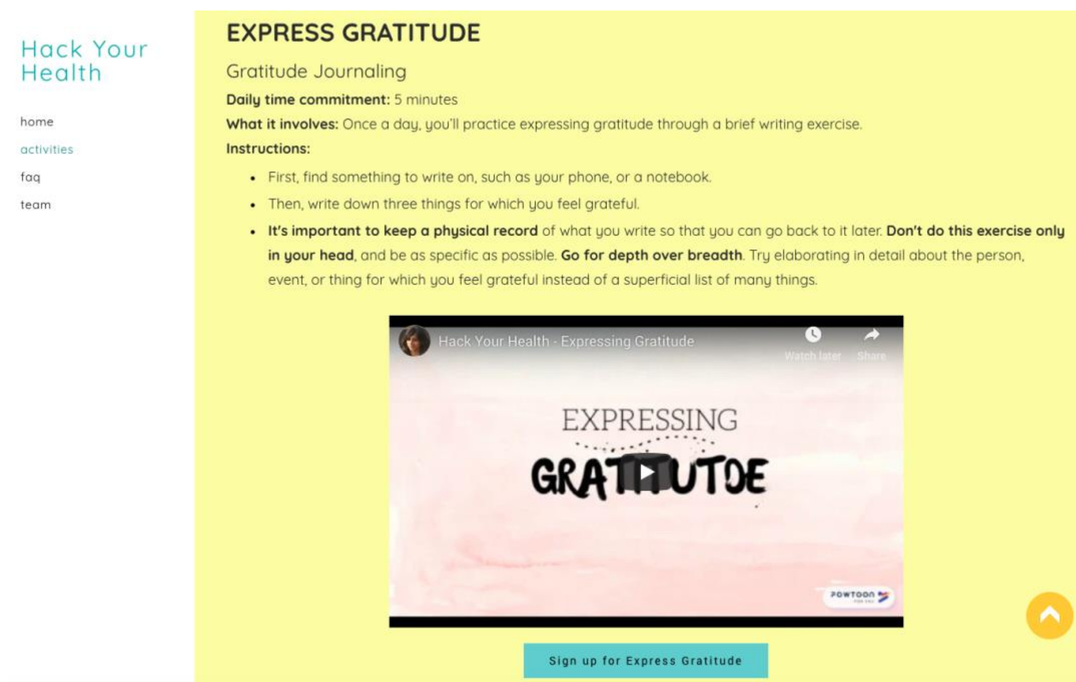
The study website contains written and video-based instructions on how to do the activity, along with other information such as the required daily commitment (Figure 4.1). Once they sign up, participants are also provided links to instructions for their specific activity. While in the experiment, participants can perform the activity at a time of their choice. We included the flexibility in order to balance the needs to assess enjoyment and fit and intervention impact.



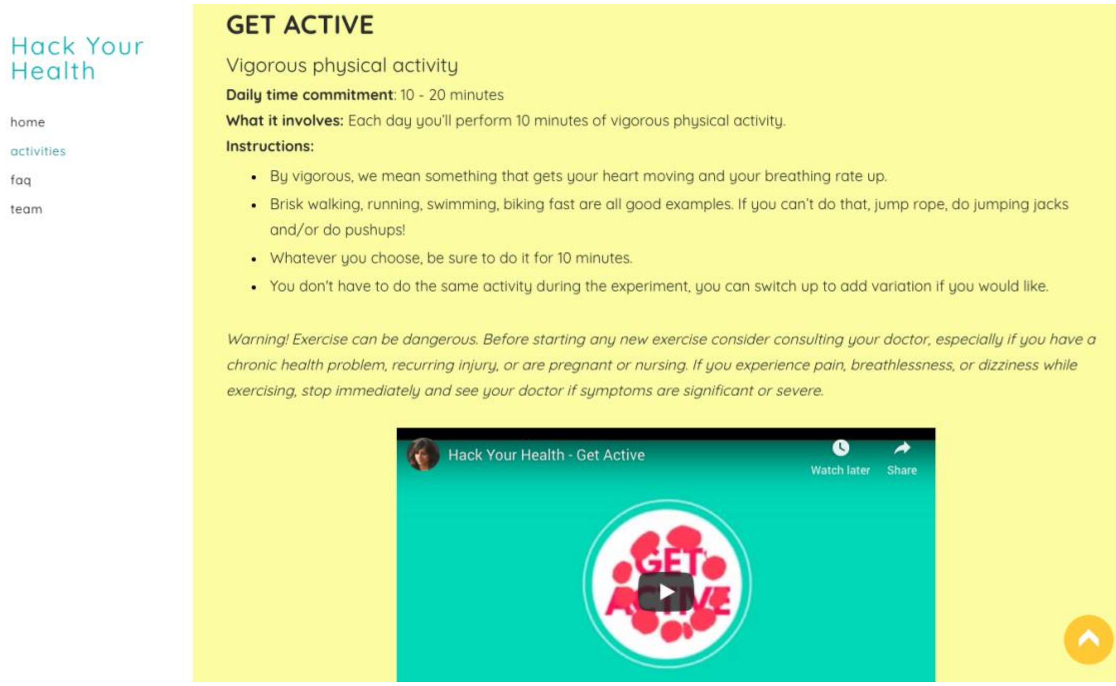
(a)


Hack Your Health
home
activities
faq
team
UNPLUG
Blocking distracting apps and websites
Daily time commitment: 3 - 8 hours
What it involves: Each day, you'll be blocking non work-related websites/apps that tend to distract you. For most people, this usually includes social media such as Facebook, Twitter, YouTube, Reddit, Pinterest, Snapchat, or even websites like Amazon, Pinterest, news outlets, blogs, etc.
Preparation before your experiment:
List all the websites that you think tend to distract you.
Select a time period of your preference, between 3 to 8 hours that you can commit to blocking those websites and apps. T
Come up with a plan to block those apps (see tips below).
Instructions:
On days that you are prompted to unplug, don't visit those websites/apps for your chosen period of time! Make sure you avoid those websites on all of your devices.
Tips
1. You may want to consider using some applications on your device/devices to help you block distractions.
2. Example apps for desktops/laptops (both Windows and Mac) are: Freedom, SelfControl, Cold Turkey
3. Example for Android: AppBlock, using "Do Not
4. Example for iOS: The latest version (iOS 12) of the iPhone has an app called "ScreenTime" and settings called 'downtime' and 'app limits' that let you specify times of day when you want to block certain apps.
5. You can also use features such as "Do not Disturb" mode on Mac OS and iPhones, as well as Android phones that bl
notifications and other interruptions.
6. Decide a time of day when you'll commit to blocking distractions and set an alarm as a reminder to begin!

(b)


Hack Your Health
home
activities
faq
team
EXPRESS GRATITUDE
Gratitude Journaling
Daily time commitment: 5 minutes
What it involves: Once a day, you'll practice expressing gratitude through a brief writing exercise.
Instructions:
First, find something to write on, such as your phone, or a notebook.
Then, write down three things for which you feel grateful.
It's important to keep a physical record of what you write so that you can go back to it later. Don't do this exercise only in your head, and be as specific as possible. Go for depth over breadth. Try elaborating in detail about the person, event, or thing for which you feel grateful instead of a superficial list of many things.
Hack Your Health - Expressing Gratitude
Watch later
Share
EXPRESSING
Sign up for Express Gratitude

(c)



(d)

Figure 4.1. Screenshots of The Hack Your Health Website Displaying Information About the Four Included Activities

4.3 Outcomes measured in Hack Your Health

Psychological well-being

Aspects of psychological well-being that are measured as part of Hack Your Health include daily perceptions of focus, energy, happiness, and stress. One item per outcome was used to avoid the risk of high participant burden, since these surveys are completed daily for the entire duration of the study. As our focus is on N-of-1 studies, the concept of a "validated" measurement tool does not match with classical conceptualizations of this. This is because classical approaches to measurement validation assume an intent of doing aggregation-oriented (also called nomothetic) statistical analyses. N-of-1 implies different criteria of success for determining if it is "valid." When analyses are to be conducted by and for individuals, single items with high face validity may be more desirable because they may be more likely to capture the concept of interest than selecting a subset of items from a longer questionnaire. Measures validated for nomothetic purposes often use multiple-item questionnaires that together can be used to infer a latent concept of interest. For ecological momentary assessment and also idiographic time series purposes (i.e., what we are doing in this N-of-1 study), picking a subset of items of interest from longer existing questionnaires is accompanied with the risk of lowering content validity by using an inappropriate operationalization of the construct of interest and, also, increasing burden on the participant, thus reducing the likelihood of repeated measures, and, also, often multi-item questions avoid the exact term of greatest interest, thus reducing the likelihood of face validity and, by extension value the individual could glean from the item[103].

Recognizing these issues, our items were explicitly selected based, in part, on high face validity (see Table 4.1). Face validity has been defined as "the degree that respondents or users judge that the items of an assessment instrument are appropriate to the targeted construct and assessment objectives"[103]. When signing up for the tool, users are informed that they will be measuring their energy, focus, happiness and stress levels for the 18 days of the experiment. A second criteria we used to select items was burden. We sought to keep the length and wording of the daily surveys short, to avoid excessive response burden. Very few existing validated questionnaires measuring the outcomes of interest fulfill these criteria of being short, and low response burden. While multi-item questionnaires are common, when data is collected frequently (e.g., daily), there is precedent for using single-item measures[104,105]. Finally, and perhaps most importantly, our primary goal is to provide insights for the individual to answer "does it work for me, and is it worth it?". Based on this goal, single item questions with high face validity have, arguably, the least amount of bias for that particular purpose as the items are going to be more understandable for participants and, thus, valuable for supporting their personal decision-making.

A response scale of 0-100 was chosen based on the Bayesian statistics used (simple comparison of means when doing the intervention vs. sticking to usual routine) and the benefit of using a continuous scale in terms of statistical power.

Enjoyment and Fit

Based on the insights we obtained from the user research suggesting participants' strong interest in habit formation, prior work suggesting the importance of enjoyment and fit into a person's life, and also our desire to support decision-making for behavior change, daily survey items were added to the tool that can provide insights that could assist participants in making an informed decision about which activity to adopt as a habit. Particularly, questions about whether they reported enjoying performing the activity they were trying, whether they found it easy/difficult to fit into their routine, and whether they were able to perform the activity were included. Exact questions are displayed in Table 4.1. Questions 3-5 were the only asked on activity days.

Altogether, participants tracked up to 10 survey items every evening. These surveys are similar to our previous work using daily surveys administered to participants indicated high compliance to surveys containing up to 10 questions (90%) administered through a smartphone app[22].

Table 4.1. Daily Survey: Wave 1

Q1. What were you assigned to do today? [Activity] Stick to my usual routine
Q2. Did you do [activity] today? Yes No
If Q1 = [Activity] and Q2 = Yes **Q3. How was your experience with [activity] today?** I enjoyed it I found it okay I did not enjoy it
Q4. If 1 = Activity, and 2 = No I felt like doing [activity] today 1= Not at all, 5 = Very much
Q5. How easy or difficult was it to fit [activity] into your day today? Very easy; Somewhat easy; Somewhat difficult; Very difficult
Q6. How focused are you feeling today? **Q7. How energetic are you feeling today?** **Q8. How stressed are you feeling today?** **Q9. How happy are you feeling today?** Slider Scale: 0 = Not at all to 100 = As much as possible
Q10. Did any event significantly affect your overall well-being today? Yes _________________ No

4.4 Daily Reminders and Prompts

Overall, participants are sent two SMS every day:

1. Every morning at 7a, a reminder to perform their activity or stick to their usual routine.
2. Every evening at 7p, a link to the daily survey, and information about their next day's assignment. This was included to give participants a heads up and time to plan the activity into their schedule.

The language used in the prompts is varied over time to avoid monotony. Examples of prompts are displayed in Figure 4.1.

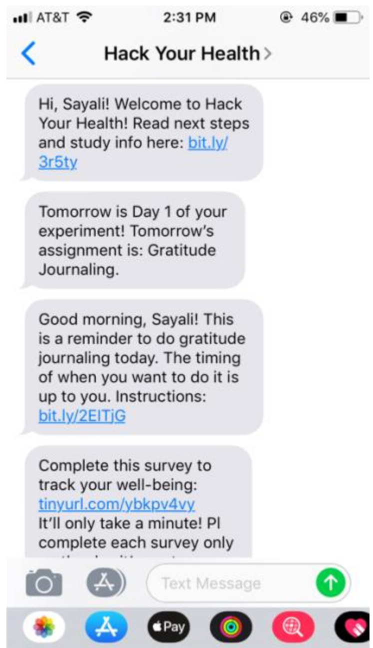


Figure 4.1. Screenshots of Daily Messages

4.5 Experimental Designs in Hack Your Health

N-of-1 studies can be carried out in multiple ways. Hack Your Health uses 'crossover designs', a category of experimental designs commonly used in clinical settings[11] in which participants are exposed to each intervention included in the study (or in this case, one intervention and baseline/usual routine). The effect of the treatment is estimated by comparing outcomes within person under each condition (A/B; explained below)[28,30]. Crossover designs can be designed in numerous ways. For this tool, the specific crossover designs and a study length we chose were based on the desire to balance scientific rigor, participant burden, and participant engagement. As mentioned earlier, N-of-1 crossover studies usually consist of an "A" phase, which can either be a baseline phase or intervention 1, and a “B” phase, or the intervention phase (or intervention 2; see formative work in Chapter 3, on why we chose a baseline to intervention comparison). An appropriate cycle length (number of days spent in each phase) can be chosen based on the intervention’s characteristics, such as potential for carry-over effects, etc[33]. Carryover effects are the effects of the intervention on outcomes of interest that are retained even beyond the crossover (when one phase ends and the next one starts).

A minimum of one phase each of intervention and baseline is required to make any basic causal claim, with the inference getting stronger with more cross-over periods[30]. Studies with at least 3 crossovers are considered strong[30]. A length of 18 days and a cycle length of 3 days was chosen to reduce participant burden and maintain engagement (for example, a length of 7 days would have led to 7 days of no activity during the baseline phase, which we hypothesized would increase the likelihood of reduced engagement with the tool). A length of 7-days would also have risked aliasing the impact with days of the week. For a cycle length of 3 days, and a study with 3 phases each for baseline and intervention (leading to a total of 18 days), 20 combinations of A and B are possible. Out of all possible combinations, we chose a subset of 12 sequences that include at least 2 crossovers and excluded sequences that include a very long baseline (e.g., AAABBB) or those with insufficient crossover (e.g., AABBBA).

Participants are not aware of the precise experimental design and are informed of their assignment on the prior evening. Since the potential carry-over effects and required washout periods for most of our intervention-outcome pairs are relatively understudied and not known a priori, the same subset of intervention schedules and same cycle length are used for all interventions. Each participant is randomly assigned to one of the sequences presented in Table 4.2.

Table 4.2. Experimental Schedules Included in Hack Your Health

1	A	B	A	A	B	B
2	A	B	A	B	A	B
3	A	B	A	B	B	A
4	A	B	B	A	A	B
5	A	B	B	A	B	A
6	B	A	A	B	A	B
7	B	A	A	B	B	A
8	B	A	B	A	A	B
9	B	A	B	A	B	A
10	B	A	B	B	A	A
11	B	B	A	A	B	A
12	B	B	A	B	A	A

A = Baseline, B = Intervention; each phase lasts for 3 days

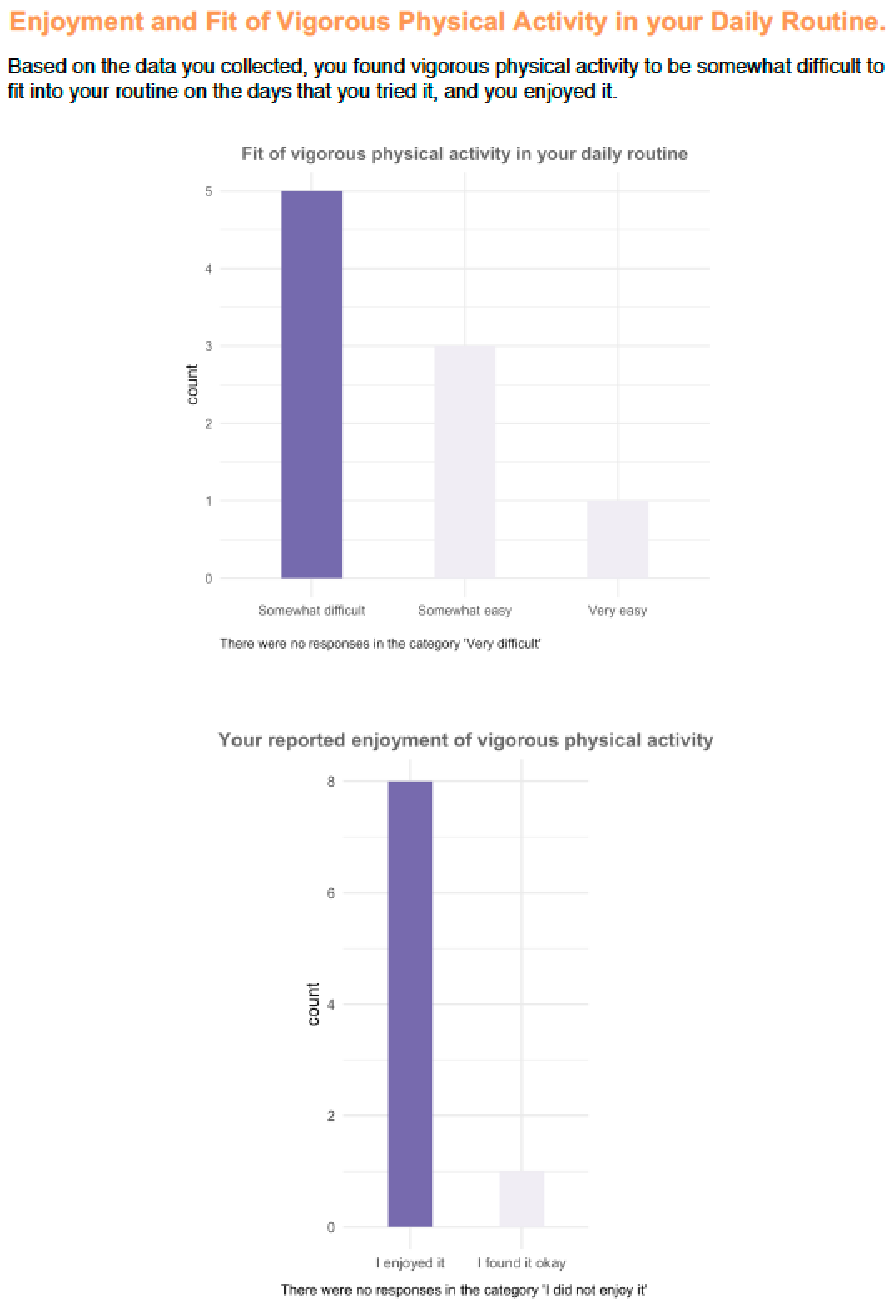
4.6 Statistical Analyses in Hack Your Health

At the end of each N-of-1 trial, each individual's trial data is analyzed in an idiographic (person-specific) manner. The data are collected via Qualtrics (Qualtrics, Provo, UT) and then passed to package nof1[106] in R[107] to estimate the treatment effect. These results are then provided to participants in simpler, easy-to-understand formats consisting of text and figures. The main comparison for each N-of-1 study is a simple comparison of means under the two conditions: intervention (B phase) and participant's usual routine (A phase). This comparison is carried out for each of the four outcomes.

Bayesian generalized linear models are fitted to provide the posterior quantiles of the estimated difference between the responses in the intervention phase vs. usual routine. A coefficient for the comparison between the 2 conditions is included in the model as the linear predictor. Models incorporate the appropriate distributions for different outcome scales and appropriate link functions connecting the expected outcome to linear predictors (in this case, normal distribution) and identity link for continuous variables. Since reliable information about possible changes in outcomes in 18 days of the study was unavailable for most intervention-outcome pairs, we used uninformative priors to the parameters in the model[33]; specifically, the prior for the intercept was N(0.01, 1000), that for the coefficient for the comparison is N(0, 1000), and that for σ^2 is an inverse Gamma distribution with the shape and the scale parameter both being 0.001.

