Self-E: Smartphone-Supported Guidance for Customizable Self-Experimentation

Nediyana Daskalova*
Spotify

Eindra Kyi†
Smith College

Kevin Ouyang†
Brown University

Arthur Borem
Brown University

Sally Chen
Brown University

Sung Hyun Park
Grinnell College

Nicole Nugent
Warren Alpert Medical School

Jeff Huang
Brown University

ABSTRACT

The ubiquity of self-tracking devices and smartphone apps has empowered people to collect data about themselves and try to self-improve. However, people with little to no personal analytics experience may not be able to analyze data or run experiments on their own (self-experiments). To lower the barrier to intervention-based self-experimentation, we developed an app called Self-E, which guides users through the experiment. We conducted a 2-week diary study with 16 participants from the local population and a second study with a more advanced group of users to investigate how they perceive and carry out self-experiments with the help of Self-E, and what challenges they face. We find that users are influenced by their preconceived notions of how healthy a given behavior is, making it difficult to follow Self-E's directions and trusting its results. We present suggestions to overcome this challenge, such as by incorporating empathy and scaffolding in the system.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; *Empirical studies in ubiquitous and mobile computing*.

KEYWORDS

self-experiments, personal informatics, self-tracking

ACM Reference Format:
Nediyana Daskalova, Eindra Kyi, Kevin Ouyang, Arthur Borem, Sally Chen, Sung Hyun Park, Nicole Nugent, and Jeff Huang. 2021. Self-E: Smartphone-Supported Guidance for Customizable Self-Experimentation. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3411764.3445100

*Work was conducted while author was at Brown University.
†These authors contributed equally.

CHI '21, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445100

1 INTRODUCTION

Significant aspects of our life such as sleep [6, 17, 19] and nutrition [13, 14, 36, 76] can be tracked with the help of technology, and research has shown that 69% of U.S. adults already engage in self-tracking practices [27]. Numerous studies have focused on developing personal informatics systems to help people collect and review personally relevant information [42]. Some people self-track as a step towards behavior change [42], in order to implement a generally recommended behavior that they have heard such as "meditate to lower stress" [58].

However, since traditional health studies emphasize the average person's response to a given treatment [8], even interventions widely believed to be beneficial and harmless, such as meditation, may not be advisable for some individuals [24]. In contrast, to discover what works specifically for them, people can conduct "self-experiments," which allow them to vary aspects of their lifestyle in a controlled way and uncover potential causal relationships [53, 70].

Self-experimentation, however, can be challenging for people as it entails systematically collecting and analyzing data. Even "extreme users," as defined by Choe et al. [10], with experience in self-tracking, encounter difficulties in rigorous self-experimentation.

Existing studies of self-experimentation systems have focused either (1) on the *self-tracking* aspect, which helps little in determining potential causation and generating actionable goals [51], or (2) on domain-specific [1, 18, 36, 68] or population-specific applications [2, 10]. Evaluations of these systems have shown a gap between what is available and what users desire: a general-purpose self-experimentation system that balances guidance with freedom of choice, so they can incorporate the experiment into their daily lives and conduct experiments across multiple domains with the help of a single tool [2, 37]. Accordingly, in this work, we pose the following research questions:

- **RQ1**: What functionality should a *general* self-experimentation tool contain so that people can use it to investigate potential causal relationships in their daily lives across multiple domains?
- **RQ2**: What lessons can we learn from the way people use such a system for their experiments?

To answer **RQ1**, we designed, implemented, and deployed Self-E: an app and server combination that serves as a general-purpose tool for self-experimentation. Our system uses scaffolding to guide

users through the steps of designing and conducting a practical experiment on their own. Simultaneously, it allows users to set goals and procedures for data collection that can be incorporated most easily into their lives. Thus, we define "practical" as experiments in real circumstances where users might not adhere to the recommendation sometimes, or drop out after uneventful periods.

To answer **RQ2**, we evaluated the system with two studies. First, we conducted in-person semi-structured interviews with 16 participants from the local population, supplemented by a two-week diary study during which they used the Self-E app. Next, we redesigned the system to include the ability to create custom experiments. We then asked a set of users with experimental design experience to download and use the app while conducting custom experiments.

We found that the most common use of self-tracking data among participants was as a motivator to enact behavior change. When designing their self-experiments, participants also instinctively started with a general goal in mind, so the additional guidance was beneficial in narrowing down the scope and setup of their experiments. We also identified issues that cause potential mistrust in the system, as well as reasons people sometimes disregard its instructions. We conclude by presenting critical considerations for the design and development of future general-purpose self-experimentation systems, including suggestions about how to better match users' mental models of self-experiments and build trust in the app.

2 RELATED WORK

Self-experiments have been of interest to a diverse group of research communities. We first review relevant work done on self-tracking in personal informatics, which provides the infrastructure for self-experiments. We then discuss single-subject research design, a departure from traditional clinical research studies, which has influenced work in self-experiments. Finally, we summarize previous self-experimentation systems and build on their findings.

2.1 Self-Tracking in Personal Informatics

Due to the ubiquity of tracking devices and smartphones, people can collect various data about themselves. Previous studies of self-tracking have addressed areas such as food intake [11, 13, 14], personal fitness [28], multiple sclerosis [2], mindfulness [3], migraines [61, 69], menstrual tracking [21], personal finance [38], mental wellness [39], and productivity [34, 44, 45, 64, 80]. Prototypes for manual and automatic self-tracking of general factors in one's life have been developed by Kim et al. [46].

Research has shown that self-tracking can encourage *reactivity* in people, meaning that they can monitor a behavior and decide if they want to change or maintain it [10, 47, 57]. However, it has been observed that even experienced users fail to make the most of their personal data even if they desire to do so [9]. Interpreting the collected data is challenging, so people often turn to health providers for help [10, 52, 69]. It is important to note that there is nuance in prior work: while challenges with paper-based self-tracking approaches have been identified (such as inaccurate and incomplete data [31, 36]), such practices also lead to mindful and joyful experiences [75]. While self-tracking tools can help people make their own interpretations about their data [2], that agency alone does not lead to actionable changes [51]. However, such tools can be useful in gathering the appropriate data to then determine causal relationships for effective lifestyle interventions [16, 36]. These tools can be complemented by previous research in persuasive technology to develop encouraging and trustworthy software [7, 25].

Li et al. have developed a five-stage model of personal informatics that includes *preparation*, *data collection*, *data integration*, *reflection*, and *action*, through which people can choose how they will act on "their newfound understanding of themselves" [52]. Self-experimentation technologies incorporate elements of all stages. For example, previous work has aimed to minimize user burden by providing semi-automation in the *preparation* and *integration* stages [17, 36]. Both Li et al. [52]'s and Epstein et al. [22]'s models for personal informatics note similarities with the transtheoretical model of behavior change [62], highlighting how often users turn to the mental model of using self-tracking for behavior change.

2.2 Single-Subject Research Design

Randomized controlled trials (RCTs) are a traditional method in clinical studies involving numerous subjects [12]. Due to inherent limitations, RCTs are not ideal study designs for self-experimentation. For example, most findings from RCTs are based on the responses of average persons in a study; therefore, they cannot directly inform an individual about their specific case, as the study sample may not be representative of this individual [54, 56, 79].

In contrast, individuals in single-case study designs (SCDs) serve as their own controls, which can reduce the inferential errors of group analysis in RCTs [40]. SCDs allow the empirical testing of whether an intervention is effective for an individual. This feature makes them more suitable for self-experiments, as they provide personalized interventions and more flexibility than RCTs [36, 53].

Self-experiments in the form of SCDs have been conducted by academics from medicine [41] and psychology [63], as well as by non-academics in areas involving well-being in QuantifiedSelf and its practitioners [10]. The most common SCD is the AB phase design, in which the A phase represents the baseline condition, and the B phase represents the intervention. To mitigate the commonly cited limitations of SCDs such as internal validity [30, 36], one can apply the phases at random, as we did in the Self-E system [33, 36]. We further evolved our system by using Bayesian techniques in the data analysis, which aim to increase rigor while maintaining practicality [70].

It is important to note that self-experiments and SCDs in general suffer from numerous potential downsides and limitations, such as "the inability of the experimenter to be objective, [and] the problematic aspects of self-reported data," among others [74]. Self-experiments, however, do not have to aim to match the high level of scientific rigor of RCTs. In our work, we are focused on finding ways to incorporate simple self-experiments in a practical fashion into a person's daily life, but we acknowledge the limitations that the nature of these experiments presents.