4.7 Communication of Results

Participants are provided with an estimate of the difference between both conditions along with the 95% Bayesian credibility intervals. Results are shared with participants in the form of a PDF. Results are displayed numerically and graphically, in easy-to-understand language. An example is provided in Figure 4.2a and 4.2b


Enjoyment and Fit of Vigorous Physical Activity in your Daily Routine.
Based on the data you collected, you found vigorous physical activity to be somewhat difficult to fit into your routine on the days that you tried it, and you enjoyed it.
Fit of vigorous physical activity in your daily routine
count
0
1
2
3
4
5
Somewhat difficult
Somewhat easy
Very easy
There were no responses in the category 'Very difficult'
Your reported enjoyment of vigorous physical activity
count
0
2
4
6
8
I enjoyed it
I found it okay
There were no responses in the category 'I did not enjoy it'

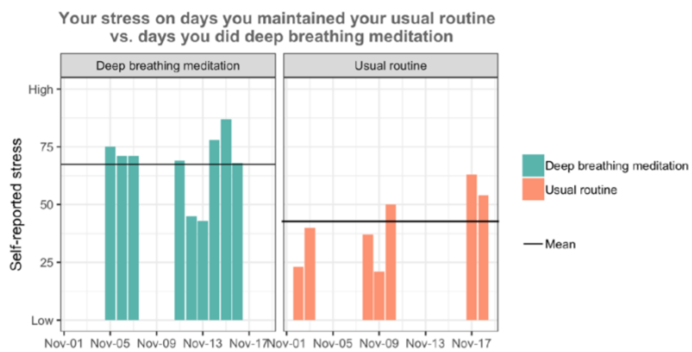
(a)

Deep Breathing Meditation and Daily Stress

Based on our analysis of your data, our best guess is that on average, your stress increased by 26 points (on a scale of 0-100) on days when you did deep breathing meditation compared to days when you stuck to your usual routine. While that number is our best guess, it's very likely that the increase in stress falls somewhere between 11 to 41 points.

Average stress on days you did deep breathing meditation: 67 points

Average stress on days that you stuck to your usual routine: 41 points



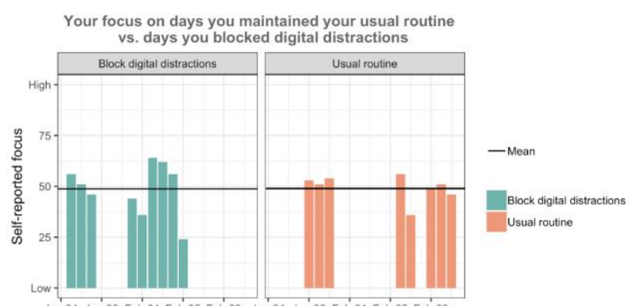
(b)

Blocking Digital Distractions and Daily Focus

Based on our analysis of your data, our best guess is that on average, your focus did not change on days when you were assigned to blocked digital distractions compared to days when you stuck to your usual routine.

Average focus on days that you stuck to your usual routine: 49 points

Average focus on days you were assigned to block distractions: 49 points



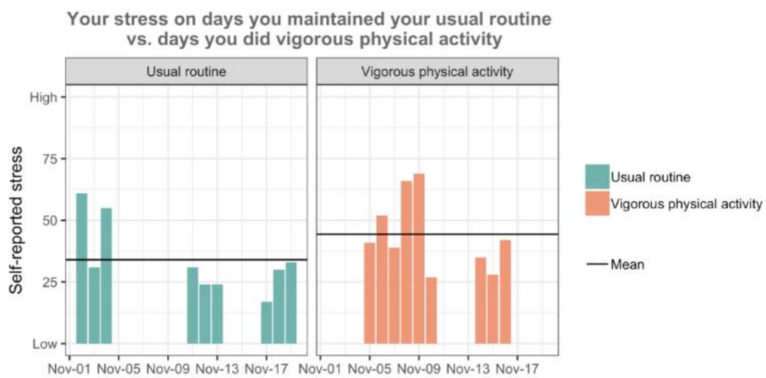
(c)

Vigorous Physical Activity and Daily Stress

Based on our analysis of your data, our best guess is that on average, your stress increased by 11 points (on a scale of 0-100) on days when you did vigorous physical activity compared to days when you stuck to your usual routine. While that number is our best guess, it's very likely that the effect could be an increase in stress of as much as 25 points, or even a decrease in stress of up to 4 points.

Average stress on days you did vigorous physical activity: 44 points

Average stress on days that you stuck to your usual routine: 33 points



(d)

Figure 4.2. Example of Results Shared with Participants

CHAPTER 5

EVALUATION OF HACK YOUR HEALTH

To be able to thoroughly examine participant experiences and perspectives to such self-experimentation, it is ideal that users first go through the entire experience of using the tool. As part of this study, users participated in an 18-day cross-over self-experiment to try an activity of their choice. For all 18 days, they tracked their psychological well-being, enjoyment and fit of activity into their routine. Once their experiment ended, we shared their personalized results with them. Once the experiment was over, participants were invited to provide qualitative and quantitative feedback in the form of follow-up surveys and interviews.

We used an iterative approach to design and evaluate this tool, wherein we first tested and obtained feedback on the version of the tool described in Chapter 4 with an initial wave of participants (Wave 1). Based on the feedback and themes identified from interviews with those participants, we modified the tool. The modified version was then deployed to a second, smaller wave of participants (Wave 2).

Specifically, in Wave 1 (N = 20), all participants tracked the same health outcomes (energy, focus, stress, happiness, enjoyment of activity, and fit of activity into their routine; Table 4.1) in a quantitative manner. Based on insights and user feedback from Wave 1, we revised the daily surveys to incorporate more reflective and personalized components, specifically, qualitative tracking of participant experience with the activity, and tracking of self-selected outcomes that are personally relevant to each person (Wave 2).

This chapter describes the methods and results for both waves. I first provide an overview of the study approach, description of participants, recruitment and setting, followed by separate sections for Waves 1 and 2 describing the measures, analyses, and findings. The next chapters include a detailed discussion of the interpretations of the results, the strengths and limitations of the approach, along with a description of the contributions to existing literature.

5.1 Wave 1

5.1.1 Overview of Research Approach

The evaluation was conducted using a mixed methods approach. Eligible participants were invited to sign up for Hack Your Health and try an intervention of their choice through an 18-day self-experiment (to test that intervention against their usual routine). Participants completed a baseline questionnaire where relevant demographic and other baseline information was collected prior to beginning their experiment. At the end of the experiment, they received personalized results (simplified interpretation of results from N-of-1 analyses, in PDF format), and were invited to provide feedback through follow-up questionnaires and semi-structured interviews.

Project aims

The overarching research question that we aimed to address through this work was: *How might we design tools that can help individuals make decisions about which behavior change strategy to adopt, in a scientific and systematic way that also meets their personal needs and motivations?*

We addressed this question through the following specific aims:

Aim 1: *Examine* ***user burden*** *and* ***adherence*** *to an 18-day self-experiment focused on assessing the impact of a behavioral intervention on the user's health and wellbeing*

Aim 2: *Evaluate the* ***usefulness of Hack your Health*** *to* support a person's decision-making related to continuing or not with simple behavioral interventions

Aim 3: *Examine individual's* ***conceptual understanding of n-of-1 study methods*** *for self-experimentation*

Aim 4: *Examine* ***the heterogeneity of individual response to interventions*** *in terms of psychological well-being, enjoyment, and fit into their life*

Why mixed methods?

A mixed-methods approach was chosen for this work for a number of reasons. Firstly, our aim was not only to evaluate the tool in terms of 'WHAT' happened, but to understand 'WHY', and 'HOW'.

We used quantitative methods when appropriate, and also complemented it with qualitative insights when available and feasible. For example, for Aim 1 we assessed the *user burden* of the system via a validated questionnaire, the User Burden Scale[98] (UBS) which is designed to be used with mobile and web-based systems. The User Burden Scale was administered at follow-up to help us uncover and quantify the type and magnitude of burden that performing an experiment using Hack Your Health placed on the participants. The *perceived usability* (Aim 2) of the system was examined quantitatively via a widely used questionnaire developed to measure usability of web-based systems, the System Usability Scale[108] (SUS). When possible, we asked participants about their extreme responses (primarily, negative responses on the UBS and SUS) during follow-up interviews to gain an understanding of the reasons behind those responses. *Perceived* *utility* (Aim 2) was examined through surveys as well as semi-structured interviews at follow-up. *Conceptual understanding of N-of-1 methods* (Aim 3) was also assessed during interviews at follow-up.

The SUS and UBS, while useful to quantitatively evaluate the overall usability and user-burden of the tool, provides only closed responses and thus, limited insights about why or how Hack Your Health was perceived to be useful/not useful. When the aim of the research is to explore and describe participant experiences, and little is known a-priori, such as in our case, a qualitative approach is a valuable complement to gain a rich, contextual understanding of participant experiences in the real-world, and through direct conversations with the users[109]. Using semi-structured interviews also allows for flexibility, such as modification of the interview guide based on insights obtained and themes that emerge during the data collection process[110].

The research method for the qualitative aspects of this study (interviews) was a descriptive case-study approach[111]. Yin[111] describes a case study as an enquiry "that investigates a contemporary phenomenon within its real-life context, especially when the boundaries between phenomenon and context are not clearly evident". This approach is appropriate for this work, because participant experiences are not expected to occur without interacting with Hack Your Health, and are expected to be highly contextual based on how it is used by the diverse sample of participants, necessitating a case study approach[111,112]. This work is considered descriptive, because it is used to describe the phenomenon (perceived utility of Hack Your Health) in the real-life context in which it was used[111]. In addition, Yin[111] also suggests that a case-study strategy may be appropriate for situations in which the intervention that is being evaluated does not have clearly defined outcomes, as in our case.

5.1.2 Participants, Screening and Recruitment

Screening Process

The target participants for this study were adults, aged 18 years of age and above, able to read and write in English, living anywhere in the United States, and using a smartphone that has regular internet access.

The activities that were offered as part of Hack Your Health are: (a) performing 10 minutes of vigorous activity, (b) writing three things you're grateful for, (c) meditating or deep breathing for 5 minutes, (d) blocking digital distractions. Most of these activities are low risk activities that could be performed by most individuals without needing supervision. Additionally, there is limited existing literature about how experiences might differ based on other demographic characteristics, warranting excluding certain individuals when exploring the study aims. Moreover, one of the overarching aims of this work is also to make the tool scalable to be used by a diverse group of individuals. With that in mind, a diverse sample may be more valuable. Hence, there were no overall exclusion criteria for this study. However, people were instructed to select an activity that they are not currently doing or have recently tried. They were also instructed to not to engage in any behaviors that would endanger their health and well-being, and warnings accompanied any activities that could be associated with increased risk for particular people (e.g., performing physical activity). Warnings were also explicitly stated in the consent form. In addition, those who selected exercise as the activity were provided a link to Physical Activity Readiness Questionnaire to help them self-assess whether they need a physician's approval before trying it.

Recruitment

Recruitment was carried out via convenience sampling using advertisements and posts on social media platforms such as Facebook and Twitter, and by contacting those who completed our online survey (as part of user research described in Chapter 3) and indicated an interest in pilot testing the tool. The posting included a brief description of Hack Your Health and this study, and directed those interested to the study website (hack-your-health.org) that provided detailed information about the study procedures. Figure 5.1 shows an example of the flyer used for recruitment.

On the website, participants could read more detailed information about the tool and the different activities currently offered. If interested, they proceeded to complete the sign-up procedures for their selected activity. Recruitment was carried out on a rolling basis. Participants were not compensated for participation in any phase of the study. This study was approved by the Institutional Review Board at Arizona State University.

Participants

A total of 22 participants consented and began their experiment. Two participants discontinued their experiment. One of the participants dropped out because she felt that the tool did not meet her expectations in terms of motivating her to actually do the deep breathing meditation. The other participant wanted to try physical activity as her intervention, but since she was already doing it, the baseline survey suggested she try another activity. She chose meditation, and within a few days of starting her experiment realized that she did not like it and hence wanted to drop out.

Twenty participants completed their self-experiments. Most participants (n= 19) completed the follow-up questionnaires and a subset signed up for interviews (n= 13).

More participants signed up to do deep breathing meditation and gratitude journaling as compared to vigorous physical activity and blocking digital distractions. Tables 5.1 and 5.2 display the breakdown of sample size by aspect of study and activity type. Participant demographics are presented in Tables 5.3 and 5.4, and in Figure 5.2.

Figure 5.1. Recruitment Flyer

Table 5.1. Sample Sizes for Different Aspects of Wave 1 (N = 22)

	Interviews	**SUS**	**UBS**	**N=1 analysis**
Completed self-experiment (n =20)	13	19	19	17*
Dropped out (n=2)	N/A	2	2	N/A

*3 participants had substantial missingness/backfilled data and so their experiments were not analyzed

Table 5.2. Sample Size by Activity (N = 20; excluding dropouts)

	n	**Interviews (n)**	**SUS (n)**	**UBS (n)**	**N=1 experiment analysis (n)**
Deep breathing meditation	8	6	8	8	7
Gratitude journaling	6	3	5	5	5
Blocking digital distractions	3	3	3	3	2
Vigorous physical activity	3	1	3	3	3

The baseline questionnaire asked participants a simple question about their current self-tracking habits, and whether they had done self-experimentation prior to the experiment. 11 out of the 20 participants reported that they do some form of self-tracking, with physical activity being the most common (n=6). Most (n=11) indicated that they hadn't tried self-experimentation prior to trying Hack Your Health.

Table 5.3. Participant Occupations

	n	Occupations
Employed, full-time	14	Software engineer; Researcher; Postdoctoral Researcher (2); Elementary School Assistant Principal; Developer; Professor; Program Coordinator; Sr. Manager; Program Manager, Education; Archivist
Employed, part-time	2	Waitress; Teaching/Research Assistant; Attorney, Dog trainer, Teacher
Self-employed, part-time	2	Actress, Nanny
Student	4	3 graduate, 1 undergraduate
Partially-retired	2	Attorney, Dog trainer, Teacher; Tasting Room Associate

Table 5.4. Participant demographics (N=20)

	N(%)	M ± SD
Age, years	-	43.28 ± 14.01
Gender		-
Female	14	
Male	6	
Education		-
Advanced degree	11	
Bachelor's degree	8	
Highschool	1	
Race		-
White	15	
African/African American	1	
Asian/Asian American	3	
Biracial	1	

Figure 5.2. Participant Locations

5.1.3 Measures

In this section, I describe the measures used to address each of the study aims.

Measures for Aim 1

User Burden

User burden was measured using the User Burden Scale (UBS) developed by Suh et al[98]. Participants completed the UBS after they completed their 18-day experiment. Those that dropped out were also asked to complete the questionnaire. The UBS measures 6 aspects of burden: (1) difficulty of use; (2) physical burden; (3) time and social burden; (4) mental and emotional burden; (5) privacy burden; (6) financial burden.

Items on financial burden were removed as they were not relevant to Hack Your Health. The UBS has been previously validated, and yielded good internal consistency for most subscales, convergent validity and concurrent validity. Although it is a fairly new questionnaire, it has been used in a few studies since its conception in 2016, and is one of the very few questionnaires that focuses specifically on user burden.

Participants were required to answer each question. For items that were assumed to be not relevant (in our case, financial burden), the response was assumed to be 0. When possible, qualitative feedback about burdensome aspects (based on extreme negative responses on UBS) was obtained during interviews at follow-up.

Adherence

Adherence was assessed in terms of completion rates for daily surveys and adherence to the experimental protocol.

Measures for Aim 2

According to Nielson[99], usability is the quality attribute that assesses how easy user interfaces are to use, while utility refers to the design's functionality, and whether the system does what the users need. Usability and utility together make up 'usefulness' of the system.

The ***usability*** of the Hack Your Health system was measured using the System Usability Scale (SUS) developed by Brooke[108], which is a 10-item scale that gives a global subjective assessment of the usability of a system. In general, it measures the effectiveness of the system, which is the ability of users to complete tasks using the system and the quality of output of those tasks; efficiency -- which is the level of resource consumed in performing tasks; and satisfaction -- users' subjective reactions to using the system. During the follow-up interviews, participants were also asked to elaborate on some of their responses on the SUS where they indicated low usability to dig deeper into the reasons behind those responses.

Perceived utility was assessed through semi-structured interviews conducted at follow-up. Based on our user research and existing literature, we identified four broad domains related to perceived utility relevant to Hack Your Health to explore via the interviews:

1) *Decision making about chosen activity*: Did the tool meet their expectations and needs in terms of influencing their decision-making process about continuing to incorporate, or not, their target behavior in their life?
2) *Aid process of habit formation*: For those were interested in habit formation, did participants view the tool as a useful step in the process of forming a habit?
3) *Tensions between statistical analyses and lived experience/intuition*: What tensions, if any, emerged between statistical analyses and users' lived experience and expectations (intuitions)?
4) *Outcomes*: Were the outcomes measured as part of the experience relevant? What kind of outcomes would they be interested in measuring?

The aspect of tensions between analyses and participants' intuition about what happened (#3 above) were also explored through surveys administered pre and post-experiment. The pre-experiment survey (Figure 5.3) asked participants to predict the effect they expect their chosen activity to have on all four aspects of well-being they tracked during their experiment. Specifically, they predicted their hunches on enjoyment of activity, fit of activity into their daily routine, directionality and magnitude of effect (including an option to say "no effect"), and their level of confidence in that hunch. At follow-up, after they had completed their experiment but before we shared results with them, they filled out the same questionnaire, this time providing their hunch on the kind of effect they thought the intervention had on the same outcomes. This allowed us to examine the discrepancies between participant perceptions and what the statistical analyses suggest in a quantitative manner.

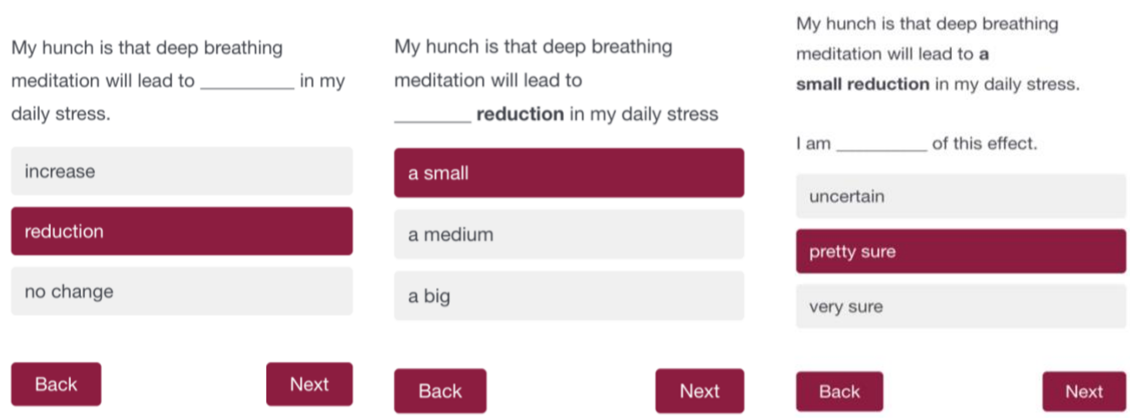


Figure 5.3. Example of Pre-Experiment Survey Asking Participants to Provide Their Hunch on The Impact of The Intervention on Outcomes Tracked in The Experiment

Interview Protocol

Interviews were conducted over the telephone and audio recorded if participants consented. Interviews lasted up to 46 minutes. The different parts of the interview protocol were designed to explore the domains outlined above as well as to explore user burden/usability issues highlighted through the follow-up surveys. A semi-structured interview structure was deemed appropriate for this work because it enables us to conduct the interview in a conversational manner. It allows flexibility in terms of the ordering and how the questions are asked[113]. I conducted all the interviews, along with taking notes during and after each interview. When necessary, the interview protocol was modified based on insights from initial interviews. This practice is recommended, as often, the first few interviews provide insights on issues with the initial protocol.

Measures for Aim 3

Through this aim, we were interested in examining whether participants have a conceptual understanding of self-experimentation as a method for generating evidence at the individual level. This aim was explored during the semi-structured interview at follow-up. Dimensions of self-experimentation that were examined were:

1) *Experimental design* (e.g., active vs. rest days; What was the purpose of rest days as you understand it?)
2) *Tracking outcomes* (e.g., Why is it important to track outcomes even on rest days?)

Measures for Aim 4

Results from all N-of-1 experiments that had enough data to analyze were used to examine intervention impact and response variability across all participants who completed the experiment. Throughout their experiment, participants also completed a daily (optional) open-ended question asking them to provide information about any event that they think may have significantly influenced their well-being that day (other than the assigned activity). Since we measured outcomes only once a day, all the experiments had a risk of being influenced by confounding variables (for example, a particularly stressful meeting or receiving some significant good news) that are difficult to control. Capturing such contextual information provided a richer understanding of the data and were used to complement the examination of the observed heterogeneity.

5.1.4 Analyses

Analyses for Aim 1

User Burden

The UBS consists of 20 items and uses two 5-point Likert scales (ranging from 0 to 4). A higher score on the scale indicates higher user burden. Overall user burden score is calculated as the sum of scores from all items, and ranges from 0 (lowest possible score) to 80 (highest possible score). User burden was calculated for the different subscales to examine the different aspects of user-burden and be able to tease apart the least and most burdensome aspects of the system. User burden across the sample was calculated in terms of mean score, median score, and standard deviations, and range of obtained scores.

Adherence

Overall adherence was examined in terms of the means, standard deviation of the means, and median of:

1. Completion rates for daily surveys (%, and days) across sample
2. Compliance to experimental protocol (%, and days)

Analyses for Aim 2

Usability

The SUS consists of 10 items and uses a 5-point Likert scale (from 0-4). The SUS score yields a single number that represents the overall usability of the system, and the score can range from 1 to 100. First, the sum of scores from individual items is calculated. For items 1,3,5,7, and 9, the score contribution is considered to be the scale position minus 1. For items 2,4,6,8, and 10, the contribution is 5 – the scale position. The sum of scores is then multiplied by 2.5 to obtain the overall system usability score. The overall usability of Hack your Health was calculated in terms of mean scores, median score, standard deviations and range across all participants.

Utility

Within this case study approach explained in section 5.1.1, a thematic analysis approach was used to identify key themes and categories emerging from each interview. Thematic analysis is defined as "a method for identifying, analyzing, and reporting patterns (themes) within data"[101,102]. This work made realist assumptions -- that the accounts provided by the participants offer insights into their actual experiences with Hack Your Health and the self-experiment[102,110]. Thematic analysis was chosen as the analysis method in place of other methods that seek to describe patterns across qualitative data because this work is not theoretically bound (although we have identified dimensions of interest to guide the data collection through user research and based on prior literature on self-experimentation)[102]. The four identified aspects described in section 5.1.3 (under Aim 2) were used to define the research as well as interview questions.

All interviews were transcribed verbatim. Data were analyzed using MAXQDA software (MAXQDA, VERBI Software – Consult – Sozialforschung GmbH, Berlin, Germany). Data analysis was initiated simultaneously with data collection. This method is recommended because it helps the researcher to get fully immersed in the data set, and begin identifying important emerging themes that can be further explored in subsequent interviews and refine the data collection process[113,114]. All interviews were initially labelled with 'codes', which are short words or short phrases that assigns meaning to a portion of the interview[102]. This process helps deconstruct the data. All identified codes, in turn, represent the larger themes present in the data. Portions of interviews that deviate from questions of interest were left un-coded. In-vivo, structural, as well as process coding were used[110,112]. In-vivo coding uses phrases from the interviews, i.e., participant's own language, for codes. Structural coding assigns conceptual phrases or content-based phrases that represent the topic of inquiry (research question). Process coding uses gerunds to code the data[115].

Data were coded two separate times and reviewed together. A codebook with all identified codes was created and organized under different identified domains.
Each interview (participant) was then examined for presence or absence of all themes identified under each domain. Data was organized in terms of count frequency of each theme under each domain.

Analyses for Aim 3

Understanding of N-of-1 methods was also explored via semi-structured interviews following the same process as explained above.

Analyses for Aim 4

Heterogeneity of response was examined based on N-of-1 analyses from the Hack your Health system using descriptive statistics, such as the directionality of response to the different interventions, reported enjoyment, and reported fit of the activity across all participants. Heterogeneity was also examined in light of the qualitative data from daily surveys.

5.1.5 Findings

In this section, I describe the findings and insights gained by organizing them under the different aims of the study. When reporting and discussing results, we refer to participants in different activities as:

Gratitude journaling: G1-G6

Deep breathing meditation: M1-M8

Vigorous physical activity: P1-P3

Blocking digital distractions: B1-B3

Findings: Aim 1

User Burden

Results from the User Burden Scale (UBS) indicated that most participants did not find Hack Your Health burdensome, and reported no burden on most items on all subscales. The mean, standard deviation, and median within each subscale of the UBS are displayed in Table 5.5.

In semi-structured interviews at follow-up, it was clear that few participants were confused by the source of the burden the scale was asking about, i.e., whether the scale was referring only to the technological aspects of Hack Your Health or the experience of the activity they tried. For example, P1's response to one of the items of the Physical subscale of the UBS, *Use of Hack Your Health is too physically demanding*, was, "a little bit of the times". When probed about it, she said: "*I think what I, because some of the questions, I could not figure out if it was asking about the set-up being demanding, or the activity. Uh sometimes, I felt that the activity was demanding, because sometimes even the 10 minutes was not easy on some days to find time for it. That's what I meant by demanding.*"

Table 5.5. Scores for Hack Your Health on UBS Subscales

UBS Subscale	Mean	SD	Median	Grade
Difficulty of Use	0.14	0.14	0	A
Physical	0.02	0.03	0	A
Time and Social	0.05	0.11	0	A
Mental and Emotional	0.03	0.03	0	A
Privacy	0.07	0.08	0	A

Additionally, one of the major problems reported by participants was related to the response scale used for the items assessing their perceived stress, energy, focus and happiness. The response scale we used was a 100-point scale (e.g., 0=Not at all stressed to 100=As much as possible). As mentioned earlier, these particular questions and response scales were used for their properties of having high face validity and being continuous (to increase statistical power). However, participants found it "*hard to quantify*" their state on that 100-point scale. B3, for example, said he found it difficult to differentiate between "*72% happy today vs. 78% happy the next day*" and that it felt somewhat "*arbitrary*". For some, it was using a 100-point slider-scale on their phone screen that made it difficult to be accurate.

Adherence

Adherence and completion rates are reported in Table 5.6. Overall, the average survey completion rate was 88.89% (SD = 12.13, Median = 88.88), which is an average of ~16 days of completed surveys out of 18. Four participants had a 100% completion rate.

The average adherence rate (sticking to the experimental protocol) was 69.44% (SD = 27.42, Median = 77.28) with 4 participants who had a 100% adherence rate. This was equivalent to participants performing the activity as assigned on 6-7 of the 9 activity days. Other aspects of adherence are presented in Table 5.6. At follow-up, 37% of participants reported that they found it somewhat easy to stick to the day's assignment, 21% found it very easy, 26% found it somewhat difficult, and 16% found it very difficult (Figure 5.4).

Most (95%) participants reported some significant event affecting their well-being on at least one of the days of the experiment. The average number of days with reported significant events affecting their well-being was 4.6 days (SD = 2.11, Median = 4.5).

Visualization of the missingness in data over time (Figure 5.5) across all participants suggested that missingness may have increased slightly over time, around days 13/14. It is important to note that the high missingness seen on Day 1 could be due to a technical error with surveys that occurred at the beginning of the experiment (majority of participants started their experiments on the same date).

Table 5.6. Adherence and Compliance Rates

	Mean	**SD**	**Median**	**Range**
Missingness (days)	2.15	2.18	2	0-7
Missingness, activity days	1.25	1.71	1	0-6
Completion Rate (days)	15.85	2.18	16	11-18
Completion Rate (%)	88.89	12.13	88.88	61-100
Adherence to experimental protocol (days)	6.25	2.47	7	1-9
Adherence to experimental protocol (%)	69.44	27.42	77.78	11.11-100
Days with a significant event	4.6	3.38	4.5	0-11
Activity days with significant events	2.55	2.11	2	0-6

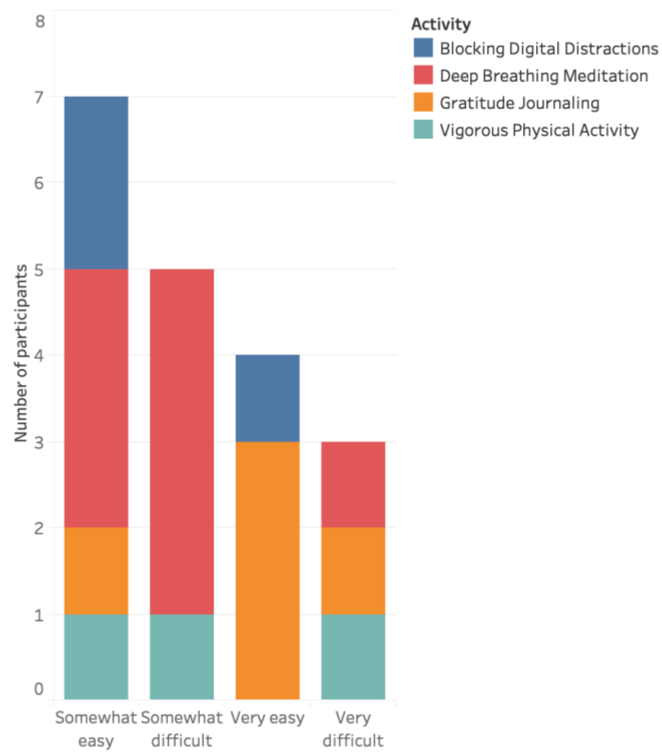


Figure 5.4 Perceived Ease/Difficulty of Sticking to the Day's Assignment

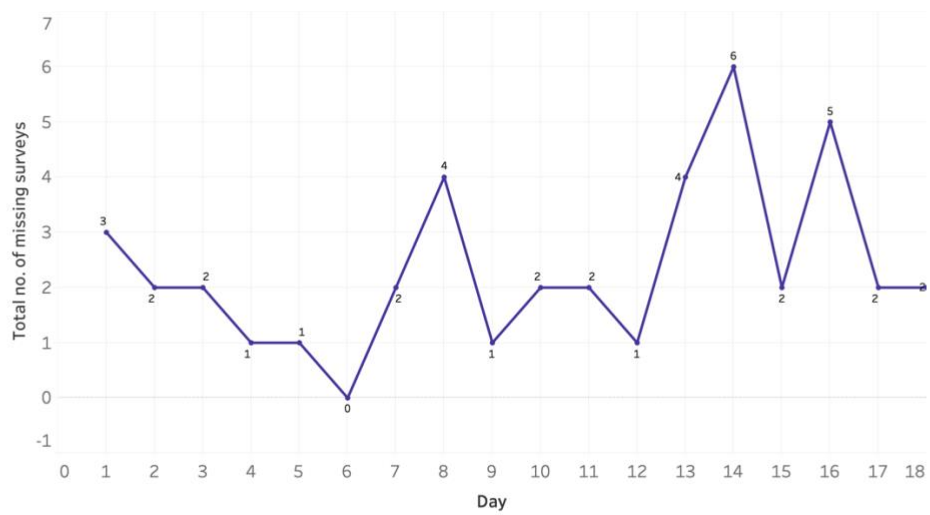


Figure 5.5. Missingness Over Time

Findings: Aim 2

Usability

System usability, measured using the SUS, indicated a mean score of 90.12 out of 100 (SD = 10.37), and a median score of 93 out of 100, indicating high usability overall.

Out of the items on the SUS, lower usability scores were on the items "I think I would like to use Hack Your Health frequently", and "I found the various functions of Hack Your Health were well integrated". When probed during interviews, participants noted difficulties with the sliding scale in the daily survey, and the survey links not working correctly on a couple days.

Participant reasons for not wanting to use it frequently were as would be expected. Many participants interpreted that item as referring to frequently using Hack Your Health to try the same activity they already tried, and their response indicated that they wouldn't go through a self-experiment for the same activity again. For example, B2, when probed about their response, said, "*Umm, so when I thought about using Hack Your Health, I thought about the 6-hour blocking of apps, that's what I interpreted it as. So, I don't know if I'd go as far as to block my apps for 6 hours outside of the experimentation period".* Most participants said they'd use it if they want to experiment with something else, or if there's a different behavior they are interested in.

Utility

In this section, I describe the results from follow-up interviews (n=13), and supplement this information with baseline and other participant information when relevant. Insights from interviews are summarized under four domains based on themes that were identified in the interviews. Due to the semi-structured nature of the interview, at times, not all questions were asked to each participant. Frequency counts are provided for the different themes; however, the denominator should not be assumed to be 13.

Themes were organized under the following domains:

1. Motivations behind wanting to try selected activity
2. Experiment structure
3. Outcomes
4. Results and Decision-Making

Domain 1: Motivations behind wanting to try selected activity

Participants' initial motivations behind trying their selected activity were assessed at baseline using a multiple-choice questionnaire with 8 options to choose from (options shown in Table 5.7), where participants were asked to select all options that apply to them. These options were selected based on insights from user research (Chapter 3). Most frequent motivations included wanting to experiment and see if it improves their well-being (N=20), wanting to make it a habit (n =14), not having been able to adopt the activity in spite of multiple tries (n = 10), because research says it works (n=9), and because they haven't tried it before and want to see how they feel (n=8). Motivations for all participants are presented in Table 5.7.

Similar to what was indicated through the baseline assessment, interviews also suggested that participants used this tool not only as a way to figure out intervention impact but also as a source of accountability to help them get back into an activity, make it part of their routine, or to initiate an activity they had been wanting to adopt or try.

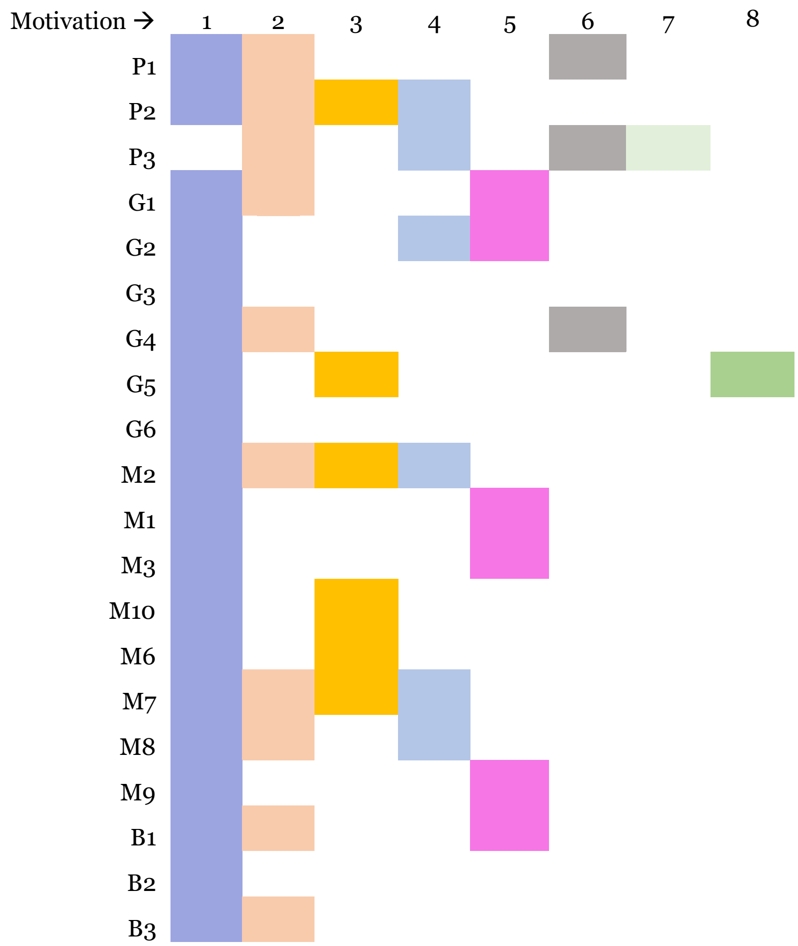
a) Intervention impact

A small portion of participants used the self-experiment primarily to figure out if the activity was impacting their well-being (n=5). For example, participant B1 who believed she uses apps and websites on her phone to distract herself from the work she should be doing, said that blocking digital distractions through an experiment was a way to "*test whether it really was adding stress and adding distractions in a meaningful way to my daily life*."

b) Accountability

While all interview participants reported that they were trying the activity to test and see if it improves their well-being when signing up, follow-up interviews indicated that was not the primary goal for all. Participants also saw Hack Your Health **as a tool that can help them get back into an activity they had previously engaged in** (n=3), **initiate a behavior they had been wanting to incorporate in their life** (n=3), or to **make the activity a part of their routine** (n=5). For example, G4, who had tried gratitude journaling before and remembered liking it, when talking about why the tool interested her, said, "*I was trying to see if I could get back into it, and I was like wow this is like the perfect opportunity for me to try to get back into it because it did help when I tried it the first time*". B3, who had been wanting to spend less time on social media saw Hack Your Health as *"good extrinsic motivation to supplement my intrinsic motivation".*

Table 5.7. Baseline Motivations Behind Trying Selected Activity (N=20)



LEGEND

Motivation	Key	Freq.
I want to experiment and test if it improves my well-being	1	19
I want to make it a habit	2	10
I haven't been able to adopt this habit in spite of multiple tries	3	6
Research says that it works	4	6
I haven't tried it before, want to see how I feel	5	6
I remember liking it and want to do it again	6	3
Some other reason	7	1
Somebody asked me to try it	8	1

Domain 2: Experiment structure

a) Reminders and surveys

Most participants **reported really liking the daily SMS that reminded them to do the activity** (n=11), and many **reported liking the experiment especially because of the accountability** through the reminders and daily surveys (n=11):

"*I think that doing this experiment was great because of the accountability of it, and like, if you're putting a new habit in your life, and you want to have some kind of change, having that text message to check-in every morning or every afternoon, it keeps you accountable to committing to that change and supports you, and eventually may be you wouldn't need a check-in because it's a part of your life."* (G5)

"*Best thing about it was getting the message every morning to remind me to do it, and then knowing that at the end of the day I would have to fill out the survey about how I did that day."* (M9)

"I felt, having the experiment and having the sort of accountability of having to fill out the surveys at the end of the day every day, umm, made me more committed to actually sticking to it, whereas, I've never tried this seriously before, but like, other times where I have tried to limit any way, if I don't have that sort of responsibility to report back, there is no consequence for not doing it, it's harder to stick with." (B1)

Participants also mentioned liking the **daily surveys** they completed at the end of each day **as a check-in and "*outlet to reflect*"** (G5) and to track how they were feeling (n=5). G4 said that the daily surveys "*just put it into perspective for me, like what events were happening that day and how it did affect my well-being*."

b) Experimental Design

Hack Your Health uses a cross-over N-of-1 experimental design where participants perform the activity for half of the days and stick to their usual routine for half of the days. Each participant was randomly assigned to one of the 12 experimental schedules. Most participants did not note any issues with the experimental design.