2.3 Existing Self-Experimentation Studies

The goal of self-experiments in the context of personal informatics is to find knowledge about oneself that is individually meaningful [10, 51]. Previous systems have explored self-experimentation in specific domains: e.g., SleepCoacher and SleepBandits for sleep [17,

18], TummyTrials for IBS management [36], Trackly for multiple sclerosis [1], or Trialist for chronic pain management [4]. QuantifyMe, another self-experimentation app, asked users to follow a rigid experimental schedule that only allowed 1 of the 13 participants to finish their experiment [68]. These studies contributed to self-experimentation literature through collecting and analyzing qualitative data [20], providing a rationale for SCD experiments [36], and improving quantitative data evaluation methods [16, 36, 37]. Two systems, Paco [23] and Galileo [77], exist to help people conduct experiments outside of a study setting. However, they are not optimized to help people conduct self-experiments with only their own data.

Previous work has identified challenges that people face in self-experiments, such as tracking fatigue and flawed experimental designs. Even "extreme users," with above-average background experience in self-tracking and experimentation, face common pitfalls, such as tracking too many things, having non-actionable and underspecified goals, and not knowing what to track nor how to analyze or interpret data to extract insights [51]. Karkar et al.'s TummyTrials system is based on the researchers' framework that could be applied to the design of a general self-experimentation tool [37]. Daskalova et al. built on this framework to outline guidelines for self-experiments that can be helpful for novices [16]. Lee et al. developed a prototype for self-experiments on pen and paper [51]. Most recently, Daskalova et al.'s open-source SleepBandits system was focused on guiding users through the steps of self-experiments in the domain of sleep. While we build on existing work, we are taking a more general approach: rather than focus on a specific domain, we explore how a system can guide people through general self-experiments in a practical and less burdensome way, and how it can help them create their own customized experiments without the help of professionals.

Overall, recent work in HCI has highlighted the need for and importance of n-of-one studies for supporting people's health [43, 55]. Our study contributes to the literature by evaluating a different approach toward conducting self-experiments, one which is aimed at increasing *flexibility* and user *agency* in order to conduct more *practical* self-experiments, while minimizing user burden in accordance with lived informatics principles. Self-E is a guided self-experimentation system that aids people in data-driven self-discovery so that they can make better-informed decisions.

3 SELF-E DESIGN AND IMPLEMENTATION

In this section, we present the initial implementation of Self-E, followed by the diary study we conducted to evaluate it in the next section. In this original version, Self-E presents users with a list of pre-configured experiments and lets them select one while tweaking small aspects of it. Next, in Section 5, we present a redesigned implementation of the app that lets users create their own customizable experiments in addition to the ones from the pre-configured list.

Self-E first asks users to select an experiment, which is comprised of two conditions: on any given day, the user either conducts the given behavior (e.g., meditate) or not. Then, it asks the user which condition they conducted that day and how they would rate the effect of that behavior on their well-being. After a few days of data collection, Self-E automatically computes the probability that the new behavior improves the user's well-being, as well as the size of the effect of this behavior change.

When designing Self-E, we focused on creating a tool that uses self-tracked data to deliver individualized health and behavioral insights to the user. Our exploration of prior work led us to identify several persistent issues among self-experimentation systems. First, challenges at various stages of self-experimentation can reduce adherence rates, particularly in unstructured evaluation environments [18, 68], which leads to failure to obtain results [10, 16]. Second, users require significant amounts of guidance and restriction during setup to prevent poor experimental design [37]. A third theme, in tension with the previous, is user desire for greater freedom, whether in making minor changes such as adjusting timing or altering the entire purpose of the experiment [36]. Lastly, a consideration that arises specifically within the context of a general-purpose self-experimentation system is that experimental design, scaffolding, and guidance should generalize to the different self-experiments that users are likely to run.

3.1 Prototype Iteration

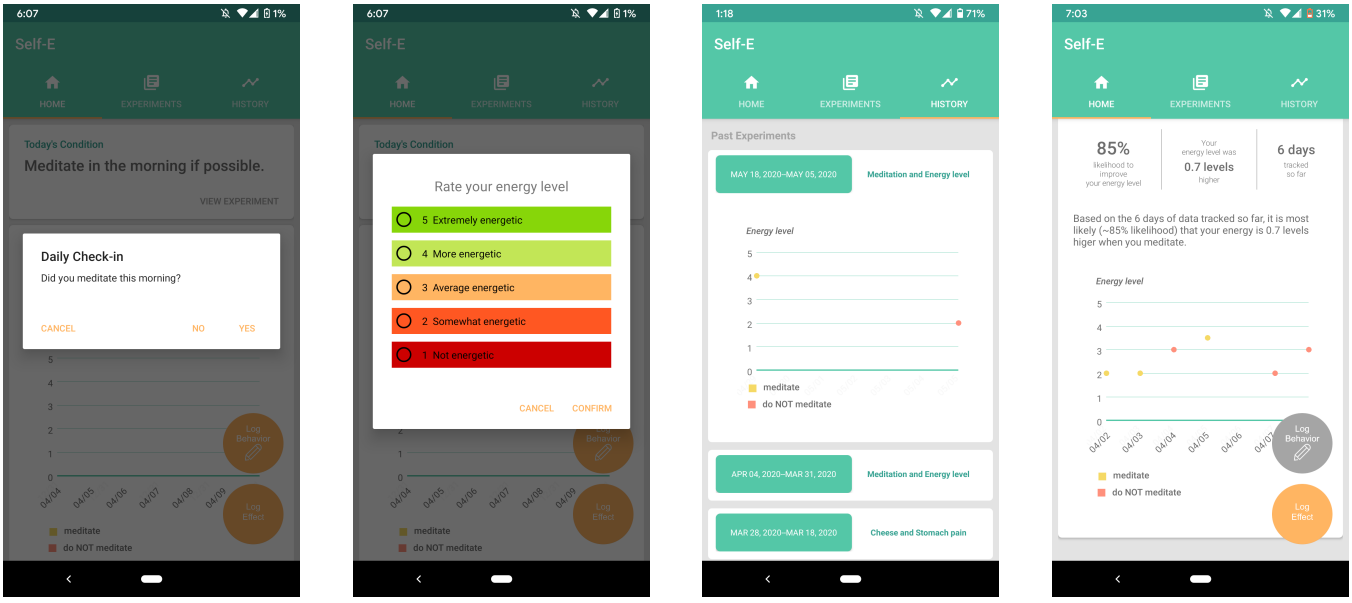
Our design goals were to provide guidance and flexibility during the experiment setup and to reduce friction for accurate self-reporting during the experiment. To tackle these challenges, we iterated upon several prototypes of varying fidelity to gather feedback from users in a nearby coffee shop. Some insights that we gathered from these iterations were that many users approached the experimentation process with an objective in mind first ("I want to improve my productivity") before developing a behavior modification ("meditate"). Additionally, technical concepts such as "independent variable" created significant confusion. Thus, we replaced jargon with simpler keywords which we further contextualized with illustrative examples (e.g., "independent variable" is referred to as "cause" or "intervention" and "dependent variable" is "effect" or "outcome"). Lastly, similar to findings from SleepBandits [18], we learned that users prefer having some preset examples of common self-experiments to choose from for their initial experiments.

3.1.1 Guidance at Different Levels of Expertise. Another design consideration we addressed in Self-E was to strike a balance between providing guidance and not sacrificing user agency so that people of different experience levels could always learn something through the app. Self-E minimizes user burden by employing design strategies such as notifications to schedule interventions (Figure 2b), data abstraction, and automatic analysis of the results to simplify aspects that are otherwise challenging even for experienced users [10]. To lower the barrier for users at different levels of learning, we incorporate both guidance as well as opportunities for more fine-tuned customization.

3.1.2 Improving Quality of Self-reported Data. Self-E aims to strengthen the quality of data collected by presenting the option to use the Experience Sampling Method (ESM) to gather dependent variable data (Figure 2b). ESM is commonly used to assess affective state and technology usage [78] since it better captures variables that may fluctuate throughout the day, in addition to reducing reliance on human memory by asking participants to reflect on shorter periods

Experimental Design	Examples	Customizable?
Intervention/**cause**	Napping, step count	Choose from pre-configured list
Goal amount	20-minute nap, 10,000+ steps	Customizable, with defaults
Outcome/**effect**	Energy level, appetite at dinner	Choose from pre-configured list
User-report scale	Exhausted ... alert, not hungry ... famished	Customizable, with defaults
Effect sampling interval	Random between 11am-5pm, daily at 6pm	Customizable, with defaults
Effect sampling frequency	3 times a day, once per day	Customizable, with defaults
Cause sampling time	Daily at 6pm, daily at 8pm	Customizable, with defaults

Table 1: Experimental design attributes with examples and the level of customization offered in the initial implementation of Self-E.



a b c d

Figure 1: (a) User prompt for the independent variable (*"the cause"* / the intervention). (b) User prompt for dependent variable (*"the effect"* / the outcome). (c) History of past experiments: each experiment's graph is moved to this tab when the user switches to a new experiment. (d) Home screen explaining that meditating in the morning leads to 0.8 increase in energy levels with a 73% likelihood based on 8 days of data. Also, a new point appears on the graph for each day of tracking.

of time [49]. This results in more holistic data about aspects such as productivity, mood, and energy level. Fixed scheduling may be more appropriate when the aspect should be measured at a consistent time every day or when assessing aspects that do not vary throughout the day.