For some, the **resulting irregularity in performing the activity negatively affected their experience** (n=3). G5 who did not like being randomly assigned to the activity, nor having it be easily scheduled said, "*For me, if it was a regular thing I did every day, I think it would have been much more, long term effects for me. I think it would have been like, oh, this is my moment of reflection and then I could have that feeling carry through. But because it was sometimes so rushed and because it was sporadic, I don't think I had that long-term effect.*" M9 felt that the sporadic nature of doing deep breathing meditation "*made it difficult for me to feel like I was getting started.*"

Participants were notified the previous evening of their next day's assignment. Some noted that the **receiving information about the random assignment the evening before made it difficult for them to do the activity and fit it into their day** (n=3). G2 also mentioned not picking physical activity to try because "*I kind of have to plan ahead when I'm going to do that. So that's part of why I didn't pick that one*". M3, who had a low adherence to the experimental protocol suggested that he would have liked a step after signing up where he could put the schedule in his calendar "*I wonder if having an initial step where you put it in your calendar, so you know, okay, this is the* *time that I'm going to do it every evening so once it's in your calendar, at least that's how I work, I set that time aside, and I'll try not to schedule around it."*

G5 also felt that her enjoyment of the activity was lower because of it not being part of her schedule: "*I do remember seeing, 'did you enjoy it', and it's like, you know, it's hard, because I would have enjoyed it, I think. I believe that I would have enjoyed it had I made it a routine that I actually was sticking to. But I didn't enjoy it when I was stressed out, and I was like, oh crap, I gotta do this gratitude thing, you know. It goes back to not scheduling it and not making it a priority. It became a chore instead of something that is an intentional moment of being connected, it became a chore*."

Some participants felt that the experiment duration was not long enough (n=4). For some, it was because of wanting to make it a part of their routine. Like P1, who wanted to get back into physical activity: "*My hope was that it somehow made that switch in my head that yes I need to separate time for this and do it, but eventually at the end of these days that I participated in your study, after finishing that, I didn't have that implemented in my [routine], it didn't change my behavior, I think that's what I was hoping to but it did not*."

M3 wanted a longer duration because it would better incorporate potentially missing data: "*I think I would want to be over a longer period of time, just because I know I will miss days, just the nature of my day to day life".*

Domain 3: Outcomes

Most participants said that overall, outcomes that were tracked as part of Hack Your Health were relevant to them. But for some, **only a subset of the outcomes tracked seemed relevant** (n=8). For some, it was because they felt that other aspects of their life affected certain outcomes much more than the activity. For example, M10 felt that happiness wasn't as important of an outcome for her to measure while trying deep breathing meditation *because, "I feel like there were so many other factors that were going into that on a daily basis that I wasn't sure that it would be affected by meditation."*. For others, it was because it was not an outcome that they felt needed improvement: *"I never had an issue dealing with stress."* (P1)

Participants also noticed differences (as a result of the activity) in outcomes other than the ones measured in the experiment (n=6). For example, M10, when describing effects that she noticed when she did deep breathing meditation, said, "*I would say the other area that I noticed a difference in is sometimes I have like, this off and on chronic pain issue, and sometimes on days when I'm more stressed, it's even worse. So, I could notice the difference in the pain too with the deep breathing and that's something that I was like really excited to find and something that I would definitely use moving forward for that."* P1 said: "*I believe I slept better on days I was doing vigorous physical activity. I also felt I needed to go to bed earlier on those days."*

In some cases, **the outcome that the person cared about was the activity itself** (n=3). For example, for B2 and B3, the outcome was to spend less time on their phones. B3, when talking about the outcomes he cared about said "*biggest thing is the autonomy with blocking the distractions"*.

Some of the outcomes participants noticed differences in may have also been more proximal and the effect may have been short-lived (n=7). M3 said, "*It felt like, hitting a reset button type thing. I guess I just felt more relaxed, less stressed. Yeah, I think there are sort of two parts to that, one is like the... so the effect was immediate and may be it was short lived."*

"I did feel at least right afterwards like, okay I've got the rest of the day, I can go through and do the rest of my stuff here today now." (M7)

"*I definitely noticed, you know it would kind of slow my heart rate down, and especially on days when I was already stressed, and I kind of didn't realize until I started the meditation, like oh, my heart is beating faster than normal because I'm feeling a lot of anxiety. And then after I did the meditation, it forced me to not have that physiological anxiety response that I think also helped me to calm down a little bit.*" (M10)

Domain 4: Results and Decision-Making

An important thing to note was that **5 of the 13 participants had not gone over the PDF of results that we shared with them before their interviews**. When we shared results with participants, we had not explicitly asked them to go over results before their follow-up interview. It was an assumption that we made that participants would go over their results, but quickly realized that that was not the case, and an important insight in itself.

a) Intuitions about activity's impact did not match results

This aspect was explored through surveys administered pre- and post-experiment (explained in section 5.1.3; Figure 5.3). The pre-experiment survey asked participants to predict the effect they expect their chosen activity to have on all four aspects of well-being they would be tracking during their experiment. Specifically, they predicted their hunches on the directionality and magnitude of effect (including an option to say "no change"), and their level of confidence in that hunch. At follow-up, after they had completed their experiment but before we shared results with them, they filled out the same questionnaire, this time providing their hunch on the kind of effect they thought the intervention had on the same outcomes. This allowed us to quantitatively examine participant expectations, subjective experience, and the discrepancies between participant perceptions of the impact and what the statistical analyses suggest.

Table 5.8 displays the comparison between participant intuition about directionality and the certainty of the activity's impact after they completed the experiment, and the directionality as indicated by the results. Overall, participants were 'pretty sure' or 'very sure' about their hunch on the effect they thought that the activity had on all outcomes (Stress: 15/17, Focus: 12/17, Happiness: 12/17; Energy: 14/17). In the table, red pairs under each outcome indicate a mismatch between directionality of participant intuition and results, while green pairs indicate a match. Overall, for most participants, their intuition about the effect did not match with results (Stress: 3/17 matched; Focus: 6/17 matched; Happiness: 7/17 matched; Energy: 2/17 matched). Table 5.9 displays the comparison between participant predictions about directionality before they began their experiment, and intuition about what happened after they completed the experiment. The table also displays intuition certainty post the experiment. Overall, participants' own prediction and post-experiment intuitions matched more times as compared to post-experiment intuition vs. results. In the table, red pairs indicate a mismatch between directionality of participant prediction (prior to experiment) and participant intuition about what happened (after the experiment) (Stress: 11/17 matched; Focus: 10/17 matched; Happiness: 8/17 matched; Energy: 5/17 matched).

Table 5.8. Comparison of Intervention Effects: Participants' Post-experiment Intuition vs. Results*

	Stress			Focus			Happiness			Energy		
PID	**Certainty Intuition**	**Direction Intuition**	**Direction Results**	**Certainty Intuition**	**Direction Intuition**	**Direction Results**	**Certainty Intuition**	**Direction Intuition**	**Direction Results**	**Certainty Intuition**	**Direction Intuition**	**Direction Results**
22	uncertain	↓	↑	uncertain	↑	↑	uncertain	↑	↑	pretty sure	—	↑
21	pretty sure	—	↑	pretty sure	—	↓	pretty sure	—	—	pretty sure	—	↑
210	pretty sure	↓	↓	uncertain	↑	—	uncertain	—	↑	pretty sure	—	↑
25	pretty sure	↑	↑	pretty sure	—	↓	pretty sure	—	↓	pretty sure	—	↓
27	pretty sure	↓	↑	pretty sure	↑	—	pretty sure	↑	—	uncertain	—	—
28	pretty sure	—	↑	pretty sure	—	↓	pretty sure	—	↓	pretty sure	—	↓
29	pretty sure	↓	—	uncertain	↑	↑	uncertain	↑	—	pretty sure	—	↑
11	pretty sure	↓	↑	pretty sure	—	—	pretty sure	↑	↑	pretty sure	—	↑
12	pretty sure	—	↑	pretty sure	—	↓	pretty sure	—	—	pretty sure	—	↑
13	pretty sure	↓	↑	very sure	—	↓	pretty sure	—	—	pretty sure	—	↑
14	pretty sure	↓	—	very sure	—	—	very sure	—	↓	very sure	—	↓
15	pretty sure	—	↑	pretty sure	—	↑	pretty sure	—	↓	pretty sure	—	↑
41	uncertain	↓	↓	uncertain	↑	↑	uncertain	—	—	pretty sure	—	↓
42	very sure	↓	↑	very sure	↑	—	very sure	↑	—	very sure	↑	↓
31	pretty sure	↓	↑	pretty sure	—	—	uncertain	↑	↑	uncertain	—	↑
32	pretty sure	↓	↑	uncertain	—	↓	pretty sure	↑	—	uncertain	—	↓
33	pretty sure	↓	—	pretty sure	—	↑	pretty sure	—	↑	very sure	↑	↑

*Red pairs indicate a mismatch between directionality of participant intuition and results, while green pairs indicate a match. Certainty of the intuition is how certain participants about the effect the intervention had on the respective outcomes.

Table 5.9. Comparison of Intervention Effects: Participants' (pre-experiment) Prediction vs. Post-experiment intuition**

	Stress			Focus			Happiness			Energy		
PID	Certainty Intuition Post	Direction Pre	Direction Post	Certainty Intuition Post	Direction Pre	Direction Post	Certainty Intuition Post	Direction Pre	Direction Post	Certainty Intuition Post	Direction Pre	Direction Post
22	uncertain	↓	↓	uncertain	↑	↑	uncertain	↑	↑	pretty sure	—	—
21	pretty sure	↓	—	pretty sure	↑	—	pretty sure	↑	—	pretty sure	—	—
210	pretty sure	↓	↓	uncertain	↑	↑	uncertain	↑	—	pretty sure	—	—
25	pretty sure	—	↑	pretty sure	—	—	pretty sure	—	—	pretty sure	—	—
27	pretty sure	↓	↓	pretty sure	↑	↑	pretty sure	—	↑	uncertain	—	—
28	pretty sure	↓	—	pretty sure	↑	—	pretty sure	↑	—	pretty sure	↑	—
29	pretty sure	↓	↓	uncertain	↑	↑	uncertain	—	↑	pretty sure	—	—
11	pretty sure	↓	↓	pretty sure	↑	—	pretty sure	↑	↑	pretty sure	—	—
12	pretty sure	↓	—	pretty sure	↑	—	pretty sure	↑	—	pretty sure	—	—
13	pretty sure	↓	↓	very sure	↑	—	pretty sure	↑	—	pretty sure	↑	—
14	pretty sure	↓	↓	very sure	—	—	very sure	↑	—	very sure	—	—
15	pretty sure	↓	—	pretty sure	—	—	pretty sure	↑	—	pretty sure	—	—
41	uncertain	↓	↓	uncertain	↑	↑	uncertain	—	—	pretty sure	—	—
42	very sure	—	↓	very sure	↑	↑	very sure	—	↑	very sure	—	↑
31	pretty sure	↓	↓	pretty sure	↑	—	uncertain	↑	↑	uncertain	↑	—
32	pretty sure	↓	↓	uncertain	—	—	pretty sure	↑	↑	uncertain	↑	—
33	pretty sure	↓	↓	pretty sure	↑	—	pretty sure	↑	—	very sure	↑	↑

**Red pairs indicate a mismatch between directionality of participant prediction (prior to experiment) and participant intuition about what happened (after experiment). Certainty of the intuition is how certain participants about the effect the intervention had on the respective outcomes.

b) Decision based on intuition, not results

Data suggest that **the experience overall helped some participants figure out whether the activity is worth continuing with** (n=7). M5, who felt an increase in anxiety as a result of deep breathing meditation and said "*I want to try different kinds of meditation. I feel like it should work. I like the idea of meditation, and I have so many friends that are like, you know, they meditate and it's great. I'm like, okay, I'd like to try that, I'd like to feel those benefits. So, I'm going to keep trying. But I think that, my conclusion is, just the sitting and breathing didn't really work for me."* This may not have been true for those participants who already knew they liked the activity or that it works (n=3).

However, the decision was not always based on outcomes they measured via Hack Your Health or the results we shared once their experiment concluded. Although the process seemed to have helped in decision-making, **participants' decision about whether the activity is worth continuing to do or not, was often based on what they thought happened and on subjective experience rather than the results** we shared (n=11).

P1, whose results indicated very little or no change in all outcomes tracked, when asked about whether the results impact how she feels about physical activity said, "*No it does not. If I had the time and motivation I would still do it even if it doesn't have a huge impact that I would expect it to have*."

B3 had not looked at the results we shared, and said that he "*had a very good idea that it would help*" and results at the end would not make a difference. He said "*Yes I definitely found that the days when I was conscious about blocking my apps, I felt like I had more autonomy over my time and the decisions I made during that time. So, I felt more focused, I definitely felt less stressed, more happy in general*."

In line with what the results from the surveys comparing participant intuition and results suggest, **some participants reported being surprised at their results when they did not match their intuition** (n=5). M7 felt that meditation helped when he did it in the middle of the day: "*I definitely felt more relaxed, I definitely felt more focused or at least was able to look at the rest of my calendar for that day and go, oh okay I can get this done, I can get this done, and this is probably going to have to go on my tomorrow's list to be done first and that kind of stuff. So, I probably did the 5 minutes of breathing and meditation and then spent 2-3 minutes looking at my schedule and making some adjustments for the next 24 hours because I had that time to do that and had a better sense of what I was going to be able to accomplish for the day, at least related to work.*",

He was surprised when looking at the results: "*I looked through it briefly and I was a little surprised at my numbers seemed the same across a lot of the things, I was like okay, well maybe that means that it didn't help, or maybe I wasn't good at reporting myself accurately too, I don't know.*" At the time of the interview, he said "*I kept doing it. I've pretty much have been doing it every day even though the experiment stopped a week ago.*"

G2, who had discontinued gratitude journaling after the experiment, said, *"I didn't necessarily notice that it changed anything on the days that I did it versus not, umm, it did make me do it so that was good I guess."* She was surprised that results suggested an impact on focus and energy, and tried to think of reasons other than gratitude journaling that might be related to the difference: "*the results also show basically no change or may be higher ratings when I didn't do it, which is why I thought maybe it was a time, maybe I just started giving higher ratings over time and at the end I wasn't doing it, but I don't know if that's true.*"

c) Participants know things data don't

The daily survey also asked participants about any other event that significantly impacted their well-being that day. Overall, participants had 4-5 days (median = 4.5 days) out of 18 where they reported that something significant had affected their well-being. These data and the interviews shed light on many aspects of participants' context and their experience with the study that likely affected their data and results but was not accounted for in the analyses. The analyses we used did not factor in covariates, and neither was the study duration long enough to control for outliers.

Some participants (n=4) noted that at times**, they were prompted to do the activity because of the evening SMS** (with the daily survey), and if they had not done it by then, they did it in a hurry because they wanted to do it before the survey, because the survey would be asking them about whether they did the activity. Not surprisingly, this seemed to be the case especially with meditation and gratitude journaling, both of which potentially require less time as compared to vigorous physical activity and blocking distractions. Additionally, overall, data suggest that the **time of day that participants did the activity tended to be different over the duration of the experiment, while the survey was administered at the same time each day**. Two participants also reported feeling **stressed because of the pressure of having to fit the assigned activity** (M7 and M8) into their day. Two of the three (B2 and B3) **participants who tried blocking digital distractions as their activity also reported subconsciously reducing their use of their phone and social media even on control days**. B2, when talking about days she was assigned to maintain her usual routine said, "*I was like, yesterday I was able to do it, so why should I do anything different today*."

Findings: Aim 3

Most participants (n=11) correctly stated what they thought the purpose of the self-experiment was. M9 said that he thought the purpose of the experiment was to find out "*whether or not it was beneficial to do meditation on a regular basis*."

"So for me, I was doing the deep breathing meditation and then afterwards getting feedback on how that went in your daily life, changing different health related outcomes." (M10)

Two participants thought of this a tool to help them form a habit "*I would say that, I thought was an experiment to kind of use self-tracking in sort of, and also little bit of accountability elsewhere when you're like reporting back at least with like the survey function in order to start building a habit that you're interested in forming*."

"*To help people improve their lifestyle and get rid of bad habits but usually, I mean I've been wanting to do this for a while and I just didn't have the incentive and this gave me the incentive to do it. So, I think that's that, what my sense is, that's what this tool would do, help people get rid of bad habits that they have, or at least develop good habits*." (B2)

Most had a sense of the purpose behind days when they were asked to stick to their usual routine: "*I guess it was to test out a new behavior that might help with health or well-being -- on days you do it, do you feel better than days you don't do it.*" (G2).

While participants seemed to have a conceptual understanding of what they were trying to do through the self-experiment, other behaviors that were observed, such as issues with compliance to experimental protocol, maintaining the fidelity of the intervention (e.g., doing gratitude journaling in a hurry), and backfilling also suggest that they may not have a clear understanding of effects of such behaviors on the fidelity and validity of the experiment.

Findings: Aim 4

Results from the individual analyses were used to examine the directionality of effect for all interventions across all participants (displayed in Table 5.10; arrows represent directionality of change: increase or decrease, and flat line represents no change). Overall, **results indicate that there was considerable heterogeneity across participants for the different intervention-outcome pairs, suggesting that not all interventions resulted in improvement for all participants**. For example, gratitude journaling led to increase in happiness for 2 participants, and decrease in happiness for 2 participants and resulted in no change for 1 participant. Intervention response for other intervention-outcome pairs is displayed in Table 5.10. Additionally, individual-level results also suggest that the same intervention could have both, a positive and a negative impact for a given person depending on the outcome (one could look at Table 5.8 that illustrates this). However, in light of the short duration of the experiment and other limitations of the experimental design described earlier, these insights should be interpreted with caution. Additionally, all interventions had a low sample size, and vigorous physical activity and blocking digital distractions had only 2 and 3 participants each. As mentioned in the section describing findings for Aim 2, the experiment did not control for confounding variables and contextual factors that affected peoples' experience with the activity.

Overall perceived enjoyment and fit for the same intervention varied across participants; not all participants enjoyed the interventions or found it easy to fit their routine. Enjoyment and fit across all interventions are reported in Table 5.11. While these values are the most frequently reported category, participants' reported enjoyment and fit also varied over time.

Table 5.10. Directionality of Intervention Effect Across Participants

	Stress			Energy			Happiness			Focus		
	↑*	↓	—	↑	↓	—	↑	↓	—	↑	↓	—
Gratitude journaling	4	1	0	4	1	0	1	2	2	1	2	2
Deep breathing meditation	5	1	1	4	2	1	2	2	3	3	3	1
Blocking digital distractions	1	1	0	0	1	1	0	0	2	1	0	1
Vigorous physical activity	1	1	1	2	1	0	2	0	1	1	1	1

*Arrows indicate directionality of effect. Flat line represents 'no change'.

Table 5.11. Perceived Enjoyment of Interventions

	Enjoyed it	Enjoyed it/ Found it okay*	Found it okay	Did not enjoy it
Gratitude journaling	2	3	0	0
Deep breathing meditation	2	3	0	2
Blocking digital distractions	0	0	2	0
Vigorous physical activity**	2	0	0	0

*equal number of responses in both categories
**One participant did not do vigorous physical activity on most assigned days and reported not feeling like doing it

Table 5.12. Perceived Fit of Interventions into Participants' Routine

	Very easy	Somewhat easy	Somewhat easy/Somewhat difficult*	Somewhat difficult	Very difficult
Gratitude journaling	2	2	0	0	1
Deep breathing meditation	1	2	1	2	1
Blocking digital distractions	2	0	0	0	0
Vigorous physical activity	0	0	0	2	1

*equal number of responses in both categories

5.2 Wave 2

Based on insights gathered in Wave 1, we made a few changes to the tool and deployed the modified version to a smaller, second wave of participants. As mentioned in previous chapters, the first iteration of Hack Your Health indicated that participants' unquantified subjective experiences with the activity were important in their decision about continuing or not with the activity. Data also suggested that these effects might be proximal and short-lived/acute, and that participants' timing of performing the activity might have differed substantially from day to day. Participants also noticed differences in outcomes other than ones we measured in the experiment, and that not all outcomes were relevant to them. Additionally, many participants suggested that the evening survey time of 7p often interrupted their evening activities. Data also suggested that the irregularity in performing the activity may have negatively affected participants' experience with the activity. These were the insights that informed the modifications to the tool in Wave 2.

This section describes the specific changes we made to the tool, overall aims and approach used, followed by sections describing the findings. The protocol and methods used in this study were almost the same as those used for Wave 2. Methods are described in detail only when they differ substantially.

5.2.1 Changes to the Hack Your Health system and protocol

Changes were limited only to a few aspects of the daily surveys. For this exploration, we decided to make minimal changes to the original version of the Hack Your Health system so that it would also allow us to comparatively examine themes and other findings from Wave 1. The changes that were made were:

1. ***Timing of daily surveys***: Based on feedback from multiple participants, the evening SMS was sent to participants at 8p instead of 7p.
2. ***Survey questions****:* The final survey in Wave 2 included four questions in addition to those included in Wave 1. The final questions are displayed in Table 5.13. Questions 4,5,7 and 13-14 are the ones added in Wave 2. Two *types* of questions were added to the daily survey:
   i) ***Tracking self-selected and personally relevant outcomes***: At baseline, when signing up for the study, participants were asked to list two aspects of their well-being that they personally care about to track during the experiment (Figure 5.6). We then developed questions and response formats to track the outcomes chosen by participants and asked them to confirm whether those questions seemed appropriate. Once participants confirmed, questions were added to their daily surveys and their experiment was initiated.
   ii) ***Open-ended reflection about experience with activity***: On days that participants were assigned to do the activity and did it, they were asked to reflect on whether they thought it impacted their well-being that day (Q5 in Table 5.13). If they did not do it, they were asked to reflect on what kept them from doing the activity that day (Q7 in Table 5.13).

In addition to the outcomes you just answered questions about, we would also like to track more aspects of your health that you personally care about and are relevant to you.

Please list two aspects of your well-being that you'd like to track during your experiment to see if [ACTIVITY] affects them.

Your chosen outcomes can be something that is an observable action, or psychological state, whatever you prefer.

1. **Observable -** such as "getting through my to-do list for the day" or "spending more time with family" or "hours of sleep"
2. **Psychological -** such as sadness, fatigue, serenity, anxiety, etc.

Don't worry about HOW they can be measured. Leave that to us. We'll do our best to come up with a simple question to track it.

1st aspect of well-being _____________
2nd aspect of well-being_______________

Figure 5.6. Question from Baseline Questionnaire Asking Participants to List to Personally Relevant Questions

5.2.2 Overview of Research Approach

Beyond Aims 1-3 from Wave 1, we also sought to examine the implications of including participant-chosen outcomes and open-ended reflection on participant experience with Hack Your Health. As mentioned earlier, for this study, we only made these minimal changes to the version of the tool used in Wave 1 so that it would also allow us to comparatively examine themes and other findings from Wave 1. Specifically, the added aims for this study were:

Aim 5**:** Examine participant experience of tracking self-selected outcomes and of open-ended reflection of perceived impact of activity

Aim 6: Examine perceived utility of the 'usual routine' days

Table 5.13. Daily Survey: Wave 2

Q1. What were you assigned to do today? Activity Stick to my usual routine
Q2. Did you do gratitude journaling today? Yes No
If Q1 = Activity and Q2 = Yes **Q3. How was your experience with [activity] today?** I enjoyed it I found it okay I did not enjoy it
If Q1 = Activity and Q2= Yes **Q4. Approximately what time did you do [activity] today?** __________
Q5. Did you feel that doing [activity] affected your well-being today? In what way? _______________________________
Q6. If Q1 = Activity, and Q2 = No: I felt like doing gratitude journaling today 1= Not at all, 5 = Very much
Q7. What kept you from doing [activity] today? _______________________________
Q8. How easy or difficult was it to fit [activity] into your day today? Very easy; Somewhat easy; Somewhat difficult; Very difficult
Q9. How focused are you feeling today? **Q10. How energetic are you feeling today?** **Q11. How stressed are you feeling today?** **Q12. How happy are you feeling today?** Slider Scale: 0 = Not at all to 100 = As much as possible
Q13 and Q14: Questions on two personally chosen outcomes
Q15. Did any event significantly affect your overall well-being today? Yes ________________ No

5.2.3 Method

Besides the changes stated in section 5.2.1, the target participants, overall protocol and recruitment process for Wave 2 was identical to Wave 1: The evaluation was conducted using a mixed methods approach. Eligible participants were invited to sign up for Hack your Health and try an intervention of their choice through an 18-day self-experiment (to test that intervention against their usual routine). Participants completed a baseline questionnaire where relevant demographic and other baseline information was collected prior to beginning their experiment. At the end of the experiment, they received personalized results (simplified interpretation of results from N-of-1 analyses, in PDF format), and were invited to provide feedback through follow-up questionnaires and a semi-structured interview. Additional interview questions were included to obtain participant feedback about the changes to the tool.

5.2.4 Participants

A total of 9 participants consented and started their experiment. One participant discontinued their study (was traveling without phone access for the first 5 days of the study and wasn't able to complete surveys, hence dropped out).

Eight participants completed their self-experiments. Almost all participants (n=7) completed the follow-up questionnaire and interviews. One participant did not have enough data to analyze their self-experiment and the same participant did not complete the follow-up questionnaire. Sample sizes for the different aspects of the study are provided in Tables 5.14 and 5.15. Three of the participants reported that they had done some form of self-experimentation before this study, and three reported that they currently do some form of self-tracking. Other participant demographics are presented in Tables 5.16 and 5.17.

Table 5.14. Sample Sizes for Different Aspects of the study (N = 9)

	Interviews	SUS	UBS	N=1 experiment analysis
Completed self-experiment (n =8)	7	7	7	7*
Dropped out (n=1)	N/A	1	1	N/A

*1 participant had substantial missingness and their experiment was not analyzed

Table 5.15. Sample Size by Activity (N = 8; excluding dropout)

	Total (n)	Interviews (n)	SUS (n)	UBS (n)	N=1 experiment analysis (n)
Deep breathing meditation	3	3	3	3	3
Gratitude journaling	2	1	1	1	1
Blocking digital distractions	2	2	2	2	2
Vigorous physical activity	1	1	1	1	1

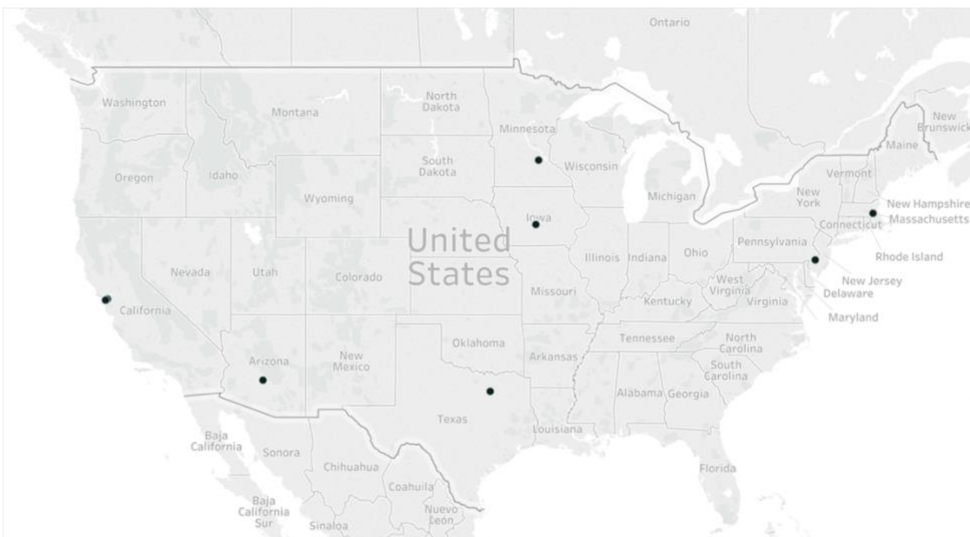


Figure 5.7. Participant Locations: Wave 2

Table 5.16. Participant Occupations: Wave 2

	N	**Occupation**
Employed, full-time	6	Forecasting, Program Evaluator, Recruiter, Speech Therapist, Cartographer, Psychologist
Self-employed, full-time	1	Architect
Student	1	Graduate

Table 5.17. Participant Demographics (N=8): Wave 2

	N(%)	M ± SD
Age, years	-	38.82 ± 8.05
Gender		-
Female	6	
Male	2	
Education		-
Advanced degree	5	
Bachelor's degree	3	
Race/Ethnicity		
White	8	-
Hispanic	1	
Household income		
>100,000	6	
50,000 – 59,999	1	
20,000 – 29,999	1	

5.2.5 Analyses

Analyses for Aims 1-3

Data was analyzed to examine user burden, adherence and usability using identical procedures from Wave 1. Qualitative data from interviews were examined for themes in alignment with Wave 1, with a particular focus on examining if saturation had occurred or if any new themes emerged in Wave 2.

Analyses for Aims 6 and 7

Through email interactions with participants while developing the personally relevant questions, and through follow-up interviews, we explored usefulness of tracking those outcomes, as well as issues and challenges of incorporating personally chosen outcomes and developing appropriate questions to track those outcomes during the experiment.

Participant responses to the open-ended question (asked only on activity days) about whether they thought that the activity impacted their well-being was examined to understand subjective experiences and factors that impact their decision related to the activity. Participants were also asked for feedback on the open-ended question during the follow-up interviews.

Perceived utility of usual routine days was examined through thematic analysis of the semi-structured interviews at follow-up. The same protocol and procedures for interviews and thematic analysis as explained in Chapter 5.1.4 were used to address both aims 6 and 7.

5.2.6 Findings

Findings: Aims 1-3

Adherence

Overall, adherence in Wave 2 was similar (in most aspects, slightly better) to that in Wave 1. The average completion rate was 93.06% (SD = 11.01, Median = 94.44%), slightly higher than the average rate of 88.89 ± 12.13% that was observed in Study 1. Put differently, the average number of missing days for Wave 2 was 1.25 ± 1.98 days, compared to 2.15 ± 2.18 days in Wave 2. One of the participants had substantial missing data as compared to the other participants, and likely skewed the result. Three participants had a 100% completion rate, and 4 participants had only 1 day of missing data (94.44% completion rate). The average adherence rate (sticking to experimental protocol) was 73.61% (SD = 19.64, Median = 83.33%). When asked about perceived ease/difficulty of sticking to the day's assignment, four participants reporting finding it somewhat easy, two reported it being somewhat difficult, and one reported it being very easy (Figure 5.6). Missingness over time is illustrated in Figure 5.7, and just the visualization suggests that missingness may not have increased over time.

As was observed in Wave 1, most participants reported some significant event affecting their well-being on at least one of the days of the experiment. The average number of days with reported significant events was 5.75 days (SD = +/= 4.98, median = 5 days).

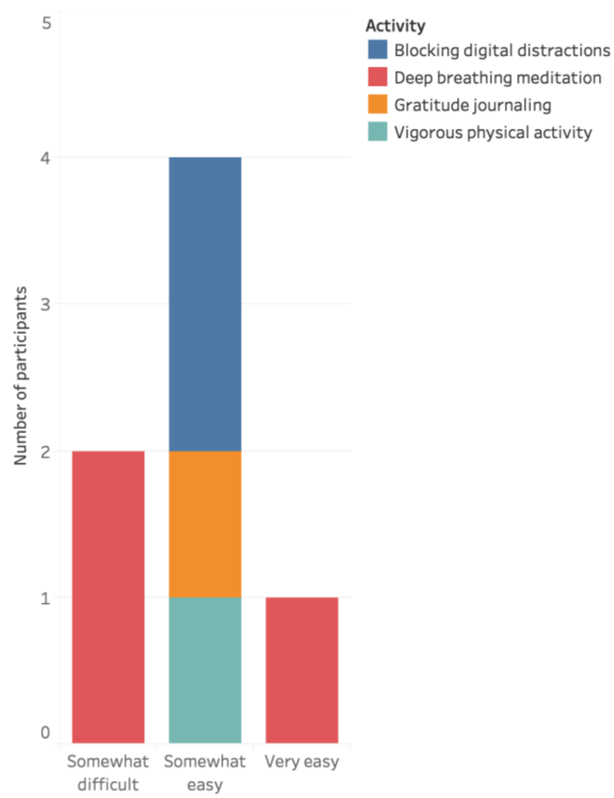


Figure 5.6 Perceived Ease/Difficulty of Sticking to the Day's assignment (Wave 2)

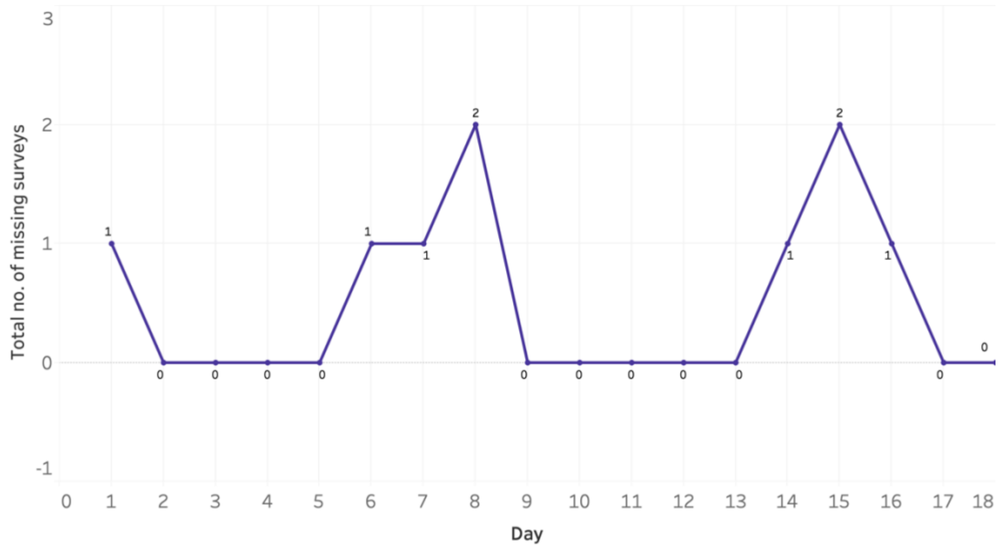


Figure 5.7. Missingness Over Time in Wave 2

Table 5.18 Adherence and Compliance Rates: Wave 2

	Mean	SD	Median	Range
Number of missing days	1.25	1.98	1	0-6
Completion Rate (days)	16.75	1.98	17	12-18
Completion Rate (%)	93.06	11.01	94.44	66.66-100
Number of missing days that were activity days	0.63	0.74	0	0-4
Number of days adhered to experiment	6.63	1.78	7.5	4-8
Percentage of days adhered to experiment	73.61	19.64	83.33	44.44-88.89
Number of days with a significant event	5.75	4.98	5	0-7
Number of activity days with significant events	2.88	2.53	2	0-7

User Burden

Results from the User Burden Scale (UBS) were similar to those in Wave 1, and indicated that most participants did not find Hack Your Health burdensome, and reported no burden on most items on all subscales. The comparison between scores on the UBS subscales for Wave 1 and Wave 2 are displayed in Table 5.19.

Table 5.19 Comparison of UBS Scores in Waves 1 and 2

	Wave 1	**Wave 2**
UBS Subscale	**Mean ± SD**	**Mean ± SD**
Difficulty of Use	0.17 ± 0.36	0.14 ± 0.14
Physical	0.05 ± 0.08	0.02 ± 0.03
Time and Social	0.11 ± 0.14	0.05 ± 0.11
Mental and Emotional	0.07 ± 0.14	0.03 ± 0.03
Privacy	0.10 ± 0.08	0.07 ± 0.08

Usability

Overall, there was a mean score of 82.86 out of 100 on the SUS (SD = 13.64), indicating good usability overall, and a median score of 87.5. Overall, scores observed were lower than Wave 1. Similar to Wave 1, lower scores were primarily on the items "I think I would like to use Hack Your Health frequently", and "I found the various functions of Hack Your Health were well integrated".

Since the sample size was low, we took a closer look at the data, which indicated that the lowest scores were from participant B4, who gave it lower scores on the items "I would imagine that most people would learn to use Hack Your Health very quickly" and

“I needed to learn a lot of things before I could get going with Hack Your Health”. Probing during the follow-up interview revealed that she was referring to the effort that was needed to effectively block the websites and apps on her devices as part of her experiment.

Utility

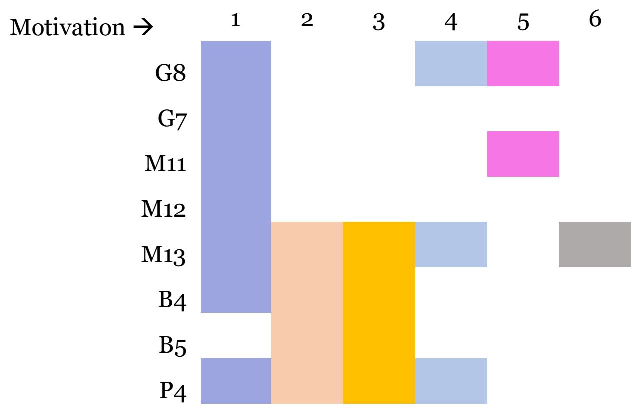
Baseline motivations behind trying selected activity were similar to that observed in Wave 1. Most frequent motivations included wanting to experiment and seeing if it improves their well-being (n=7), wanting to make it a habit (n=4), not having been able to adopt the activity in spite of multiple tries (n=4), and because research says it works (n=3). Motivations are displayed in Table 5.20.

No new themes were identified in interviews conducted as part of Wave 2 (beyond those that emerged in Wave 1; Chapter 5.1.6), which is supportive of achieving thematic saturation, which is valuable in qualitative research[116,117].

Understanding od N-of-1 study methods

No new themes were identified beyond those that emerged in Wave 1, thus suggesting thematic saturation**.**

Table 5.20. Baseline Motivations: Wave 2



LEGEND

	KEY	FREQ.
I want to experiment and test if it improves my well-being	1	7
I want to make it a habit	2	4
I haven't been able to adopt this habit in spite of multiple tries	3	4
Research says that it works	4	3
I haven't tried it before, want to see how I feel	5	2
I remember liking it and want to do it again	6	1
Some other reason	7	0
Somebody asked me to try it	8	0

Findings: Aim 5

Self-selected outcomes

In some cases, participants' desired target outcomes (collected via the question displayed in Figure 5.6) were too broad, or not specific enough; based on this, I needed to clarify/provide options on possible questions to track those outcomes, with support from EBH. For example, at sign up, M13 indicated wanting to track being 'disciplined' but did not state which aspect of his life he wanted to be more disciplined in. When I emailed him to clarify, it turned out that he was more interested in being mindful about unhealthy habits. G7 wanted to track 'serenity'. I emailed her with two options (single items chosen from an existing multi-question validated scale: NIH Toolbox Positive Affect Age 18+ v2.0) – 'feeling peaceful', and 'feeling content', so she could choose one that seemed most relevant for her. All but one participant listed outcomes that they thought would be affected by the activity they were trying. G7, on the other hand, chose two outcomes she cared about *in general*. At times, participants chose outcomes that would be best measured in the morning ("How rested did you feel when you woke up this morning?") instead of evenings, when the Hack Your Health survey was taken. Participants' self-selected outcomes and questions used to track those are provided in Table 5.21.