Self-E imposes a maximum of 5 check-ins per day between 6:00am and 11:45pm. Each check-in measures the effect using only a 1–5 rating questionnaire (Figure 1b). These decisions were based on prior research that suggested self-reported data quality is negatively impacted by lengthy questionnaires and tracking fatigue [78], a common occurrence for self-experimenters who perform multiple check-ins per day [16].

3.2 Architecture

The Self-E system is comprised of a backend server built in Python and a mobile client built in Android (Self-E is now also available on iPhones, but was not during the user studies). The open source Self-E system is available online at https://selfe.cs.brown.edu/.

User profiles are created upon registration with an email and are stored in the backend server. Any configurations made or data tracked are sent to the server throughout the use of the application. Daily check-ins are sent to users' phones via app notifications from the backend server.

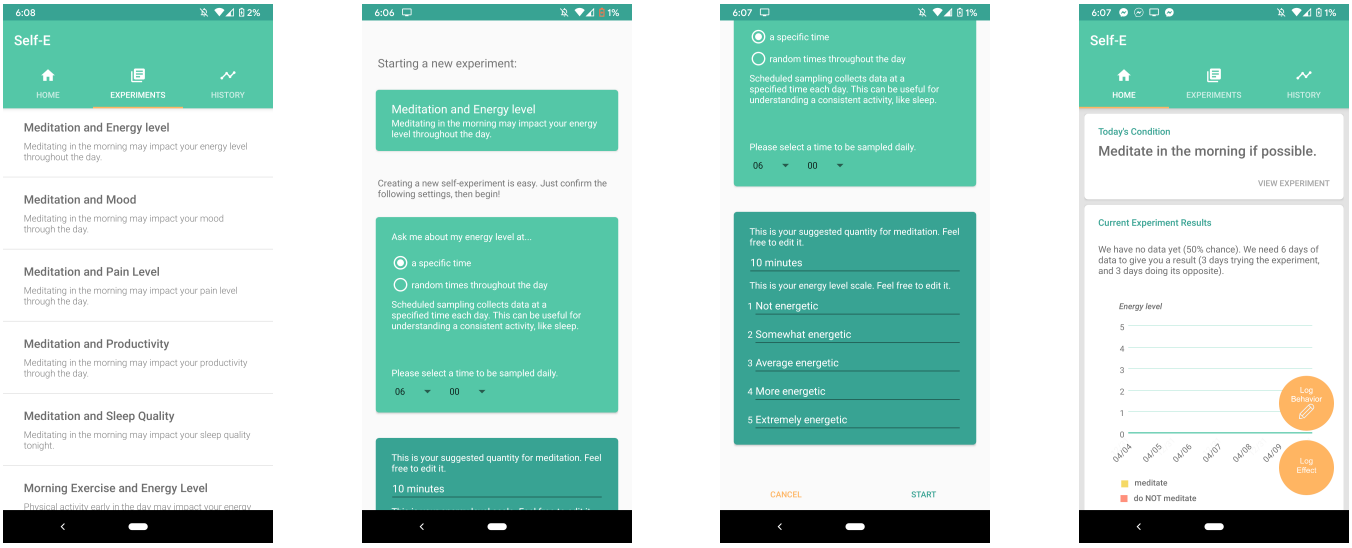


Figure 2: Self-E screens. (a) Experiment selection: users first select an experiment during onboarding, but then are free to change it at any point. (b) Experiment Setup: choose a time to be prompted and edit the goal amount. (c) Revise the labels of the scale. (d) Home: tonight's condition on top and current results below.

3.3 User Interface

3.3.1 Experiment Setup. When a new user begins using Self-E, they are taken through an on-boarding process that introduces self-experimentation and its benefits. Upon completing registration, users are required to select an experiment from a list of pre-configured experiments. These are meant to be easy starting points, as they are each configured by our team using recommended settings backed by research. To curate our list, we drew from self-tracking literature to determine commonly-tracked aspects [65] and then consulted with clinicians to generate interventions that would be both interesting for single-case experimentation and viable for our specific experimental design (e.g., unlikely to have carryover effect). The list contains 24 experiments, each of which is a combination of one of the five independent variables (meditation, physical activity, food & drink, walking, and hours slept) and one of the five dependent variables (energy level, mood, pain level, productivity, and sleep quality). The only combination that we did not include in the list was hours slept and pain level due to the lack of background research to support the viability of such experiments.

3.3.2 Experiment Operationalization. After selecting an experiment, users are taken to a settings page (Figure 2b) where they can customize other aspects of the experiment, as summarized in Table 1: *the independent variable/intervention* check-in time, *the dependent variable* check-in time window, the check-in style (fixed sampling or ESM), and the amount of the intervention they will be applying (e.g., 10 minutes of meditation). This flexibility was intended to allow users to fit self-experimentation into their schedules and to encourage better adherence. As previously discussed, users can also customize the labels of the scale used to rate the dependent variable because the ranges of experience with something like pain can vary among individuals (Figure 2c). By default, labels are populated with the commonly accepted scale for a given variable (except for "Productivity," which does not have a widely accepted scale and so is measured from "very productive" to "very unproductive").

Once the user sets up the experiment, they are taken to the home screen of the app (Figure 2d). Users check in daily with the app at a fixed time for the intervention they are tracking. We chose 8:00pm as it was most appropriate for our list of experiments. Check-ins are initiated via a notification from the app, which takes the user to a pop-up dialog in the app that asks a "yes" or "no" adherence question for the intervention (Figure 1a), or a rating scale for the effect (Figure 1b). Should a user miss the notification or close the pop-up, they can log their data manually via the "Log Behavior"/"Log Effect" buttons on the home screen (Figure 2d).

As a user continues to use Self-E, their data is displayed on a graph (Figure 1c). Self-E requires at least three data points in each condition (e.g., "meditate" and "don't meditate") to calculate a result. This length is based on the minimum length required by the single-case intervention research design standards [48]. Self-E applies an "as-treated" analysis [32], meaning that it only considers the condition users actually followed on a given day, rather than whether they adhered to what they were instructed to do by the app. This type of analysis is recommended when adherence rates are low [32]. With the understanding that users will not always consistently check their phones or go into the app, we elected to prioritize overall adherence over rigidity. This choice exemplifies the balance we attempted to achieve between experimental rigor and practicality.

3.3.3 Experiment Conclusion and Reflection. Results in Self-E include a likelihood percentage and an effect size (Figure 1d). Presenting a likelihood percentage rather than p-value has been shown to be more understandable for users not well-versed in statistics [18]. Past experiment results and data can be viewed in the "History" tab (Figure 1c), providing further opportunity for user reflection. Once results are attained, users may opt to continue tracking data or to start a new experiment.

Similarly to the approach in the open-source SleepBandits system [18], Self-E implements Thompson Sampling, in order to estimate the likelihood that a given condition is helpful. Overall, the goal of the algorithm shown in Equation 1 is to find the action (in this case the condition) that is most likely to lead to an improvement based on the data that has been collected so far [66]. With each new data point, the algorithm updates the beta distributions for each condition, which are based on the α and β shape parameters. These parameters are calculated from the prior probabilities of the given condition and the number of data points so far that either lead to improvement or not, over the running average of the intervention's effect [66].

$$x_t \leftarrow argmax_{x \in \chi} \mathbb{E}_{q_{\hat{\theta}}}[r(y_t)|x_t = x] \quad (1)$$

Once we have the beta distributions of the two conditions, we sample 1,000 times from each and keep track of which distribution returned a higher probability (meaning that condition has a higher likelihood to improve the dependent variable). After 1,000 samplings, we count how many times each condition was returned as the better one, which we transform into the percentage likelihood that the condition is helpful. Figure 1d shows an example result where the likelihood is estimated to be 73%: "Based on the 8 days of data tracked so far, it is most likely (about 73% likelihood) that your energy level is 0.8 levels higher when you meditate in the morning."

4 DIARY STUDY AND FINDINGS

To understand how people use the Self-E system and to draw implications for the design of future self-experimentation systems, we conducted an IRB-approved diary study with participants from the local population.

4.1 Procedure

The study was conducted in three parts: (1) an initial semi-structured interview to gather the participant's background and experience with self-tracking and self-experimentation, (2) a 2-week diary study with daily voicemails, (3) and a semi-structured exit interview to discuss the user's experiences with the Self-E app.

After the initial interviews, we emailed participants a list of the questions to answer in the voicemails. We set up a Google Voice number, and asked them to leave voicemails regardless of how they used the app that day (their pay depended only on leaving the voicemail). For the voicemails, participants were asked to share any challenges they encountered with the app that day, and whether they changed experiments. We list the complete set of questions that were asked in our diary study in the supplementary materials.