Table 5.21 Self-selected Outcomes and Final Questions Included in Experiments

Participant/ Activity		**Outcome**	**Question**
Gratitude	1	Dissertation Process	How productive were you today? 0 = not at all, 1 = a little bit, 2 = somewhat, 3 = quite a bit, 4 = very much
	2	Feeling hopeless	Today I felt motivated 0 = not at all, 1 = a little bit, 2 = somewhat, 3 = quite a bit, 4 = very much
Gratitude	1	Serenity	Today I felt peaceful 0 = not at all, 1 = a little bit, 2 = somewhat, 3 = quite a bit, 4 = very much
	2	Motivation	Today I felt motivated 0 = not at all, 1 = a little bit, 2 = somewhat, 3 = quite a bit, 4 = very much
Meditation	1	Be more efficient at work so I can get more done in a day	How efficient did you feel at work today? 0 = not at all, 1 = a little bit, 2 = somewhat, 3 = quite a bit, 4 = very much

	2	Feel less overwhelmed by things I need to do	How overwhelmed did you feel by things you need to do? 0 = not at all, 1 = a little bit, 2 = somewhat, 3 = quite a bit, 4 = very much
Meditation	1	Waking up feeling rested	How rested did you feel when you woke up this morning? 0 = not at all, 1 = a little bit, 2 = somewhat, 3 = quite a bit, 4 = very much
	2	Serenity	Today, I felt peaceful 0 = not at all, 1 = a little bit, 2 = somewhat, 3 = quite a bit, 4 = very much
Meditation	1	Handle stress better	How would you rate your ability to handle stress today? 0=poor, 1=fair, 2=good, 3=excellent
	2	Be more disciplined	How mindful did you feel today? 0 = not at all, 1 = a little bit, 2 = somewhat, 3 = quite a bit, 4 = very much
Blocking digital distractions	1	Catching train in morning	Were you able to catch the train this morning? 0 = No, 1 = Yes, 2 = Not applicable
	2	Getting through to-do list	How much of your to-do list did you get through today? 0 = none, 1 = Very little, 2 = Half, 3 = Most of it, 4= All of it
Blocking digital distractions	1	Being more productive	How productive were you today? 0 = not at all, 1 = a little bit, 2 = somewhat, 3 = quite a bit, 4 = very much
	2	Being more content with my life vs. comparing myself to others	Today, I felt content with my life 0 = not at all, 1 = a little bit, 2 = somewhat, 3 = quite a bit, 4 = very much
Vigorous physical activity	1	Energy levels	How active have you been feeling today? 0 = not at all, 1 = a little, 2 = somewhat, 3 = quite a bit, 4 = extremely
	2	Less stress, able to focus on family while home	Were you able to focus on family time while at home today? 0 = not at all, 1 = a little bit, 2 = somewhat, 3 = quite a bit, 4 = very much

At follow-up, **all participants expressed liking the tracking of self-selected outcomes**. For example, when providing general reactions to the daily survey, said, M12 said, "*I liked that we could choose an aspect to track. I liked that part of the self-experimentation*."

Most (n=6) participants indicated that they would like to track a combination of self-selected and researcher-selected outcomes. P5 also liked having to reflect on outcomes that she did not select: "*The ones I chose, because I chose them, they didn't take as much thought, maybe I already had preconceived notions of how to respond to them because it's something that is already on my mind. The happiness and stress, I don't normally put a number to. Those were outside of comfort zone and made me think a little bit more.*"

B5 on the other hand, felt that if given a choice, she would choose the outcomes that she selected because they felt most relevant to what she wanted to achieve with blocking digital distractions.

Two participants realized during the experiment that their questions needed to be modified. G7 felt that the outcome she chose were "*vague and open-ended*" and felt that, "*it probably would be helpful if I had given specific questions and specific things I was thinking about. As I was answering them, I was applying them to my own situation*".

A few days into the experiment, B4 realized that the outcomes she chose were not really measuring what she was truly interested in, and felt the need to refine them to be specific enough. She tracked being able to catch her train to work every morning as her outcome, but in hindsight realized that "*I do catch the train most mornings. Well really, I guess I don't want to be running to the train*".

Open-ended reflection

Participant responses to the open-ended reflection provided information about their experiences with the activity that would otherwise not be captured through the quantitative questions, including how those experiences varied from day to day. For example, it was clear that **most participants felt a positive impact on some days versus others** (n=7). For example, in one of the daily surveys, M13 reported: "*Not really, doing the breathing in the afternoon didn't have as much of an effect as a morning session*", and G7 reported: "*Not quite, it was late and quick, less reflective than yesterday*".

Open-ended reflection also provided information about reasons behind non-compliance by providing more context of the participant's day and how it affected the way the interacted with the intervention. One example is provided in Table 5.22. This participant did not block her distractions on one of the days because she "*had a minor operation today and thought having distractions would be nice*".

All but one participant (M11) **liked having open-ended questions that allowed them to reflect on their experience with the activity, although they also expressed concerns**. M12 felt that the open-ended questions "*helps you reflect a bit more about your own experience. I think that did help me to sit and think about if there was anything that was affected*" Although she liked them, she also expressed concern that **the open-ended reflection may not have been as practical**, and that she "*may have written shorter phrases instead of longer because it was daily*".

G7 felt similarly about the open-ended questions: "*I felt like I was able to give more meaningful information and context in the open-ended question, but then it was also like, not ideal to type on the phone. So it was helpful that I could give more information but I also did not want to type full sentences*."

B5 felt that while she liked the open-ended reflection, she "*didn't often have something to add. I felt like I was repeating myself throughout the experiment*."

M11, who felt that the experience helped her figure out that the particular deep breathing meditation that was included in Hack Your Health was not working for her, retrospectively felt that that the open-ended questions were useful because "*writing it out made me think about it*" but she also believed that she "*probably would have had that thought anyhow*".

Findings: Aim 6

Insights from Wave 1 indicated that the irregular nature of performing the activity negatively affected the experience for some participants. The same theme was observed in Wave 2 (n=3). During the follow-up interviews, we asked participants if they thought having days when they maintained their usual routine were useful for them, apart from the need for baseline days to enable comparison as part of the experiment. Overall, **participants had mixed reactions towards usual routine days**:

B4 did not like the short cross-over periods, and said that *"I thought it would have been more helpful if I had planned week-over- week assignments rather than getting a random assignment daily"*. She liked having usual routine days as comparison, but felt that the cross-over periods needed to be longer for her to really experience blocking her distractions and "*break the habit*". She also found the usual routine days helpful because "*it made me a little more aware of my habits, and that was positive"*.

Table 5.22 Example of Qualitative Data Obtained from One Participant on Days They Were Assigned to Do the Activity. This Participant Tried Blocking Digital Distractions as Their Activity.

Did you block your digital distractions today?	**Did you feel that blocking your digital distractions affected your well-being today? In what way?**	**What kept you from blocking your digital distractions today?**
No		“I unblocked the distractions during my morning commute and forgot to reapply them during the work day. That said, I did actively avoid the apps I wanted to so it didn't really matter.”
Yes	“I was prompted to do other, occasionally productive things when I reached for my blocked apps and remembered I couldn't access them.”	
Yes	“Yes, I started reading a new book”	
Yes	“I don't think it affected my well-being”	
No		“I didn't have that much to do at work today so I thought there would be no harm in unblocking things early. I did block them for a few hours though.”
No		“It's the weekend and I didn't care”
No		“I thought I could get through my to do list without formally blocking apps”
Yes	“I don't know if it did or not”	
No		“I had a minor operation today and thought having distractions would be nice”

B5 echoed that, and felt that the usual routine days were good for her because "*it was interesting to see on those days that I didn't actually block the apps, to see how mindless I am with that activity*". This also suggests that the rest days may not have been a true baseline for both these participants. The same theme was observed in participants trying that activity in Wave 1 (through interviews).

M12 liked the breaks because she felt that "*If I had to do the meditation every single day then it would have burnt me out, so having the break was nice*".

G7 felt that that because she did the gratitude journaling irregularly, it "*did not feel like it was lasting as long*" and expressed that she would like to do it more regularly to really feel the difference. She noted that she understands the purpose that it serves for the experiment, "*but if I was going to be doing it as sort of an intervention for myself in the long term trying to establish a pattern, I'd do it all the time and not have the days off*".

P5 also said that she understands the need for such days in terms of data for the experiment, but if she had a choice, she would not have the usual routine days so that she can establish a routine. However, she also noted that "*looking back, I do appreciate that I was able to differentiate how I was feeling at the end of the day, I really do appreciate that time to reflect on it.*"

CHAPTER 6

DISCUSSION

In the following sections, I summarize key findings, organized by aims from both waves of the study. I then offer possible lessons learned, some of which were not anticipated and fall outside of the scope of the targeted aims. Based on these two points, I then discuss this work in relation to similar recent work, and then offer plausible design implications for future work that researchers interested in building technology support for self-experimentation might want to consider. Following this, I summarize key limitations of this work and offer suggested areas of future work related to self-experimentation in the context of behavioral interventions.

6.1 Summary of Conclusions

The formative evaluation revealed several interesting insights relevant to self-experimentation within the realm of behavioral interventions. The following subsection summarizes the conclusions for each aim from both study waves.

Aim 1: Overall, the tool was perceived as low burden. However, participants were also confused about the source of the burden (the activity they tried vs. the tool). The amount of missingness was ~2 days on average, with some participants having much more missing days than others. On average, participants performed the activity on ~6 of the 9 assigned days over both studies, suggesting somewhat low compliance with the protocol.

Aim 2: Hack Your Health had high usability overall. Participant motivations for trying the activity they chose were not only to test its impact on their well-being, but also because they were looking for accountability to help them get back into the activity or make it a part of their routine (a theme that was also observed in user research described in Chapter 3). In line with that, participants reported liking the daily reminders and surveys. Findings suggest that participants found the experience useful to test if the intervention helped them, but their decision about the intervention was often based on intuition over results. Overall, participant intuition about what happened often did not match results, and when that was the case, participants seemed to rely on intuition over results. This could mean that the aspects of the experience that helped them were process-related., i.e., structure, accountability, and means of self-reflection that the experiment provided than the specific experimental design and the product of the experiment, i.e., the results. Data also indicated several issues with the experimental design and how the short phase length and random assignment with only a day's warning may have negatively impacted certain aspects of participant experience, such as their compliance, perceived enjoyment, and perceived fit. Data also provided insight into these and other confounding factors that likely impacted the fidelity of the experiment. Outcomes impacted by the activity could have also been more proximal and short-lived, and possibly not captured in the daily surveys. All these factors could also explain some of the mismatch between their intuition and results.

Aim 3: Participants understood the concept of self-experimentation, and that they were doing the experiment to test the effect of the intervention on their well-being. However, insights from other aspects of the study which highlight participant behaviors such as low compliance to protocol, backfilling, and not maintaining fidelity of the experiment (e.g., not blocking apps for the selected duration each time they were assigned, doing it in a hurry) also suggest that while they may have a conceptual understanding, they may not have a theoretical understanding of experimentation or of the effect of those behaviors on the validity and fidelity of their experiment. It should be noted that the latter insight is not surprising, since we did not include any educational materials on this topic as part of the protocol.

Aim 4: Results from the multiple N-of-1 experiments revealed that there was considerable heterogeneity across people in terms of intervention impact, and perceived enjoyment and fit. Additionally, it's possible that the same intervention had a positive and negative impact for a given person, depending on the outcome.

Aim 5: Overall, participants reported liking the tracking of self-selected outcomes, and most said they would like to track a combination of researcher and participant-selected outcomes. Some of the outcomes participants listed were not specific enough and we needed to clarify before creating a relevant question for them to track. Some participants also listed outcomes that were close to, but not exactly what they were interested in, and outcomes that would be better measured at a time of day different than the one used in this study. These insights indicated that tracking self-selected outcomes could be valuable, but participants would likely need more scaffolding to be able to list measurable, specific and relevant outcomes to track.

Participants also liked the open-ended reflection in the daily surveys, but noted that it could be impractical and burdensome to do it frequently via typing on their phone. Participants' survey responses over the 18-days to the open-ended question provided unique information about their experience with the activity that was not captured in quantified data. These qualitative reflections illustrate how participant experiences with the activity were impacted by their day's/life's context and the perceived impact varied over time (e.g., positive impact on some days vs. others).

Aim 6: Participants had mixed reactions towards the utility of the usual routine days. While they understood the utility in terms of enabling comparison of the experimental conditions, frequently having such days seemed to be at odds with their desire to do the activity for longer durations and making it a part of their routine. Essentially, insights suggest possible tensions between participants' process of trying/testing such activities vs. how they were trying it through the crossover experiment we used. However, it should also be noted that some participants also appreciated having days without the activity because it provided an opportunity to reflect on how the activity might be affecting them, or just as a break to avoid getting burnt out by doing the activity too often.

6.2 Lessons Learned

Experimental Design and Set-up

Data indicate that the experimental design had several limitations that need to be addressed. On average, participants missed ~2 of the 18 daily surveys, and performed the activity on ~6 of the 9 assigned days. The current analyses did not factor in days that participants did not comply with protocol and assumed them to be activity days. That, coupled with the short duration of the study likely reduced statistical power, suggesting that the experimental design needs to be flexible enough to account for missingness and non-compliance to protocol. This is in line with previous studies on self-experimentation with similar experimental durations, that suggest that missingness is inevitable even with shorter experiments and tools need to be designed with that in mind[31,118].

In the current design, each outcome was measured once a day to avoid increasing participant burden. This experimental design inherently assumes that if the overall day's well-being (measured once a day at the end of the day) is not affected, the intervention did not have an impact. If the effect was highly acute and short-lived i.e., proximal to the time the activity was performed, it may not have been captured in the daily surveys. Insights from interviews, however, suggest that proximal/short-lived effects may have been important and driven participant decision-making, as opposed to how they felt that day overall. Additionally, participant feedback also indicated that many factors affect their psychological well-being on any given day, and often, more than the activity they tried. In future iterations, it may be important to explore the use of a longer experimental duration, to account for covariates, and increase measurement frequency to improve the statistical power of the analyses, as well as to capture effects that may be highly acute/proximal. A way to capture proximal effects without increasing measurement frequency could also be to trigger the survey right after participants perform the activity.

Lived experience of the structured experiment

In terms of the experimental design, some participants expressed that the sporadic nature of performing the activity (in phases of 3 days each) negatively affected their experience. A few of the participants also said they would have preferred to have the experimental schedule instead of being randomly assigned so that they can plan the activity in their calendar. The random assignment may have affected their assessment of fit and enjoyment, since they did not get to choose when to perform the activity as they would outside of the experiment. Future iterations could consider sharing the experimental schedule with participants. Knowledge of the schedule may bias results, but may offer a more optimal experience for participants in terms of assessing whether the activity "works for them".

There seems to be a tension between assessing fit and enjoyment, and systematically assessing impact of the activity using the same experimental set-up. For example, we did not restrict the time of day that participants could perform the activity while in the experiment. Data suggest that without the restriction, participants chose to perform the activity at varying times of day throughout the experiment. We maintained the flexibility so as to balance the needs of assessing enjoyment and fit of the activity into their day. However, the time inconsistency likely affected the data and analyses. So, it may be important to keep time of day consistent for stronger analyses – but that in turn might affect and bias the assessment of fit and enjoyment, as well as overall participant experience with the activity.

Overall, there seem to be two positive aspects of the usual routine days. One, as a break from the activity, and two, as serving the intended purpose of enabling comparison. However, although participants seemed to appreciate having the usual routine days in terms of the intended theoretical purpose, they did not find them as useful personally, because the usual routine days induced irregularity in performing activities that would want to perform for longer durations so that it feels like a part of their routine to get a sense of how it's affecting their well-being. Future work should explore ways of balancing these three needs. One way of doing that could be by increasing the experimental duration and length of the phase (longer than 3 days), and having different lengths for activity and usual routine days (activity days phase length could be longer than usual routine days) to avoid long breaks from activity but still enable comparison. Additionally, break days could be inserted as appropriate, also with consideration to the activity itself. For example, it may not be advisable to assign someone to do vigorous physical exercise for many days in a row, and a "break" may be important as well as desirable. These changes would have to be made with consideration to their impact on the strength of the experimental design and validity of the results.

Participants liked tracking and reflecting on the outcomes that they personally chose. In this study, the research team worked with the participants to clarify and then devise questions to track the outcomes. In future work, this process could be automated to make it scalable. Further, if tracking of self-selected outcomes is to be included in such tools, it may be necessary to guide participants so that their questions are measurable, and specific enough to track. Participants could enter an outcome of their choice and be given a range of response scales to choose from. While in the current study, we only provided a simple visualization of the data back to participants (did not perform any statistical analyses), future studies could also explore how these data could be analyzed as part of the experiment, and whether such analyses (hence, going beyond just facilitating reflection) are something that is of value and interest to participants.

Insights related to the open-ended reflection provided a more nuanced look at participants' subjective experience about the perceived impact of the activity and how that varied from day to day based on their life's context. These data suggested that such open-ended reflection within the context of a structured experiment could also be valuable in terms of decision-making in a way that is complementary to quantitative data and analyses. Additionally, these potentially time-varying relationships could also explain the discrepancy between results and lived experience. In this work, we only provided participants a list of their listed subjective experiences in their results document. Future work could look into how these data could be leveraged in a way that is useful to participants or converted to more actionable insights.

While most participants reported liking the open-ended reflection in this short study, and the completion rates for surveys were high, they also expressed concerns in terms of practicality of typing over the phone and it becoming burdensome over time. Additionally, an important point to note is that responses differed in terms of depth across participants. Future work could explore ways of facilitating such self-reflection in a less-burdensome manner.

The qualitative survey data and the varying within-person positive/negative/neutral responses over time also suggest that it is possible that instead of testing *whether* the intervention worked on average, it might be more important to help people figure out *in what context* the intervention works for them, they enjoy it, and it fits in their life.

It may also be important to explore ways in which such experimentation could be made more iterative, and account for factors such as discontinuing the experiment for participants who are not enjoying the activity. A more iterative approach might mean that they can discontinue that activity and restart their experiment with either a new activity or a different operationalization of the same activity (e.g., try guided meditation instead of deep breathing).

Decision-Making

Results suggested that participants' decision-making about the activity was often influenced by factors other than the quantitative self-tracked data and results of their experiment. These factors included subjective experiences that were not tracked, such as proximal perceived effects likely not captured in the experiment, or noticing differences in outcomes other than those tracked in the experiment. Beyond this, data also suggest a mismatch between results (as defined by the experimental protocol) and intuitions of effects based on overall lived experience of the participants. Critical for decision-making, when a mismatch between the results from the experiment vs. overall lived experience was present, participants appeared to rely more on their intuitions than on the results of the experiment when making decisions about future use of the intervention.

There are many plausible explanations for these results that require further examination. For example, this result could be due to expectancy bias or confirmation bias as the activities were self-selected. Further, the limitations of the study design itself, as already discussed, could partially explain this mismatch. As several participants did not go over results at all and yet also had an opinion and intuitive sense on the value of an intervention, a third explanation is that the process and structure offered may not, in itself, be usable or robust enough, as operationalized in this study, to surpass the influence of intuitions on a person's decision-making process. As this last point highlights, it is important to note that these findings should be interpreted only within the context of the structure of the current experiment, especially the measurement frequency of 1x/day, and short duration of the experiment. These results may have been different if the study used a more rigorous protocol for providing a causal inference that would have less susceptible to contextual disruptions, which were common and likely influenced the robustness of statistical results.

6.3 Insights from This Work in Relation to Similar Recent Work

A few recent studies have designed digital tools to support the process of self-experimentation in the health domain. TummyTrials by Karkar et al.[31] guides patients with Irritable Bowel Syndrome through self-experiments to detect whether a certain food triggers particular symptoms. Recent work that is most similar to our work is a study by

Taylor et al[118], where they developed a self-experimentation app called QuantifyMe, which guides users through self-experiments to optimize behaviors they already do (by answering questions such as "How does my nightly sleep affect my productivity?"). Both studies uncovered important challenges when designing tools to support self-experimentation. Our work adds to this work by Karkar et al. and Taylor et al.

TummyTrials utilized a version of alternating treatment designs where they randomized individuals to treatment/control every day. TummyTrials was developed using the self-experimentation framework developed by Karkar et al.[35] which focuses on self-experimentation in clinical settings or for health conditions. Their framework requires a more rigid experimental design, and that the independent variable (intervention) must be well-specified in terms of 'amount', and time of day, etc. It is also desirable that the independent variable is applied in a reliable manner each time. While that makes good sense and may be feasible in a clinical setting where the intervention may be elimination or consumption of a particular food, or taking a particular medication, I believe that it may not be advisable *as the first step in the process* in the case of self-guided behavioral interventions

Our underlying assumption that drove the design of the tool was that not only is intervention effect important, but an intervention that "works" for a person is also one that is enjoyable and fits into their routine. With that assumption, we purposely kept the protocol flexible so that people could perform the activity in a way that suits their lifestyle. In line with that, our findings (discussed in previous sections) suggest that in this context of common healthy behaviors (specifically, outside of a well-defined clinical context, where there are specific outcomes that participants want to improve), applying an even rigorous design might not provide an optimal experience for participants. Taylor et al.[118], through their work on QuantifyMe showed similar findings. Participants found it difficult to comply with rigid experimental conditions that the app assigned, suggesting that in this context, it may be important to retain some flexibility in terms of the independent variable. Additionally, phase-based designs that let people try the activity for longer durations may be better-suited to this context as compared to designs that use daily randomization such as the ones used in TummyTrials.

To balance these needs for scientific rigor with participant needs when it comes to daily health-related behaviors, a guided and iterative self-exploration approach such as the one used by Lee et al[39], could be explored. In their work, users iteratively self-experimented with behavior change plans *without using N-of-1 experimental designs*. As also suggested by our work, the process of performing the activity in a structured manner (without specific experimental design) could in itself be a valuable first step when trying an activity.

Building on these insights and the work by Lee et al., one way to balance the different needs could be by doing this process in phases: Users could first engage in structured self-exploration (i.e., scaffolding provided via reminders and surveys for self-reflection *without using single-case designs*) with an intervention of their choice to assess whether they find value in it, like it, enjoy it and when it might fit in their routine. Insights from that phase could then be used to clearly specify a hypothesis, and operationalization of the intervention (including factors such as time of day, duration, etc.) in order to inform a more rigorous experiment to assess intervention effect that also fits with participants' own routine and needs.

6.4 Design Implications

In its current form, Hack Your Health used a simple website, SMS, and an online survey platform for all experiment-related communication. Aspects of the current tool that worked well were the daily surveys and reminders sent via SMS. We purposely kept the daily communication with participants at only 2 SMS sent per day to reduce participant burden (except the first two days in experiment, when more than 2 SMS were sent to share important experiment-related information). Participants found this experience low-burden overall, liked the accountability it provided, liked having SMS as a medium of communication, getting reminders to perform the activity as well as to stick to their usual routine, having activity instructions in written as well video format, and the previous evening's heads up that informed them of the next day's assignment.

A few key design implications emerged from this work. Digging deeper into participant perceptions related to the experimental design indicated that **experiments designed for rigor may not match with participants' mental models related to engaging with such behavioral interventions.** For example, when people think of these activities in life, they may want to try them for an extended time to truly assess how they feel. While a design with multiple cross-overs might increase statistical power and internal validity, it came at the cost of negatively affecting participant experience for some. Future tools should explore strategies to balance participant needs with those of achieving scientific rigor.

Protocols supporting such self-experimentation need to be flexible and account for non-compliance and missingness. Beyond features such as clear instructions and reminders, another way to improve participant compliance and fidelity could be to include educational components that provide them the theoretical rationale behind aspects of the experiment, such as sticking to the experimental protocol and tracking outcomes throughout the experiment.

Tools designed to support such self-experimentation should explore strategies to **incorporate participants' lived experiences with the activity, and outcomes**

important to the person that the tool may not track. This could help provide more reliable and context-specific insights about the effect of the intervention on their well-being. Beyond that, future work in this domain should account for ingrained beliefs that participants may have about the effects of popular interventions.

We have used some of the key insights and directions highlighted in this and previous sections to inform the next iteration of Hack Your Health (currently in development; study being led by one of the collaborators of this project, Dr. Richard Kravitz from UC Davis) which will be deployed to a larger sample of individuals. Specifically, we are now testing a longer experimental duration (30 days) with a longer phase length (5 days), which would let individuals try the intervention for a longer period of time overall, as well as in each phase. Additionally, participants will now be provided their experimental schedule when they sign up, giving them more time to plan it in their schedule.

6.5 Limitations and Methodological Considerations

There are certain limitations to this work that should be noted. Firstly, those related to the Hack Your Health system. The outcomes measured in Hack Your Health are self-reported and hence accompanied with biases that affect self-reported data. However, each individual's data is only used for their own analyses. Additionally, while ideally, N-of-1 experimentation can be more rigorous when both participant and researcher are blinded to the treatment, this is impossible when the treatment is a behavioral intervention, and the researcher is also the subject of the experiment. This increases the risk of expectance bias[34]. However, with the use of multiple randomized crossovers, we can help reduce that risk. Additionally, the length of experiments is short (18 days), and the interventions are expected only to capture acute effects, if any. As mentioned in previous sections, the models currently used in Hack Your Health do not handle missingness or factor in days when participants don't stick to their assignment. Additionally, these models performed a simple comparison of means across both conditions, and did not factor in the effect of time or potential correlation. Some of the insights obtained in the study could be due to these and other limitations of the experimental design (such as low measurement frequency) and should be interpreted within that context.

Relevant to the aims of this work, due to the limited existing literature on self-experimentation in this context, we had a limited understanding of potential underlying covariates, such as if previous self-tracking experience matters or not, that may be associated with differential experiences across individuals, which might necessitate qualitative analyses also being carried out at the sub-group level. This work was also accompanied by the risk of biases associated with qualitative work, especially those pertaining to issues with credibility and trustworthiness of findings. Relevant and feasible measures (such as creating thorough interview protocols, note-taking, using multiple cycles of coding) were taken when possible at every step to mitigate such risks.

The sample size in different intervention groups was small and unequal. As we had expected based on user research, more people signed up to try gratitude journaling and deep breathing meditation, so qualitative data is driven by individuals who tried those two interventions. Nonetheless, the limited data from other interventions indicated that there might indeed be differences in participant experience depending on the intervention. Future work should take measures for more targeted recruitment to examine these differences more comprehensively.

Responses about outcomes people noticed differences in were elicited in the interviews. So, it is possible that they may have reflected on it due to the interview conversation. There's no way of knowing whether people thought about it on their own. However, the daily qualitative tracking in Wave 2 captured those thoughts. Participants across all the studies were highly educated, many had advanced degrees or bachelor's degrees. It is possible that this concept of self-experimentation seems interesting to a particular group of people. Additionally, participants were predominantly white females. Findings may not generalize outside the context of this demographic, or outside the context of a short self-experiment using similar experimental designs.

6.6 Future Work

In this section, I share some thoughts that I haven't yet touched upon, or touched upon in enough detail in the previous sections and offer some food for thought for future areas of work in this domain.

Subjective experiences are important in decision-making

Much of the work in self-experimentation has focused on and given more importance to quantitative tracking and statistical analyses over participant intuition. Indeed, that is the assumption we began with, as stated in the introduction: "*The focus [of such self-experimentation] is not on gaining generalizable knowledge, but to aid self-knowledge and self-discovery in a scientifically grounded way,* ***beyond relying on intuition"***[34–36] . In the case of psychological well-being, we avoid relying on intuition and "individual subjective" data because it can be misleading and a biased form of knowledge[119], and affected by issues such as recollection and expectance biases. On the other hand, we strive to collect "individual objective data" (i.e., quantification of a person's subjective experience of factors like stress or happiness, as converted into a number) via experiments because, as the assumption of scientists go, this can help to control for these biases and provide a more reliable, "objective" and quantified form of knowledge. When we initially developed the research question, "*How might we design tools that can help individuals make decisions about which behavioral intervention to adopt, in a scientific and systematic way that also meets the person's personal needs and motivations?"* we assumed that the quantified experiences as systematically collected through the experiment and self-report measures would have a large impact on decision-making.

However, as observed in this work, intuition and subjective experience were important in participant decision-making. While one response to this could be to double-down on quantification via making the studies longer and the like, as suggested in previous sections, another possible way to look at this is to think more deeply about what is lost when quantification is used to understand a complex phenomenon, such as a person's experience with trying a behavioral intervention to determine if it works for them. As demonstrated in this study, qualitative reflection can help capture subjective experiences that are important to participants that quantitative data does not. This could be due to a number of experiment-related (e.g., low measurement frequency, data not accounting for confounding factors, not capturing proximal effects, or effects that the participant cares about, etc.) as well as participant-related factors (self-selection bias, expectancy bias, prior beliefs, etc.). Lived experience could also be important in itself as this information, particularly how lived experience can uniquely capture insights about how a person's personal history and context can influence outcomes, are arguably essential for achieving the high-level aspirations of developing valid causal inferences of interventions. As the many contextual examples of factors that likely impacted inference imply (e.g., knowledge of days of particularly high stress or busyness related to other life event, illness, spending a day outdoors with family and not needing any intervention that day, noticing that the activity doesn't seem helpful if done at night, noticing that the activity helps them in certain states such as in periods of high work load or high stress) there is logical reason and justification to assert that, for this specific person, these lived experience qualitative assertions illustrate some very plausible factors that, causally, influenced the study outcomes. As discussed in agile science[56], future methodological work is needed to further understand and unpack how concrete operations, which, like this, offer clear and plausible insights around issues like causation, need to be advanced.

If we are to assume that participant subjective experience is valuable, then in addition to increasing rigor of the experimental design, we could also explore strategies to make the subjective qualitative data more reliable. This could be achieved via more frequent qualitative reflection such as the daily reflection used in this study to help avoid biased recollections at the end of the experiment. Such daily reflection could also be made more reliable and low burden by providing scaffolding for the self-reflection (such as guiding users through the self-reflection process via specific questions)[120].

Ultimately, we want to support people in self-learning, and empower them to navigate and tune their health. Before getting fixated on the kind of data used in such tools, it may be important to take a few steps back (and looking back, something we should have examined in our user research) to understand, outside the context of an experiment, how do people go about such decision-making? What kind of evidence and how much evidence do people need for decision-making? Answers to these questions could help us get closer to understanding the reasons behind the mismatch between participant experience and results, and its implications on the design of such tools.

CHAPTER 7

CONCLUSION

7.1 Summary

In this dissertation, we designed and evaluated a tool that provides the structure for people to try a simple, self-guided behavioral intervention via an N-of-1 experiment to test if the activity improves their well-being while simultaneously also tracking their perceived enjoyment and fit of the activity into their routine.

In partnership with Elaine Chen from WNYC (now at the New York Times) and several others (see acknowledgements), we began designing this tool with an aim to make it scalable enough to be used by a diverse group of ~15,000 users. We conducted user research to understand participant motivations, and interventions and outcomes of interest to a diverse audience. Insights from user research were used to develop a testable prototype of this tool, Hack Your Health. The initial design let users try one of four available activities (deep breathing meditation, vigorous physical activity, blocking digital distractions, and gratitude journaling) to test the impact of the activity on their energy, focus, stress, and happiness, along with assessing its perceived enjoyment and fit into their life. In Hack Your Health, the experiment lasts for 18 days and uses a cross-over design (with intervention and baseline phases consisting of 3 days each). At the end of the experiment, each participant's data is analyzed and shared in simple language (and using plots) in the form of a PDF document.

In this work, we conducted a formal evaluation of the tool in an iterative manner, using mixed methods, via two waves of participants. In the first wave, all users tracked the same outcomes (mentioned above) in a quantitative manner. Insights suggested that participant intuition about the effect of the activity drove their decision-making and they noticed differences in outcomes not tracked in the experiment, as well as effects that may have been acute/proximal to when they performed the activity. These findings motivated the changes we made to the tool, and as such, in the second wave, the tool was modified to support more qualitative tracking of experience with activity and to include tracking of personalized and participant-selected outcomes.

The formative evaluation revealed several interesting insights relevant to self-experimentation in the realm of behavioral interventions. The tool had high usability and low-burden overall. Participants understood the concept of self-experimentation, and that they were doing it to test the effect of the intervention on their well-being. Participants reported liking the daily reminders, and the survey at the end of the day. Findings suggest that participants found the experience useful to figure out if the intervention helped them, but it was possibly through the structure, accountability, and means of self-reflection that the experiment provided than the precise experimental design and results.

7.2 Contributions

This dissertation explores the use of self-experimentation to help people in decision-making about healthy activities that work in general and/or are often associated with heterogeneity in terms of response. It highlights the need for expanding the conceptualization of tools that support behavior change. While many existing tools support behavior change related to various aspects of life, from finance management to meditation, the focus is largely either on self-tracking or habit formation of specific behaviors. Very few tools take a step back and help individuals empirically test, through self-experimentation, *whether* an intervention has the desired effect in the first place while also providing insights on if it fits into their lives and if they enjoy them. Participant feedback indicated that they liked the structure the self-experiment provided. This highlights the value of this approach overall, including of the lived experience of the study beyond the study results itself, in supporting decision-making about healthy behaviors. Put differently, the process of performing the activity in a structured manner (without specific experimental design) could in itself be a valuable first step when trying an activity.

By including four distinct activities for participants to try, we were also able to begin exploring how user experience might differ in the context of different interventions. For example, those trying blocking digital distractions found themselves being more aware of their phone usage and reducing phone usage even on usual routine days, suggesting that a frequent-crossover design may not be appropriate for such an intervention. Participants trying gratitude journaling and meditation at times did it in a hurry (potentially because those two require lower time and resource commitment) when they hadn't done it by the time the received the evening survey, likely impacting fidelity of the intervention.

The use of single-case designs has not yet been widely examined in digital tools that let users try popular behavioral interventions. Through the use of our mixed methods approach, we were able to uncover challenges with using commonly used single-case designs in the context of decision-making related to behavioral interventions, specifically, when performing an activity in an experimental set up may *not* offer the most optimal experience or be appropriate, such as when the purpose of the experience might be to assess whether the activity fits within their current context and routine. We offered suggestions on how these designs could be modified to be better suited to support such self-experimentation.

Insights from the second wave of the study, where we included qualitative tracking of experience with the activity revealed important contextual information likely important in terms of causal inference that was missed in the quantified data. These daily reflections also illustrated that participant experiences with the intervention were unique, varied over time (positive at times, neutral at times, and negative at times; as opposed to static) and were often impacted by the context of their day and life (e.g., busy day at work, illness, low motivation, etc.). This suggested that self-experimentation tools may need to go beyond just assessing intervention effect on average, and account for such time-varying experiences and help participants gain context-specific insights about the intervention effect (e.g., meditation seems to calm you down on days when you report being particularly stressed). These data also highlighted the importance of incorporating lived experiences into the process of inference and supporting decision-making.

Additionally, by using pre- and post-experiment assessment of participants' prediction and intuition about the intervention's effect on their own well-being, and through interviews at follow-up, we uncovered discrepancies between participants' lived experience and statistical analyses that such tools will likely need to account for in their design. The pre-experiment assessment of participants' hunches on intervention effect indicated that participants may have ingrained beliefs about the effects of popular interventions (surveys that indicated participants were often 'pretty sure' of the effect they predicted an intervention would have) they tried and that intuitions might be an important factor to examine and consider when designing such tools (participants were 'very' or 'pretty sure' how the intervention affected their well-being).

These insights were used to inform design implications for tools to support self-experimentation in this domain, such as using longer experimental duration and phase length, capturing proximal outcomes via appropriate measurement strategies, and complementing quantified tracking with qualitative reflection of participant experiences.

REFERENCES

1. Gardner, C. D. *et al.* Effect of Low-Fat vs Low-Carbohydrate Diet on 12-Month Weight Loss in Overweight Adults and the Association With Genotype Pattern or Insulin Secretion. *Jama* **319,** 667 (2018).

2. Schueller, S. M. Preferences for positive psychology exercises. *J. Posit. Psychol.* **5,** 192–203 (2010).

3. Lyubomirsky, S. & Layous, K. How Do Simple Positive Activities Increase Well-Being? *Curr. Dir. Psychol. Sci.* **22,** 57–62 (2013).

4. Schueller, S. M. & Parks, A. C. Disseminating self-help: Positive psychology exercises in an online trial. *J. Med. Internet Res.* **14,** (2012).

5. Bouchard, C. & Rankinen, T. Individual differences in response to regular physical activity. *Med. Sci. Sports Exerc.* **33,** S446–S451 (2001).

6. Younge, J. O., Gotink, R. A., Baena, C. P., Roos-Hesselink, J. W. & Hunink, M. G. M. Mind-body practices for patients with cardiac disease: A systematic review and meta-analysis. *Eur. J. Prev. Cardiol.* **22,** 1385–1398 (2015).

7. Goyal, M. *et al.* Meditation programs for psychological stress and well-being: A systematic review and meta-analysis. *JAMA Intern. Med.* **174,** 357–368 (2014).

8. Ireland, M. Meditation and Psychological health and functioning: a descriptive and critical review. *Sci. Rev. Ment. Heal. Pract.* **9,** 4–19 (2012).

9. Greenfield, S., Kravitz, R., Duan, N. & Kaplan, S. H. Heterogeneity of Treatment Effects: Implications for Guidelines, Payment, and Quality Assessment. *Am. J. Med.* **120,** 3–9 (2007).

10. Lillie, E. O. *et al.* The n-of-1 clinical trial: the ultimate strategy for individualizing medicine? *Per. Med.* **8,** 161–173 (2011).

11. Duan, N., Kravitz, R. L. & Schmid, C. H. Single-patient (n-of-1) trials: A pragmatic clinical decision methodology for patient-centered comparative effectiveness research. *J. Clin. Epidemiol.* **66,** S21–S28 (2013).

12. Conn, V. S., Hafdahl, A. R. & Mehr, D. R. Interventions to increase physical activity among healthy adults: meta-analysis of outcomes. *Am. J. Public Health* **101,** 751–758 (2011).

13. Nahum-shani, I. *et al.* Just-in-Time Adaptive Interventions (JITAIs) in Mobile Health : Key Components and Design Principles for Ongoing Health Behavior Support. *Ann. Behav. Med.* (2016). doi:10.1007/s12160-016-9830-8

14. Senf, K. & Liau, A. K. The Effects of Positive Interventions on Happiness and Depressive Symptoms, with an Examination of Personality as a Moderator. *J.*

Happiness Stud. **14,** 591–612 (2013).

15. Salmon, J., Owen, N., Crawford, D., Bauman, A. & Sallis, J. F. Physical activity and sedentary behavior: A population-based study of barriers, enjoyment, and preference. *Heal. Psychol.* **22,** 178–188 (2003).

16. Bauman, A. E., Sallis, J. F., Dzewaltowski, D. A. & Owen, N. Toward a better understanding of the influences on physical activity: The role of determinants, correlates, causal variables, mediators, moderators, and confounders. *Am. J. Prev. Med.* **23,** 5–14 (2002).

17. Hekler, E. B. *et al.* Determining who responds better to a computer-vs. human-delivered physical activity intervention: Results from the community health advice by telephone (CHAT) trial. *Int. J. Behav. Nutr. Phys. Act.* **10,** 1–10 (2013).

18. Stevens, V. J., Glasgow, R. E., Hollis, J. F. & Mount, K. Implementation and Effectiveness of a Brief Smoking-Cessation Intervention for Hospital Patients. *Med. Care* **38,** 451–459 (2000).

19. Zeng, X., Chan, V. Y. L., Liu, X., Oei, T. P. S. & Leung, F. Y. K. The Four Immeasurables Meditations: Differential Effects of Appreciative Joy and Compassion Meditations on Emotions. *Mindfulness (N. Y).* **8,** 949–959 (2017).

20. Lumma, A. L., Kok, B. E. & Singer, T. Corrigendum: "Is meditation always relaxing? Investigating heart rate, heart rate variability, experienced effort and likeability during training of three types of meditation" [Int. J. Psychophysiol. 97/1 (2015) 38–45](S0167876015001658)(10.1016/j.ijpsych. *Int. J. Psychophysiol.* **117,** 126–130 (2017).

21. Rash, J. A., Matsuba, M. K. & Prkachin, K. M. Gratitude and well-being: Who benefits the most from a gratitude intervention? *Appl. Psychol. Heal. Well-Being* **3,** 350–369 (2011).

22. Phatak, S. S. *et al.* Modeling Individual Differences: A Case Study of the Application of System Identification for Personalizing a Physical Activity Intervention. *J. Biomed. Inform.* **79,** 82–97 (2018).

23. Conroy, D. E., Elavsky, S., Hyde, A. L. & Doerksen, S. E. The dynamic nature of physical activity intentions: A within person perspective on intentions-behavior coupling. *J. Sport Exerc. Psychol.* **33,** 807–827 (2011).