Diary studies are high in ecological value since they allow in-situ remote data collection on the real experiences of users [15, 29, 35, 44, 59, 71]. We followed recommendations to limit the duration of similar studies to two weeks [78]. The diary study provided further qualitative feedback to support interview insights.

4.2 Participants

We recruited 16 participants, (P1 – P16), 4 male and 12 female, from the local area by posting flyers in public spaces such as cafes and supermarkets and by posting online in location-based communities such as Reddit and NextDoor. We specifically recruited participants who self-identified as being overall healthy, since the use of Self-E and its single-case design methodology might be inappropriate for people with severe health concerns [53]. To be eligible, they also had to (1) own an Android phone and (2) be over 18 years old. While we recruited broadly, one limitation of our study is that our sample was skewed towards undergraduate students (four research assistants, three undergraduate students, an auditor, an engineer, a librarian, and a barista, among others), likely due to the location where interviews were conducted. Participants' ages ranged from 20 to 70 (M=33.7, SD=15.4), and their familiarity with mathematical statistics ranged from none to expert or professional levels. Participants were compensated on a pro-rated basis: $10 for each interview, and $2 for each daily voicemail, for up to $30 for the duration of the study (including a $2 bonus if they completed all 14 days of the study).

4.3 Analysis

In our analysis, we first performed inductive thematic analysis [5, 67] on the voicemails and the open-ended interview questions. Then, similarly to Ye et al., we used an "iterative coding process with open and axial coding to identify emergent themes in the data" [81] because the semi-structured interviews led us to themes that we had not identified in advance (axial coding is a grounded theory method [73]). Examples of codes included 'attitude towards self-experimentation concept' and 'reasons for changing experiments.' Each coded response was reviewed by at least two authors, who wrote summaries for the emerging themes for each code.

After multiple coding meetings, we reached a consensus on the following themes: (1) mental model mismatch and goal-oriented experiment instincts, (2) using the Self-E app: substituting and disregarding instructions, (3) mistrust in the app's results, and (4) most helpful aspects of the system. The first theme helps us answer **RQ1** by summarizing what people's mental models and instincts for self-experiments are, and thus clarifying what functionality a general self-experimentation tool should contain. The rest of the themes help us answer **RQ2** by outlining how people use such a system and what attributes they find most helpful.

4.4 Mental Model Mismatch and Goal-Oriented Experiment Instincts

For this study, we recruited participants who had never used technology to conduct self-experiments. Thus, we were interested in their pre-existing views on and attitudes towards self-tracking and self-experiments.

Our participants reported that their **most common use of self-tracking data was to motivate themselves to change their behavior**. Usually, this desire meant that they would decide to enact a change such as increasing some good behavior or reducing a bad

behavior and would track their data to make sure they were following through. This finding echoes prior insights that self-tracking apps have the potential to support habit formation and behavior change [72]. The mental model of using self-tracking as a method towards *behavior change* is not a form of self-experimentation itself, but it does reveal that participants were interested in implementing changes to improve some aspects of their lives. Thus, the common mental model in participants was to start with implementing a behavior change based on an assumption about how helpful it would be. In contrast, self-experimentation in the context of Self-E was the opposite: the goal is to learn something new and *then* use that as motivation for implementing the change. Thus, although Self-E was built on the mental model of self-tracking for the purpose of experimentation, not automatic behavior change, it was going to be used by people who did not necessarily share that model of thinking, a discrepancy which led to some interesting findings, described in Section 4.5.

Furthermore, most of our participants had not heard of **the concept of "self-experiments"** before. However, they had intuitions about what self-experiments were (P10: *"doing an experiment except you do it on yourself to see if there's any changes based on what you changed in your life"*). Interestingly, two participants thought the word itself carried negative connotations: *"when I think of an experiment, I think of something that may be harmful"* (P4), and *"I'm actually a little put off by it because it's not a super intuitive way to think about it. To me it's just an attempt to get better"* (P11).

Overall, participants said that if they were to conduct a self-experiment, their **main goal would be to identify what general advice works specifically for them**. Three participants brought up the idea that their bodies were perpetually changing with time (both with age and on a monthly or seasonal basis), so that behaviors which were once helpful may no longer be as beneficial to them. As P11 explained, *"you're always just kind of readjusting parameters,"*, so it would be useful for users to learn what works for them at a given time.

We asked study participants to design their own self-experiments, which helped us gain insight into how people think such experiments should be conducted. Prior work suggests that tutorials can help people set and achieve behavioral goals throughout an unstructured form of self-experimentation [51]. We noticed that **users first state a general goal** such as "I want to improve my sleep." However, that statement in itself does not constitute an experiment, so the interviewer often had to nudge participants to specify their goals (e.g., "wake up less") and to clarify what changes they would implement and observe (e.g., "wear earplugs"). It was also necessary to further probe how long they would experiment for, how they would keep track of the variables, and how they would measure success.

Regarding **the setup of the experiment**, fifteen of the sixteen participants said that they would simply implement the change for a given period of time (between 2 weeks and 1 month). When that period passed, they would reflect whether there was an improvement relative to how they felt before the experiment. In essence, they would instinctively conjure an interrupted time-series setup for understanding the effects of their behavior change. Only three participants brought up and tried to mitigate the effects of possible confounding variables, such as how weekends would affect their experiments. This behavior is similar to how participants in previous studies instinctively set up self-experiments without considering temporal effects, and without randomizing when to implement the behavior change, which could lead to internally flawed experiments [16].

4.5 Using Self-E: Substituting and Disregarding Instructions

At the end of each pre-study interview, we asked participants to download the Self-E app and to use it every day for two weeks. They were free to select whichever experiment from the app's list that they wanted (e.g., 'Meditation and Energy Level'), as well as to change to a new one whenever they wanted. As described in Section 3.3, the app also instructed which condition to follow every given day ('Today, meditate' or 'Today, don't meditate'). On average, the compliance rate for those instructions was 73%. As a comparison, prior work has reported varying degrees of average adherence rates among similiar studies (from 22.5% in QuantifyMe [68], to 60% in Sleepbandits [18], and 95% in TummyTrials [36]).

A common trend, when following the instructions was not possible, was for participants to *"do the best [they] could,"* meaning that **they substituted the exact behavior that the app required with something more feasible**. For example, P1 exercised at night instead of in the morning because that was what her schedule allowed. Furthermore, most participants also said that they would gradually build up to a certain goal amount of the intervention. P15, for example, wanted to decrease her sugar intake, but would *"try to make it gradual rather than cold turkey—decrease per day for a while until it's down to quite a little."* This is in accordance with Fogg's behavioral model of Tiny Habits [26] in which people are advised to start with the smallest effort in order to build up to a new habit.

A novel trend we noticed was that participants willfully **disregarded the instructions when asked to follow the behavior they considered less healthy**. While previous work cites reasons for lack of adherence, this one has not been discussed in detail in the context of self-experimentation systems. P11, for example, also chose to run an experiment on the effects of decreasing the amount of sugar, so on some days, the app asked her to have some cookies and on others to avoid them. When the app asked her to have cookies, however, she said that *"I'm not going to intentionally do something bad for me."*

Overall, diary study participants who disregarded the app's instructions had preconceived notions of how helpful the new behavior would be for them, so they chose to avoid the "harmful" condition, even though they set up the experiment and goal amount themselves. We consider the implications of this finding in the Discussion section, as it presents an opportunity both for further educating users on the randomization in the experiment, and for revising the app in a way that accounts for the users' prior beliefs about the new behavior's effect.

4.6 Mistrust in the App's Results

Once each user had completed at least three days in each condition, Self-E calculated the result with the help of Thompson Sampling. The graph and the sentence summarizing the result (Figure 1d) were mostly clear and easy to understand for users. P14 said that

"what was ultimately more helpful was the result sentence" and P2 particularly liked that the app *"organized everything in my brain."*

Overall, some participants revealed that they trusted and agreed with the results that the Self-E was showing them (P2, P6, P8, P10), especially when it confirmed how they already felt (P5, P10, P15). However, it is important to focus on the three main reasons for mistrust in the results as they carry important implications for future systems.

First, some participants felt that **the effect of the intervention was too negligible** compared to the effects of other things in their daily life (P15). P11 elaborated that there were many other confounding variables that were affecting her more than the one cookie she was eating a day: *"So like reporting how my sleep was and whether I eat a cookie or not, like literally one cookie... What if I took a nap that day? Or what if I run every other day... There's so many variables that, I assume anecdotally, have a stronger effect on my sleep."*

Second, if **the results contradicted a previously held belief**, some participants expressed skepticism and found excuses for why the results were 'inaccurate.' P15, for example, thought the result was just calculated too early: *"I was thinking maybe it'll be lower energy for the first couple of days, and then overall higher energy. But it just said it was overall less which is a little bit unexpected to me. But maybe it's just something that needed more time."* P14, on the other hand, said that *"If it weren't for the extreme data points, I would have trusted it."*

Third, as suggested by prior work [18], **the high likelihood percentage or its drastic fluctuation** over the course of the experiment raised suspicion. P12 and P15 were skeptical because the result either fluctuated too quickly (from 66% to 58% to 97%) or was too high on the first day (96%). P5, P9, and P11 noted that the fluctuations were likely due to unclear distinctions between the two experimental conditions. P13 thought that he was not varying his hours slept enough to lead to a difference, and P5 said that on the days she was not meditating, she was reading before bed, but she realized over time that *"it had the same effect on sleep."*

4.7 Most Helpful Aspects of Self-E

Overall, twelve participants (75%) found Self-E useful for their self-experimentation needs. Ten participants said they would recommend Self-E to someone without self-experimentation experience who is interested in exploring if an intervention works for them.

In line with previous work [18], six participants specifically appreciated **the list of suggested experiments** which was meant to guide new users. P14, for example, said *"I really like that it had setup experiments because it gives you a place to start if you're like 'I want to improve my health.' But if you just Google 'improve your health,' it'll tell you 1,200 different ways to do that. And that's not particularly helpful."* Some participants expressed the desire to make their own experiments where they have the freedom to choose their own variables, yet stressed that the default list should be kept for structural guidance.

Nine participants said that **the scaffolding aspect of the app** was particularly helpful in running a self-experiment because it guided them through: (1) the choice of the experiment, (2) the process of what to do and when to do it, and (3) the input and analysis of their data to provide *"credible results"* (P10). P9 elaborated that *"I don't fully trust myself to design a rigorous self-experiment."* P12 and P15 expressed how they would not have incorporated aspects such Experience Sampling Methods (ESM) for collection of user data or randomizing what condition to follow on a given day.

Additionally, most participants appreciated the **low level of effort** required to add data points. Most participants (60%) liked how brief the data entry questions were, and thought the scales for reporting the dependent variables did not need any modification. Six participants expressed that while they might be capable of conducting such experiments without the app, it would be too tedious or challenging. P14 pointed out that it *"provides a minimal amount of structure what would still give you some flexibility to play with, while making it drastically easier to track it all."*

5 CUSTOMIZED EXPERIMENTS REDESIGN AND FINDINGS

In our diary study, six participants expressed a desire for more flexibility in choosing their own independent and dependent variables, so we iterated on Self-E's design to allow custom experiments from scratch while maintaining the balance of practicality and usability. In setting up a custom experiment, users can determine the independent variable, dependent variable (Figure 3b), two conditions to compare (Figure 3c), and experiment length, on top of existing configurations that can already be made (such as rating scale labels (Figure 3d), check-in time, etc). Unlike existing research systems, the custom experiments feature presents the user with an unprecedented amount of freedom in self-experimentation, allowing them to run a vast variety of potential experiments.

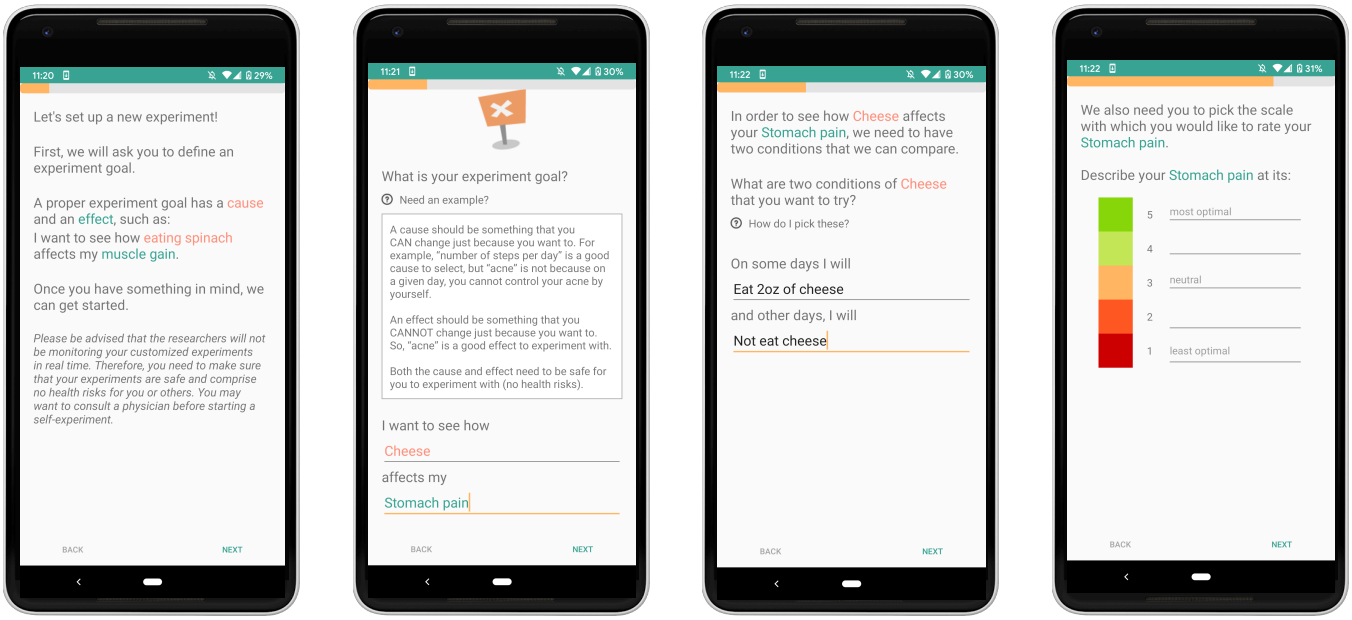
5.1 Customized Experiments Study Method

To learn how potential users create customized experiments on their own with this more configurable version of Self-E, we sought to have more experimentally-minded users, a sort of extreme group, with the idea that if more advanced users are not able to benefit from Self-E, then it is likely a more general population would not. We asked 16 students (S1 – S16) who had received experimental design instruction in a Human-Computer Interaction research seminar course to download and use the app. They were tasked with running two iterations of a self-experiment: the first using their own methods to track and record data and the second using the Self-E application. They were also required to use statistical methods other than those present in Self-E to measure the significance of their findings. While they were encouraged to run experiments relating to mood, many also opted to measure productivity or focus levels. They were asked to keep a journal recording every decision they made and action they pursued.

We conducted thematic analysis on the students' journals and identified two main themes: (1) need for iteration on the experiment setup and, similar to the local population, (2) a mistrust in the app's results.

5.2 Need for Iteration

Participants in the diary study were encouraged to change their experiments whenever they wanted. However, they were not specifically instructed to iterate on their experiments, so most of them did



a b c d

Figure 3: Self-E customized experiment flow. (a) First screen explaining the need for a cause and effect. (b) The user can define their own *cause* and *effect* and the guiding text only appears if they tap on the "Need an example?" (c) The user has to pick two conditions by filling in the blanks. (d) In addition to selecting scheduled or randomized sampling, and the length of their experiment, users can also revise the labels for their "effect" scale.

not modify their experiment for the duration of the two-week study. The seminar students, however, were specifically asked to run two iterations. The first iteration of the self-experiment was intended to give participants practice with self-experimentation and identify and later address issues and challenges during the second iteration. However, some issues remained, primarily around bias and validity, and new issues arose with the introduction of Self-E in the second iteration.

Ten participants altered their dependent variables to account for newly-discovered confounding factors. S14, for example, removed the measurement of a productivity scale since it was *"heavily dependent on other activities during the day."* Others removed dependent variables that caused validity issues, such as S12 who *"found that eliminating the to-do list significantly mitigated the carryover effect."* Half of the participants found that mood and other broad qualifiers were too challenging to pin down and therefore narrowed the scope or merged dependent variables (S7) to more specific indicators, such as irritability, engagement, and sleep quality.

Despite the adjustments made after the first iteration of the self-experiment, some participants still struggled with their dependent variables. S2 had trouble distinguishing between positive and negative moods and S15 had trouble *"scoring"* these moods on a fixed scale. Additionally, many of these participants questioned the construct validity of their measurements and dependent variables, wondering, such as S5, if their chosen measure, *"roughly [equated]"* to their dependent variable.

Lastly, the class study was conducted during Spring 2020, when the COVID-19 global pandemic disrupted usual routines. Three participants reported constrained timelines in addition to highly unusual emotional circumstances as debilitating to the smooth running of the experiment, primarily those related to attention, productivity, or focus. To combat the unexpected burdens of the pandemic and other confounds, these experiments tended to have controls for sleep and wake up times while experiments that involved food or drinks tended to have controls for the quantity and frequency in which these were consumed.