24. Fitzsimmons, P. T. *et al.* A daily process analysis of physical activity, sedentary behavior, and perceived cognitive abilities. *Psychol. Sport Exerc.* **15,** 498–504 (2014).

25. Beltz, A. M., Wright, A. G. C., Sprague, B. N. & Molenaar, P. C. M. Bridging the Nomothetic and Idiographic Approaches to the Analysis of Clinical Data. *Assessment* (2016). doi:10.1177/1073191116648209

26. Sackett, D. L. Evidence-based medicine. *Semin. Perinatol.* **21,** 3–5 (1997).

27. Barr, C. *et al.* The PREEMPT study - evaluating smartphone-assisted n-of-1 trials in patients with chronic pain: Study protocol for a randomized controlled trial. *Trials* **16,** 1–11 (2015).

28. Guyatt, G. H. *et al.* N of 1 randomized trials for investigating new drugs. *Control. Clin. Trials* **11,** 88–100 (1990).

29. Guyatt, G. H. *et al.* The N of 1 randomized controlled trial: clinical usefulness. Our 3 year experience. *Ann. Intern. Med.* **112,** 293–299 (1990).

30. William, N. Bidirectional Case-Crossover Designs for Exposures with Time Trends. *Biometrics* **54,** 596–605 (1998).

31. Karkar, R. *et al.* TummyTrials: A Feasibility Study of Using Self-Experimentation to Detect Individualized Food Triggers. *Proc. 2017 CHI Conf. Hum. Factors Comput. Syst. - CHI '17* 6850–6863 (2017). doi:10.1145/3025453.3025480

32. Daskalova, N. *et al.* SleepCoacher: A Personalized Automated Self-Experimentation System for Sleep Recommendations. *Proc. 29th Annu. Symp. User Interface Softw. Technol.* 347–358 (2016). doi:10.1145/2984511.2984534

33. Duan, N. *et al. Design and Implementation of N-of-1 Trials: A User's Guide*. (AHRQ Publication No. 13(14)-EHC122-EF. Agency for Healthcare Research and Quality, 2014).

34. Neuringer, A. Self-Experimentation : A Call for Change. *Behaviorism* **9,** 79–94 (1981).

35. Karkar, R. *et al.* A framework for self-experimentation in personalized health. *J. Am. Med. Informatics Assoc.* **23,** 440–448 (2016).

36. Choe, E., Lee, N. & Lee, B. Understanding quantified-selfers' practices in collecting and exploring personal data. *Proc. 32nd* … (2014).

37. Davidoff, F. Heterogeneity Is Not Always Noise: Lessons From Improvement. *JAMA* **302,** 2580–2586 (2009).

38. Lee, J. & Walker, E. Understanding Users ' Creation of Behavior Change Plans with Theory- Based Support. 2301–2306 (2015). doi:10.1145/2702613.2732870

39. Lee, J. *et al.* Self-experimentation for behavior change: Design and formative evaluation of two approaches. *Dl.Acm.Org* (2017). doi:10.1145/3025453.3026038

40. Lupton, D. Quantifying the body: monitoring and measuring health in the age of health technologies. *Crit. Public Health* **23,** 393–403 (2013).

41. Rapp, A. *et al.* New frontiers of quantified self. *Proc. 2015 ACM Int. Jt. Conf.*

Pervasive Ubiquitous Comput. Proc. 2015 ACM Int. Symp. Wearable Comput. - UbiComp '15 969–972 (2015). doi:10.1145/2800835.2807947

42. Li, I., Dey, A. & Forlizzi, J. Understanding My Data, Myself: Supporting Self-Reflection with Ubicomp Technologies. *Proc. 13th Int. Conf. Ubiquitous Comput.* 405–414 (2011). doi:10.1145/2030112.2030166

43. Karkar, R. Opportunities and Challenges for Self-Experimentation in Self-Tracking. in *Ubicomp/ISWC, Adjunct,* 991–996 (ACM, 2015).

44. Lupton, D. The diverse domains of quantified selves: self-tracking modes and dataveillance. *Econ. Soc.* **45,** 101–122 (2016).

45. Renfree, I., Harrison, D., Marshall, P., Stawarz, K. & Cox, A. Don ' t Kick the Habit : The Role of Dependency in Habit Formation Apps. in *CHI'16 Extended Abstracts* (ACM, 2016).

46. Katarzyna, S., Cox, A. L. & Blandford, A. Beyond Self-tracking and reminders: Designing smartphone apps that support habit formation. in *CHI* (ACM, 2015).

47. Chu, L. C. The benefits of meditation vis-à-vis emotional intelligence, perceived stress and negative mental health. *Stress Heal.* **26,** 169–180 (2010).

48. Bauman, A. *et al.* Updating the Evidence for Physical Activity : Summative Reviews of the Epidemiological Evidence , Prevalence , and Interventions to Promote " Active Aging " Updating the Evidence for Physical Activity : Summative Reviews of the Epidemiological Evidence , . doi:10.1093/geront/gnw031

49. Davis, D. E. *et al.* Thankful for the little things: A meta-analysis of gratitude interventions. *J. Couns. Psychol.* **63,** 20–31 (2016).

50. Dickens, L. R. Using Gratitude to Promote Positive Change: A Series of Meta-Analyses Investigating the Effectiveness of Gratitude Interventions. *Basic Appl. Soc. Psych.* **39,** 193–208 (2017).

51. Mark, G., Czerwinski, M. & Iqbal, S. T. Effects of Individual Differences in Blocking Workplace Distractions. (2018).

52. Marotta, V. & Acquisti, A. Online Distractions, Website Blockers, and Economic Productivity: A Randomized Field Experiment. 1–27 (2017).

53. Pressman, S. D. & Cohen, S. Does positive affect influence health? *Psychol. Bull.* **131,** 925–971 (2005).

54. Danner, D. D., Snowdon, D. A. & Friesen, W. V. Positive emotions in early life and longevity: Findings from the nun study. *J. Pers. Soc. Psychol.* **80,** 804–813 (2001).

55. Diener, E. Subjective well-being: The science of happiness and a proposal for a

national index. *Am. Psychol.* **55,** 34–43 (2000).

56. Klasnja, P. *et al.* Agile science: creating useful products for behavior change in the real world. *Transl. Behav. Med.* **6,** 317–328 (2016).

57. Glasgow, R. E., Vogt, T. M. & Boles, S. M. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. *Am J Public Heal.* **89,** (1999).

58. Glasgow, R. E., Lichtenstein, E. & Marcus, A. C. Why Don't We See More Translation of Health Promotion Research to Practice? Rethinking the Efficacy-to-Effectiveness Transition. *Am. J. Public Health* **93,** 1261–1267 (2003).

59. Glasgow, R., Davidson, K., Dobkin, P., Ockene, J. & Spring, B. Practical behavioural trials to advance evidence-based behavioural medicine. *Ann Behav Med* **31,** 5–13 (2006).

60. Green, L. W., Glasgow, R. E., Atkins, D. & Stange, K. Making Evidence from Research More Relevant, Useful, and Actionable in Policy, Program Planning, and Practice. Slips 'Twixt Cup and Lip'. *Am. J. Prev. Med.* **37,** 187–191 (2009).

61. Bouchard, C. *et al.* Familial aggregation of VO2 max response to exercise training : results from the HERITAGE Family Study. *J. Appl. Physiol.* **87,** 1003–1008 (1999).

62. Bouchard, C. *et al.* Adverse metabolic response to regular exercise: Is it a rare or common occurrence? *PLoS One* **7,** (2012).

63. Smith, P. J. *et al.* Aerobic exercise and neurocognitive performance: a meta-analytic review of randomized controlled trials. *Psychosom. Med.* **72,** 239–252 (2011).

64. Reed, J. & Ones, D. S. The effect of acute aerobic exercise on positive activated affect: A meta-analysis. *Psychol. Sport Exerc.* **7,** 477–514 (2006).

65. De Vibe, M., Bjørndal, A., Hammerstrøm, K. T., Kowalski, K. & Tipton, E. Mindfulness Based Stress Reduction (MBSR) for Improving Health, Quality of Life, and Social Functioning in Adults. *Campbell Syst. Rev.* (2012).

66. Sears, S. & Kraus, S. I Think Therefore I Om: Cognitive Distortions and Coping Style as Mediators for the Effects of Mindfulness Meditation on Anxiety, Positive and Negative Affect, and Hope. *J. Clin. Psychol.* **65,** 561–573 (2009).

67. J.D., L. & J.E., S. Brief meditation training can improve perceived stress and negative mood. *Altern. Ther. Health Med.* **13,** 38–44 (2007).

68. Jacobs, T. L. *et al.* Intensive meditation training, immune cell telomerase activity, and psychological mediators. *Psychoneuroendocrinology* **36,** 664–681 (2011).

69. Otis, L. S. Adverse effects of transcendental meditation. in *Meditation: Classic and contemporary perspectives* 201–108 (1984).

70. Lazarus, A. A. & Mayne, T. J. Relaxation: Some limitations, side effects, and proposed solutions. *Psychotherapy* **27,** 261–266 (1990).

71. Lustyk, M. K. B., Chawla, N., Nolan, R. S. & Marlatt, G. A. Mindfulness meditation research: issues of participant screening, safety procedures, and researcher training. *Adv. Mind. Body. Med.* **24,** 20–30 (2009).

72. Layous, K. & Lyubomirsky, S. The how, why, what, when, and who of happiness: Mechanisms underlying the success of Positive Activity Interventions. *To appear in: The light and dark side of positive emotions* (eds. Gruber, J. & Moscowitz, J.) 1–23 (Oxford University Press, 2012).

73. Schueller, S. M. & Parks, A. C. The science of self-help: Translating positive psychology research into increased individual happiness. *Eur. Psychol.* **19,** 145–155 (2014).

74. Mark, G., Iqbal, S. & Czerwinski, M. How blocking distractions affects workplace focus and productivity. *Proc. 2017 ACM Int. Jt. Conf. Pervasive Ubiquitous Comput. Proc. 2017 ACM Int. Symp. Wearable Comput. - UbiComp '17* 928–934 (2017). doi:10.1145/3123024.3124558

75. Kardefelt-Winther, D. A conceptual and methodological critique of internet addiction research: Towards a model of compensatory internet use. *Comput. Human Behav.* **31,** 351–354 (2014).

76. Bucciol, A., Houser, D. & Piovesan, M. Temptation at Work. *PLoS One* **8,** 1–5 (2013).

77. Orleans, C. T. Promoting the maintenance of health behavior change: Recommendations for the next generation of research and practice. *Heal. Psychol.* **19,** 76–83 (2005).

78. Damschroder, L. J. *et al.* Fostering implementation of health services research findings into practice: A consolidated framework for advancing implementation science. *Implement. Sci.* **4,** 1–15 (2009).

79. Davis, J. M., Fleming, M. F., Bonus, K. A. & Baker, T. B. A pilot study on mindfulness based stress reduction for smokers. *BMC Complement. Altern. Med.* **7,** 1–7 (2007).

80. Wolin, K. Y., Bennett, G. G., Mcneill, L. H., Sorensen, G. & Emmons, K. M. Physical Activity and Intervention Uptake. **32,** 563–570 (2008).

81. King, A. C. *et al.* Determinants of physical activity and interventions in adults. *Med. Sci. Sport. Exerc.* (1992).

82. Van Hooff, M. L. M. & Baas, M. Recovering by means of meditation: The role of recovery experiences and intrinsic motivation. *Appl. Psychol.* **62,** 185–210 (2013).

83. Stutts, W. C. Physical Activity Determinants in Adults. Perceived benefits, barriers, and self-efficacy. *Aaohn J.* **50,** 499–507 (2002).

84. Dobkin, P. L., Irving, J. A. & Amar, S. For whom may participation in a mindfulness-based stress reduction program be contraindicated? *Mindfulness (N. Y).* **3,** 44–50 (2012).

85. Eysenbach, G. The law of attrition. *J. Med. Internet Res.* **7,** 1–9 (2005).

86. Lewis, B. A., Marcus, B. H., Pate, R. R. & Dunn, A. L. Psychosocial mediators of physical activity behavior among adults and children. *Am. J. Prev. Med.* **23,** 26–35 (2002).

87. Schwartz, B. The paradox of choice: Why less is more. 1–6 (2004).

88. Kohl, H. W. *et al.* The pandemic of physical inactivity: Global action for public health. *Lancet* **380,** 294–305 (2012).

89. Lillie, E. O. *et al.* The n-of-1 clinical trial: the ultimate strategy for individualizing medicine? *Per. Med* **8,** 161–173 (2011).

90. Dallery, J., Cassidy, R. N. & Raiff, B. R. Single-case experimental designs to evaluate novel technology-based health interventions. *J. Med. Internet Res.* **15,** (2013).

91. Kravitz, R. L. *et al.* What ever happened to N-of-1 trials? Insiders' perspectives and a look to the future. *Milbank Q.* **86,** 533–555 (2008).

92. Zucker, D. R., Ruthazer, R. & Schmid, C. H. Individual (N-of-1) trials can be combined to give population comparative treatment effect estimates: Methodologic considerations. *J. Clin. Epidemiol.* **63,** 1312–1323 (2010).

93. Scheibehenne, B., Greifeneder, R. & Todd, P. M. Can There Ever Be Too Many Options? A Meta-Analytic Review of Choice Overload. *J. Consum. Res.* **37,** 409–425 (2010).

94. Ellis, A. *A Guide to Rational Living*. (1861).

95. Corti, K., Reddy, G., Choi, E. & Gillespie, A. The Researcher as Experimental Subject : Using Self-Experimentation to Access Experiences , Understand Social Phenomena , and Stimulate Reflexivity. 288–308 (2015). doi:10.1007/s12124-015-9294-6

96. Roberts, S. The unreasonable effectiveness of my self-experimentation. *Med. Hypotheses* **75,** 482–489 (2010).

97. Kravitz, R. L. *et al.* Marketing therapeutic precision: Potential facilitators and barriers to adoption of n-of-1 trials. *Contemp. Clin. Trials* **30,** 436–445 (2009).

98. Suh, H., Shahriaree, N., Hekler, E. B. & Kientz, J. a. Developing and Validating the User Burden Scale: A Tool for Assessing User Burden in Computing Systems. *Proc. ACM Conf. Hum. Factors Comput. Syst. (CHI 2016)* Under Review (2016). doi:10.1145/2858036.2858448

99. Nielson, J. Usability 101: Introduction to Usability. (2012). Available at: https://www.nngroup.com/articles/usability-101-introduction-to-usability/.

100. Phatak, S. *et al.* Design of a Large-scale Self-Experimentation Tool for Scientific Self-Explorations. in *PervasiveHealth (Demos, Posters, Doctoral Colloquium)* (EAI, 2018). doi:10.4108/eai.20-4-2018.2276355

101. Fereday, J. & Muir-Cochrane, E. Demonstrating Rigor Using Thematic Analysis: A Hybrid Approach of Inductive and Deductive Coding and Theme Development. *Int. J. Qual. Methods* **5,** 80–92 (2006).

102. Braun, V. & Clarke, V. Using thematic analysis in psychology. *Qual. Res. Psychol.* **3,** 77–101 (2006).

103. Hardesty, D., Bearden, W., Hardesty, D. M. & Bearden, W. O. The use of expert judges in scale development : Implications for improving face validity of measures of unobservable constructs Implications for improving face validity of measures of. (2014). doi:10.1016/S0148-2963(01)00295-8

104. Abdel-Khalek, A. M. Measuring Happiness With a Single-Item Scale. *Soc. Behav. Personal. an Int. J.* **34,** 139–150 (2006).

105. Stone, A. A., Hufford, M. R. & Shiffman, S. Ecological Momentary Assessment. *Annu. Rev. Clin. Psychol.* **4,** 1–32 (2008).

106. Seo, Michael, Schmid, C. nof1: Analysis of N of 1 Trials. (2018).

107. R Core Team. R: A language and environment for statistical computing.

108. Brooke, J. SUS: A 'quick and dirty' usability scale. in *Usability evaluation in industry* 189–193 (1996).

109. Meissner, H., Creswell, J., Klassen, A. C., Plano, V. & Smith, K. C. Best Practices for Mixed Methods Research in the Health Sciences. *Methods* **29,** 1–39 (2011).

110. Cassell, C. & Symon, G. Essential Guide to Qualitative Methods in Organizational Research. *Athenaeum Stud. Period. Di Lett. E Stor. Dell Antich.* 388 (2004).

111. Yin, R. K. Case Study Reserach - Design and Methods. *Clin. Res.* **2,** 8–13 (2006).

112. Baxter, P. & Jack, S. Qualitative Case Study Methodology : Study Design and

Implementation for Novice Researchers Qualitative Case Study Methodology : Study Design and Implementation. **13,** 544–559 (2008).

113. Fylan, F. Semi-structured interviewing. in *A handbook of research methods for clinical and health psychology* 65–78 (2005).

114. Thorne, S. Data analysis in qualitative research. *Evid. Based Nurs.* **3,** 68–70 (2000).

115. Saldana, J. *Qualitative researcher's companion*. (2009).

116. Hennink, M. M., Kaiser, B. N. & Marconi, V. C. Code Saturation Versus Meaning Saturation. *Qual. Health Res.* **27,** 591–608 (2016).

117. Francis, J. J. *et al.* What is an adequate sample size? Operationalising data saturation for theory-based interview studies. *Psychol. Heal.* **25,** 1229–1245 (2010).

118. Taylor, S., Sano, A., Ferguson, C., Mohan, A. & Picard, R. W. Quantifyme: An open-source automated single-case experimental design platform. *Sensors (Switzerland)* **18,** (2018).

119. Fischbein, E. *Intuitions in Science and Mathematics: An Educational Approach*. (Springer Science and Business Media, 1987).

120. Baur, D., Choe, E. K., Riche, N. H., Zhu, H. & Lee, B. Understanding self-reflection. 173–182 (2018). doi:10.1145/3154862.3154881

APPENDIX A

PERMISSION STATEMENT

The work included in Chapter 3 was previously as a poster titled "Design of a Large-scale Self-Experimentation Tool for Scientific Self-Explorations" at Pervasive Health Conference 2018, and published as an extended abstract in the EU Digital Library. I hereby attest that the co-authors of that paper, Eric Hekler, Elaine Chen, Richard Kravitz, Christopher Schmid and Ida Sim have granted their permission to include that work as the third chapter in this dissertation.

Sincerely,

Sayali Phatak

APPENDIX B

IRB APPROVAL

APPROVAL: EXPEDITED REVIEW

Matthew Buman

Exercise Science and Health Promotion

602/827-2289

Matthew.Buman@asu.edu

Dear Matthew Buman:

On 5/12/2018 the ASU IRB reviewed the following protocol:

Type of Review:	Initial Study
Title:	N-of-1 Self Experimentation Tool Pilot Study
Investigator:	Matthew Buman
IRB ID:	STUDY00008266
Category of review:	(6) Voice, video, digital, or image recordings, (7)(b) Social science methods, (7)(a) Behavioral research
Funding:	None
Grant Title:	None
Grant ID:	None
Documents Reviewed:	• Consent form for 7-day prototype test, Category: Consent Form; • Protocol, Category: IRB Protocol; • Consent for interview audio recording, Category: Consent Form; • Consent for full study, Category: Consent Form; • Consent for 1-hour prototype testing, Category: Consent Form; • Recruitment Material, Category: Recruitment Materials;

The IRB approved the protocol from 5/12/2018 to 5/11/2019 inclusive. Three weeks before 5/11/2019 you are to submit a completed Continuing Review application and required attachments to request continuing approval or closure.

If continuing review approval is not granted before the expiration date of 5/11/2019 approval of this protocol expires on that date. When consent is appropriate, you must use final, watermarked versions available under the “Documents” tab in ERA-IRB.

In conducting this protocol you are required to follow the requirements listed in the INVESTIGATOR MANUAL (HRP-103).

Sincerely,

IRB Administrator

cc: Sayali Phatak

Sayali Phatak