5.3 Mistrust in App's Results

As part of the experimental requirements, students in the class used statistical methods other than those present in Self-E to measure their results. Ten participants preferred their own method to Self-E's (including two stating their own method and Self-E's method are more preferable in different situations), regardless of whether their results aligned with Self-E's. Four participants thought Self-E's Thompson Sampling was insufficiently detailed, unable to capture *"actual observed effect of the condition"* (S7) or take into account *"external circumstances"* (S9). Three also believed that there were insufficient data points for Thompson Sampling to perform well.

Five participants' own calculations or feelings towards the experiment did not align with Self-E results. Amongst them, four were more convinced by their own results. Justifications were attributed to learning new things about themselves or noticing changes in themselves with regard to the interventions. S8, for example, realized that the intervention unexpectedly made them *"want to end the day earlier"* and thus the high probability of sleeping earlier found by Self-E might have been *"just based on chance."* In the end,

nine of the sixteen participants classified their interventions as effective and ten received Self-E results that aligned with their own calculations or perceptions.

6 OPPORTUNITIES FOR SELF-EXPERIMENTATION APPS

These studies helped us identify three main opportunities for improving an app-based approach to practical self-experiments. They focus on the themes from our findings and highlight the need to build trust in what the app is doing in collaboration with the user, so that the app is more than a data entry and statistical calculator.

6.1 Even More Guidance

Participants expressed a desire for more guidance from the app in three key areas. First, they kept bringing up confounding variables that might have affected their experiments and at the same time participants themselves did not always comply with the randomized schedule from the app. The practical approach could be improved by **explaining why randomization is important** and how it could address problems raised by the confounding variables. Second, some participants were not sure how to go about applying the suggested interventions in their lives. In the diary experiment, P2 wanted the ability to read more about the interventions before changing experiments, and P11 wanted to learn more about the science behind self-experimentation in general. So it would have been helpful, for example, to either have **a short tutorial** on how to meditate or to point people to a specific app that can help them with meditation. Third, participants said that they wished the app would **nudge them to change experiments** once they received their calculated result or would just suggest any other relevant ones to try.

6.2 Revising the Data

Contextualizing the data entered and shown was an important consideration for participants, but the existing app did not provide enough flexibility to account for the data after-the-fact. Some participants mentioned that their experimental data had outliers, which led to a distrust of the results due to the potential influence of these outliers. P13 wanted to restart his experiment because he noticed that some of his data were faulty, so he changed to an entirely different experiment, then immediately reverted to the original. P6 wanted to be able to revise the timing of the sampling after her experiment began.

Meanwhile, P4, P10, and P9 asked for a clearer indication of what has been logged so far to be able to verify the recorded data. Participants proposed that the app should allow users to edit data directly at a later time or label them as outliers so they would be interpreted differently during the analysis. Another suggestion was to let users **add extra comments and text to data points** purely for illustrative purposes (P5).

6.3 In-App Motivation

P4 wanted a more visual and interactive interface of the app that would motivate her, give her rewards, and push her limits. She was hoping the app would be more like a coach/personal trainer: *"let's make a plan, let's push your boundaries, let's do another mile today."* For this, we believe the concept of "streaks," where users are encouraged to follow the recommendation multiple days in a row, could be helpful. Streaks would map well to self-experiments, both as a reward for compliance as well as a motivator to collect data daily, while increasing the experimental validity of the results.

Additionally, most participants were enthusiastic about the idea of **sharing and comparing their results with friends and family.** Interestingly, P14 and P5 saw self-experiments as a potential *"bonding experience"* to see how a given intervention affected people they know: *"I would compare my result, to see the difference between two people doing the same thing... With my brother, we don't live in the same place"* (P5). P14 wanted to share her results with online communities of people who are interested in improving the same dependent variable (pain level). She said that *"People tell us so many garbage things that don't work. I cannot tell you how many times people told me to do yoga."* Other participants did not want to compare or share results because *"everyone has their own rate"* of exercising (P6) and they would not be surprised if other people had different results (P15, P16).

7 DISCUSSION

The study revealed insights about how people intuitively design self-experiments, and how they interact with an app that aimed to guide them through the steps of a practical self-experiment. There were some overlapping themes, despite one population having more experimental design training.

7.1 Instinctive versus Scientific Experimental Design

Most participants in our diary study used their self-experiments as a way to start implementing a behavior change that they had been thinking about for a while. People's mental models did not match Self-E's attempts to nudge them to implement the intervention on some days, and to avoid it on others. The participants saw the experiment as a catalyst that finally helped them act on the new behavior, and they were reluctant to stop doing it. It was simply easier to continue doing a behavior than to have to check and switch between experimental conditions.

On a related note, most participants brought up the notion of building up to a goal amount of the intervention they were trying to implement. This mental model is in accordance with Fogg's Tiny Habits behavioral model, in which one should start a behavior change with the smallest increment possible [26]. However, existing self-experimentation systems such as TummyTrials, QuantifyMe, SleepCoacher, and even Self-E, are not designed to handle this behavior [17, 36, 68].

In order to bridge the two paradigms, future self-experimentation systems could leverage the motivational and gamifying aspects of self-tracking systems as powerful tools for helping people conduct and finish their self-experiments. Future systems could give themselves the role of providing the initial inertia for users to commence a behavior change. Self-tracking tools could, in turn, benefit from richer data analysis to help users gain actionable insights. Larger block sizes may be a useful parameter for users who wish to balance between the convenience of inertia, and the efficiency of the number times the condition is randomized to gain more probabilistic confidence in the causal outcome.

7.2 Need for Iteration

It is important to note that fully customizable experiments pose risks as people might setup ones that cause them harm. With the help of our ethics review board, we decided on including a disclaimer on the first page of the setup process (Figure 3a) encouraging users to consult a physician.

As our findings from the study with the seminar students showed, even advanced users struggle with the design and implementation of self-experiments. This is in accordance with prior research that suggests running iterations of self-experiments is helpful for users to conduct higher quality self-experimentation [16]. Thus, such systems for self-experimentation should be designed with iteration in mind and allow users to easily revise and restart their experiments when necessary. For example, as mentioned in the previous point, Self-E users often wanted to start with a small amount of the new behavior and build their way up. However, while we allowed participants to revise the default goal amount at the beginning of the study, they were not able to change it without restarting the experiment. Therefore, future systems for self-experimentation should take this into consideration. In addition to that, as noted in our findings, some participants had to be nudged to specify their goals as they were under-specified and not actionable, which resonates with prior studies [51]. This suggests that there is a need for scaffolding around the setup of the self-experiment to educate the user about why it is necessary to have a specific and measurable goal in mind [50] rather than a broad one ("improve sleep" is too broad, but "wake up less" is more actionable) in order to make meaningful science [60].

7.3 Need for Empathy

Our findings related to labeling and handling data outliers, acknowledging and accounting for result preconceptions, and building user trust in the system, all demonstrate the need both for more empathy to be built into the system and for better communication between the user and the app. For example, we see that participants are aware that their data sometimes is faulty or contains outliers, so they want to be able to exclude certain data points from analysis. Thus, future systems could let users label extreme data points and then take that additional information into account when calculating the result.

Additionally, most participants in our study also had preconceived notions about whether an experiment would be helpful to them before they even started it. Our findings suggest that they thought the intervention would be beneficial, so when Self-E presented results that contradicted that belief, users found excuses for confounding factors that could have interfered with the data. A more empathetic app might explain, "we know you think this is bad, but let's test once just to confirm."

Another way to account for this bias while calculating the results is to allow the participants to feed these assumptions as the priors in the Thompson Sampling approach. That way, if participants are very certain something might be helpful for them, the more helpful condition will have a high prior likelihood than the other condition, which in turn will be reflected in how often they are asked to try each condition. However, we need to be cautious because feeding assumptions as priors might lead to potential biases in the results.

7.4 Future Research Directions

We released the Self-E app to the Google Play Store on March 1, 2020, for anyone to download and use for free. Users who download the app are asked to sign an IRB-approved informed consent form if they wish to continue using the app. Users are free to use or delete Self-E whenever they wish. With Self-E widely available on the app store, we plan on analyzing the way people are creating customized experiments on their own and the challenges they face at the different stages and with different experiments. Additionally, we hope to explore what levels of scaffolding in the customized experiments process are most applicable for different populations.

A future iteration of Self-E could also educate users and give advanced ones more agency to conduct complex experiments with different data measures and study designs with multiple conditions. Future research will also aim to answer whether incorporating social features such as sharing and commenting on other people's experiment designs mitigates some of the negative aspects of customized experiments.

7.5 Limitations

As with other empirical and diary studies, ours has limitations that should be acknowledged while interpreting our findings. Our studies are limited by their small sample size and duration, as a broader sample might have been able to tease apart what health conditions or variables self-experimentation is most beneficial for. Furthermore, we recruited a generally healthy population sample, who are not representative of users who might be conducting self-experiments with the goal of ameliorating a specific health condition. While participants were encouraged to use the app as they normally would, their participation in this study might have biased their natural use of the system. Future work can explore a more naturalistic approach which would better reflect how people adopt Self-E in the wild.

8 CONCLUSION

This work presents Self-E, an app that guides users through the steps of a self-experiment. We conducted a two-week diary study and interviews with 16 participants who used Self-E to conduct general self-experiments in their own lives. Our qualitative study sought to expand on the field's understanding of how people with little experience in personal analytics perceive self-tracking and self-experimentation with the help of mobile tools. Because users bring in self-conceived notions about what affects them, the constant challenge is for the app to build trust with the user, that it is interpreting the data within the user's context. Based on those findings, we redesigned the system to include completely customizable experiments and asked a group of advanced users for feedback. Overall, we find that the instinctive way users conduct self-experiments does not match the implementations in existing systems, so future research can explore ways to help people conduct such experiments in more intuitive ways.

ACKNOWLEDGMENTS

We thank Diana Lee for the editing help. This research is funded in part by the Brown University Seed Award and National Science Foundation IIS-1656763.

REFERENCES

[1] Amid Ayobi, Paul Marshall, and Anna L Cox. 2020. Trackly: A Customisable and Pictorial Self-Tracking App to Support Agency in Multiple Sclerosis Self-Care. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, 1–15.

[2] Amid Ayobi, Paul Marshall, Anna L Cox, and Yunan Chen. 2017. Quantifying the body and caring for the mind: self-tracking in multiple sclerosis. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 6889–6901.

[3] Amid Ayobi, Tobias Sonne, Paul Marshall, and Anna L Cox. 2018. Flexible and Mindful Self-Tracking: Design Implications from Paper Bullet Journals. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 28.

[4] Colin Barr, Maria Marois, Ida Sim, Christopher H Schmid, Barth Wilsey, Deborah Ward, Naihua Duan, Ron D Hays, Joshua Selsky, Joseph Servadio, et al. 2015. 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 (2015), 67.

[5] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[6] Colleen E Carney, Daniel J Buysse, Sonia Ancoli-Israel, Jack D Edinger, Andrew D Krystal, Kenneth L Lichstein, and Charles M Morin. 2012. The consensus sleep diary: standardizing prospective sleep self-monitoring. *Sleep* 35, 2 (2012), 287–302.

[7] Ting-Ray Chang, Eija Kaasinen, and Kirsikka Kaipainen. 2012. What influences users' decisions to take apps into use?: A framework for evaluating persuasive and engaging design in mobile Apps for well-being. In *Proceedings of the 11th International Conference on Mobile and Ubiquitous Multimedia*. ACM.

[8] Connie Chen, David Haddad, Joshua Selsky, Julia E Hoffman, Richard L Kravitz, Deborah E Estrin, and Ida Sim. 2012. Making sense of mobile health data: an open architecture to improve individual-and population-level health. *Journal of medical Internet research* 14, 4 (2012), e112.

[9] Eun Kyoung Choe, Bongshin Lee, Haining Zhu, Nathalie Henry Riche, and Dominikus Baur. 2017. Understanding self-reflection: how people reflect on personal data through visual data exploration. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare*. ACM, 173–182.

[10] Eun Kyoung Choe, Nicole B Lee, Bongshin Lee, Wanda Pratt, and Julie A Kientz. 2014. Understanding quantified-selfers' practices in collecting and exploring personal data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1143–1152.

[11] Chia-Fang Chung, Elena Agapie, Jessica Schroeder, Sonali Mishra, James Fogarty, and Sean A Munson. 2017. When personal tracking becomes social: Examining the use of Instagram for healthy eating. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 1674–1687.

[12] John Concato, Nirav Shah, and Ralph I Horwitz. 2000. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *New England journal of medicine* 342, 25 (2000), 1887–1892.

[13] Felicia Cordeiro, Elizabeth Bales, Erin Cherry, and James Fogarty. 2015. Rethinking the mobile food journal: Exploring opportunities for lightweight photo-based capture. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 3207–3216.

[14] Felicia Cordeiro, Daniel A Epstein, Edison Thomaz, Elizabeth Bales, Arvind K Jagannathan, Gregory D Abowd, and James Fogarty. 2015. Barriers and negative nudges: Exploring challenges in food journaling. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1159–1162.

[15] Mary Czerwinski, Eric Horvitz, and Susan Wilhite. 2004. A diary study of task switching and interruptions. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 175–182.

[16] Nediyana Daskalova, Karthik Desingh, Alexandra Papoutsaki, Diane Schulze, Han Sha, and Jeff Huang. 2017. Lessons learned from two cohorts of personal informatics self-experiments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 46.

[17] Nediyana Daskalova, Danaë Metaxa-Kakavouli, Adrienne Tran, Nicole Nugent, Julie Boergers, John McGeary, and Jeff Huang. 2016. SleepCoacher: A personalized automated self-experimentation system for sleep recommendations. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. ACM, 347–358.

[18] Nediyana Daskalova, Jina Yoon, Yibing Wang, Cintia Araujo, Guillermo Beltran Jr, Nicole Nugent, John McGeary, Joseph Jay Williams, and Jeff Huang. 2020. SleepBandits: Guided Flexible Self-Experiments for Sleep. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, 1–13.

[19] Orianna DeMasi, Sidney Feygin, Aluma Dembo, Adrian Aguilera, and Benjamin Recht. 2017. Well-being tracking via smartphone-measured activity and sleep: cohort study. *JMIR mHealth and uHealth* 5, 10 (2017), e137.

[20] Markéta Dolejšová and Denisa Kera. 2017. Soylent Diet Self-Experimentation: Design Challenges in Extreme Citizen Science Projects. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 2112–2123.

[21] Daniel A Epstein, Nicole B Lee, Jennifer H Kang, Elena Agapie, Jessica Schroeder, Laura R Pina, James Fogarty, Julie A Kientz, and Sean Munson. 2017. Examining menstrual tracking to inform the design of personal informatics tools. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 6876–6888.

[22] Daniel A Epstein, An Ping, James Fogarty, and Sean A Munson. 2015. A lived informatics model of personal informatics. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 731–742.

[23] Bob Evans. 2019. PACO: The Personal Analytics Companion. https://pacoapp.com/

[24] Miguel Farias and Catherine Wikholm. 2016. Has the science of mindfulness lost its mind? *BJPsych bulletin* 40, 6 (2016), 329–332.

[25] Brian J Fogg. 2002. Persuasive technology: using computers to change what we think and do. *Ubiquity* 2002, December (2002), 5.

[26] Brian J Fogg. 2019. *Tiny Habits: The Small Changes That Change Everything*. Houghton Mifflin Harcourt.

[27] Susannah Fox and Maeve Duggan. 2013. Tracking for Health. https://www.pewinternet.org/2013/01/28/tracking-for-health/

[28] Daniel Harrison, Paul Marshall, Nadia Bianchi-Berthouze, and Jon Bird. 2015. Activity tracking: barriers, workarounds and customisation. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 617–621.

[29] Eiji Hayashi and Jason Hong. 2011. A diary study of password usage in daily life. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2627–2630.

[30] Steven C Hayes. 1981. Single case experimental design and empirical clinical practice. *Journal of consulting and clinical psychology* 49, 2 (1981), 193.

[31] Reetta Heinonen, Riitta Luoto, Pirjo Lindfors, and Clas-Håkan Nygård. 2012. Usability and feasibility of mobile phone diaries in an experimental physical exercise study. *Telemedicine and e-Health* 18, 2 (2012), 115–119.

[32] Miguel A Hernán and Sonia Hernández-Díaz. 2012. Beyond the intention-to-treat in comparative effectiveness research. *Clinical Trials* 9, 1 (2012), 48–55.

[33] Mieke Heyvaert and Patrick Onghena. 2014. Randomization tests for single-case experiments: State of the art, state of the science, and state of the application. *Journal of Contextual Behavioral Science* 3, 1 (2014), 51–64.

[34] Alexis Hiniker, Sungsoo Ray Hong, Tadayoshi Kohno, and Julie A Kientz. 2016. Mytime: Designing and evaluating an intervention for smartphone non-use. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 4746–4757.

[35] Tero Jokela, Jarno Ojala, and Thomas Olsson. 2015. A diary study on combining multiple information devices in everyday activities and tasks. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 3903–3912.

[36] Ravi Karkar, Jessica Schroeder, Daniel A Epstein, Laura R Pina, Jeffrey Scofield, James Fogarty, Julie A Kientz, Sean A Munson, Roger Vilardaga, and Jasmine Zia. 2017. Tummytrials: a feasibility study of using self-experimentation to detect individualized food triggers. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 6850–6863.

[37] Ravi Karkar, Jasmine Zia, Roger Vilardaga, Sonali R Mishra, James Fogarty, Sean A Munson, and Julie A Kientz. 2015. A framework for self-experimentation in personalized health. *Journal of the American Medical Informatics Association* 23, 3 (2015), 440–448.

[38] Joseph Jofish Kaye, Mary McCuistion, Rebecca Gulotta, and David A Shamma. 2014. Money talks: tracking personal finances. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 521–530.

[39] Christina Kelley, Bongshin Lee, and Lauren Wilcox. 2017. Self-tracking for mental wellness: understanding expert perspectives and student experiences. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 629–641.

[40] John M Kelley and Ted J Kaptchuk. 2010. Group analysis versus individual response: the inferential limits of randomized controlled trials. *Contemporary clinical trials* 31, 5 (2010), 423–428.

[41] Ian Kerridge. 2003. Altruism or reckless curiosity? A brief history of self experimentation in medicine. *Internal medicine journal* 33, 4 (2003), 203–207.

[42] Elisabeth T Kersten-van Dijk, Joyce HDM Westerink, Femke Beute, and Wijnand A IJsselsteijn. 2017. Personal informatics, self-insight, and behavior change: A critical review of current literature. *Human–Computer Interaction* 32, 5-6 (2017), 268–296.

[43] Julie A Kientz. 2019. In Praise of Small Data: When You Might Consider N-of-1 Studies. *GetMobile: Mobile Computing and Communications* 22, 4 (2019), 5–8.

[44] Young-Ho Kim, Eun Kyoung Choe, Bongshin Lee, and Jinwook Seo. 2019. Understanding Personal Productivity: How Knowledge Workers Define, Evaluate, and Reflect on Their Productivity. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 615.

[45] Young-Ho Kim, Jae Ho Jeon, Eun Kyoung Choe, Bongshin Lee, KwonHyun Kim, and Jinwook Seo. 2016. TimeAware: Leveraging framing effects to enhance

personal productivity. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 272–283.

[46] Young-Ho Kim, Jae Ho Jeon, Bongshin Lee, Eun Kyoung Choe, and Jinwook Seo. 2017. OmniTrack: A Flexible Self-Tracking Approach Leveraging Semi-Automated Tracking. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 67.

[47] Judy Kopp. 1988. Self-monitoring: A literature review of research and practice. In *Social Work Research and Abstracts*, Vol. 24. Oxford University Press, 8–20.

[48] Thomas R Kratochwill, John H Hitchcock, Robert H Horner, Joel R Levin, Samuel L Odom, David M Rindskopf, and William R Shadish. 2013. Single-case intervention research design standards. *Remedial and Special Education* 34, 1 (2013), 26–38.

[49] Reed Larson and Mihaly Csikszentmihalyi. 2014. The experience sampling method. In *Flow and the foundations of positive psychology*. Springer, 21–34.

[50] Gary P Latham. 2003. Goal setting: A five-step approach to behavior change. *Organizational Dynamics* (2003), 309–318.

[51] Jisoo Lee, Erin Walker, Winslow Burleson, Matthew Kay, Matthew Buman, and Eric B Hekler. 2017. Self-experimentation for behavior change: Design and formative evaluation of two approaches. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 6837–6849.

[52] Ian Li, Anind Dey, and Jodi Forlizzi. 2010. A stage-based model of personal informatics systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 557–566.

[53] Elizabeth O Lillie, Bradley Patay, Joel Diamant, Brian Issell, Eric J Topol, and Nicholas J Schork. 2011. The n-of-1 clinical trial: the ultimate strategy for individualizing medicine? *Personalized medicine* 8, 2 (2011), 161–173.

[54] David Mant. 1999. Can randomised trials inform clinical decisions about individual patients? *The Lancet* 353, 9154 (1999), 743–746.

[55] Sean A Munson, Jessica Schroeder, Ravi Karkar, Julie A Kientz, Chia-Fang Chung, and James Fogarty. 2020. The Importance of Starting With Goals in N-of-1 Studies. *Frontiers in Digital Health* 2 (2020), 3.

[56] Gina Neff and Dawn Nafus. 2016. *Self-tracking*. MIT Press.

[57] Rosemery O Nelson and Steven C Hayes. 1981. Theoretical explanations for reactivity in self-monitoring. *Behavior Modification* 5, 1 (1981), 3–14.

[58] Doug Oman, Shauna L Shapiro, Carl E Thoresen, Thomas G Plante, and Tim Flinders. 2008. Meditation lowers stress and supports forgiveness among college students: A randomized controlled trial. *Journal of American College Health* 56, 5 (2008), 569–578.

[59] Leysia Palen and Marilyn Salzman. 2002. Voice-mail diary studies for naturalistic data capture under mobile conditions. In *Proceedings of the 2002 ACM conference on Computer supported cooperative work*. ACM, 87–95.

[60] Vineet Pandey. 2018. Creating Scientific Theories with Online Communities using Gut Instinct. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 109–112.

[61] Sun Young Park and Yunan Chen. 2015. Individual and social recognition: challenges and opportunities in migraine management. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 1540–1551.

[62] James O Prochaska and Wayne F Velicer. 1997. The transtheoretical model of health behavior change. *American journal of health promotion* 12, 1 (1997), 38–48.

[63] Seth Roberts and Allen Neuringer. 1998. Self-experimentation. In *Handbook of research methods in human operant behavior*. Springer, 619–655.

[64] John Rooksby, Parvin Asadzadeh, Mattias Rost, Alistair Morrison, and Matthew Chalmers. 2016. Personal tracking of screen time on digital devices. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 284–296.

[65] John Rooksby, Mattias Rost, Alistair Morrison, and Matthew Chalmers. 2014. Personal tracking as lived informatics. In *Proceedings of the 2014 CHI Conference on Human Factors in Computing Systems*. ACM, 1163–1172.

[66] Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, Zheng Wen, et al. 2018. A tutorial on Thompson sampling. *Foundations and Trends® in Machine Learning* 11, 1 (2018), 1–96.

[67] Johnny Saldaña. 2015. *The coding manual for qualitative researchers*. Sage.

[68] Akane Sano, Sara Taylor, Craig Ferguson, Akshay Mohan, and Rosalind W Picard. 2017. QuantifyMe: An Automated Single-Case Experimental Design Platform. In *International Conference on Wireless Mobile Communication and Healthcare*. Springer, 199–206.

[69] Jessica Schroeder, Chia-Fang Chung, Daniel A Epstein, Ravi Karkar, Adele Parsons, Natalia Murinova, James Fogarty, and Sean A Munson. 2018. Examining self-tracking by people with migraine: goals, needs, and opportunities in a chronic health condition. In *Proceedings of the 2018 on Designing Interactive Systems Conference 2018*. ACM, 135–148.

[70] Jessica Schroeder, Ravi Karkar, James Fogarty, Julie A Kientz, Sean A Munson, and Matthew Kay. 2019. A Patient-Centered Proposal for Bayesian Analysis of Self-Experiments for Health. *Journal of healthcare informatics research* 3, 1 (2019), 124–155.

[71] Timothy Sohn, Kevin A Li, William G Griswold, and James D Hollan. 2008. A diary study of mobile information needs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 433–442.

[72] Katarzyna Stawarz, Anna L Cox, and Ann Blandford. 2015. Beyond self-tracking and reminders: designing smartphone apps that support habit formation. In *Proceedings of the 33rd annual ACM conference on Human Factors in Computing Systems*. ACM, 2653–2662.

[73] Anselm Strauss and Juliet Corbin. 1998. *Basics of qualitative research techniques*. Sage publications Thousand Oaks, CA.

[74] Melanie Swan. 2013. The quantified self: Fundamental disruption in big data science and biological discovery. *Big data* 1, 2 (2013), 85–99.

[75] Jakob Tholander and Maria Normark. 2020. Crafting Personal Information-Resistance, Imperfection, and Self-Creation in Bullet Journaling. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[76] Eric Topol. 2019. The A.I. Diet. *The New York Times* (Mar 2019). https://www.nytimes.com/2019/03/02/opinion/sunday/diet-artificial-intelligence-diabetes.html

[77] San Diego University of California. 2019. Galileo: Design and Run Experiments with people from around the world. https://galileo-ucsd.org/galileo/home

[78] Niels van Berkel, Jorge Goncalves, Peter Koval, Simo Hosio, Tilman Dingler, Denzil Ferreira, and Vassilis Kostakos. 2019. Context-Informed Scheduling and Analysis: Improving Accuracy of Mobile Self-Reports. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 1–12.

[79] Heather P Whitley and Wesley Lindsey. 2009. Sex-based differences in drug activity. *American family physician* 80, 11 (2009), 1254–1258.

[80] Steve Whittaker, Vaiva Kalnikaite, Victoria Hollis, and Andrew Guydish. 2016. 'Don't Waste My Time': Use of Time Information Improves Focus. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 1729–1738.

[81] Hanlu Ye, Meethu Malu, Uran Oh, and Leah Findlater. 2014. Current and future mobile and wearable device use by people with visual impairments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 3123–3132.